Tavatar: Topology-Aware Gaussian Attribute Derivation for Animatable Human Avatars

Hailin Luo[1*] Yifan Yang[5,6*] Jiazhi Shu[1*] Zixiong Huang[1] Qi Chen[2†] Qing Du[1†] Mingkui Tan[1,3†]
[1]South China University of Technology [2]Adelaide University
[3]Pazhou Lab [5]Electric Power Research Institute, CSG
[6]Guangdong Provincial Key La-boratory of Power System Network Security
seluohailin@mail.scut.edu.cn, qi.chen04@adelaide.edu.au; mingkuitan@scut.edu.cn

Abstract

*Reconstructing high-fidelity, animatable human avatars from monocular videos remains a critical challenge. Existing 3DGS-based human animation methods constrain Gaussian parameters but exclude scale, which we argue is crucial for adapting human poses to challenging out-of-distribution poses. To achieve robust animation under unseen poses, we propose **Tavatar**, which derives key parameters such as scale, rotation, and other geometric attributes directly from the local mesh geometry, instead of learning them through unconstrained optimization. This paradigm shift enforces topological consistency by design, as each Gaussian is analytically anchored to the local mesh geometry, inheriting its spatial structure and deformation behavior. Specifically, we bind Gaussians to mesh faces and vertices, deriving their scales and orientations from triangle properties and local edge lengths to ensure coherent surface coverage. To ensure the stability of this analytical mapping, we introduce a crucial equilateral regularization term that preserves mesh integrity. Extensive experiments demonstrate that Tavatar achieves superior animation robustness on challenging out-of-distribution poses, reducing normal error by 13.8% on X-Avatar and 17.9% on PeopleSnapshot against the best baseline, while maintaining competitive rendering quality. Project Page is https://hailin545.github.io/tavatar/.*

1. Introduction

The creation of animatable 3D human avatars holds transformative potential for immersive technologies like virtual reality, gaming, and film production. Recent methods have integrated neural rendering techniques (e.g., NeRF [18], 3DGS [11]) with parametric human models (e.g., SMPL [17]) to enable drivable avatar reconstruction from monocular video. While NeRF-based approaches [7, 8, 26] demonstrate high-quality results, they suffer from prohibitive computational costs due to the intensive volume rendering process. 3D Gaussian Splatting (3DGS) offers a more efficient explicit geometry representation and real-time rendering, leading recent methods [5, 14, 23, 34] to adopt 3DGS for human reconstruction.

However, existing 3DGS-based methods typically treat Gaussians as free-floating entities with independently optimized geometric attributes (rotation and scale). This independence, while offering modeling flexibility, fundamentally lacks topological coherence. During animation, the clothed human surface undergoes complex stretching and compression that are constrained by the underlying mesh topology. Without explicit topological constraints, these freely optimized Gaussians decouple from the underlying mesh's deformation—they overfit to training poses (e.g., simple rotations) and fail for out-of-distribution (OOD) poses (e.g., complex gestures), resulting in Gaussian detachment (Fig. 2) or surface holes that noticeably degrade immersion. Even recent improvements [23, 34] only constrain Gaussian rotation but leave scale unregulated, still causing surface inconsistencies. Such a partial constraint is insufficient: without topology-aware scale attribute derivation, Gaussians often fail to follow local surface deformation, inevitably leading to artifacts under novel poses.

To resolve this topological incoherence, we introduce Tavatar, a geometry-driven paradigm for animatable human avatar reconstruction. Instead of freely optimizing Gaussian attributes, we analytically derive them from a deformable mesh, ensuring topological consistency and robust generalization across poses. Concretely, Gaussians are attached to a clothed mesh predicted from a canonical SMPL model and animated via SMPL Linear Blend Skinning (LBS), which provides the stable topological structure required for analytical computation. Building on this foundation, we detail how Gaussian attributes are derived and how mesh quality is maintained during animation.

* Authors contributed equally, † Corresponding author.

Analytical Gaussian Attribute Derivation. Unlike existing methods that only regulate Gaussian orientation, we analytically determine both orientation and scale from local mesh geometry. For face-bound Gaussians, scales are computed using the inscribed-circle radii of triangular faces for structured surface coverage. For vertex-bound Gaussians, scales are derived from local edge lengths, enabling sensitivity to local geometric detail while preserving topological consistency. This geometry-driven derivation keeps Gaussian attributes coherent with mesh deformation, avoiding the instability of free-form optimization.

Equilateral Regularization. Since the stability of this analytical mapping depends on mesh quality, we introduce an equilateral regularization term to prevent triangle degeneracy during training and animation. By minimizing edge-length variance and angular distortion, this regularizer preserves mesh integrity and ensures stable Gaussian scaling and complete surface coverage across poses.

In summary, our main contributions are:

- We propose Tavatar, a novel topology-aware framework that shifts from optimization-based to geometry-driven determination of Gaussian geometric attribute. In contrast to prior methods that partially rely on optimization to infer Gaussian attributes, our approach determines them directly from mesh structure, yielding topologically coherent representations.
- We introduce a novel Gaussian binding mechanism where geometric attributes are analytically derived from local mesh topology, ensuring deformation coherence and preventing Gaussian detachment under complex poses.
- We design an equilateral regularization that maintains mesh quality by minimizing edge length variance and angular distortion, ensuring stable analytical mapping throughout optimization and animation.

2. Related Work

Mesh-based Human Avatars Traditional 3D human modeling often employs parametric human models, such as SMPL [17] and SMPL-X [25]. These methods capture fundamental posture and overall shape by optimizing model parameters, which can be predicted through regression [10, 22, 24] or optimization [4]. Some approaches [13, 16, 19] directly regress 3D vertices. While these methods generate animatable meshes, they are limited in reconstructing intricate geometric details like clothing wrinkles. Although some research extends parametric models with deformation layers [2, 3] to simulate clothing, achieving precise reconstruction of high-frequency details while maintaining topological consistency remains a significant challenge, especially with sparse observational data.

NeRF-based Human Avatars Neural Radiance Fields (NeRF) [6, 18] learn an implicit function for scene geometry and appearance, enabling high-quality novel view synthesis. In human reconstruction, NeRF has been integrated with parametric human models like SMPL [17] in works such as Neural Body [27], MonoHuman [37], and Human Nerf [35] to create drivable virtual humans by modeling deformation as a warp field. Although NeRF-based methods offer remarkable visual fidelity, they are computationally intensive, leading to slow training and inference speeds. Their generalization capabilities are also limited for out-of-training poses, often causing rendering artifacts. Efforts to accelerate processing, such as Instant Avatar [7] and Instant-NVR [9], exist, but high memory consumption and stability in complex dynamics persist as challenges.

3DGS-based Human Avatars 3D Gaussian Splatting (3DGS) [11, 15] has emerged as a leading technique for dynamic human reconstruction due to its exceptional training and rendering efficiency. Existing 3DGS-based avatar methods can be categorized by their approach to topological constraints: (1) *Unconstrained Optimization*: Many methods [5, 14, 29] directly optimize free-form Gaussians alongside SMPL deformations. Although recent works [20, 31] improve rendering fidelity or pose robustness through SMPL-X alignment or pose-driven deformations, the unconstrained Gaussian optimization often leads to geometrically incoherent structures, causing floating artifacts and surface discontinuities on out-of-distribution poses. (2) *Partial Topological Constraints*: To address this, some works introduce limited constraints. IHuman [23] constrains Gaussian orientations via mesh normals but leaves their scale unregulated, causing surface inconsistencies. GomAvatar [34] binds Gaussians to the SMPL mesh surface but lacks mesh quality regularization, leading to unstable scaling when mesh triangles are distorted. Other methods use mesh priors for initialization [28, 32] but struggle to capture fine details and maintain consistency during animation.

3. Preliminaries

Skinned Multi-Person Linear model (SMPL). The SMPL model [17] is a parametric framework that defines a human body mesh using shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ and pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{3K}$ for K joints. It provides a canonical template mesh, $\mathcal{M}_c = (\mathbf{V}_c, \mathbf{F})$, which can be deformed to match subject-specific shape and pose. The animation is achieved via Linear Blend Skinning (LBS), which transforms a given set of vertices $\mathbf{V}$ into a posed configuration $\mathbf{V}_p$. For each vertex $\mathbf{v} \in \mathbf{V}$, its posed counterpart $\mathbf{v}_p$ is:

$$\mathbf{v}_p = \sum_{k=1}^{K} w_k \left(\mathbf{R}_k(\boldsymbol{\theta}) \mathbf{v} + \mathbf{t}_k(\boldsymbol{\theta}) \right), \tag{1}$$

where w_k is the predefined skinning weight for the vertex with respect to the k-th joint, and $\mathbf{R}_k(\boldsymbol{\theta}) \in SO(3)$ and

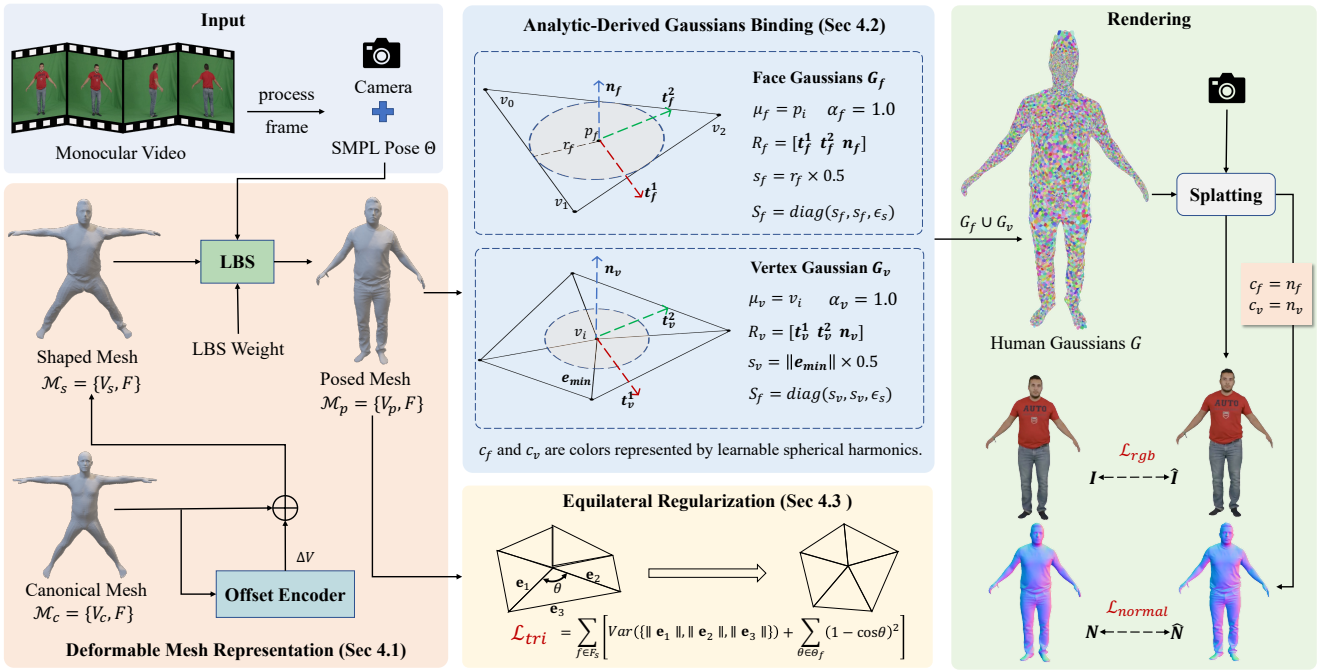

Figure 1. **Method overview.** We propose Tavatar, a geometry-driven paradigm that reconstructs high-quality animatable human avatars by analytically deriving Gaussian attributes from a deformable mesh. Our approach includes: **Analytical Gaussian Attribute Derivation**: All Gaussian positions, scales, and orientations are computed directly from mesh topology yielding structurally correct Gaussian placement, improved surface coverage, and pose-consistent animation across challenging motions. **Equilateral Geometry Regularization**: An equilateral constraint enforces stable Gaussian binding on the mesh, preventing degeneration and ensuring robust reconstruction quality, especially under large deformations and in fine-detail regions such as hands and clothing folds. *Indices in formulas are omitted for brevity.*

$\mathbf{t}_k(\boldsymbol{\theta}) \in \mathbb{R}^3$ are the joint's rotation and translation derived from the pose parameters $\boldsymbol{\theta}$.

3D Gaussian Splatting. 3DGS [11] represents scenes as a collection of anisotropic 3D Gaussians, $\mathcal{G} = \{g_i\}$. Each Gaussian g_i is defined by a position $\boldsymbol{\mu}_i \in \mathbb{R}^3$, a covariance matrix $\boldsymbol{\Sigma}_i \in \mathbb{R}^{3\times3}$, an opacity value $\alpha_i \in [0, 1]$, and color, which is represented by Spherical Harmonics coefficients $\mathbf{sh}_i \in \mathbb{R}^{(n+1)^2\times3}$ of degree n. To ensure positive semi-definiteness, the covariance matrix is parameterized as:

$$\boldsymbol{\Sigma}_i = \mathbf{R}_i \mathbf{S}_i \mathbf{S}_i^T \mathbf{R}_i^T, \tag{2}$$

where $\mathbf{R}_i \in SO(3)$ is a rotation matrix represented by a quaternion, and $\mathbf{S}_i = \text{diag}(\mathbf{s}_i)$ is a diagonal scaling matrix derived from a scaling vector $\mathbf{s}_i \in \mathbb{R}^3$. For rendering, each pixel's color C is computed through differentiable α-blending of Gaussians sorted by depth:

$$C = \sum_{i \in \mathcal{N}} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{3}$$

where $\mathcal{N}$ denotes the set of Gaussians ordered front-to-back along the viewing ray, and c_i is the view-dependent color decoded from $\mathbf{sh}_i$. The Gaussian parameters are optimized via photometric reconstruction loss. During training, 3DGS adaptively refines the representation through periodic densification and pruning operations, ensuring optimal spatial distribution for high-quality scene representation.

4. Topology-Aware Gaussian Avatar

Overview. As illustrated in Fig. 1, our approach consists of three key components: (**i**) We construct a personalized mesh representation (Sec. 4.1) by deforming a canonical SMPL template [17] with learned vertex offsets, which is then animated via Linear Blend Skinning (LBS). (**ii**) We introduce an analytical Gaussian binding mechanism (Sec. 4.2) that derives all geometric attributes directly from local mesh geometry, establishing inherent topological coherence. (**iii**) We propose an equilateral regularization (Sec. 4.3) to maintain mesh quality and ensure stable analytical mapping throughout optimization and animation. This unified framework enables joint optimization of mesh geometry and Gaussian appearance, yielding avatars with both photorealistic rendering and robust pose generalization.

4.1. Deformable Mesh Representation

Mesh parameterization. Our method builds upon a deformable mesh that provides the topological scaffold for Gaussian binding. We initialize with a canonical SMPL model [17] in T-pose, denoted as $\mathcal{M}_c = (\mathbf{V}_c, \mathbf{F})$, where $\mathbf{V}_c = \{\mathbf{v}_i^c\}_{i=1}^N \in \mathbb{R}^{N\times 3}$ represents N vertices and $\mathbf{F} = \{f_j\}_{j=1}^M$ denotes M triangular faces with fixed topology.

To capture subject-specific details such as clothing wrinkles and body shape variations, we introduce a neural shape encoder $E_s : \mathbb{R}^3 \to \mathbb{R}^3$ implemented via multi-resolution hash encoding [21]. This encoder predicts per-vertex offsets $\Delta\mathbf{V} = \{\Delta\mathbf{v}_i\}_{i=1}^N$ from canonical vertex positions, producing a personalized mesh:

$$\mathbf{V}_s = \mathbf{V}_c + E_s(\mathbf{V}_c) = \{\mathbf{v}_i^s = \mathbf{v}_i^c + \Delta\mathbf{v}_i\}_{i=1}^N, \quad (4)$$

where $\mathcal{M}_s = (\mathbf{V}_s, \mathbf{F})$ retains the canonical topology while adapting to subject-specific geometry.

Pose-driven deformation. To animate the personalized mesh into arbitrary poses, we apply standard Linear Blend Skinning (LBS) as defined in Eq. (1), driven by SMPL pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{23\times 3}$. This yields the final posed mesh $\mathcal{M}_p = (\mathbf{V}_p, \mathbf{F})$ with $\mathbf{V}_p = \{\mathbf{v}_i^p\}_{i=1}^N$, where each personalized vertex $\mathbf{v}_i^s$ is transformed to its posed position $\mathbf{v}_i^p$ via the skinning weights and joint transformations.

Motivation for regularization. LBS may introduce mesh distortion (e.g., stretched triangles near joints), which propagates to Gaussian attributes and causes artifacts. This motivates our equilateral regularization (Sec. 4.3) to preserve mesh quality.

4.2. Topology-Aware Gaussian Derivation

The core innovation lies in establishing a *topology-aware* representation where Gaussian geometric attributes are analytically derived from human mesh, ensuring intrinsic coherence during animation. We design two complementary Gaussian types: Face Gaussians for coarse-grained surface coverage, and Vertex Gaussians for fine-detail preservation and junction blending. Both types' attributes are analytically derived from mesh topology, ensuring deformation coherence.

Gaussian parameterization. Each Gaussian is parameterized as $\mathcal{G} = \{\boldsymbol{\mu}, \mathbf{R}, \mathbf{s}, \alpha, \mathbf{c}\}$, where $\boldsymbol{\mu} \in \mathbb{R}^3$, $\mathbf{R} \in SO(3)$, and $\mathbf{s} \in \mathbb{R}_+^3$ are the center, rotation, and scale, while $\alpha \in [0, 1]$ and $\mathbf{c} \in \mathbb{R}^{(L+1)^2\times 3}$ denote opacity and L-order SH coefficients. The covariance is $\boldsymbol{\Sigma} = \mathbf{R}\text{diag}(\mathbf{s}^2)\mathbf{R}^\top$.

Face Gaussians. For each triangular face $f_i \in \mathbf{F}$ with vertices $\{\mathbf{v}_{i,1}^p, \mathbf{v}_{i,2}^p, \mathbf{v}_{i,3}^p\}$, we derive Face Gaussian $\mathcal{G}_f^i$ by analytically computing its geometric attributes from local face geometry.

Position. The center is placed at the incenter, weighted by opposite edge lengths $l_j = \|\mathbf{v}_{i,(j+1) \bmod 3}^p - \mathbf{v}_{i,(j+2) \bmod 3}^p\|$:

$$\boldsymbol{\mu}_f^i = \frac{l_1\mathbf{v}_{i,1}^p + l_2\mathbf{v}_{i,2}^p + l_3\mathbf{v}_{i,3}^p}{l_1 + l_2 + l_3}. \quad (5)$$

Orientation. The rotation $\mathbf{R}_f^i = [\mathbf{t}_1^i, \mathbf{t}_2^i, \mathbf{n}_f^i] \in SO(3)$ aligns with the face's local coordinate frame, constructed from the face normal:

$$\mathbf{n}_f^i = \frac{(\mathbf{v}_{i,2}^p - \mathbf{v}_{i,1}^p) \times (\mathbf{v}_{i,3}^p - \mathbf{v}_{i,1}^p)}{\|(\mathbf{v}_{i,2}^p - \mathbf{v}_{i,1}^p) \times (\mathbf{v}_{i,3}^p - \mathbf{v}_{i,1}^p)\|}, \quad (6)$$

and orthogonal tangent vectors:

$$\mathbf{t}_1^i = \frac{\mathbf{v}_{i,2}^p - \mathbf{v}_{i,1}^p}{\|\mathbf{v}_{i,2}^p - \mathbf{v}_{i,1}^p\|}, \quad \mathbf{t}_2^i = \mathbf{n}_f^i \times \mathbf{t}_1^i. \quad (7)$$

Scale. The in-plane scale is tied to the inradius $r_i = A_i/s_i$ (where A_i is face area and $s_i = (l_1 + l_2 + l_3)/2$ is the semi-perimeter), making the Gaussian adapt to face size:

$$s_{f,x}^i = s_{f,y}^i = \beta \cdot r_i, \quad s_{f,z}^i = \epsilon, \quad (8)$$

with $\beta = 0.5$ and $\epsilon = 10^{-3}$ ensuring flat shapes. The opacity is set to $\alpha_f^i = 1.0$ to ensure complete mesh coverage.

Vertex Gaussians. To capture fine details and ensure seamless blending at face junctions, we introduce Vertex Gaussians $\mathcal{G}_v = \{\mathcal{G}_v^j\}_{j=1}^N$, one per vertex.

Position. The center is anchored directly at the vertex: $\boldsymbol{\mu}_v^j = \mathbf{v}_j^p$.

Orientation. The rotation aligns with the vertex normal, computed by averaging normals of adjacent faces:

$$\mathbf{n}_v^j = \frac{\sum_{f\in\mathcal{N}(j)} A_f\mathbf{n}_f}{\|\sum_{f\in\mathcal{N}(j)} A_f\mathbf{n}_f\|}, \quad (9)$$

where $\mathcal{N}(j)$ denotes adjacent faces. To construct an orthonormal frame $\mathbf{R}_v^j = [\mathbf{t}_1^j, \mathbf{t}_2^j, \mathbf{n}_v^j] \in SO(3)$, we select a reference axis $\mathbf{r} = \arg\min_{\mathbf{u}\in\{\mathbf{e}_x,\mathbf{e}_y,\mathbf{e}_z\}} |\mathbf{u}^\top\mathbf{n}_v^j|$ that is most perpendicular to $\mathbf{n}_v^j$ (maximizing cross-product stability), then compute orthogonal tangents:

$$\mathbf{t}_1^j = \frac{\mathbf{r} \times \mathbf{n}_v^j}{\|\mathbf{r} \times \mathbf{n}_v^j\|}, \quad \mathbf{t}_2^j = \mathbf{n}_v^j \times \mathbf{t}_1^j. \quad (10)$$

Scale. The scale adapts to local mesh density by measuring the minimum edge length to one-ring neighbors:

$$s_{v,x}^j = s_{v,y}^j = \gamma \cdot \min_{\mathbf{v}_k\in\mathcal{N}_1(j)} \|\mathbf{v}_k^p - \mathbf{v}_j^p\|, \quad s_{v,z}^j = \epsilon, \quad (11)$$

where $\mathcal{N}_1(j)$ denotes one-ring neighbors, $\gamma = 0.5$, $\epsilon = 10^{-3}$, and $\alpha_v^j = 1.0$.

Complete Gaussian Representation. Our complete representation consists of $M + N$ Gaussians: $\mathcal{G} = \{\mathcal{G}_f^i\}_{i=1}^M \cup \{\mathcal{G}_v^j\}_{j=1}^N$, where all geometric attributes $(\boldsymbol{\mu}, \mathbf{R}, \mathbf{s})$ are analytically derived from mesh $\mathcal{M}_p$ and only color is optimized. This ensures Gaussian deformations remain coherent with mesh animation, enabling robust pose generalization without per-pose optimization. Unlike standard 3DGS, our method intentionally forgoes adaptive densification to maintain strict correspondence between Gaussians and the underlying mesh, which is crucial for stability in animation.

4.3. Optimization Framework

We optimize the model end-to-end by learning: (**i**) the parameters $\boldsymbol{\psi}$ of the shape encoder E_s and (**ii**) the SH coefficients $\{\mathbf{c}_f^i, \mathbf{c}_v^j\}$ of all Gaussians. Critically, all geometric attributes $(\boldsymbol{\mu}, \mathbf{R}, \mathbf{s})$ are *not optimized* but remain analytically determined by the mesh, enforcing topological coherence.

Photometric and geometric losses. We supervise both RGB and normal map rendering to ensure appearance fidelity and geometric consistency. The photometric loss $\mathcal{L}_{\text{rgb}}$ and normal loss $\mathcal{L}_{\text{normal}}$ are defined as:

$$\mathcal{L}_{\text{rgb}} = \lambda_{\text{SSIM}} \cdot \mathcal{L}_{\text{SSIM}}(\hat{\mathbf{I}}, \mathbf{I}) + (1 - \lambda_{\text{SSIM}}) \|\hat{\mathbf{I}} - \mathbf{I}\|_1, \tag{12}$$

$$\mathcal{L}_{\text{normal}} = \lambda_{\text{SSIM}} \cdot \mathcal{L}_{\text{SSIM}}(\hat{\mathbf{N}}, \mathbf{N}) + (1 - \lambda_{\text{SSIM}}) \|\hat{\mathbf{N}} - \mathbf{N}\|_1, \tag{13}$$

where both losses combine an L_1 term and an SSIM term $\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(\cdot, \cdot)$ with $\lambda_{\text{SSIM}} = 0.2$. For the normal loss, we supervise the rendered normal map $\hat{\mathbf{N}}$ against pseudo-ground-truth normals $\mathbf{N}$ from a pre-trained estimator [12]. The rendered normals are obtained by replacing the SH coefficients with mesh normals during rendering. This pseudo-supervision provides geometric guidance during training, while our geometric evaluation (Sec. 5.2) employs predictor-independent metrics to ensure fairness.

Equilateral regularization. Since Gaussian scales are analytically tied to local mesh (inradius for faces, edge length for vertices), LBS may introduce mesh distortion (e.g., stretched triangles near joints), which propagates to Gaussian attributes and causes artifacts, degenerate triangles cause extreme scale values and rendering instability. To maintain well-conditioned mesh geometry, we regularize the personalized mesh $\mathcal{M}_s$ by penalizing deviations from equilateral geometry:

$$\mathcal{L}_{\text{tri}} = \sum_{f \in \mathbf{F}_s} \left[\text{Var}(\{\|\mathbf{e}_1\|, \|\mathbf{e}_2\|, \|\mathbf{e}_3\|\}) + \sum_{\phi \in \Theta_f} (1 - \cos\phi)^2 \right], \tag{14}$$

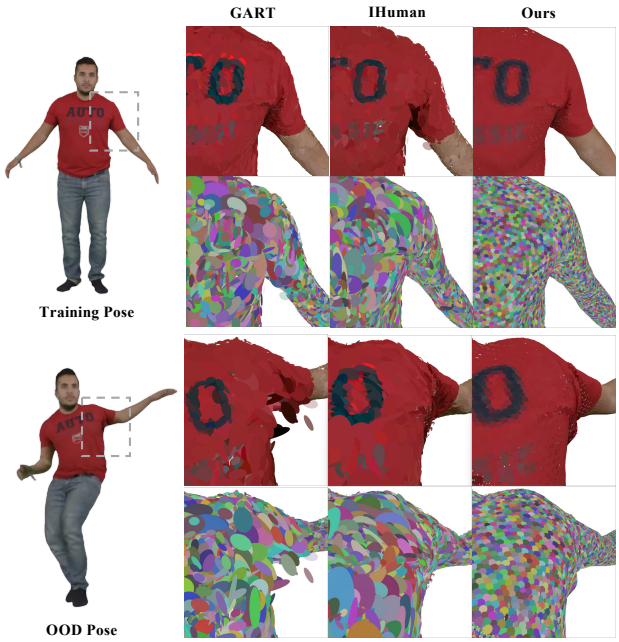

Figure 2. **Gaussian distribution under OOD poses on PeopleSnapshot.** Our topology-aware binding maintains structured, mesh-coherent Gaussian layouts across pose variations, while GART and IHuman exhibit floating Gaussians and geometric artifacts, demonstrating the necessity of analytical attribute derivation for robust animation.

where $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ are the three edges of face f, and $\Theta_f = \{\phi_1^f, \phi_2^f, \phi_3^f\}$ are the interior angles. The equilateral regularization consists of two terms: (1)edge length variance minimizes deviations from uniform edge lengths, and (2)angular term penalizes deviations from 60°, jointly preventing triangular degeneracy that would distort Gaussian scale/rotation derived from mesh geometry.

Mesh smoothness. To prevent high-frequency noise in the personalized mesh, we apply standard geometric regularization:

$$\mathcal{L}_{\text{mesh}} = \mathcal{L}_{\text{lap}} + \mathcal{L}_{\text{nc}}, \tag{15}$$

where $\mathcal{L}_{\text{lap}}$ enforces Laplacian smoothness on vertex positions and $\mathcal{L}_{\text{nc}}$ penalizes normal inconsistency across adjacent faces.

Overall objective. The complete loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rgb}} + \lambda_n \mathcal{L}_{\text{normal}} + \lambda_m \mathcal{L}_{\text{mesh}} + \lambda_t \mathcal{L}_{\text{tri}}, \tag{16}$$

where λ_n, λ_m, and λ_t are hyperparameters that balance the contributions of normal supervision, mesh regularization, and equilateral constraints, respectively. This joint optimization yields avatars with high visual quality, stable geometry, and robust pose generalization.

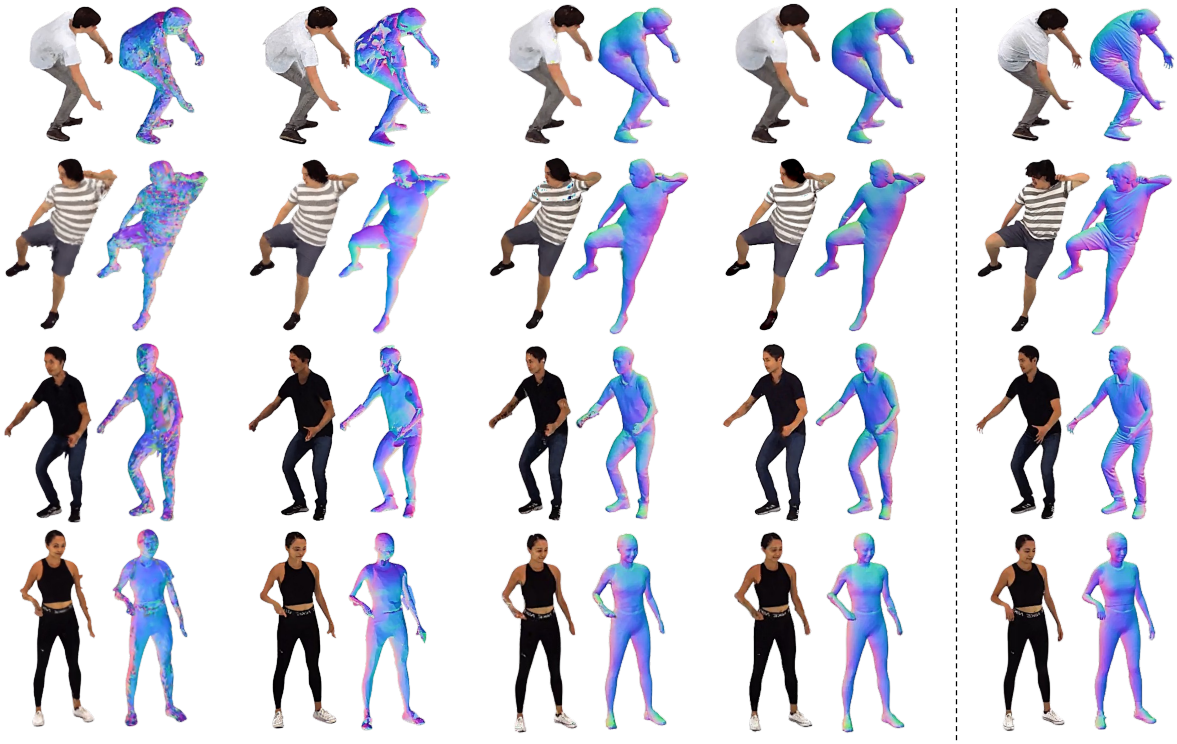

Figure 3. **Qualitative comparison on X-Avatar dataset.** Our method achieves superior RGB rendering and accurate normal maps across challenging OOD poses. Topology-aware Gaussian binding preserves high-fidelity surface details with anatomically plausible deformations, while baselines exhibit geometric artifacts and surface inconsistencies, validating the effectiveness of our method.

Table 1. Quantitative comparison on PeopleSnapshot dataset. The best scores are **bold**, and the second best scores of ours are underlined.

Methods	male-3-casual			male-4-casual			female-3-casual			female-4-casual		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
InstantAvatar	26.73	0.9411	0.0395	24.80	0.9312	0.0648	22.62	0.9381	0.0512	25.33	0.9354	0.0387
GomAvatar	28.27	0.9697	0.0228	25.31	0.9570	0.0388	24.13	0.9528	0.0361	27.15	0.9655	0.0239
IHuman	26.31	0.9362	0.0202	24.32	0.9263	0.0363	21.27	0.9135	0.0544	25.72	0.9358	0.0260
GART	**30.21**	0.9702	0.0359	27.55	0.9590	0.0589	**26.14**	**0.9671**	0.0465	**28.14**	0.9640	0.0352
Ours	28.93	**0.9705**	**0.0168**	**27.56**	**0.9621**	**0.0261**	25.27	0.9611	**0.0323**	27.84	**0.9661**	**0.0224**

5. Experiments

5.1. Experimental Setup

Datasets. We conduct comprehensive evaluations on two challenging datasets to validate both reconstruction quality and generalization capability.

PeopleSnapshot [1] comprises 24 subjects with diverse clothing styles performing simple rotating motions in front of a camera. Following established protocols, we evaluate on four representative subjects (female-3-casual, female-4-casual, male-3-casual, male-4-casual) to assess in-distribution reconstruction quality. The experimental setup follows [7].

X-Avatar [30] features 12 subjects executing complex actions with diverse poses. The dataset includes ground-truth meshes. Critically, training and testing poses exhibit significant distribution gaps, making this dataset ideal for evaluating **out-of-distribution (OOD)** pose generalization—the core strength of our topology-aware approach. We evaluate on four subjects (00016, 00017, 00018, 00019) with challenging motions and varying body shapes.

Evaluation Metrics. We evaluate both photometric quality and geometric accuracy comprehensively. For render-

Table 2. Quantitative comparison on X-Avatar. The best scores are **bold**, and the second best scores of ours are underlined.

Methods	00016			00017			00018			00019		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
IHuman	27.34	0.9560	0.0311	23.75	0.9453	0.0356	25.17	0.9688	0.0259	27.00	0.9742	0.0203
GomAvatar	28.86	0.9594	0.0222	24.87	0.9527	**0.0225**	26.77	0.9743	0.0187	27.86	0.9772	0.0140
GART	28.08	0.9615	0.0278	**25.24**	0.9560	0.0296	25.69	0.9720	0.0235	25.66	0.9692	0.0250
Ours	**29.03**	**0.9705**	**0.0247**	**25.24**	**0.9614**	0.0250	**26.82**	**0.9784**	**0.0174**	**28.11**	**0.9813**	**0.0135**

Table 3. Geometric accuracy comparison on PeopleSnapshot and X-Avatar datasets. Our method significantly outperforms existing approaches in geometry quality. The best scores are **bold**.

Methods	PeopleSnapshot	X-Avatar		
	Normal↓	Normal↓	CD↓	P2S↓
GART	4.263	3.531	0.141	0.154
GomAvatar	3.876	3.344	0.155	0.169
IHuman	2.055	2.056	0.132	0.126
Ours	**1.687**	**1.772**	**0.111**	**0.107**

ing quality, we report **PSNR**, **SSIM** [33], and **LPIPS** [38]. For geometric fidelity, we measure surface normal error(**Normal** [36]) on both datasets using pseudo-ground-truth normals predicted by Sapiens [12], which provides consistent evaluation across all methods. Additionally, on X-Avatar (which provides ground-truth meshes), we compute mesh-based metrics including Chamfer Distance (**CD**) and Point-to-Surface Distance (**P2S**) [36] between reconstructed and ground-truth mesh geometries, offering predictor-independent geometric validation. Normal errors are scaled by 10^3 for clarity. These metrics jointly assess our method's capability to maintain photorealistic appearance while preserving accurate geometric structure—essential for robust animation.

Implementation Details. We implement our framework in PyTorch and train on a single NVIDIA RTX-3090 GPU for 2,000 iterations per subject using the Adam optimizer (learning rate: 10^{-3}). The shape encoder E_s employs multi-resolution hash encoding [21] with 16 levels. Loss weights are set as: $\lambda_{\text{SSIM}} = 0.2$, $\lambda_n = 0.05$, $\lambda_m = 0.01$, $\lambda_t = 0.01$. All Gaussians have fixed opacity $\alpha = 1.0$ with learnable second-order spherical harmonics for color. The out-of-plane scale factor $\epsilon = 10^{-3}$ enforces disk-like shapes, while in-plane scales use $\beta = \gamma = 0.5$ for Face and Vertex Gaussians. Ground-truth normals are extracted using Sapiens [12].

Baselines. We compare against state-of-the-art methods for monocular video-based avatar reconstruction: three 3DGS-based approaches—**GART** [14], **IHuman** [23], **GomAvatar** [34]—and one NeRF-based method—**InstantAvatar** [7]. All baselines reconstruct articulated avatars from single RGB videos for novel view and pose synthesis. Unlike these methods with free-floating or partially constrained Gaussians, our method achieves superior geometric fidelity and OOD pose robustness.

5.2. Comparison with State-of-the-Art

We conduct comprehensive comparisons on PeopleSnapshot and X-Avatar datasets, evaluating both photometric quality and geometric accuracy.

Geometric Accuracy. Tab. 3 demonstrates that our method achieves superior geometric fidelity across all metrics (Normal, CD, P2S). Compared to the best baseline, we reduce normal error by **13.8%** on X-Avatar and **17.9%** on PeopleSnapshot, directly validating our core hypothesis: *analytically deriving Gaussian attributes from mesh topology ensures geometric coherence during animation*. We note that all compared methods use identical Sapiens-predicted normals for supervision during training to ensure fair comparison. Fig. 2 provides visual evidence—while GART and IHuman exhibit floating Gaussians and surface holes under OOD poses, our Gaussians maintain structured, topology-aware distributions that strictly follow mesh deformation. The normal map comparisons in Fig. 3 further confirm accurate surface orientation recovery, demonstrating the practical effectiveness of our geometry-driven paradigm for robust avatar animation.

Rendering Quality. Tab. 1 and 2 show that our method achieves state-of-the-art photometric quality on the challenging X-Avatar dataset and competitive performance on PeopleSnapshot. Notably, while GART slightly outperforms on PeopleSnapshot's simple rotations (where free-floating Gaussians can overfit easily), our method excels on X-Avatar's OOD poses, demonstrating the critical advantage of topology-aware constraints for generalization. This trade-off validates our design principle: *sacrificing minor fitting flexibility for substantial robustness gains*. Fig. 3 presents qualitative comparisons on X-Avatar, highlighting

Table 4. Quantitative ablation result on X-Avatar subject 00019.

Settings	PSNR↑	SSIM↑	LPIPS↓	Normal↓	CD↓
w/o FG	26.89	0.9721	0.0189	2.143	0.128
w/o VG	25.67	0.9654	0.0231	2.687	0.145
w/o ER	27.83	0.9798	0.0142	1.834	0.115
Full	**28.11**	**0.9813**	**0.0135**	**1.772**	**0.111**

our superior RGB rendering and surface detail preservation across challenging poses. Our topology-aware binding maintains anatomically plausible deformations without visual artifacts, while baselines suffer from surface inconsistencies and geometric breakdown during animation.

5.3. Ablation Study

We validate each component of our topology-aware framework through systematic ablations on X-Avatar subject 00019, removing Face Gaussians (**FG**), Vertex Gaussians (**VG**), and Equilateral Regularization (**ER**). Tab. 4 and Fig. 4 reveal the necessity of each design choice.

Component Analysis. We conduct qualitative analysis to dissect the contribution of each key component, with visual results presented in Fig. 4. **(w/o FG)** When removing Face Gaussians, we make the scale of Vertex Gaussians learnable (akin to IHuman) to compensate. Nevertheless, the model produces a sparse representation with significant rendering gaps, demonstrating the necessity of our analytically-derived Face Gaussians for establishing coarse geometric integrity. **(w/o VG)** Conversely, ablating Vertex Gaussians, which target fine-grained features, leads to a loss of surface detail and visible discontinuities at mesh junctions. This highlights their complementary role in representing intricate geometry. **(w/o ER)** Disabling the Equilateral Regularization leads to a degraded mesh parameterization. This breakdown disrupts our analytical mapping, causing chaotic and misaligned Gaussian distributions. This outcome underscores a core principle of our method: the stability of the analytical attribute derivation is directly dependent on the quality of the underlying mesh geometry.

Quantitative Validation. Tab. 4 quantitatively confirms our design's efficacy. Combining Face and Vertex Gaussians (FG and VG) substantially improves all metrics (PSNR: 27.83, Normal: 1.834), demonstrating their synergistic contribution to accuracy. Adding Equilateral Regularization (ER) provides further significant gains by stabilizing the analytical mapping from a well-formed mesh. These results validate our dual-primitive design and the critical role of mesh quality preservation for robust topology-aware binding. More visualization results, experiment setup and comprehensive performance evaluations are provided in the supplementary material.

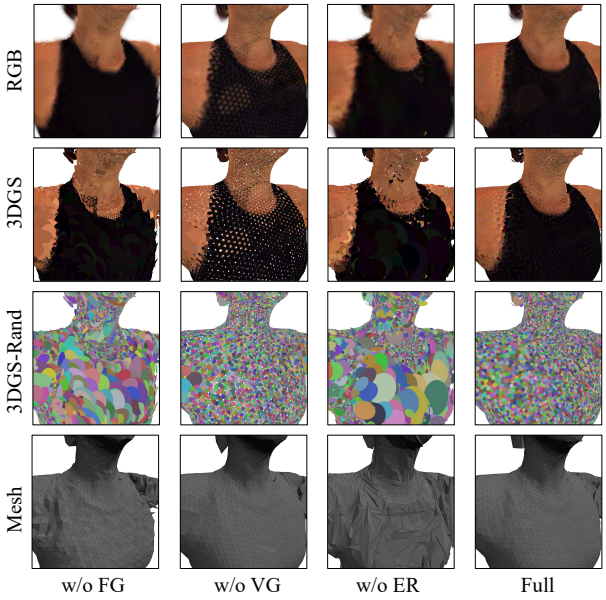

Figure 4. Qualitative ablation result on X-Avatar subject 00019.

6. Conclusion

We introduced Tavatar, a paradigm-shifting approach to animatable human avatar creation that imposes geometric structure onto the traditionally unstructured 3D Gaussian Splatting representation. By analytically deriving Gaussian attributes from an underlying mesh topology, we move beyond simple point-based optimization and establish an explicit, coherent link between geometry and appearance. This core innovation not only guarantees topological consistency but also unlocks robust, high-fidelity animation under novel poses where prior methods often degrade. Our geometry-driven methodology, further stabilized by an equilateral regularization, sets a new state-of-the-art in both geometric accuracy and animation quality. More importantly, it redefines the pathway toward creating physically-grounded and reliable digital humans, bridging the gap between explicit mesh-based representations and the expressive power of implicit rendering.

Limitations and Future Work. The performance of Tavatar, like other state-of-the-art methods in this domain, is inherently linked to the fidelity of the underlying parametric body model. As such, its accuracy is coupled with the quality of the initial body model fitting. This shared dependency highlights a broader opportunity for community-wide advancements in pose and shape estimation to benefit all avatar creation techniques. Future work will extend Tavatar to dynamic clothing by integrating physics-based garment simulation, leveraging our explicit mesh connectivity for enhanced realism.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (U23B2013 and 62276176).

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8387–8397, 2018. 6

[2] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 311–329. Springer, 2020. 2

[3] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. *Advances in Neural Information Processing Systems*, 33:12909–12922, 2020. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, pages 561–578. Springer, 2016. 2

[5] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang 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 (CVPR)*, pages 634–644, 2024. 1, 2

[6] Zixiong Huang, Qi Chen, Libo Sun, Yifan Yang, Naizhou Wang, Qi Wu, and Mingkui Tan. G-nerf: Geometry-enhanced novel view synthesis from single-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10117–10126, 2024. 2

[7] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16922–16932, 2023. 1, 2, 6, 7

[8] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision (ECCV)*, pages 402–418. Springer, 2022. 1

[9] Yuheng Jiang, Kaixin Yao, Zhuo Su, Zhehao Shen, Haimin Luo, and Lan Xu. Instant-nvr: Instant neural volumetric rendering for human-object interactions from monocular rgbd stream. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 595–605, 2023. 2

[10] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 2

[11] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1, 2, 3

[12] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision (ECCV)*, pages 206–228. Springer, 2024. 5, 7

[13] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4501–4510, 2019. 2

[14] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19876–19887, 2024. 1, 2, 7

[15] Haiyi Li, Qi Chen, Denis Kalkofen, and Hsiang-Ting Chen. Ougs: Active view selection via object-aware uncertainty estimation in 3dgs. *Computer Graphics Forum (CGF)*, 2025. 2

[16] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12939–12948, 2021. 2

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 1, 2, 3, 4

[18] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[19] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, pages 752–768. Springer, 2020. 2

[20] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *ECCV*, 2024. 2

[21] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 4, 7

[22] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 International Conference on 3D Vision (3DV)*, pages 484–494. IEEE, 2018. 2

[23] Pramish Paudel, Anubhav Khanal, Danda Pani Paudel, Jyoti Tandukar, and Ajad Chhatkuli. ihuman: Instant animatable digital humans from monocular videos. In *European Conference on Computer Vision (ECCV)*, pages 304–323. Springer, 2024. 1, 2, 7

[24] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2018. 2

[25] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2

[26] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14314–14323, 2021. 1

[27] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei 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 (CVPR)*, pages 9054–9063, 2021. 2

[28] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20299–20309, 2024. 2

[29] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu 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 (CVPR)*, pages 1606–1616, 2024. 2

[30] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16911–16921, 2023. 6

[31] Geonhee Sim and Gyeongsik Moon. PERSONA: Personalized whole-body 3D avatar with pose-driven deformations from a single image. In *ICCV*, 2025. 2

[32] Joanna Waczyńska, Piotr Borycki, Sławomir Tadeja, Jacek Tabor, and Przemysław Spurek. Games: Mesh-based adapting and modification of gaussian splatting. *arXiv preprint arXiv:2402.01459*, 2024. 2

[33] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[34] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G Schwing, and Shenlong 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 (CVPR)*, pages 2059–2069, 2024. 1, 2, 7

[35] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira 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 (CVPR)*, pages 16210–16220, 2022. 2

[36] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, 2022. 7

[37] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16943–16953, 2023. 2

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 7