SurfAvatar: Versatile Human Avatar with Meshified Surfel Gaussians

Zijian Wu[1] Jiawei Zhang[1] Yanwen Wang[1] Yao Yao[1]
Siyu Zhu[2] Xun Cao[1] Hao Zhu[1†]

[1]Nanjing University [2]Fudan University

Figure 1. SURFAVATAR generates high-fidelity 3D human avatar. As shown in figure above, each avatar standing on his/her corresponding attribute maps for *Meshified Surfel Gaussians*.

Abstract

Generating high-fidelity, animatable 3D human models is crucial for applications in the metaverse, telepresence, digital games, and film production. Traditional 3D human avatar modeling methods are limited by high costs and complexity, whereas 3D human generative models offer a more accessible approach. In this paper, we address the limitations of existing methods, which suffer from degraded appearance after pose manipulation. A novel approach is introduced for learning a riggable, high-quality 3D human generation model, utilizing a dataset comprising unstructured static 3D human models. We present Meshified Surfel Gaussians, a unique fusion of Gaussian and mesh representations specifically designed for avatar modeling. This innovation establishes explicit connections among Gaussian points, facilitating connectivity-based optimization regularizers. Our method surpasses baseline approaches and accommodates a range of downstream tasks, rendering it highly versatile for high-fidelity human generation and practical applications. The code and data will be publicly

released upon publication.

1. Introduction

Generating 3D human models is a pivotal endeavor in computer vision and computer graphics, with extensive applications spanning various industries, including metaverse, telepresence, digital games, and film production. The capability to efficiently create lifelike, animatable 3D human models is essential for advancing digital avatar technology, improving user interaction experiences, and expanding the frontiers of creative production.

The research into 3D human avatar modeling boasts a rich history. Initially, high-quality 3D models of the human body were crafted through manual design by skilled artists and the utilization of multi-view light field [11] or range field [56] reconstruction systems. Nevertheless, these methods are constrained by prohibitively high hardware and labor costs, thereby limiting their practical applications. In stark contrast, 3D human generative models have emerged as a markedly more easy-to-use approach to 3D human avatar modeling, garnering widespread interest and research attention.

The core idea of generative models lies in establishing a conditional probability prediction model on a dataset, modeling the intrinsic structure and generative mechanism of the data. Vision generative models have achieved success in tasks such as 2D image generation [18, 42] and 3D face generation [4, 46]. However, research on generative models for 3D human has progressed slowly. This is primarily due to the fact that the human body is a complex articulated non-rigid structure consisting of numerous joints, making the underlying generative mechanisms more challenging to learn. Consequently, previous state-of-the-art works [62] have exhibited issues such as overly smooth geometries, blurred textures in generated 3D avatars, and prominent artifacts in animated 3D avatars.

In this paper, we define our task as learning a riggable, high-quality 3D human generation model based on a dataset of 526 unstructured static 3D human models, similar to E^3Gen. E^3Gen maps 3D Gaussian models onto the UV space and utilizes 2D generative networks to predict Gaussian parameters. Experimental results reveal that the generation quality of this approach is relatively limited, especially after pose manipulation, where the appearance of the 3D human models significantly degrades. Our insight is that the key to enhancing the performance of 3D human generation models lies in introducing a more efficient representation method and its corresponding learning framework tailored for 3D human generation.

Specifically, we introduce an innovative integration of Gaussians with mesh representations, explicitly tailored for avatar modeling, which we have dubbed **Meshified Surfel Gaussians**. Drawing inspiration from the ExAvatar [32], we propose to place the center of each Gaussian at the vertices of a predefined parametric human mesh. This methodology establishes explicit connectivity among the Gaussian points by leveraging the vertex-to-vertex connections inherent in the mesh. Consequently, this approach facilitates the utilization of connectivity-based optimization regularizers, such as Laplacian regularization, enabling the generation of avatars with exceptional geometric fidelity. Our method surpasses other baselines and extensively supports various downstream tasks, including animation (especially facial animation), and texture/geometry editing. This makes our approach versatile for high-fidelity human generation and practical real-world applications.

To summarize, our main contributions are three-fold:

- We introduce *Meshified Surfel Gaussians*, a novel mechanism that binds 3D Gaussian primitives with a parametric mesh model of a riggable 3D human, thereby enhancing the representation capability of 3D human generative models.
- Connectivity-based regularization is introduced in the model training to ensure that the 3D Gaussian primitives are more accurately aligned with the semantic UV space.
- Our method achieves cutting-edge rendering quality, particularly with a significant enhancement in the quality of human renderings animated into unseen poses. Remarkably, this outstanding performance is attained through training on just a few hundred static human models with unconstrained poses.

2. Related Work

2.1. 3D Neural Representation

When describing 3D scenes and objects, different 3D representations exhibit distinct preferences in expressing geometry and appearance. Traditional graphics pipelines include Voxel [25, 30, 59], Point cloud [38, 39], and Mesh [48, 51], which represent 3D structures using discrete elements such as grids, unstructured points, or polygonal surfaces. These representations efficiently capture geometric details and have been widely used in computer graphics and 3D modeling. More recently, neural representations have emerged as an alternative. The neural field [5, 31, 33] is an implicit function that takes a 3D position and viewing direction as input and outputs color and density, enabling photorealistic rendering through volumetric ray casting. In contrast, Gaussian Splatting [15, 19, 58] is a representation that models a 3D scene as a collection of anisotropic Gaussian kernels, where each Gaussian is defined by its position, covariance, and appearance. By leveraging tile-based differentiable rasterization, Gaussian Splatting enables efficient and high-speed rendering while maintaining flexibility.

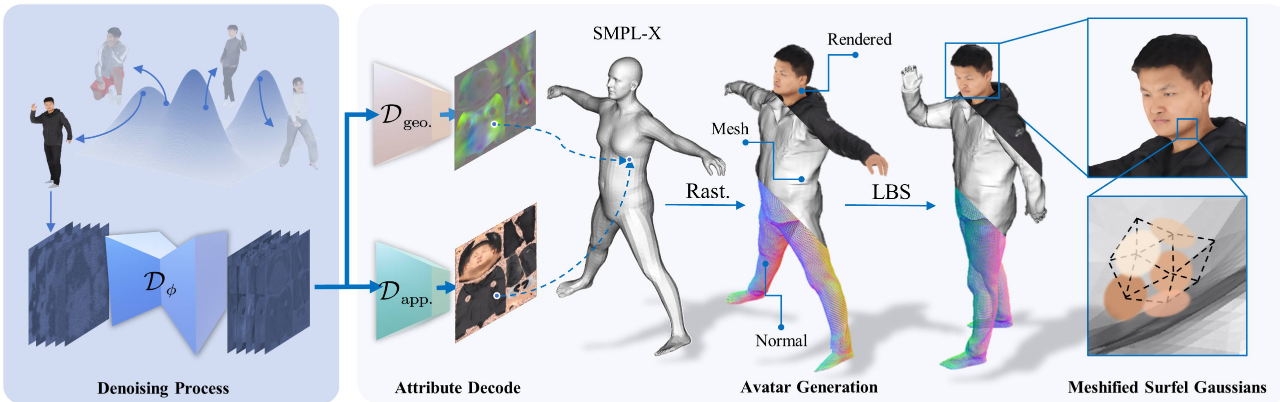


Figure 2. **Pipeline Overview**. Our method learns the underlying distribution of latent UV feature from the dataset, then obtains color and offset attribute maps through two separate decoders. Subsequently, Gaussians are anchored to the SMPL-X model by sampling from the two attribute maps. Leveraging differentiable rasterization and our proposed meshified surfel Gaussians, the rendered results achieve both photorealism and high-quality geometry (*e.g.*, mesh and normals).

2.2. 3D Generative Model

In 2D, generative models such as GANs [18], diffusion models [42], and autoregressive models [22] have made remarkable strides in the general tasks [64, 65] of text-image synthesis and image-conditioned generation. In the realm of 3D, two widely explored paradigms for 3D generation are 3D-aware GANs and multi-view diffusion models. 3D-aware GANs [2–4, 21] utilize a well-designed 2D generator backbone to synthesize a triplane representation of the scene, which is then decoded by a small decoder to produce radiance field information. On the other hand, multi-view diffusion models [24, 27, 44, 49] control a heavy diffusion backbone using multi-view images and camera embeddings. This allows the model to generate images from multiple viewpoints, capturing the scene from different angles. The generated images are then used as textures, which are mapped onto meshes generated by other models [47, 52, 54, 61], creating coherent 3D representations with detailed textures.

2.3. 3D Avatar Generation

Extensive efforts have been made to create 3D human avatars [8, 9, 23, 28, 36, 37, 41, 43, 53, 60, 66, 67]. Early approaches [7, 50] primarily relied on scanned datasets to model human shapes. With the emergence of large-scale 2D human image datasets, 3D-aware GAN has been widely explored for human generation. AG3D [68] extends 3D-aware GAN by explicitly modeling the pose space, enabling more controllable human synthesis. GETAvatar [63] introduces a disentangled representation of geometry and texture. EVA3D [14] models articulated humans in a part-based manner, achieving high-resolution image synthesis without additional super-resolution. PrimDiffusion [10] adopts a new primitive that incorporates radiance and kinematic information. With the advancement of 3DGS, recent work explores generating human by Gaussian attribute maps in UV space. GSM [1] constructs a hierarchical shell-map-like structure that includes multiple sets of attribute maps. E3Gen [62] optimizes a set of random feature maps while jointly training the decoder to produce attribute maps. These approaches highlight the evolution of 3D human generation, transitioning from explicit modeling to learning-based representations with improved realism and efficiency.

3. Method

We propose SurfAvatar, a generative method for learning the generation of animatable avatars from a multiview human dataset. The overview of our pipeline is illustrated in Fig. 2. In this section, we first provide a brief introduction to prior knowledge about SMPL-X [35] and 3D Gaussian Splatting [19] in Sec. 3.1, then introduce avatar modeling with our novel Meshified Surfel Gaussians in Sec. 3.2. Then, the integration of our designed Gaussians into existing generative frameworks is explained in Sec. 3.3. Lastly, Sec. 3.4 explains our design of loss functions.

3.1. Preliminary

SMPL-X [35] is a widely used parametric human body model that extends the SMPL model by incorporating expressive details for hands and faces. By representing body shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$ through low-dimensional parameter space, SMPL-X enables robust and flexible deformation of a 3D human mesh. With given facial expressions $\boldsymbol{\psi}$, the deformed template mesh in the canonical (T-pose) space is defined as:

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \bar{T} + B_S(\boldsymbol{\beta}) + B_P(\boldsymbol{\theta}) + B_E(\boldsymbol{\psi}), \quad (1)$$

where the $\bar{T}$ is the mean template mesh, and $B_S(\boldsymbol{\beta})$, $B_P(\boldsymbol{\theta})$ and $B_E(\boldsymbol{\psi})$ represent the blend shapes of shape, pose and expression respectively. Once the targeted shape, pose and expressions are obtained, the final posed mesh M is computed via Linear Blend Skinning (LBS):

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \mathrm{LBS}\Big(T(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, W\Big). \quad (2)$$

where $J(\boldsymbol{\beta})$ is the joint regressor, and W is the skinning blend weights.

3D Gaussian Splatting [19] (3DGS) is a rendering technique that represents a scene using a sparse set of 3D Gaussian ellipsoids. Each Gaussian is defined by its position $\boldsymbol{\mu} \in \mathbb{R}^3$, covariance $\boldsymbol{\Sigma}$ (which is stored as scaling $\boldsymbol{s} \in \mathbb{R}^3$ and quaternion $\boldsymbol{q} \in \mathbb{R}^4$), color $\boldsymbol{c} \in \mathbb{R}^3$, and opacity α. The rendered color $\boldsymbol{C}$ of a pixel is computed by blending all Gaussians overlapping this pixel:

$$\boldsymbol{C} = \sum_{i=1}^{N} \boldsymbol{c}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j). \quad (3)$$

where $\boldsymbol{c}_i$ is the color of each Gaussian, and α_i is the blending weight derived from the opacity and probability density.

3.2. Meshified Surfel Gaussians

In this paper, we propose a novel integration of Gaussians with mesh representations for avatar modeling, which we term **Meshified Surfel Gaussians**. Inspired by ExAvatar [32], we position the center of each Gaussian at the vertices of the predefined template (*e.g.*, SMPL-X mesh). This approach creates explicit connectivity among Gaussians by mapping the vertex-to-vertex connections on mesh, which in turn enables the use of connectivity-based optimization regularizers (*e.g.*, Laplacian regularization) to generate avatars with high geometric fidelity without the need for additional geometric supervision. Building on this, we devise Gaussian attributes tailored for representations suitable for avatar modeling as follows.

Position and color. We observe that native Gaussians, due to their flexibility and anisotropy, often produce artifacts such as spiking or floating during avatar animation. To mitigate this, we constrain the degrees of freedom of Gaussians to color and position. Given the relatively simple surface of human avatars, we do not account for complex reflection or refraction effects. Instead, we simplify the color representation $\boldsymbol{c}$ to three RGB channels and set the opacity as 1. The position is determined by the sum of the previously mentioned vertices and a learnable offset $\Delta\boldsymbol{\mu}$.

Scaling and rotation. Previous works constrain Gaussians to isotropic spheres or simply use a fixed value for scaling. However, such oversimplifications often prevent Gaussian primitives from adequately covering the entire mesh, resulting in noticeable artifacts. Inspired by the design of surfel primitives [13, 15], we propose that flattened Gaussians can more accurately represent the surface of an avatar. In our implementation, for each Gaussian located at a mesh vertex, the third axis is defined along the vertex normal direction and its value is defined as a small constant σ. The other two axes lie in the plane perpendicular to the vertex normal, with each assigned a magnitude δ times the average distance $\bar{d}_i$ between the vertex and its adjacent vertices. The scaling $\boldsymbol{s_i}$ is calculated as follows:

$$\bar{d}_i = \frac{1}{K_i} \sum_{j=1}^{K_i} \|\boldsymbol{p_i} - \boldsymbol{k_j}\|_2, \quad (4)$$

$$\boldsymbol{s_i} = [\delta \cdot \bar{d}_i, \delta \cdot \bar{d}_i, \sigma]. \quad (5)$$

where $\boldsymbol{p_i}$ is the position of each Gaussian, $\boldsymbol{k_j} \in \mathcal{K}(\boldsymbol{p_i})$, indicates the neighboring Gaussians of $\boldsymbol{p_i}$, K_i is the number of its neighboring Gaussians. δ is set to 0.75 and σ is set to $1e^{-5}$. This design eliminates the need for additional prediction networks or further optimization for pose-dependent scaling, while allowing the scaling to adapt naturally to the underlying deformed mesh geometry. The resulting disc-shaped Gaussians form a smooth Gaussian surface proximate to the original mesh.

Therefore, consider an upsampled SMPL-X template mesh in a canonical pose with vertices $V_{init} \in \mathbb{R}^{N\times 3}$ and faces $\mathcal{F} \in \mathbb{R}^{F\times 3}$. Each vertex is anchored with meshified Gaussians that possess color $\boldsymbol{c}$ and offset $\Delta\boldsymbol{\mu}$. The canonical Gaussians position $\boldsymbol{\mu}_{can}$ will be obtained by adding the offset $\Delta\boldsymbol{\mu}$ to the predefined canonical densified SMPL-X template V_{init}. Given a targeted SMPL-X shape $\boldsymbol{\beta}$, pose $\boldsymbol{\theta}$, and expression $\boldsymbol{\psi}$ coefficients, we apply LBS to yield the deformed Gaussians position $\boldsymbol{\mu}$ using Eq. (2). With the deformed position, we can obtain scaling $\boldsymbol{s}$ as described in Eq. (5). With face index $\mathcal{F}$, we can calculate the normal $\boldsymbol{n}$ directly and subsequently obtain rotation $\boldsymbol{q}$ of Gaussians.

3.3. Avatar Generation

The bridge connecting Gaussians with grid-based generators is UV sampling. In avatar generation, a common strategy is to assign attributes to primitives anchored at corresponding positions. This is done by leveraging the UV layout of the human template to sample from the attribute maps produced by the generator. Formally, given an attribute map $\mathcal{A}(u, v)$, Gaussian attributes $\mathcal{G}$ with predefined UV coordinates $\boldsymbol{z}_{uv}$ are obtained via bilinear interpolation:

$$\mathcal{G} = \textsc{GridSample}(\mathcal{A}, \boldsymbol{z}_{uv}). \quad (6)$$

In our scheme, $\mathcal{G}$ involves the offset $\Delta\boldsymbol{\mu}$ in canonical space and color $\boldsymbol{c}$. To obtain these two attribute maps, we adopt a single-stage diffusion scheme following previous methods [6, 62] in an end-to-end manner. Formally, the overall loss function is expressed as:

$$\mathcal{L} = \lambda_{\text{fit}} \mathcal{L}_{\text{fit}}(\chi_i; \varphi) + \lambda_{\text{diff}} \mathcal{L}_{\text{diff}}(\chi_i; \phi), \quad (7)$$

where χ_i is a latent feature, ϕ is the parameters of the denoising U-Net $\mathcal{D}_\phi$ used in our diffusion process, and φ is the parameters of our decoders $\mathcal{D}_{\text{app.}}$ and $\mathcal{D}_{\text{geo.}}$, which decode the latent feature into the corresponding Gaussian attributes map. $\mathcal{L}_{\text{fit}}$ and $\mathcal{L}_{\text{diff}}$ represent the loss function of the training for the fitting and diffusion process, respectively. The loss function will be detailed in Sec. 3.4.

In the training phase, we feed a batch of avatars exhibiting diverse poses, with each scene randomly sampled from different camera views to provide RGB observations. The latent feature of the corresponding scene χ_i is randomly initialized and optimized by minimizing both the fitting loss $\mathcal{L}_{\text{fit}}$ and the diffusion loss $\mathcal{L}_{\text{diff}}$, while the model parameters ϕ and φ are updated simultaneously.

3.4. Loss Function

Diffusion Loss. The diffusion model regularizes the latent features by learning a denoising process. A noisy latent feature is generated at a diffusion time step t as:

$$\chi_i^{(t)} = \xi(t)\,\chi_i + \tau(t)\,\epsilon, \tag{8}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise, and $\xi(t)$ and $\tau(t)$ are schedule functions that control the noise level. The denoising U-Net $\mathcal{D}_\phi$ is trained to predict the original latent feature χ_i from $\chi_i^{(t)}$ using the following loss:

$$\mathcal{L}_{\text{diff}}(\chi_i; \phi) = \mathbb{E}_{t,\epsilon}\left[\frac{1}{2} w(t) \left\| \phi(\chi_i^{(t)}, t) - \chi_i \right\|^2\right], \tag{9}$$

where $w(t)$ is an empirically designed weighting function that emphasizes particular time steps during training.

Fitting Loss. To ensure consistency between the rendered images and the ground truth, we incorporate two photorealistic losses in the fitting loss:

$$\mathcal{L}_{\text{fit}}(\chi_i; \varphi) = \lambda_{\text{L2}}\mathcal{L}_{\text{L2}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}}, \tag{10}$$

where the $\mathcal{L}_{\text{L2}}$ is defined as the L2 distance between the ground truth images and the rendered outputs of our fitted avatars. Additionally, the perceptual loss $\mathcal{L}_{\text{perc}}$ [17] is incorporated, computed from the feature maps of both the ground truth images and the rendered outputs extracted via a pre-trained VGG network [45].

Hand Consistency. In addition, we incorporate an L1 loss $\mathcal{L}_{hands}$ that measures the difference between the colors of Gaussians in the hand region and the average color of Gaussians in the cheek region.

$$\mathcal{L}_{\text{hands}} = \left|\boldsymbol{c}_{\text{hands}} - \text{mean}\left(\boldsymbol{c}_{\text{cheeks}}\right)\right|, \tag{11}$$

This loss term ensures that the generated hand colors align with the overall skin tone.

Part-aware Laplacian. Laplacian regularization [12, 16, 32, 34, 40] enforces smoothness in mesh deformation by keeping vertices near the average position of their neighbors. Using the target mesh's connectivity, it preserves local geometry and ensures a coherent deformation. Since our Gaussians are anchored at the mesh vertices, we can naturally apply Laplacian regularization to optimize their positions. Considering the Gaussian in position $\boldsymbol{p}_i$ and its neighbors $\boldsymbol{k}_j \in \mathcal{K}(\boldsymbol{p}_i)$, the Laplacian distance $\boldsymbol{\zeta}_i$ is acquired via:

$$\boldsymbol{\zeta}_i = \boldsymbol{p}_i - \frac{1}{K_i}\sum_{j=1}^{K_i} \boldsymbol{k}_j, \tag{12}$$

and the Laplacian loss we optimized is:

$$\mathcal{L}_{\text{lap}} = \sum_{i=1}^{N} \omega_i \cdot \left\| \boldsymbol{\zeta}_i - \boldsymbol{\zeta}_i' \right\|_2^2, \tag{13}$$

where $\boldsymbol{\zeta}_i$ is the Laplacan distance in template mesh and ω_i is the regularization strength. We set the different regularization strength for different regions, 200 for the facial area, 50 for the hand area, and 30 for the ear area.

Overall, the total loss function is summarized as follows:

$$\begin{aligned}\mathcal{L} = \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{fit}}\mathcal{L}_{\text{fit}} \\ + \lambda_{\text{hands}}\mathcal{L}_{\text{hands}} + \lambda_{\text{lap}}\mathcal{L}_{\text{lap}}.\end{aligned} \tag{14}$$

4. Experiments

Experiment Settings. In our experiments, we employed the THuman2.0 Dataset [57] as the primary training source. This dataset comprises 526 textured 3D scans captured with a high-density DSLR rig, offering a wide range of challenging poses. Each scan is provided with corresponding SMPL-X parameters. For data pre-processing, we rendered 500 identities from the THuman2.0 dataset, generating 54 camera views per identity.

To enhance facial rendering precision, we performed facial data augmentation by rendering an additional 54 images per identity, capturing views around the head. We didn't use any explicit 3D supervision such as ground true normals or 3D meshes. We adopt the Fréchet Inception Distance (FID) metric and Kernel Inception Distance (KID) as our metric to evaluate the quality and diversity of our method.

Implement Details. We perform two rounds of upsampling on the SMPL-X mesh to obtain approximately 160K vertices. Given that the SMPL-X UV layout covers roughly 74% of the available UV space, we employ a 512×512 resolution UV Gaussian attribute map and a 512×512 resolution UV latent feature map. Our method is trained on 8 NVIDIA L20 GPUs for approximately 3 days. The UV latent feature map consists of 32 channels; when input to the decoders, the first 16 channels are fed into the geometry decoder to predict offsets, while the remaining 16 channels are fed into the appearance decoder to predict colors. Our geometry decoder and appearance decoder are implemented

Figure 3. **Comparison Results**. Our method produces high-quality and realistic human generation results compared to EVA3D [14], GETAvatar [63], PrimDiffusion [10], and E3Gen [62]. We also compare our results to a mesh rendering implement of our training framework to show that our meshified surfel Gaussians achieve a higher render quality.

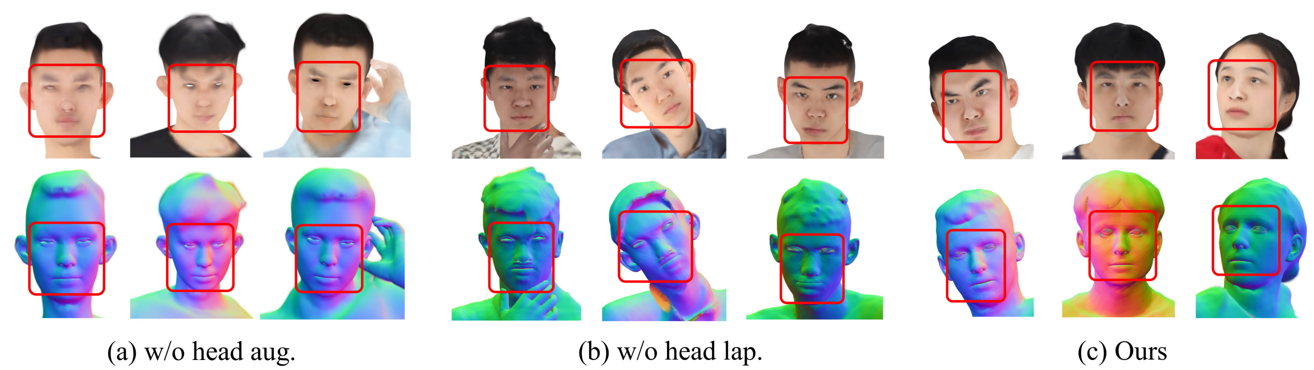


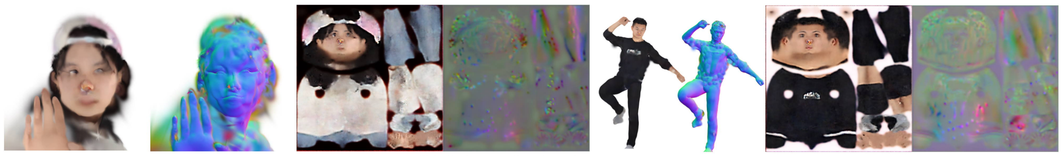
Figure 4. **Face Ablation**. Ablation study about the scheme of facial region. Without head dataset augmentation, the rendered quality is poor. Added head dataset augmentation but without facial part heavier Laplacian regularization, misalignment of rendered image and underlying geometry will appear. Our method achieve both high rendered quality and faithful facial geometry.

as shallow convolutional neural networks, with each decoder comprising two convolutional layers. The color prediction head uses a sigmoid activation function. In contrast, for the offset prediction head, no activation function is applied; the weights are initialized from a uniform distribution $\mathcal{U}(-1 \times 10^{-5}, +1 \times 10^{-5})$ and initial biases are set to zero.

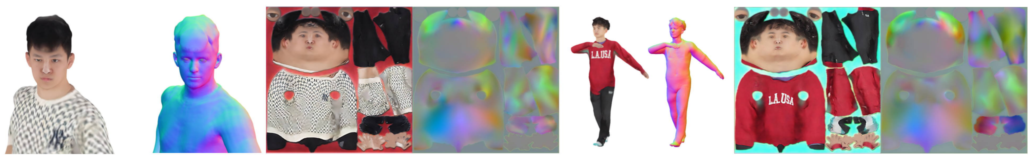
4.1. Evaluation of Generated Avatars.

Comparison with SOTA Baselines. We compare our method with four state-of-the-art approaches: EVA3D [14], GETAvatar [63], E3Gen [62], and PrimDiffusion [10]. Additionally, we conduct an experiment replacing our meshified surfel Gaussians with the traditional textured mesh as

(a) without regularization

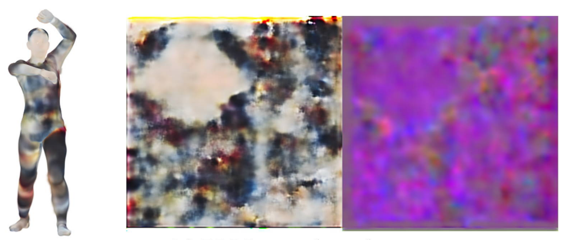


(b) L2 regularization

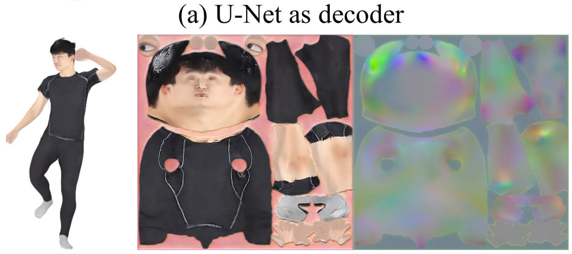


(c) Laplacian regularization (Ours)

Figure 5. **Geometry Regularization**. To demonstrate its effectiveness, we perform two ablation studies: removing Laplacian regularization and replacing it with L2 regularization.



(a) U-Net as decoder



(b) Ours

Figure 6. **Decoder Choice**. We show the performance when the decoder becomes more complex (*e.g.* U-Net).

	FID↓	KID↓
EVA3D	184.88	237.18
GETAvatar	17.91	40.67
PrimDiffusion	80.71	76.24
E3Gen	17.19	34.26
mesh implement	17.45	34.34
Ours	16.25	31.46

Table 1. Qualitative comparisons on THuman 2.0 dataset.

	FID↓	KID↓
U-Net as decoder	68.49	129.73
L2 reg.	17.82	34.57
w/o reg.	21.04	41.99
w/o head lap.	17.06	32.63
w/o head aug.	23.58	46.94
Ours	16.25	31.46

Table 2. Qualitative evaluation of ablation study.

the rendering primitive. The comparison results are presented in Fig. 3 and quantitative metrics are summarized in Table. 1. As demonstrated, EVA3D performs suboptimally on the THuman2.0 dataset, a complex multi-view dataset with challenging poses, primarily due to the instability inherent in GAN training. GETAvatar, which relies on ground truth normals as input, fails to generate plausible facial and hand details, while PrimDiffusion, which is a two-stage diffusion method employing a Mixture of Volumetric Primitives [26] for representation, similarly struggles to produce consistent fine details in these regions. Moreover, E3Gen, a generative Gaussian-based approach, suffers from the anisotropic nature of Gaussians; on datasets like THuman2.0, where pose and identity are difficult to decouple, the generated avatars are prone to artifacts such as spiking and floating Gaussians and do not generalize well to novel poses. In contrast, our method consistently achieves high-quality rendering across various poses.

Comparison with Mesh-based Rendering. In contrast, our design leverages the inherent connectivity of the mesh vertices, enabling the Gaussians to inherit a structured geometric prior. Conventional textured mesh rendering typically relies on high-resolution texture maps (*e.g.*, 1024×1024 or higher) to capture details, considerably increasing computational and memory overhead. To further demonstrate the efficacy of our method relative to traditional mesh-based representations, we implemented a base-

Figure 7. **Facial Driven and Gestures Control**. Due to our meshified surfel Gaussians, our method enables facial and hand gesture animation without artifacts.

line using a differentiable mesh renderer [20]. Experimental results indicate that our approach achieves superior performance in both generation quality and rendering fidelity.

4.2. Ablation Study

Head Augmentation. We conduct ablation studies on head augmentation and head Laplacian regularization. As shown in Fig. 4, head augmentation primarily improves the texture details in the facial region, while Laplacian regularization ensures alignment between geometry and texture.

Geometry Regularization. We further conduct ablation on Laplacian regularization. Fig. 5 highlights the geometric and texture details of the full body and face under different settings. When Laplacian regularization is removed, the attribute maps become noisy, and the normal maps start to blur. When Laplacian regularization is replaced with standard L2 regularization, the normal maps deteriorate significantly, and the attribute maps fail to capture fine details.

Decoder Design. We conduct an ablation study on the design of the attribute map decoder. One alternative approach is to replace the current shallow CNN with a more complex U-Net. As shown in Fig. 6, the use of complex decoder does not lead to performance gains; instead, it causes the attribute maps to deteriorate. We argue this to the increased difficulty in jointly optimizing the diffusion and fitting processes when employing a more complex decoder.

4.3. Application

Avatar Animation. Our method effectively anchors Gaussians onto the mesh surface, enabling freeform animation. Fig. 7 demonstrates facial and hand gesture animation driven by TalkSHOW [55] sequence, while Fig. 8 presents full-body animation driven by motion sequences from AMASS [29]. Credited by carefully designed Gaussians, our method achieves high-quality rendering while mitigating artifacts.

Adapting to different body shapes. Since our method defines Gaussian scales based on vertex distance, we can

Figure 8. **Body Driven**. We demonstrate the performance of our method on challenging motion sequences.

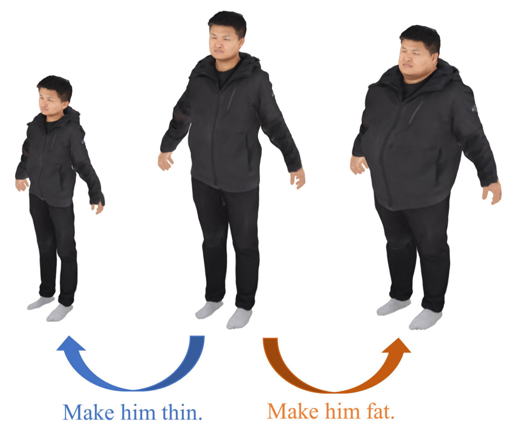


Figure 9. **Different body shape adapting**. Our method enables effortless adapting of body shape.

freely modify SMPL-X parameters to edit different body shapes without introducing cumbersome artifacts.

5. Conclusion

In this paper, we propose a novel method for learning a riggable, high-quality 3D human generation model. We introduce Meshified Surfel Gaussians, an integration of Gaussian and mesh representations tailored for avatar modeling, which establishes explicit connectivity among Gaussians and enables connectivity-based regularizers. Our approach outperforms baseline methods and supports various downstream tasks, making it versatile for human generation and practical applications.

Our method still has some limitations. Because the representation model is tightly integrated with SMPL-X, it struggles with modeling loose garments. Furthermore, constrained by the diversity of appearances in the training dataset, the generated model shows a noticeable bias and demonstrates limited generalization capacity.

References

[1] Rameen Abdal, Wang Yifan, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, and Gordon Wetzstein. Gaussian shell maps for efficient 3d human generation. In *CVPR*, 2024. 3

[2] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *CVPR*, pages 20950–20959, 2023. 3

[3] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2021.

[4] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 2, 3

[5] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *ECCV*, 2022. 2

[6] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *ICCV*, 2023. 4

[7] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars. *CVPR*, 2022. 3

[8] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 3

[9] Yushuo Chen, Zerong Zheng, Zhe Li, Chao Xu, and Yebin Liu. Meshavatar: Learning high-quality triangular human avatars from multi-view videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 3

[10] Zhaoxi Chen, Fangzhou Hong, Haiyi Mei, Guangcong Wang, Lei Yang, and Ziwei Liu. Primdiffusion: Volumetric primitives diffusion for 3d human generation. In *NeurIPS*, 2023. 3, 6

[11] Paul Debevec. The light stages and their applications to photoreal digital actors. *ACM TOG*, 2(4):1–6, 2012. 2

[12] Mathieu Desbrun, Mark Meyer, Peter Schröder, and Alan H. Barr. Implicit fairing of irregular meshes using diffusion and curvature flow. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, page 317–324, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 5

[13] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *CVPR*, 2024. 4

[14] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *ICLR*, 2022. 3, 6

[15] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. 2, 4

[16] 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. *arXiv preprint arXiv:2403.11589*, 2024. 5

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016*, pages 694–711, Cham, 2016. Springer International Publishing. 5

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 3

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

[20] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 8

[21] Heyuan Li, Ce Chen, Tianhao Shi, Yuda Qiu, Sizhe An, Guanying Chen, and Xiaoguang Han. Spherehead: Stable 3d full-head synthesis with spherical tri-plane representation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 3

[22] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *NeurIPS*, 2024. 3

[23] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *CVPR*, 2024. 3

[24] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 3

[25] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. In *NeurIPS*, 2019. 2

[26] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Trans. Graph.*, 40(4), 2021. 7

[27] Yuanxun Lu, Jingyang Zhang, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, Xun Cao, and Yao Yao. Direct2.5: Diverse text-to-3d generation via multi-view 2.5d diffusion. *CVPR*, 2024. 3

[28] Shengjie Ma, Yanlin Weng, Tianjia Shao, and Kun Zhou. 3d gaussian blendshapes for head avatar animation. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 3

[29] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, 2019. 8

[30] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IROS*, pages 922–928, 2015. 2

[31] 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. In *ECCV*, 2020. 2

[32] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *ECCV*, 2024. 2, 4, 5

[33] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 41(4):102:1–102:15, 2022. 2

[34] Andrew Nealen, Takeo Igarashi, Olga Sorkine, and Marc Alexa. Laplacian mesh optimization. In *Proceedings of the 4th International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia*, page 381–389, New York, NY, USA, 2006. Association for Computing Machinery. 5

[35] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 3

[36] 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 *ICCV*, 2021. 3

[37] 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 *CVPR*, 2021. 3

[38] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CVPR*, 2016. 2

[39] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017. 2

[40] Shenhan Qian. Vhap: Versatile head alignment with adaptive appearance priors, 2024. 5

[41] 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*, pages 20299–20309, 2024. 3

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3

[43] 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 *CVPR*, 2024. 3

[44] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 3

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 5

[46] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *CVPR*, 2023. 2

[47] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, , Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 3

[48] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018. 2

[49] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 3

[50] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. In *NeurIPS*, 2021. 3

[51] Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2mesh++: Multi-view 3d mesh generation via deformation. In *ICCV*, 2019. 2

[52] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. In *NeurIPS*, 2024. 3

[53] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *CVPR*, 2024. 3

[54] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 3

[55] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *CVPR*, 2023. 8

[56] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *CVPR*. IEEE, 2018. 2

[57] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, 2021. 5

[58] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. *CVPR*, 2024. 2

[59] A. Khosla F. Yu L. Zhang X. Tang J. Xiao Z. Wu, S. Song. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, 2015. 2

[60] Jiawei Zhang, Zijian Wu, Zhiyang Liang, Yicheng Gong, Dongfang Hu, Yao Yao, Xun Cao, and Hao Zhu. Fate: Full-

head gaussian avatar with textural editing from monocular video, 2024. 3

[61] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics*, 2024. 3

[62] Weitian Zhang, Yichao Yan, Yunhui Liu, Xingdong Sheng, and Xiaokang Yang. e^3gen: Efficient, expressive and editable avatars generation. *ACM MM*, 2024. 2, 3, 4, 6

[63] Xuanmeng Zhang, Jianfeng Zhang, Chacko Rohan, Hongyi Xu, Guoxian Song, Yi Yang, and Jiashi Feng. Getavatar: Generative textured meshes for animatable human avatars. In *ICCV*, 2023. 3, 6

[64] Xinchen Zhang, Ling Yang, Yaqi Cai, Zhaochen Yu, Kaini Wang, Jiake Xie, Ye Tian, Minkai Xu, Yong Tang, Yujiu Yang, and Bin Cui. Realcompo: Balancing realism and compositionality improves text-to-image diffusion models. *NeurIPS*, 2024. 3

[65] Xinchen Zhang, Ling Yang, Guohao Li, Yaqi Cai, Jiake Xie, Yong Tang, Yujiu Yang, Mengdi Wang, and Bin Cui. Itercomp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. In *ICLR*, 2025. 3

[66] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *CVPR*, 2022. 3

[67] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *ACM TOG*, 42(4), 2023. 3

[68] Jinlong Yang Michael J.Black Otmar Hilliges Andreas Geiger Zijian Dong, Xu Chen. AG3D: Learning to generate 3D avatars from 2D image collections. In *International Conference on Computer Vision (ICCV)*, 2023. 3