F3G-Avatar : Face Focused Full-body Gaussian Avatar

Willem Menu Erkut Akdag Pedro Quesado Yasaman Kashefbahrami Egor Bondarev
AIMS Group, Department of Electrical Engineering, Eindhoven University of Technology
{ w.j.menu, e.akdag, p.quesado.dos.santos, y.kashefbahrami, e.bondarev}@tue.nl

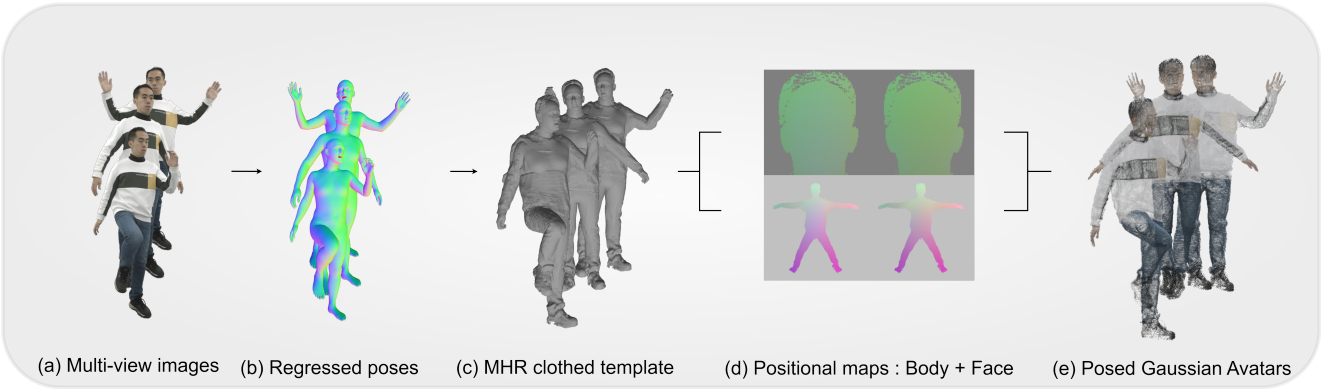

Figure 1. Framework of F3G-Avatar. Multi-view images and regressed poses are used to generate an MHR clothed template, which is encoded into body and face positional maps and subsequently rendered as posed Gaussian avatars.

Abstract

Existing full-body Gaussian avatar methods primarily optimize global reconstruction quality and often fail to preserve fine-grained facial geometry and expression details. This challenge arises from limited facial representational capacity that causes difficulties in modeling high-frequency pose-dependent deformations. To address this, we propose F3G-Avatar, a full-body, face-aware avatar synthesis method that reconstructs animatable human representations from multi-view RGB video and regressed pose/shape parameters. Starting from a clothed Momentum Human Rig (MHR) template, front/back positional maps are rendered and decoded into 3D Gaussians through a two-branch architecture: a body branch that captures pose-dependent non-rigid deformations and a face-focused deformation branch that refines head geometry and appearance. The predicted Gaussians are fused, posed with Linear Blend Skinning (LBS), and rendered with differentiable Gaussian splatting. Training combines reconstruction and perceptual objectives with a face-specific adversarial loss to enhance realism in close-up views. Experiments demonstrate strong rendering quality, with face-view performance reaching PSNR/SSIM/LPIPS of 26.243/0.964/0.084 on the Avatar-ReX dataset. Ablations further highlight contributions of the MHR template and the face-focused deformation. F3G-Avatar provides a practical, high-quality pipeline for realistic, animatable full-body avatar synthesis. The code is available at https://github.com/wjmenu/F3G-avatar.

1. Introduction

Photorealistic, animatable human avatars are the key enabling technology for telepresence, virtual/augmented reality, digital entertainment, and human-computer interaction. The central goal is to capture both the visual appearance and geometric structure of a person in a representation that can be efficiently rendered from novel viewpoints and driven by motion.

Parametric human body models, most notably the Skinned Multi-Person Linear model (SMPL) [17] and related variants [14, 24], have become a standard representation for human avatar modeling. They enable recovery of shape, pose, and expression from images or videos through a low-dimensional parameterization of a deformable mesh. SMPL models are animated by adjusting shape and pose

parameters and applying Linear Blend Skinning (LBS) [2] to obtain the posed mesh. Many approaches extend these models by incorporating displacement fields to represent clothing, but still struggle with complex geometry and high-frequency detail (e.g., loose garments or fine hair), due to limited topology and texture resolution.

Implicit approaches [5, 13, 25, 39], particularly Neural Radiance Fields (NeRFs) [18], model humans as pose-conditioned neural fields learned from RGB videos. However, these methods typically depend on coordinate-based MLPs that are known to suffer from a low-frequency bias. As a result, NeRFs struggle to accurately capture high-frequency details, even when enhanced with learned feature grids or local conditioning. More recently, 3D Gaussian Splatting (3DGS) [10] has emerged as an efficient explicit alternative, delivering high-quality rendering while significantly improving both the visual quality and rendering speed of prior approaches.

The explicit point-based nature of 3DGS further enables parameterizing appearance and deformation in 2D spaces derived from body template models. This allows the use of powerful 2D backbones for better human avatar modeling. Existing approaches exploit this property by: (i) predicting pose-dependent deformation maps from orthographic front/back projections of a canonical body template [16, 31], or (ii) using a 2D parameterization of the underlying human mesh surface in UV space [6, 7]. In both cases, posed 2D maps are processed with 2D CNNs to predict canonical-space deformations and Gaussian attributes. The obtained Gaussians are then posed via Linear Blend Skinning (LBS) and then visualized by a Gaussian renderer. During training, the Gaussians are optimized to minimize image-based reconstruction losses between the rendered outputs and the corresponding ground-truth camera observations.

Despite achieving strong quantitative performance, these methods are primarily optimized for global, full-body reconstruction and may under-represent important regions that require fine-grained detail. This limitation is most prominent in the face volume, as it occupies only a small fraction of the full-body area. Existing methods tend to allocate insufficient capacity to facial geometry and appearance, leading to oversmoothed features and loss of fine-grained expression detail. However, facial cues play an important role in human perception of identity and realism. When key facial cues are missing or distorted, it results in significant perceptual degradation, often associated with the uncanny-valley response [21].

This observation motivates the proposed **F3G-Avatar**, a full-body avatar synthesis method that extends the conventional techniques with a dedicated face-focused deformation network. Specifically, a separate set of canonical Gaussians is generated for the head and the process is driven by additional orthographic front/back projection maps. These maps define a 2D parameter space, where a face-specific deformation network, implemented by StyleUNets [9], learns high-resolution, pose-dependent Gaussians deformation maps.

To further improve the capture of subtle facial expressions, F3G-Avatar adopts the Momentum Human Rig (MHR) parametric body model [3]. Compared to commonly used SMPL-based models, MHR provides more accurate facial articulation due to high-resolution training data and a sparse, non-linear pose corrective formulation. This leads to improved preservation of local detail and reduces the overly-smoothed or globally entangled deformations observed in conventional parametric models. Furthermore, the coarse garment geometry is modeled on top of the MHR body, enabling consistent deformation of the 3D Gaussians while maintaining alignment during body movements. This yields a clothed parametric template that retains fine-grained control over both facial expressions and body motion. In summary, F3G-Avatar makes the following contributions:

- A face-focused canonical deformation network operates alongside the body deformation branch that improves the reconstruction of facial geometry and appearance. The face-focused deformation network independently predicts a set of 3D Gaussians that are concatenated with the Gaussians obtained by the body deformation network.
- Integration of the clothed MHR body template into the 3DGS-based avatar method, leading to more accurate reconstruction of facial geometry and expressions. To the best of our knowledge, this is the first implementation of the MHR body model in the context of full-body Gaussian avatar reconstruction.
- Comprehensive experimentation that achieves strong performance on AvatarReX and THuman4.0 datasets, with face-view PSNR of 26.243/26.934, SSIM of 0.964/0.961, and LPIPS of 0.084/0.062.

2. Related Work

2.1. Parametric Human Body Models

Conventional human avatar pipelines [14, 17, 24] commonly rely on parametric body models, such as SMPL [17] or SMPL-X [24], which provide representation of human shape and pose through Linear Blend Skinning. The models offer strong priors for articulation, and are widely used for animation, pose estimation, and supervision. However, the fixed topology and limited texture resolution constrain the ability to represent complex geometry, such as loose clothing, fine hair, or subtle view-dependent appearance. As a result, many works augment the models with learned displacement or appearance fields, yet capturing high-frequency detail remains challenging.

2.2. Implicit Neural Human Representations

Implicit approaches [5, 13, 18, 25, 39] address some of the limitations, by modeling humans as continuous neural fields conditioned on pose. In particular, Neural Radiance Fields(NeRFs) [18] and animatable extensions learn view-dependent appearance directly from multi-view RGB data. While providing flexibility beyond mesh-based representations, such methods typically rely on coordinate-based MLPs that exhibit a low-frequency bias, limiting the ability to reconstruct fine details. Moreover, volumetric rendering introduces substantial computational overhead, making real-time or high-resolution applications challenging.

2.3. 3D Gaussian-Based Human Avatars

Recent advances have shifted toward explicit point-based representations, particularly 3D Gaussian Splatting (3DGS) [10], which enables efficient rendering with high visual quality. This representation allows modeling deformation and appearance in parameterized 2D space. This 2D parameterization facilitates the use of powerful 2D backbones for predicting pose-dependent Gaussian attributes. A range of approaches build upon this formulation. Animatable Gaussians [16], predicts pose-conditioned Gaussian maps from orthographic front/back projections. GaussianAvatar [6] and 3DGS-Avatar [28] demonstrate high-quality animatable avatars from monocular or multi-view inputs. SplattingAvatar [29] stabilizes deformation by embedding Gaussians within a mesh structure. UV-space formulations [7] exploit surface parameterizations to improve learning stability. Extensions, such as Human Gaussian Splatting [20] and HUGS [12], adapt 3DGS to animatable human modeling under multi-view and monocular settings, while generalizable approaches like HumanSplat [23] target single-image reconstruction. Despite these advances, existing methods predominantly optimize for full-body reconstruction quality and tend to distribute model capacity uniformly across the regions of the body. As a result, small yet perceptually critical areas (most notably the face) are often left underrepresented, leading to limited detail and diminished photorealism.

2.4. Expressive and Perceptual Avatar Modeling

Head-centered methods allocate model capacity entirely to the face and have consistently advanced facial reconstruction quality. Early approaches combine dynamic NeRFs with morphable face models to enable controllable synthesis and efficient reconstruction. Point-based methods further capture the fine-grained geometric detail through deformable representations [26, 37], while more recent works integrate 3D Gaussians with parametric face models to achieve precise expression control and high-quality sharp rendering [4, 27, 35, 38, 41]. Collectively, these studies demonstrate that spatially focused modeling improves facial detail. In contrast, full-body methods, such as AvatarRex [40], X-Avatars [30], and Expressive Human Avatars [19] incorporate expression modeling, but lack a dedicated face-focused deformation mechanism, limiting the ability to fully capture fine-grained facial detail. Perceptual studies indicate that facial appearance plays a dominant role in human judgment of realism [31]. This suggests that full-body systems can benefit from allocating disproportionate capacity to facial detail.

Motivated by this observation, we introduce a face-focused deformation network alongside the main body deformation network. The face-focused deformation network allows for higher-resolution conditioning and specialized modeling of the facial Gaussians while remaining compatible with full-body rendering. The design reflects an emerging direction toward hybrid representations that combine the efficiency of explicit point-based rendering with region-specific targeting.

3. Method

3.1. 3D Gaussian Splatting Preliminaries

3D Gaussian Splatting (3DGS) [10] represents a scene as a finite set of anisotropic 3D Gaussian primitives

$$\mathcal{G} = \{G_i\}_{i=1}^{N}. \tag{1}$$

Each primitive $G_i \in \mathcal{G}$ is parameterized as

$$G_i = (\mathbf{x}_i, \mathbf{\Sigma}_i, \alpha_i, \mathbf{f}_i), \tag{2}$$

where $\mathbf{x}_i \in \mathbb{R}^3$ denotes the 3D mean, $\mathbf{\Sigma}_i \in \mathbb{R}^{3\times3}$ the covariance matrix, $\alpha_i \in [0, 1]$ the opacity, and $\mathbf{f}_i$ the spherical-harmonics coefficients encoding view-dependent color. A 3D Gaussian distribution is defined using the squared Mahalanobis distance

$$d_i^2(\mathbf{p}) = (\mathbf{p} - \mathbf{x}_i)^\top \mathbf{\Sigma}_i^{-1}(\mathbf{p} - \mathbf{x}_i), \tag{3}$$

such that its density is

$$\mathcal{N}(\mathbf{p} \mid \mathbf{x}_i, \mathbf{\Sigma}_i) = \frac{1}{(2\pi)^{3/2}|\mathbf{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2} d_i^2(\mathbf{p})\right), \tag{4}$$

where $\mathbf{p} \in \mathbb{R}^3$ and $|\mathbf{\Sigma}_i|$ denotes the determinant of the covariance matrix.

To guarantee that $\mathbf{\Sigma}_i$ is symmetric positive semi-definite, it is parameterized via a scale vector $\mathbf{s}_i \in \mathbb{R}^3$ and a unit quaternion $\mathbf{q}_i$:

$$\mathbf{\Sigma}_i = \mathbf{R}(\mathbf{q}_i)\,\mathrm{diag}(\mathbf{s}_i^2)\mathbf{R}(\mathbf{q}_i)^\top, \tag{5}$$

where $\mathbf{R}(\mathbf{q}_i) \in SO(3)$ is the rotation matrix corresponding to $\mathbf{q}_i$.

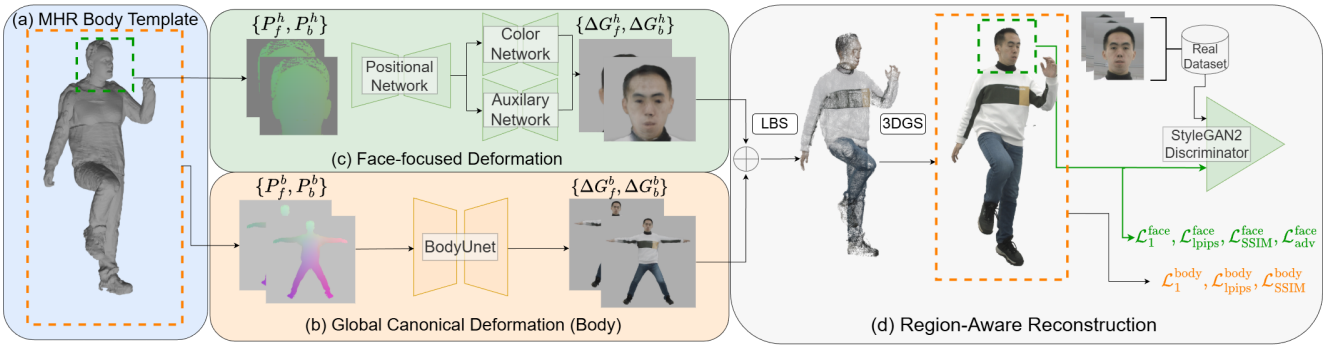

Figure 2. Overview of F3G-Avatar. (a) MHR clothed body template. (b) Global Canonical Deformation (Body): front/back body positional maps are processed by the BodyUNet to predict pose-dependent body Gaussians. (c) Face-focused Deformation: head positional maps drive three StyleUNets to predict positional, color, and auxiliary face attributes. (d) Region-aware Reconstruction : The two branches are fused, posed via LBS, rendered with 3DGS, and optimized with reconstruction losses and a face-specific adversarial loss.

Given a camera transformation and the Jacobian $\mathbf{J}$ of the projective mapping evaluated at $\mathbf{x}_i$, the covariance is projected into screen space as

$$\mathbf{\Sigma}_i' = \mathbf{J}\,\mathbf{\Sigma}_i\,\mathbf{J}^\top. \tag{6}$$

For rasterization, only the upper-left 2×2 block

$$\mathbf{\Sigma}_i'^{(2D)} = \mathbf{\Sigma}_{i,1:2,1:2}' \tag{7}$$

(i.e., the first two rows and columns) is used to define the elliptical footprint in the image plane.

Pixel colors are obtained via front-to-back alpha compositing of depth-sorted Gaussians,

$$C = \sum_{i=1}^{N} \left(\alpha_i' \prod_{j=1}^{i-1} (1 - \alpha_j') \right) c_i, \tag{8}$$

where α_i' is the effective opacity at the pixel after evaluating the projected 2D Gaussian and c_i is the view-dependent color obtained from $\mathbf{f}_i$. The parameters of $\mathcal{G}$ are optimized using image-based reconstruction losses, while the number of Gaussians is dynamically adapted through periodic densification and pruning.

3.2. Overview

Figure 2 illustrates the proposed F3G-Avatar method. Given multi-view RGB videos of a subject and the regressed pose and shape parameters, F3G-Avatar reconstructs a realistic representation of both the body and the face. The process starts from a clothed MHR body template, from which the front and back orthographic positional maps are rendered separately for the body and head regions. Next, F3G-Avatar is split into two branches: global canonical full-body deformation and face-focused deformation. In the global canonical deformation branch, BodyUNet [32] predicts pose-dependent Gaussian attribute maps in canonical space from the body positional inputs. In parallel, the face-focused deformation branch processes head-specific positional maps using three lightweight StyleGAN-based networks [9]. The parallel branches predict a set of high-resolution, pose-dependent facial Gaussian maps. The predicted maps define canonical Gaussian primitives for both body and face, which are subsequently deformed and articulated via Linear Blend Skinning (LBS) [2].

Finally, the model is trained with region-aware reconstruction. Full-body reconstruction losses provide global consistency, while additional face-specific perceptual and adversarial losses enhance fine-grained facial detail and realism.

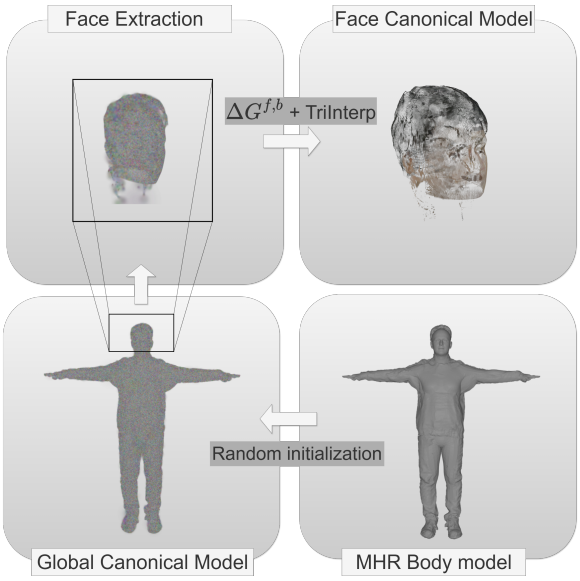

Figure 3. Visualization of the canonical face model construction from the MHR template and head positional maps.

3.3. MHR Body Template

The MHR body template is adopted as the foundation of the representation. Since most multi-view datasets provide SMPL-X parameters (regressed pose and shape parameters), conversion to the MHR representation is required. This conversion is performed by optimizing

$$\min_{\beta_{\text{mhr}},\,\theta_{\text{mhr}}} \|V_{\text{mhr}} - V_{\text{smplx}}\| + \lambda \|J_{\text{mhr}} - J_{\text{smplx}}\| \quad (9)$$

where β_{mhr} and θ_{mhr} denote the MHR shape and pose parameters, V_{mhr} and V_{smplx} represent the MHR and SMPL-X template vertices, and J_{mhr} and J_{smplx} refer to the corresponding joint locations. The reconstruction process starts from a subset of frames in which the subject is captured in a near star-like body pose (A-pose), providing maximal surface visibility across views. From these images, the full clothed-body geometry is reconstructed via implicit surface reconstruction methods [15, 22, 34, 36], for which the NeuS2 [34] is employed. For separation of non-body components (e.g., clothing and accessories), a SAM-based segmentation model [11] is applied to the input images. The segmented regions are subsequently projected and attributed onto the reconstructed body mesh using 4D-Dress [33]. To ensure consistent deformation of these non-body components, Robust Skinning Transfer [1] is applied to estimate their skinning weights. Finally, the posed MHR model is merged with the segmented non-body components to obtain the complete MHR body template.

The resulting body template is populated with 3D Gaussians, where the positions are based on the vertices of the MHR model in canonical A-pose. The other Gaussian attributes are initialized with informed random values. The canonical 3D Gaussian model is transformed into posed space through LBS. For a canonical Gaussian with position $\mathbf{p}_c$ and covariance $\mathbf{\Sigma}_c$, the transformation is given by

$$\mathbf{p}_p = \mathbf{R}\mathbf{p}_c + \mathbf{t}, \quad \mathbf{\Sigma}_p = \mathbf{R}\,\mathbf{\Sigma}_c\,\mathbf{R}^\top, \quad (10)$$

where $\mathbf{R}$ and $\mathbf{t}$ are the rotation and translation obtained from the Gaussian's skinning weights, and $\mathbf{p}_p$ and $\mathbf{\Sigma}_p$ are the Gaussian position and covariance in posed space.

3.4. Global Canonical Deformation for Body

A large StyleUNet, $\mathcal{T}(\cdot)$, is employed to capture pose-dependent, non-rigid Gaussian deformations in 2D parameter space. Given the posed MHR templates, front and back position maps $\{P_f^b, P_b^b\}$ are orthographically rendered at a resolution of 1024×1024. Each pixel in the maps corresponds to a single 3D Gaussian with position, covariance, opacity, and color. The maps, together with the camera parameters (K, R, t), are fed into the StyleUNet to predict non-rigid deformation maps $\{\Delta G_f^b, \Delta G_b^b\}$. The predicted deformation maps are added to each Gaussian's canonical attributes and then transformed to world space.

In the pretraining stage, StyleUNet is conditioned to reconstruct the input positional maps, while the remaining Gaussian attributes are supervised to match the canonical model. In the subsequent training stage, BodyUNet takes the position maps as input and predicts residual Gaussian attributes that deform the canonical representation. The deformed canonical model is then posed via LBS and rendered.

3.5. Face-Focused Deformation

3.5.1. Canonical Face Model

The canonical face model is initialized by extracting the head region from the pretrained BodyUNet template. After this, Gaussian attributes are estimated by transforming each head positional map into the canonical frame and averaging across the dataset, which is defined as

$$\mathcal{G}_{i,j}^{v} = \frac{1}{N}\sum_{k=1}^{N} \mathcal{T}\big(P_{v,k}^{h},\, K_k,\, R_k,\, t_k\big)_{i,j}\,, \quad v \in \{f, b\}. \quad (11)$$

Here, $\mathcal{G}_{i,j}^{v}$ denotes the Gaussian attributes of the head at spatial location (i, j), v indicates if the front or back map is used, $P_{v,k}^{h}$ represents the k-th head positional map in the dataset, and (K_k, R_k, t_k) define the camera calibration parameters for frame k.

To increase the face detail, we employ high-resolution positional maps zoomed in on a face. To accommodate the higher spatial resolution of the face positional maps, the canonical Gaussian grid is densified via trilinear interpolation. Formally,

$$\hat{\mathcal{G}}^{v} = \text{TriInterp}(\mathcal{G}^{v})\,, \quad (12)$$

where $\hat{\mathcal{G}}^{v}$ denotes the upsampled canonical face Gaussian representation. Figure 3 depicts the canonical face model construction. Following the pretraining strategy in 3.4, the face-focused deformation is conditioned on the head region of the pretrained body model.

3.5.2. Positional Face Maps

For face-focused modeling, it is essential to know the precise location of the head within the full-body input images. First, localized crops centered on the face region are extracted. To capture fine-grained facial details, 512×512 crops centered on the face region are extracted. After the crop-and-resize operation, the camera intrinsics must be updated accordingly by

$$f_x' = s\,f_x, \quad f_y' = s\,f_y, \quad (13)$$

$$c_x' = s\,(c_x - x_c), \quad c_y' = s\,(c_y - y_c), \quad (14)$$

which yields

$$\mathbf{K}_{\text{new}} = \begin{bmatrix} sf_x & 0 & s(c_x - x_c) \\ 0 & sf_y & s(c_y - y_c) \\ 0 & 0 & 1 \end{bmatrix}. \quad (15)$$

Here, (x_c, y_c) denotes the top-left corner of the crop, (f_x, f_y, c_x, c_y) are the original intrinsics, and (f'_x, f'_y, c'_x, c'_y) are the updated intrinsics after crop-and-resize.
With the updated K_{new}, the MHR posed positional maps are generated using orthographic rendering, resulting in front and back face maps $\{P_f^h, P_b^h\}$.

3.5.3. Face-focused Gaussian Maps

To generate Gaussian maps for the face, we employ three lightweight StyleUNets: Positional ($\mathcal{P}(\cdot)$), Color ($\mathcal{C}(\cdot)$), and Auxiliary ($\mathcal{A}(\cdot)$), which predict positional, color, and auxiliary Gaussian attributes from the head positional maps $\{P_f^h, P_b^h\}$. The positional Gaussian deformation is obtained as

$$\hat{P}_v^h = \mathcal{P}\left(P_v^h\right), \quad v \in \{f, b\}, \tag{16}$$

where $\hat{P}_v^h$ represents the deformed positional map predicted by the positional StyleUNet $\mathcal{P}$. The corresponding color and auxiliary Gaussian attributes are computed using $\mathcal{C}(\cdot)$ and $\mathcal{A}(\cdot)$, respectively. These components are then combined to form the residual Gaussian attribute map:

$$\Delta G_v^h = \mathcal{C}\left(\hat{P}_v^h, K_{\text{new}}, R, t\right) \parallel \mathcal{A}\left(\hat{P}_v^h\right) \parallel P_v^h, \quad v \in \{f, b\}. \tag{17}$$

Here, ΔG_v^h denotes the residual Gaussian attribute map for view v, corresponding to either the front (f) or back (b) of the head.

3.6. Region-Aware Reconstruction

The predictions are fused with the canonical head Gaussians from Section 3.5.1 and combined with the body Gaussian after Global Canonical Deformation. After combination, the Gaussians are transformed into posed space through LBS. The posed Gaussians are rendered to the image domain using 3D Gaussian splatting (3DGS). To enhance facial detail, a pretrained StyleGAN2 [8] discriminator is applied to rendered face crops. The discriminator provides a non-saturating adversarial loss, $\mathcal{L}_{adv}$, that is used in addition to reconstruction and perceptual losses.

4. Experiments

4.1. Evaluation Datasets

AvatarReX. The AvatarReX [40] dataset (Real-time Expressive Full-body Avatars) consists of four multi-view human performance sequences captured using 16 synchronized and calibrated RGB cameras arranged in a circular configuration. Each camera records at a resolution of 1500×2048 and 30 fps. For each frame, fitted SMPL-X parameters are provided, supplying pose, shape, and expression estimates.

THuman4.0. Similar to AvatarReX, THuman4.0 [39] provides dense multi-view supervision for animatable human reconstruction. It contains three synchronized sequences captured with 24 calibrated RGB cameras at 30 fps and a resolution of 1330×1150. The dataset includes per-frame SMPL-X registrations.

4.2. Implementation Details

Canonical template and Gaussian initialization. For each subject, the provided per-frame SMPL-X registrations are used to build a clothed MHR template in a canonical A-pose. From the canonical template, front/back concatenated position maps are rendered at a resolution of 1024×2048. The canonical model contains 320k body Gaussians and 60k face Gaussians. The initial centers come from the A-pose position map, with isotropic scales and colors sampled from a uniform distribution.
Global and Face-focused Deformation architecture. The Global Canonical Deformation employs a StyleUNet-based [9] generator to map canonical position maps to Gaussian attributes. The backbone processes a 512×512 canonical map and predicts 1024×1024 maps for color, position offsets, and additional Gaussian attributes. For face-focused modeling, a lightweight StyleUNet operates on 256×256 face crops to predict head-specific corrections, which are subsequently fused into the global Gaussian representation.
Optimization. At each iteration, a single frame-view is rendered and supervised with RGB and mask. The Global Deformation network is trained on full images, while the Face-focused Deformation network uses cropped head views with updated intrinsics. The total loss is a weighted sum of ℓ_1, LPIPS, and offset regularization term with coefficients $\lambda_{\ell_1} = 1.0$, $\lambda_{\text{LPIPS}} = 0.1$, $\lambda_{\text{off}} = 5 \times 10^{-3}$, $\lambda_{\text{adv}} = 5 \times 10^{-3}$. On AvatarReX, pretraining is performed for 5k iterations, followed by joint optimization for 400k iterations. An additional 5k-step face-only fine-tune is applied. Training is conducted on a single A100 GPU, requiring approximately 1.5 days per person.

4.3. Results

4.4. Quantitative Results

Tables 1 and 2 report quantitative comparisons on AvatarReX and THuman4.0 for novel-view synthesis, evaluated with PSNR, SSIM, and LPIPS on both full-body and head regions. On AvatarReX (Table 1), F3G-Avatar achieves competitive full-body PSNR (30.214) while obtaining the best SSIM (0.970) and LPIPS (0.032) among the SoTA methods. Although AnimatableGaussians reports a slightly higher PSNR, our approach improves structural similarity and perceptual quality, indicating better preservation of fine details and fewer rendering artifacts. For the head region, F3G-Avatar outperforms prior methods, achieving the high-

Model	Novel View (Body) PSNR↑	SSIM↑	LPIPS↓	Novel View (Head) PSNR↑	SSIM↑	LPIPS↓
AvatarRex [40]	23.248	0.957	0.065	-	-	-
3DGS-Avatar [28]	28.784	0.951	0.042	24.972	0.943	0.121
GaussianAvatar [6]	26.950	0.939	0.041	24.011	0.933	0.144
AnimatableGaussians [16]	**30.361**	0.968	0.034	25.671	0.957	0.114
Ours (F3G-Avatar)	30.214	**0.970**	**0.032**	**26.243**	**0.964**	**0.084**

Table 1. Quantitative comparison of full-body and face-focused novel-view synthesis on the AvatarReX [40] dataset.

Model	Novel View (Body) PSNR↑	SSIM↑	LPIPS↓	Novel View (Head) PSNR↑	SSIM↑	LPIPS↓
TAVA [13]	26.61	0.968	0.032	-	-	-
Ani-NeRF [25]	22.53	0.964	0.034	20.14	0.931	0.102
AnimatableGaussians [16]	**30.614**	0.980	0.029	26.434	0.953	0.071
Ours (F3G-Avatar)	30.311	**0.981**	**0.026**	**26.934**	**0.961**	**0.062**

Table 2. Quantitative comparison of full-body and face-focused novel-view synthesis on the THuman4.0 [39] dataset.

Metric	PSNR↑	SSIM↑	LPIPS↓
w/o Face-focused Deformation	25.721	0.951	0.107
w/o MHR template	26.015	0.963	0.090
w/o $\mathcal{L}_{\text{adv}}$	**27.010**	0.959	0.091
Full model	26.243	**0.964**	**0.084**

Table 3. Ablation study on AvatarReX head novel-view metrics.

est PSNR (26.243), SSIM (0.964), and a substantially better LPIPS (0.084). On THuman4.0 (Table 2), similar trends are observed. For full-body evaluation, F3G-Avatar achieves the best SSIM (0.981) and LPIPS (0.026), while maintaining PSNR (30.311) competitive to AnimatableGaussians (30.614). For the head region, our method attains the highest PSNR (26.934), SSIM (0.961) and LPIPS (0.062) indicating more accurate facial reconstruction. Overall, the results across both datasets demonstrate that decoupling global and face-specific Gaussian deformations enables improved perceptual quality.

4.5. Qualitative Results

Figure 4 presents qualitative comparisons of rendered avatars on the AvatarReX dataset. We compare F3G-Avatar with AnimatableGaussians under similar novel-view and pose conditions, showing three subjects together with the corresponding ground-truth images. Both methods produce plausible full-body renderings. However, F3G-Avatar consistently preserves sharper facial structures and more stable appearance across viewpoints.

4.6. Ablation Study

Component ablations. Table 3 reports face-focused ablations on AvatarReX. Removing the Face-focused Deformation reduces LPIPS and SSIM, while omitting the MHR template lowers PSNR/SSIM. Disabling the adversarial term yields competitive PSNR but worse LPIPS, suggesting the face-specific loss improves perceptual quality.
Face-network capacity and input resolution. Table 4 portrays the effect of both input resolution and face-network capacity. Increasing the input resolution from 128×128 to 256×256 significantly improves reconstruction quality, raising PSNR from 24.554 to 26.774 and SSIM from 0.939 to 0.956, while reducing LPIPS from 0.101 to 0.086 at the expense of higher runtime.

We further vary the StyleGAN-style mapping depth ($n_{\text{mlp}} \in \{2, 4\}$) and channel multiplier ($\text{cm} \in \{1, 2\}$) in the face sub-networks, while keeping the body network fixed. Increasing the mapping depth from $n_{\text{mlp}} = 2$ to 4 improves

Metric	PSNR↑	SSIM↑	LPIPS↓	Time (ms)
Input Resolution				
(128 x 128)	24.554	0.939	0.101	24.34 ± 1.12
(256 x 256)	26.774	0.956	0.086	61.41 ± 0.87
Model Variations				
$(n_{\text{mlp}}, \text{cm})$: (2, 1)	25.340	0.941	0.087	47.43± 1.11
$(n_{\text{mlp}}, \text{cm})$: (4, 1)	26.243	0.964	0.084	56.47± 1.16
$(n_{\text{mlp}}, \text{cm})$: (4, 2)	26.774	0.956	0.086	61.41± 0.87

Table 4. Ablation study on model variants with runtime Face-focused Deformation (Time).

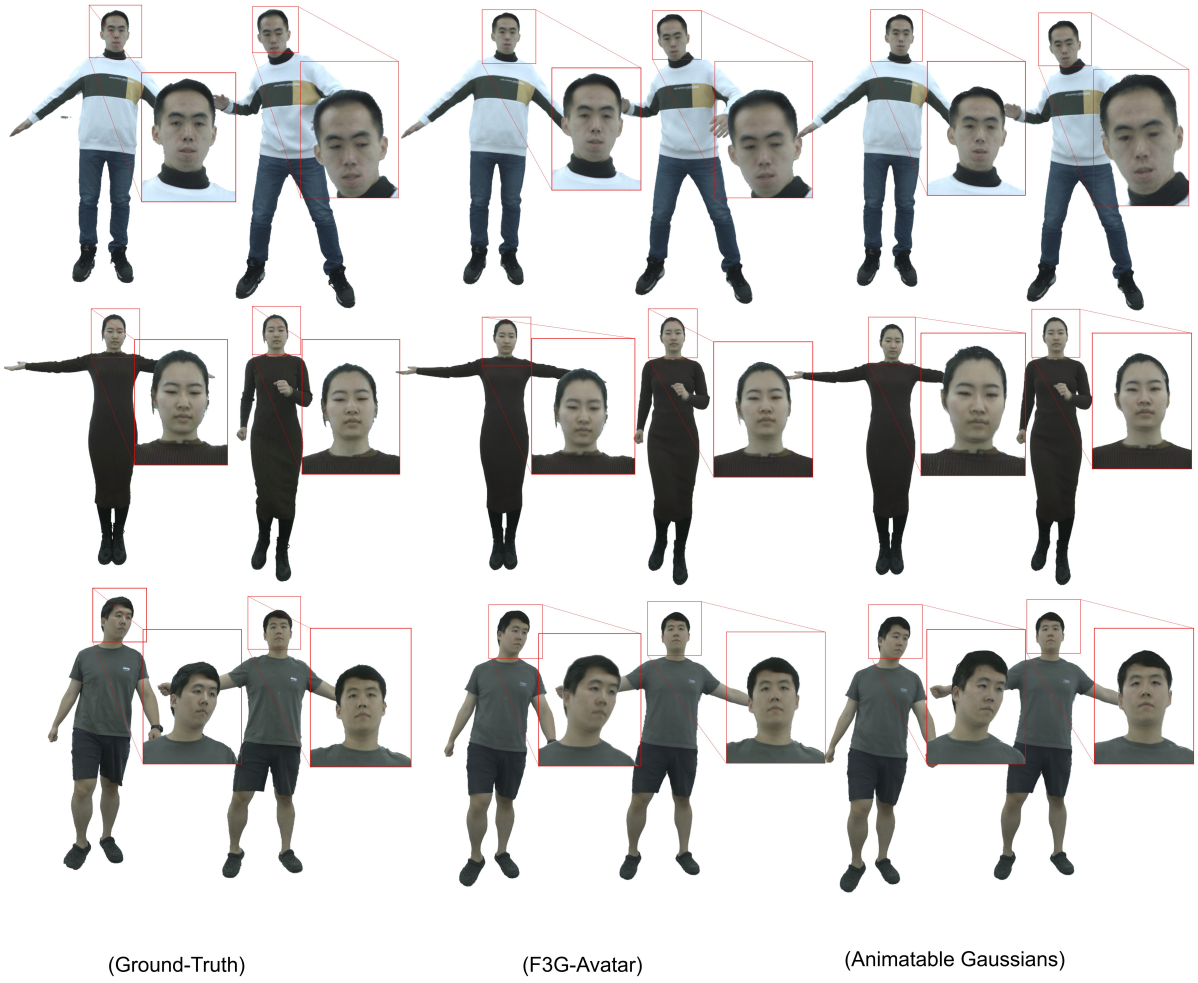

Figure 4. F3G-Avatar displays state-of-the-art rendering quality by delivering improved facial details.

reconstruction quality, increasing PSNR and SSIM while slightly lowering LPIPS. Increasing the channel multiplier from 1 to 2 provides a modest PSNR gain (26.24 $\rightarrow$ 26.77). These changes also increase runtime, from 47.43 ms for the smallest configuration to 61.41 ms for the widest variation.

5. Conclusion

The proposed F3G-Avatar model demonstrates the impact of coupling a clothed canonical template with explicit Gaussian rendering for realistic avatar synthesis. The MHR body template provides a global structure, while pose-conditioned Gaussian deformations capture fine details and maintain view consistency. The body and face branches operate in a complementary manner: the body branch models global non-rigid motion, while the face branch focuses on high-frequency features critical for close-up perception. Quantitative results show consistent improvements across PSNR, SSIM, and LPIPS for body and head views, where F3G-Avatar improves the SoTA results on the AvatarReX and THuman4.0 benchmarks.

Potential Social Impact. F3G-Avatar can synthesize life-like, animatable full-body digital humans with realistic facial details, enabling the generation of fabricated 3D content or 2D videos. Therefore, responsible use of this technology is essential.

6. Acknowledgments

This work is supported by the ELEVATION Xecs 2023022 project on cloud-based Systems-of-Systems for high-end security and broadcast applications.

References

[1] Rinat Abdrashitov, Kim Raichstat, Jared Monsen, and David Hill. Robust skin weights transfer via weight inpainting. In *SIGGRAPH Asia 2023 Technical Communications*, New York, NY, USA, 2023. Association for Computing Machinery. 5

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM siggraph 2005 papers*, pages 408–416. 2005. 2, 4

[3] Aaron Ferguson, Ahmed AA Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, David Otte, Eric Vignola, Fabian Prada, Federica Bogo, et al. Mhr: Momentum human rig. *arXiv preprint arXiv:2511.15586*, 2025. 2

[4] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2021. 3

[5] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023. 2, 3

[6] 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*, pages 634–644, 2024. 2, 3, 7

[7] Yujiao Jiang, Qingmin Liao, Xiaoyu Li, Li Ma, Qi Zhang, Chaopeng Zhang, Zongqing Lu, and Ying Shan. Uv gaussians: Joint learning of mesh deformation and gaussian textures for human avatar modeling. *Knowledge-Based Systems*, 320:113470, 2025. 2, 3

[8] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021. 6

[9] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021. 2, 4, 6

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 3

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 5

[12] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 505–515, 2024. 3

[13] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision*, pages 419–436. Springer, 2022. 2, 3, 7

[14] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 1, 2

[15] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8456–8465, 2023. 5

[16] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19711–19722, 2024. 2, 3, 7

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

[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. 2, 3

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

[20] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 788–798, 2024. 3

[21] Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, 19(2):98–100, 2012. 2

[22] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 5

[23] Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[24] 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*, pages 10975–10985, 2019. 1, 2

[25] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Animatable neural radiance fields for human body modeling. *arXiv preprint arXiv:2105.02872*, 2(3):5, 2021. 2, 3, 7

[26] Stylianos Ploumpis, Haoyang Wang, Nick Pears, William AP Smith, and Stefanos Zafeiriou. Combining 3d morphable models: A large scale face-and-head model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10934–10943, 2019. 3

[27] 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. 3

[28] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu 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. 3, 7

[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)*, 2024. 3

[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*, pages 16911–16921, 2023. 3

[31] Zhiyu Tao, Yanyan Liu, Junsheng Qiu, and Shengwei Li. Impact of virtual avatar appearance realism on perceptual interaction experience: a network meta-analysis. *Frontiers in Psychology*, 16:1624975, 2025. 2, 3

[32] Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. Styleavatar: Real-time photo-realistic portrait avatar from a single video. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–10, 2023. 4

[33] Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4d-dress: A 4d dataset of real-world human clothing with semantic annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5

[34] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3295–3306, 2023. 5

[35] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2024. 3

[36] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 5

[37] Zhongyuan Zhao, Zhenyu Bao, Qing Li, Guoping Qiu, and Kanglin Liu. Psavatar: A point-based shape model for real-time head avatar animation with 3d gaussian splatting. *arXiv preprint arXiv:2401.12900*, 2024. 3

[38] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21057–21067, 2023. 3

[39] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15893–15903, 2022. 2, 3, 6, 7

[40] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG)*, 42(4): 1–19, 2023. 3, 6, 7

[41] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4574–4584, 2023. 3