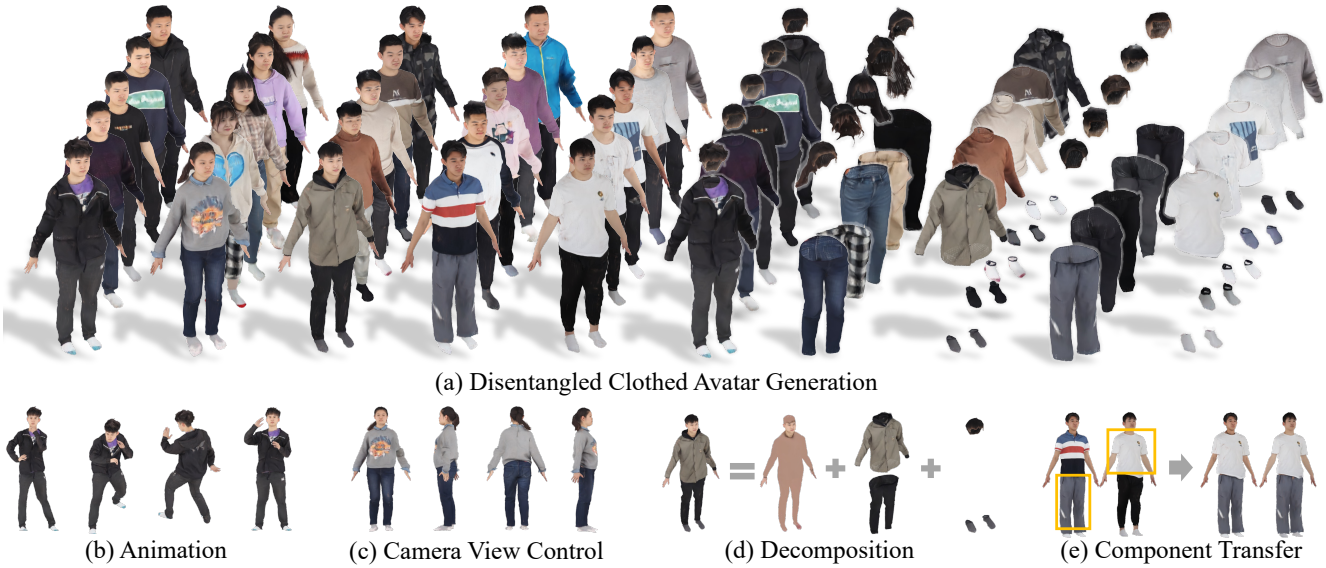
Disentangled Clothed Avatar Generation with Layered Representation

Weitian Zhang[1] Yichao Yan[1†] Sijing Wu[1] Manwen Liao[2] Xiaokang Yang[1]
[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]The University of Hong Kong
[1]{weitianzhang, yanyichao, wusijing, xkyang}@sjtu.edu.cn
[2]{manwen}@connect.hku.hk



(a) Disentangled Clothed Avatar Generation

(b) Animation (c) Camera View Control (d) Decomposition (e) Component Transfer

Figure 1. We propose **LayerAvatar** to efficiently generate diverse clothed avatars with components fully disentangled. The generated avatars can be animated and synthesized in novel views. They can also be decomposed into body, hair, and clothes for component transfer.

Abstract

Clothed avatar generation has wide applications in virtual and augmented reality, filmmaking, and more. While existing methods have made progress in creating animatable digital avatars, generating avatars with disentangled components (e.g., body, hair, and clothes) has long been a challenge. In this paper, we propose LayerAvatar, a novel feed-forward diffusion-based method capable of generating high-quality component-disentangled clothed avatars in seconds. We propose a layered UV feature plane representation, where components are distributed in different layers of the Gaussian-based UV feature plane with corresponding semantic labels. This representation can be effectively learned with current feed-forward generation pipelines, facilitating component disentanglement and enhancing details of generated avatars. Based on the well-designed representation, we train a single-stage diffusion model and introduce constrain terms to mitigate the severe occlusion issue of the innermost human body layer. Extensive experiments demonstrate the superior performances of our method in generating highly detailed and disentangled clothed avatars. In addition, we explore its applications in component transfer. The project page is available at https://olivia23333.github.io/LayerAvatar.

1. Introduction

The creation of digital avatars has various applications [2, 13] in virtual and augmented reality, filmmaking, and more. Traditional graphics-based pipelines require extensive effort from 3D artists to construct a single digital avatar. To reduce tedious manual labor and facilitate mass production, learning-based methods aiming at generating digital avatars

† Corresponding author.

automatically have been widely explored recently.

Recent learning-based methods [3, 16, 77] mainly combine 3D representations [28, 39, 55, 64] with generation pipelines (*e.g.*, 3D-aware GANs [4, 5, 43] and diffusion models [47, 52]) to create digital avatars. However, most of these methods often ignore the compositional nature of digital avatars and represent the human body, hair, and clothes as a whole, which limits their capabilities in digital avatar customization such as cloth transfer. Neural-ABC [7] and SMPLicit [14] provide parametric model with disentangled clothes and human body. However, modeling texture are leaved an un-explored problem. HumanLiff [23] proposes a layer-wise generation process that first generates clothed avatars in minimal clothes, then generates digital avatars wearing the next layer of clothing conditioned on the current layer. Nevertheless, the human body and clothes are not fully disentangled, which makes it difficult to extract the components of each layer, thus reducing the editing ability. In addition, some methods [15, 18, 59, 63, 66] follow the trend of DreamFusion [49] to achieve disentangled clothed avatar generation by learning each component of digital avatars separately through the prior knowledge of 2D diffusion models [52] in an optimization manner. These methods can generate clothed avatars with each component disentangled, however, they take hours to generate a single digital avatar, and the optimization time will increase linearly according to the number of components.

In this paper, we propose **LayerAvatar**, a novel feed-forward diffusion-based method that achieves (1) *component disentanglement*, enabling seamless transfer of individual components such as clothes, hair, and shoes; (2) *high-quality results*, with generated avatars exhibiting intricate facial details, distinct fingers, and realistic cloth wrinkles; and (3) *efficiency*, requiring only seconds to generate a single avatar. We choose 3D Gaussians [28] as the underlying representation due to its high-quality rendering results for intricate details, and strong representation capability for diverse cloth types. However, naively representing the disentangled clothed avatar using 3D Gaussians is impractical due to its unstructured nature that is incompatible with most current feed-forward generation pipelines [5, 52]. Therefore, we introduce a Gaussian-based UV feature plane, in which 3D Gaussians are projected into a predefined 2D UV space shared among subjects. The attributes of each 3D Gaussian are encoded as local geometry and texture latent features, which can be obtained from the 2D feature plane via bilinear interpolation. Furthermore, to achieve full disentanglement of avatar components (hair, shoes, upper cloth) and higher generation quality, we represent avatar components in separate layers of the UV feature plane which provides neighboring components with distinctive features from different layers to facilitate decomposition.

To generate the layered representation in a feed-forward manner, we elaborately train a single-stage diffusion model [6] from multi-view 2D images. To fully disentangle each component and ensure plausible avatar generation results, we employ supervision both in the individual components and the entire compositional clothed avatar. Moreover, several prior losses are utilized to constrain the smooth surface and reasonable color of the severely occluded human body.

We evaluate LayerAvatar on multiple datasets [9, 20, 74], demonstrating its superior performance in generating disentangled avatars. We also explore its application in component transfer. In summary, our main contributions are:

- We introduce LayerAvatar, a novel feed-forward clothed avatar generation pipeline with each component disentangled, enhancing the controllability of avatar generation.
- We propose a layered UV feature plane representation that enhances generation quality and facilitates the disentanglement of each component.
- Our method achieves outstanding generation results on multiple datasets and support downstream applications such as component transfer.

2. Related work

Diffusion in 3D Generation. Encouraged by the success of diffusion model [52] in 2D image generation area, researchers have attempted to extend it to 3D generation tasks. These works can be divided into two categories, feed-forward and optimization-based methods. Optimization-based methods [8, 32, 33, 49, 50, 61, 67], represented by DreamFusion [49], utilize SDS loss to distill prior knowledge of 2D diffusion model to supervise 3D scenes. These methods often suffer from oversaturation and Janus problems. Thus, improved SDS loss [67] and camera-conditional [35] or multi-view diffusion models [56] are introduced to mitigate these problems. Moreover, optimization-based methods usually take hours to generate a single object, which hinders its application in real life. On the other hand, feed-forward methods [37, 69, 75, 78, 82] directly learn diffusion model for 3D representations, such as points [42, 76], voxels [51], meshes [37, 72], and implicit neural representations [22, 41, 57]. These methods can generate 3D objects in seconds. Several attempts [12, 25, 79, 80] have been made to adapt them to the field of digital human generation. Different from most previous works, we regard digital avatars as a composition of multiple components instead of a unified whole and learn a diffusion model on the proposed layered representation.

Clothed Avatar Generation. Inspired by general 3D object generation, many methods [3, 12, 16, 21, 44, 81] introduce 3D-aware GANs [5] and diffusion models [52] for clothed avatar generation. These methods first learn clothed avatars in canonical space and then animate them to posed space using a deformation module. Following EG3D [5], some methods [77, 81] apply triplane features to represent

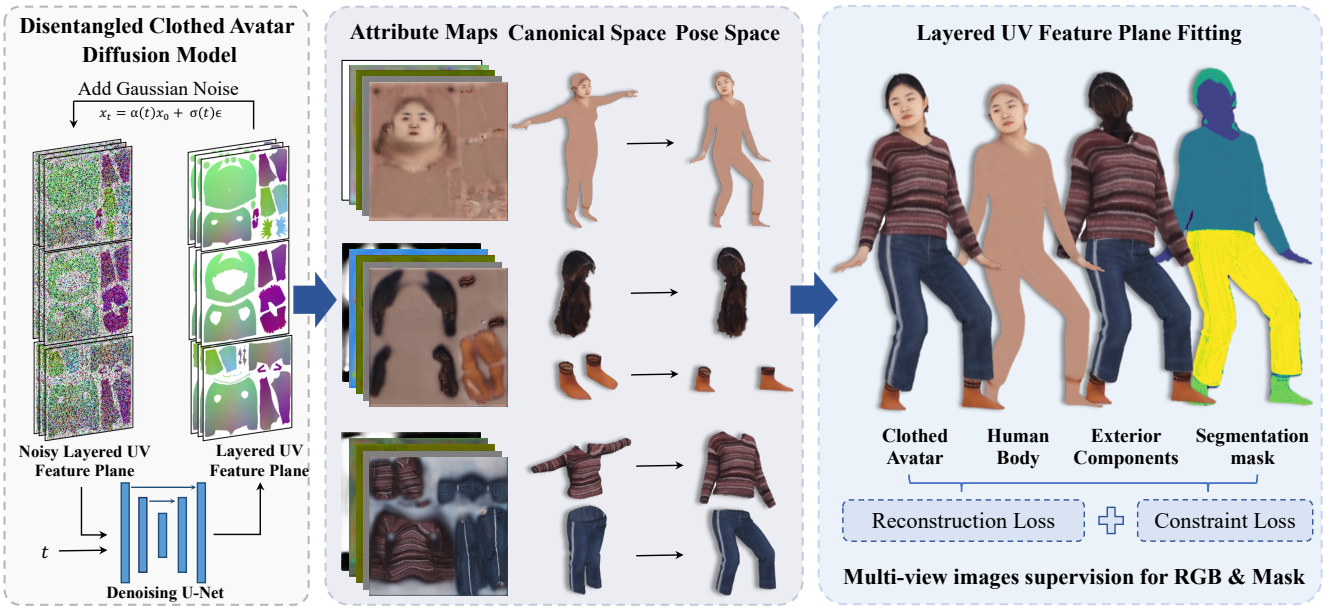


Figure 2. Method overview. LayerAvatar learns a feed-forward diffusion model to generate clothed avatars with each component disentangled. The clothed avatars are represented as layered UV feature plane where components are represented separately. After decoding the feature plane into attribute maps, we can extract 3D Gaussians from them through SMPL-X-based templates. Generated clothed avatars are then transformed into targeted pose space for further supervision. Reconstruction loss and constraint loss are both utilized to facilitate the disentanglement and handle the severe occlusion of human body layer.

clothed avatars for higher quality and utilize inverse skinning for animation. AG3D [16] introduces forward skinning technique [10, 11] to achieve robust animation including loose clothing. On the other hand, Chupa [30] and AvatarPopup [31] apply diffusion models to learn front and back view image pairs and then lift them to 3D space. Despite their impressive success, most of these methods represent clothed avatars as an entity and fail to disentangle the human body and clothes. Recently, some optimization-based methods [15, 18] achieved success in generating cloth-disentangled avatars, however, the generating process takes hours to generate a single avatar.

Compositional Avatar Representation. Instead of representing avatars [40, 68, 73] as a single entity, some methods represent clothed avatars as a combination of multiple submodules. COAP [38], Spams [45] and DANBO [58] consider human avatars as a composition of body parts, while EVA3D [21] and ENARF-GAN [44] follow this trend and utilize multiple neural networks to represent different body parts of the digital avatar, achieving more efficient and detailed generation results. Several methods [1, 23, 71] represent clothed avatars as separate layers to enable the expressiveness of various topologies. However, these methods ignore the disentanglement of human body and clothes, which makes each submodule less physically meaningful. Recently, some works [17, 34, 48, 70, 83] disentangle the human body and clothes by representing each component separately. During the rendering process, these components are combined through various compositional rendering techniques. Most of these methods are designed to optimize a single digital avatar. In this paper, we propose a novel layered UV feature plane representation that is compatible with feed-forward generation framework.

3. Method

We propose LayerAvatar, a feed-forward generative method for disentangled clothed avatar generation. The overview of our method is illustrated in Fig. 2. We provide a brief introduction for prior knowledge in Sec. 3.1. To achieve disentangled clothed avatar generation, we propose a novel layered UV feature plane representation (Sec. 3.2), which facilitates disentanglement and is compatible with current feed-forward generative pipelines. The clothed avatars are generated in canonical space and then deformed to targeted pose space via deformation module (Sec. 3.3). The training process is introduced in Sec. 3.4.

3.1. Preliminary

SMPL-X [46] is an expressive parametric human model that can produce naked human meshes $M(\beta, \theta, \psi)$ given shape parameter β, pose parameter θ, and expression pa-

rameter ψ. The producing process can be formulated as:

$$T(\beta,\theta,\psi) = T_c + B_s(\beta;s) + B_e(\psi;e) + B_p(\theta;p),$$
$$M(\beta,\theta,\psi) = LBS(T(\beta,\theta,\psi)), J(\beta), \theta, \mathcal{W}), \quad (1)$$

where a canonical human mesh T is first calculated as a combination of the mean shape template T_c and vertex displacements ($B_s(\beta;s)$, $B_e(\psi;e)$, $B_p(\theta;p)$) computed by the blend shapes s, e, p and their corresponding pose, shape, and expression parameters. The body template T is then deformed to the given pose by linear blend skinning(LBS) based on the skinning weights $\mathcal{W}$ and joint locations $J(\beta)$.

3D Gaussians [28] is a primitive-based explicit representation that combines the strengths of both previous explicit and implicit representations. It consists of a set of learnable 3D Gaussian primitive $\mathcal{G}_k$ where each contains five attributes: position $\boldsymbol{\mu}$, scaling matrix $\mathbf{S}$, rotation matrix $\mathbf{R}$, opacity α, and color $\mathbf{c}$. In practice, we employ diagonal vector $\mathbf{s} \in \mathbb{R}^3$ and axis-angle $\mathbf{r} \in \mathbb{R}^3$ to represent $\mathbf{S}$ and $\mathbf{R}$ respectively. 3D Gaussians are represented as ellipses in 3D space defined by their position μ and covariance matrix $\mathbf{\Sigma} = \mathbf{RSS}^T\mathbf{R}^T$. During the rendering process, these 3D Gaussians are projected to a 2D image plane where the pixel color $\mathbf{C}$ can be calculated as follows:

$$\mathbf{C} = \sum_{i=1}^{N} \mathbf{c}_i \sigma_i \prod_{j=1}^{i-1} (1-\sigma_j), \quad (2)$$

where $\mathbf{c}_i$ is the color of the i-th 3D Gaussian on the ray, and σ_i is the blending weight calculated with the opacity α.

3.2. Layered UV Feature Plane Representation

Previous methods mainly represent avatars as a single entity or depend on optimization-based schemes. As a result, the generated results often have difficulty with editing or are slow to create, typically taking hours to generate a single subject. To address these limitations, we propose a layered UV feature plane representation that can separate components of clothed avatars and is compatible with fast, feed-forward generation pipelines. We employ 3D Gaussians as the base representation for efficient rendering along with easy animation and editing.

To enable disentanglement, we consider clothed avatars as a composition of human body and exterior components.

$$\mathcal{G}_{\text{avatar}} = \{\mathcal{G}_{\text{body}}, \mathcal{G}_{\text{top}}, \mathcal{G}_{\text{bottom}}, \mathcal{G}_{\text{hair}}, \mathcal{G}_{\text{shoes}}\}. \quad (3)$$

Each component is represented as a set of Gaussian primitives $\mathcal{G}_i$ parameterized with five attributes: 3D position $\boldsymbol{\mu}_i \in \mathbb{R}^3$, opacity $\alpha_i \in \mathbb{R}$, rotation matrix $\mathbf{R}_i$ represented by axis angle $\mathbf{r}_i \in \mathbb{R}^3$, scale matrix $\mathbf{S}_i$ represented by diagonal vector $\mathbf{s}_i \in \mathbb{R}^3$ and rgb color $\mathbf{c}_i \in \mathbb{R}^3$. Inspired by existing works [24, 80] that initialize 3D Gaussians by attaching them to SMPL parametric model for geometry and animation prior. We initialize the 3D Gaussians of each component by attaching them to self-designed templates based on SMPL-X. For each component, we design a template that maximizes coverage of the region where the component may exist. To enhance generation quality, all templates are subdivided to support densified Gaussian primitives. Then, we initialize the positions of 3D Gaussians as the center points of faces on the densified template mesh. And the initial rotations of 3D Gaussians are set as the tangent frame of the faces, which consists of the normal vector of that face, the direction vector of one edge and their cross product.

For compatibility with feed-forward generation pipelines, previous methods typically project 3D Gaussians into 2D space (*e.g.*, multi-view image space [60, 62], UV space [80]). Unlike these methods, which represent the entire subject as a single entity and then employ a post-processing step for disentanglement, we directly model components separately by mapping their templates into a three-layer UV space, combined with semantic labels. The first layer represents the innermost part of the human body, the second layer represents the hair and shoes, and the third layer represents top and bottom clothes. Following [80], all Gaussian attributes are stored as local UV features to enhance generation quality. The UV features plane is first split into two parts in a channel-wise manner and then decoded by two light-weight shared MLP decoders $\mathcal{D}_g$ and $\mathcal{D}_t$ separately. $\mathcal{D}_g$ decodes the geometry-related attributes: position offset $\Delta\boldsymbol{\mu}$ and opacity α of 3D Gaussians and $\mathcal{D}_t$ predicts texture-related attributes, color $\mathbf{c}$ and covariance-related rotation $\Delta\mathbf{r}$ and scale $\Delta\mathbf{s}$. Given the decoded attribute maps, we can extract the attributes for each 3D Gaussian $\mathcal{G}_i$ via bilinear interpolation. The opacity value α_i and color value $\mathbf{c}_i$ are obtained directly, and the other values are obtained via following formula based on their initial values:

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}_i^0 + \Delta\boldsymbol{\mu}_i, \quad \mathbf{s}_i = \mathbf{s}_i^0 \cdot \Delta\mathbf{s}_i, \quad \mathbf{r}_i = \mathbf{r}_i^0 \cdot \Delta\mathbf{r}_i, \quad (4)$$

where $\boldsymbol{\mu}_i^0$, $\mathbf{s}_i^0$, and $\mathbf{r}_i^0$ are initial position, scale, and rotation values for 3D Gaussians, respectively. $\Delta\boldsymbol{\mu}_i$, $\Delta\mathbf{s}_i$ and $\Delta\mathbf{r}_i$ are predicted residuals extracted from attribute maps. By collecting the attributes of 3D Gaussians with the same semantic label, we can obtain a canonical space representation of each component, and their composition forms the complete disentangled clothed digital avatar.

3.3. Deformation

Benefiting from the SMPL-X-based templates, our method supports deformation in body shapes and novel poses including gestures and facial expressions. To support training with multiple subjects in various body shapes, we disentangle the body shape factor by defining all the templates in a canonical space with neutral body shapes. The neutral body shape avatar and its corresponding components can be transformed into targeted body shape space via the follow-

ing warping process:

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu} + B_s(\beta, s, \boldsymbol{\mu}), \tag{5}$$

where $\bar{\boldsymbol{\mu}}$ represents the position of 3D Gaussians in the targeted β body shape space and $B_s(\beta, s, \boldsymbol{\mu})$ are corresponding body shape related offsets extracted from the SMPL-X based templates via barycentric interpolation. We further add pose-dependent offsets $B_p(\theta, p, \boldsymbol{\mu})$ and facial expression offsets $B_e(\psi, e, \boldsymbol{\mu})$ in the same way to ensure accurate animation results.

The animation of a generated avatar from the canonical T-pose to an arbitrary target pose can be regarded as transforming the 3D Gaussian attributes. During the animation process, the opacity α and color $\mathbf{c}$ of 3D Gaussians remain unchanged. Therefore, we only discuss the transformation of position $\boldsymbol{\mu}$, rotation matrix $\mathbf{R}$ and scale matrix $\mathbf{S}$ in this section. Using the LBS function, we can transform the position $\bar{\boldsymbol{\mu}}$ of 3D Gaussians as:

$$\boldsymbol{\mu}' = \sum_{i=1}^{n_b} w_i \mathbf{B}_i \bar{\boldsymbol{\mu}}, \tag{6}$$

where n_b represents the number of joints and $\mathbf{B}_i$ is the transformation matrix of the i-th joint. For 3D Gaussians on the innermost human body layer, the corresponding blend skinning weights w are obtained directly from SMPL-X-based templates through barycentric interpolation, as these regions usually undergo minimal topology changes. For 3D Gaussians representing the exterior components, we follow [16] to extract the skinning weights from a pre-computed low-resolution volumetric field of fused skinning weights, which is more stable for points that deviate significantly from the original template. $\mathbf{T} = \sum_{i=1}^{n_b} w_i \mathbf{B}_i$ is the blended transformation matrix, and the rotation matrix $\mathbf{R}$ is updated via $\mathbf{R}' = \mathbf{T}_{1:3,1:3}\mathbf{R}$, where $\mathbf{T}_{1:3,1:3}$ is the rotational part of $\mathbf{T}$. The scale matrix $\mathbf{S}$ is recalculated in the targeted pose space to fit deformed topology.

3.4. Learning Disentangled Clothed Avatar

To mitigate the impact of occlusion, we adopt a single-stage training scheme [6], which is more robust in occluded and sparse view situations. Specifically, the layered UV feature plane fitting and diffusion training process is conducted simultaneously, and the UV feature plane is jointly optimized by the fitting and diffusion loss. Similar to the SDS loss [49], the diffusion loss provides a diffusion prior for the UV feature plane, thereby facilitating the completion of unseen regions in the training images.

Layered UV Feature Plane Fitting. Given multi-view images, we optimize the layered UV feature plane and shared decoders to reconstruct avatars with disentangled components. The objective function can be divided into reconstruction and constraint part. The reconstruction loss $\mathcal{L}_{\text{recon}}$ can be formulated as follows:

$$\mathcal{L}_{\text{recon}} = \lambda_{\text{color}} \cdot \mathcal{L}_{\text{color}} + \lambda_{\text{mask}} \cdot \mathcal{L}_{\text{mask}} + \lambda_{\text{per}} \cdot \mathcal{L}_{\text{per}} + \lambda_{\text{seg}} \cdot \mathcal{L}_{\text{seg}}. \tag{7}$$

To achieve the disentanglement between exterior components and human body, we not only minimize the color loss $\mathcal{L}_{\text{color}}$ and mask loss $\mathcal{L}_{\text{mask}}$ on the overall rendering result, but also perform supervision on each component. Specifically, we first render the 3D Gaussians corresponding to each component separately to obtain multi-view images of each component. Then, inspired by Clothedreamer [36], instead of blending the rendering results of these components via estimated depth order to obtain the rendering results of clothed avatars, we directly render all the 3D Gaussians to alleviate artifacts caused by the blending process. The silhouette masks of each component and the clothed avatar are obtained similarly. The ground truth of silhouette masks is estimated based on the semantic segmentation results predicted by Sapiens [29]. We apply Huber loss [26] for both $\mathcal{L}_{\text{color}}$ and $\mathcal{L}_{\text{mask}}$ following SCARF [17], due to its robustness to the estimated noisy segmentation results. To enhance the details of generated results, we also employ a perceptual loss $\mathcal{L}_{\text{per}}$ [27] to minimize the difference between extracted features of rendered outputs and targeted images.

Components in overlapping regions may learn inverted color or opacity values due to incorrect depth ordering. To address this, we render semantic segmentation maps of the clothed avatar by assigning the segmentation label of each Gaussian as its color. We then minimize the distance to the predicted semantic segmentation map using the Huber loss, $\mathcal{L}_{\text{seg}}$, to encourage accurate depth ordering.

Due to the severe occlusion of the inner human body layer, we apply constraints on the geometry and texture of the human body to obtain reasonable results. Since the human body is always within the exterior layer, we employ the following constraints:

$$\mathcal{L}_{\text{maskin}} = \lambda_{\text{maskin}} \text{ReLU}\left(\mathcal{R}_m^b(\mathcal{G}_{\text{body}}) - M_{\text{fg}}\right), \tag{8}$$

where $\mathcal{R}_m^b(\mathcal{G}_{\text{body}})$ represents the rendered silhouette of the human body, and M_{fg} is the estimated foreground silhouette mask. When rendering $\mathcal{R}_m^b(\mathcal{G}_{\text{body}})$, we detach the opacity values and set them to 1, only optimizing offset and covariance-related attributes, preventing the model from minimizing the loss via decreasing opacity on the boundary. Utilizing the prior that the occluded skin color should be similar to the color of hands, we introduce the following texture constraint:

$$\mathcal{L}_{\text{skin}} = \lambda_{\text{skin}} \left(M_{\text{oc}} \odot \left(\mathcal{R}_m^b(\mathcal{G}_{\text{body}}) - \mathbf{C}_{\text{skin}}\right)\right), \tag{9}$$

where $\mathbf{C}_{\text{skin}}$ is the average color of pixels in the hands region and M_{oc} is the mask of the occluded region. Other regularization terms are as follows:

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{offset}} \mathcal{L}_{\text{offset}} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}}. \tag{10}$$

$\mathcal{L}_{\text{offset}} = \|\Delta_\mu\|^2$ constrains the offset from being extremely large. $\mathcal{L}_{\text{smooth}}$ is the total variational (TV) loss, which is used to minimize the average L_2 distance between neighboring pixels on attribute maps. This regularization term encourages smooth transitions between the neighboring attributes (*e.g.* offsets, rotation, and opacity), promoting the generation of reasonable texture and geometry surface.

Disentangled Clothed Avatar Diffusion Model. To generate disentangled clothed avatars, we train a diffusion model that maps Gaussian noise to the layered UV latent space. Since diffusion models generally perform better on inputs with low channel dimensions [53], we concatenate our layered UV feature plane across widths instead of stacking them across channels. During training, we first obtain destructed UV feature plane $\boldsymbol{x}_t$ by adding Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ to the layered UV feature plane $\boldsymbol{x}_0$ according to following noise schedule:

$$\boldsymbol{x}_t := \alpha(t)\boldsymbol{x}_0 + \sigma(t)\boldsymbol{\epsilon}, \tag{11}$$

where $\alpha(t)$ and $\sigma(t)$ are predefined functions that control the intensity of added noise, and t is the time step in the range of [0, 1]. To stabilize and accelerate training, we use v-prediction proposed in [54] to train the denoising UNet. The objective function for the diffusion model is:

$$\mathcal{L}_{\text{diff}} = \mathop{\mathbb{E}}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}} \left[\frac{1}{2} w(t) \left\| \hat{\boldsymbol{x}}_0 - \boldsymbol{x}_0 \right\|^2 \right], \tag{12}$$

where $w(t) = \left(\alpha(t)/\sigma(t)\right)^{2\omega}$, the ω is set to 0.5 following [6]. $\mathcal{L}_{\text{diff}}$ is used to update not only the parameters of denoising UNet, but also the UV feature plane. It promotes the UV feature plane to adapt to the learned latent space, thereby providing priors for occluded regions.

4. Experiments

Baselines. To evaluate holistic generation quality, we compare our method with the state-of-the-art methods of animatable avatar generation (EVA3D [21], StructLDM [25], and E3Gen [80]) on THuman2.0 [74] dataset. We then evaluate the decomposition capability of our method by comparing the layer-wise generation results with the disentanglement-related method HumanLiff [23] on Tightcap [9] dataset. And we further evaluate the component generation quality against optimization-based methods: LAGA [18], SO-SMPL [63], and TELA [15]. Our method is trained on a composite dataset including CustomHuman [20], THuman2.0 [74], and THuman2.1 [74].

Metrics. For holistic generation quality evaluation, we utilize FID [19] following previous works [5, 25, 80]. For layer-wise generation results, we adopt FID for overall generation quality evaluation and L-PSNR [23] for disentanglement capability evaluation, which calculates the PSNR between two layers with the adding component masked. Additionally, we conduct a user study with 20 participants to compare the generation quality and disentanglement quality of our methods with optimization-based methods.

Dataset. For THuman2.0 [74], we sample 500 scans and render each from 54 camera views to obtain multi-view images. Then, we employ Sapiens [29] to estimate segmentation masks for these images, from which per-component silhouette masks can be extracted. This approach enables our method to learn directly from multi-view images without requiring separate meshes of each component, thereby simplifying the data collection process. Tightcap [9] contains 3D scans with separate meshes for cloth and shoe, which facilitates direct rendering of multi-view images and silhouette masks for each component. For a fair comparison, we use the preprocessed version provided by HumanLiff [23], which selects 107 samples from Tightcap. Each sample is rendered from 158 camera views, providing images and silhouette masks for both the entire avatar and each component. We also construct a composite dataset consisting of 1954 selected scans from THuman2.0 [74], THuman2.1 [74], and CustomHuman [20] datasets. Each scan is processed in the same way as the THuman2.0 data. Additionally, all SMPL-X parametric models are standardized to neutral gender with refinement in body shape parameters to ensure the body layer fits underneath the cloth surface.

4.1. Evaluation of Generation Quality and Disentanglement Capability

Disentanglement and Animation Capacities. The disentanglement and animation results are shown in Fig. 1. Our method successfully generates clothed avatars with full disentanglement of components such as hair, shoes, and clothes from the human body. Moreover, our method reconstructs the inner body layer with reasonable geometry and texture under severe occlusion. Benefiting from the SMPL-X-based templates within our layered Gaussian-based UV feature plane representation, each component can be easily deformed into novel poses.

Comparisons. For holistic generation quality, we compare our method against representative feed-forward diffusion and 3D-aware GAN pipelines. The quantitative comparison results are shown in Tab. 1a. Our method outperforms all baselines on THuman2.0 [74] dataset, achieving the best FID score. The visual comparisons shown in Fig. 3 further enhance the superiority of our method. EVA3D [21] struggles with generating digital avatars with plausible texture and geometry. Compared to StructLDM [25] and E3Gen [80], our method exhibits finer details, such as distinctive fingers, realistic cloth wrinkles, and intricate faces.

To evaluate disentanglement capability, we compare layer-wise generation results and disentangled avatar generation results of our method with state-of-the-art methods. For layer-wise generation, our method surpasses other methods in both FID and L-PSNR, as shown in Tab. 1b,

Methods	FID↓
EVA3D [21]	124.54†
StructLDM [25]	25.22†
E3Gen [80]	15.78⋆
Ours	**12.50**

(a) Holistic Generation Quality.

Methods	FID↓	L-PSNR↑
EVA3D [21]	61.58*	<20*
Rodin [65]	56.57*	18.12*
HumanLiff [23]	54.39*	28.57*
Ours	**17.37**	**>40**

(b) Layer-wise Generation Quality.

Methods	Disentanglement↑	Quality↑	Time↓
SO-SMPL [63]	13.02	3.57/8.57	5h
LAGA [18]	9.77	5.00/1.43	1.5h
TELA [15]	19.07	14.29/8.57	6h
Ours	58.14	77.14/81.43	2s

(c) Component Quality and Disentanglement Evaluation.

Table 1. Quantitative comparison. Our method outperforms other methods in both holistic and layer-wise generation quality, as well as disentanglement capability. Due to the minor difference between the two masked layers, the L-PSNR value of our method is so large that we use > 40 to express it. For component comparison, we show the quality preference in *component quality preference/overall quality preference* format. $*$, † and $\star$ denote results adopted from HumanLiff [23], StructLDM [25] and E3Gen [80] respectively.

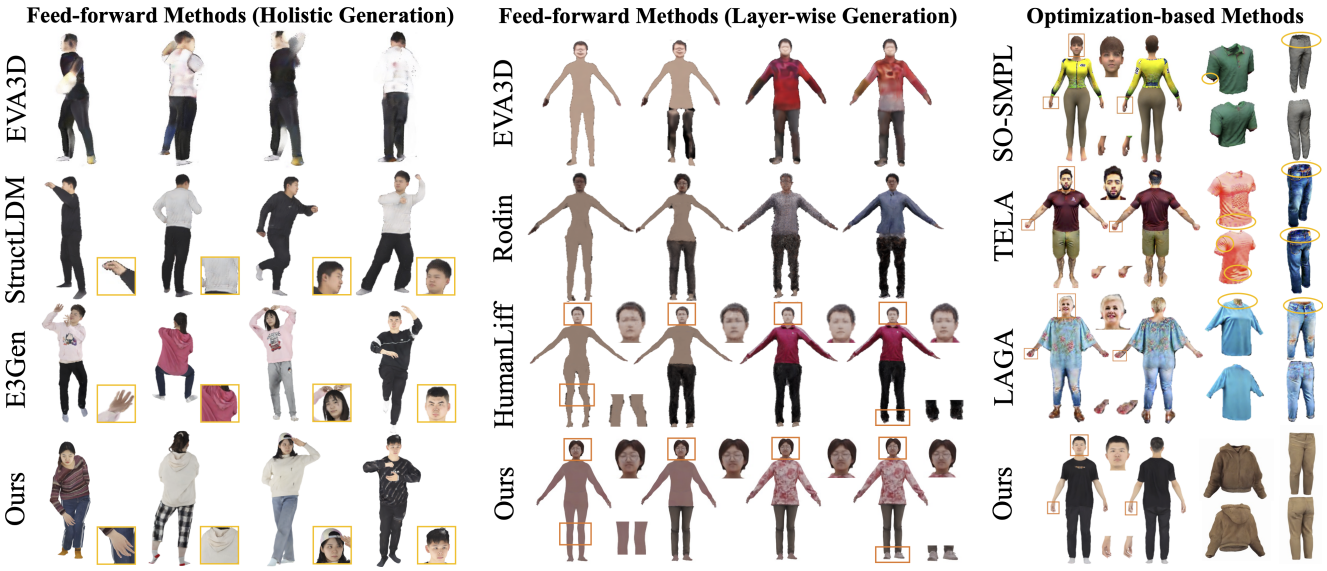


Figure 3. Qualitative comparison. The left part demonstrates our generation results on THuman2.0 dataset [74]. The middle part illustrates layer-wise generated results on Tightcap dataset [9]. Our method can generate more reasonable human bodies under severe occluded scenarios. The right part exhibits our component generation results compared with optimization-based methods.

which indicates that our method achieves higher holistic generation results and can generate disentangled avatars without facing the identity shifting problem. The elimination of identity shifting demonstrates that our method achieves full disentanglement of components, since entangled components will result in shifting problems. As shown in Fig. 3, Rodin [65] struggles to maintain identity consistency during the generation process. Although HumanLiff [23] achieves correct layer-wise generation, it produces less smooth body geometry and exhibits subtle identity shifting, shown by the enlarged face region. Our method can generate each layer without shifting problems, demonstrating more complete disentanglement.

For disentangled avatar generation, we compare our method with state-of-the-art optimization-based methods. As shown in Tab. 1c, our method generates clothed avatars in significantly less time and achieves a higher user preference. Qualitative comparisons are illustrated in Fig. 3. SO-SMPL [63] generates unrealistic, cartoonish colors and exhibits unnatural sawtooth on the border of generated components (highlighted by orange circles). LAGA [18] and TELA [15] struggle with blurry fingers, oversaturated colors, and incomplete disentanglement from the body layer. In contrast, our method generates fully disentangled avatars with distinctive fingers and realistic color.

Methods	FID↓	KID ↓
Two-stage	19.06	15.40
Pipeline (SLMO)	30.83	30.18
Pipeline (SLMN)	28.95	29.21
Full pipeline	**12.50**	**9.39**

Table 2. Ablation study on THuman2.0 Dataset. The full pipeline outperforms baselines on both FID and KID by a large margin.

4.2. Ablation Study

Layered *vs*. Single-Layer Representation.

To demonstrate the effectiveness of our layered UV feature plane, we design two baselines that use single-layer UV feature plane to generate disentangled clothed avatars.

Figure 4. Ablation study on THuman2.0 Dataset. Comparing randomly generated avatars and decomposition results, our method generates avatars with higher quality and better disentanglement.

One baseline representation is SLMO (single-layer-multi-output), which utilizes a single pair of geometry and texture decoders ($\mathcal{D}_g$ and $\mathcal{D}_t$) to predict attributes for all components. Another one is SLMN (single-layer multi-network), which employs different decoders to predict the attributes of Gaussian primitives for each component. Since they only contain a single-layer UV feature plane, attributes of different components will share one feature if they are initialized in the same location. As shown in Fig. 4, Fig.5 and Tab. 2, single-layer UV representations generate clothed avatars with lower quality both quantitatively and qualitatively. This is because our layered UV feature plane representation provides each component with an optimized UV distribution and independent feature space. Without the layered UV plane, which provides distinctive features for components in overlapping and neighboring regions, single-layer representations struggle to disentangle components, leading to blurred boundaries between components.

Single Stage *vs*. Two Stage. We also compare the single-stage training scheme with the commonly used two-stage training scheme, shown in Tab. 2 and Fig. 4. The single-stage training can generate avatars with finer details, such as intricate faces.

4.3. Applications

Component Transfer. We further explore applications such as component transfer. Thanks to the component disentanglement and shared structure provided by our method, we can directly transfer clothes and other components between generated samples. The component transfer results are shown in Fig. 6. Our method can accurately transfer components across various body shapes while maintaining high-quality details of the transferred items.

Figure 5. Zoom in comparison between E3Gen and LayerAvatar. Compared to single-layer representation, layered representation exhibits detailed faces and clear boundaries between components.

Figure 6. Component transfer. Given the generated avatars in the first column, we can transfer the upper-clothes, pants, hair, and shoes of the avatars in the second to fifth column to them. The results are shown in the rightmost column.

5. Conclusion

In this paper, we propose LayerAvatar, a novel feed-forward diffusion-based method for generating component-disentangled clothed avatars. Our method proposes a layered UV feature plane representation, which organizes 3D Gaussians into different layers, each corresponding to specific components of clothed avatars (e.g., body, hair, clothing). Leveraging this representation, we train a single-stage diffusion model to generate each feature plane, enabling the generation of fully disentangled clothed avatars. To ensure complete component disentanglement, we incorporate constraint terms into the model. Extensive experiments validate the superiority of LayerAvatar in generating fully disentangled clothed avatars and its effectiveness in component transfer tasks. Additional limitations and implementation details are discussed in the supplementary material.

Acknowledgments. This work was supported in part by NSFC (62201342), and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

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, 2023. 3

[2] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabián Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Trans. Graph.*, 40(4), 2021. 1

[3] Alexander W. Bergman, Petr Kellnhofer, Wang Yifan, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. In *NeurIPS*, 2022. 2

[4] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proc. CVPR*, 2021. 2

[5] 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, 6

[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. 2, 5, 6

[7] Honghu Chen, Yuxin Yao, and Juyong Zhang. Neural-abc: Neural parametric models for articulated body with clothes. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[8] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22246–22256, 2023. 2

[9] Xin Chen, Anqi Pang, Yang Wei, Wang Peihao, Lan Xu, and Jingyi Yu. Tightcap: 3d human shape capture with clothing tightness field. *ACM Transactions on Graphics (Presented at ACM SIGGRAPH)*, 2021. 2, 6, 7

[10] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[11] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J. Black, and Otmar Hilliges. Fast-snarf: A fast deformer for articulated neural fields. *Pattern Analysis and Machine Intelligence (PAMI)*, 2023. 3

[12] Zhaoxi Chen, Fangzhou Hong, Haiyi Mei, Guangcong Wang, Lei Yang, and Ziwei Liu. Primdiffusion: Volumetric primitives diffusion for 3d human generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[13] Hang Chu, Shugao Ma, Fernando De la Torre, Sanja Fidler, and Yaser Sheikh. Expressive telepresence via modular codec avatars. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 330–345. Springer, 2020. 1

[14] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *CVPR*, 2021. 2

[15] Junting Dong, Qi Fang, Zehuan Huang, Xudong Xu, Jingbo Wang, Sida Peng, and Bo Dai. Tela: Text to layer-wise 3d clothed human generation. *arXiv preprint arXiv:2404.16748*, 2024. 2, 3, 6, 7

[16] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. AG3D: Learning to Generate 3D Avatars from 2D Image Collections. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 5

[17] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 3, 5

[18] Jia Gong, Shenyu Ji, Lin Geng Foo, Kang Chen, Hossein Rahmani, and Jun Liu. Laga: Layered 3d avatar generation and customization via gaussian splatting. *arXiv preprint arXiv:2405.12663*, 2024. 2, 3, 6, 7

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[20] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21024–21035, 2023. 2, 6

[21] Fangzhou Hong, Zhaoxi Chen, Yushi LAN, Liang Pan, and Ziwei Liu. EVA3d: Compositional 3d human generation from 2d image collections. In *International Conference on Learning Representations*, 2023. 2, 3, 6, 7

[22] Fangzhou Hong, Jiaxiang Tang, Ziang Cao, Min Shi, Tong Wu, Zhaoxi Chen, Shuai Yang, Tengfei Wang, Liang Pan, Dahua Lin, et al. 3dtopia: Large text-to-3d generation model with hybrid diffusion priors. *arXiv preprint arXiv:2403.02234*, 2024. 2

[23] Shoukang Hu, Fangzhou Hong, Tao Hu, Liang Pan, Haiyi Mei, Weiye Xiao, Lei Yang, and Ziwei Liu. Humanliff: Layer-wise 3d human generation with diffusion model. *arXiv preprint*, 2023. 2, 3, 6, 7

[24] Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20418–20431, 2024. 4

[25] Tao Hu, Fangzhou Hong, and Ziwei Liu. Structldm: Structured latent diffusion for 3d human generation, 2024. 2, 6, 7

[26] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992. 5

[27] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[28] 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. 2, 4

[29] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024. 5, 6

[30] Byungjun Kim, Patrick Kwon, Kwangho Lee, Myunggi Lee, Sookwan Han, Daesik Kim, and Hanbyul Joo. Chupa: Carving 3d clothed humans from skinned shape priors using 2d diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15965–15976, 2023. 3

[31] Nikos Kolotouros, Thiemo Alldieck, Enric Corona, Eduard Gabriel Bazavan, and Cristian Sminchisescu. Instant 3d human avatar generation using image diffusion models. *arXiv preprint arXiv:2406.07516*, 2024. 3

[32] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6517–6526, 2024. 2

[33] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[34] Siyou Lin, Zhe Li, Zhaoqi Su, Zerong Zheng, Hongwen Zhang, and Yebin Liu. Layga: Layered gaussian avatars for animatable clothing transfer. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[35] 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. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2

[36] Yufei Liu, Junshu Tang, Chu Zheng, Shijie Zhang, Jinkun Hao, Junwei Zhu, and Dongjin Huang. Clothedreamer: Text-guided garment generation with 3d gaussians. *arXiv preprint arXiv:2406.16815*, 2024. 5

[37] Zhen Liu, Yao Feng, Michael J. Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling. In *International Conference on Learning Representations*, 2023. 2

[38] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. Coap: Compositional articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13201–13210, 2022. 3

[39] 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

[40] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. *arXiv preprint arXiv:2407.21686*, 2024. 3

[41] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4328–4338, 2023. 2

[42] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 2

[43] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[44] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *European Conference on Computer Vision*, 2022. 2, 3

[45] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela Dai. Spams: Structured implicit parametric models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12851–12860, 2022. 3

[46] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[47] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2

[48] Bo Peng, Yunfan Tao, Haoyu Zhan, Yudong Guo, and Juyong Zhang. Pica: Physics-integrated clothed avatar. *arXiv preprint arXiv:2407.05324*, 2024. 3

[49] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 5

[50] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9914–9925, 2024. 2

[51] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 6

[54] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. 6

[55] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[56] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2

[57] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20875–20886, 2023. 2

[58] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. Danbo: Disentangled articulated neural body representations via graph neural networks. In *European Conference on Computer Vision*, 2022. 3

[59] Xiaokun Sun, Zhenyu Zhang, Ying Tai, Qian Wang, Hao Tang, Zili Yi, and Jian Yang. Barbie: Text to barbie-style 3d avatars. *arXiv preprint arXiv:2408.09126*, 2024. 2

[60] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4

[61] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2

[62] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 4

[63] Jionghao Wang, Yuan Liu, Zhiyang Dou, Zhengming Yu, Yongqing Liang, Xin Li, Wenping Wang, Rong Xie, and Li Song. Disentangled clothed avatar generation from text descriptions, 2023. 2, 6, 7

[64] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 2

[65] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4563–4573, 2023. 7

[66] Yi Wang, Jian Ma, Ruizhi Shao, Qiao Feng, Yu-Kun Lai, and Kun Li. Humancoser: Layered 3d human generation via semantic-aware diffusion model. *arXiv preprint arXiv:2408.11357*, 2024. 2

[67] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[68] 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*, pages 16210–16220, 2022. 3

[69] 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. *arXiv preprint arXiv:2405.14832*, 2024. 2

[70] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, et al. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *ACM Transactions on Graphics (TOG)*, 41 (6):1–15, 2022. 3

[71] Yinghao Xu, Wang Yifan, Alexander W Bergman, Menglei Chai, Bolei Zhou, and Gordon Wetzstein. Efficient 3d articulated human generation with layered surface volumes. *arXiv preprint arXiv:2307.05462*, 2023. 3

[72] Xingguang Yan, Han-Hung Lee, Ziyu Wan, and Angel X. Chang. An object is worth 64x64 pixels: Generating 3d object via image diffusion, 2024. 2

[73] Yichao Yan, Zanwei Zhou, Zi Wang, Jingnan Gao, and Xiaokang Yang. Dialoguenerf: Towards realistic avatar face-to-face conversation video generation. *Visual Intelligence*, 2 (1):24, 2024. 3

[74] 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. 2, 6, 7

[75] Zhengming Yu, Zhiyang Dou, Xiaoxiao Long, Cheng Lin, Zekun Li, Yuan Liu, Norman Müller, Taku Komura, Marc Habermann, Christian Theobalt, et al. Surf-d: High-quality surface generation for arbitrary topologies using diffusion models. *arXiv preprint arXiv:2311.17050*, 2023. 2

[76] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[77] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. Avatargen: A 3d generative model for animatable human avatars. In *Arxiv*, 2022. 2

[78] 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 (TOG)*, 43(4):1–20, 2024. 2

[79] Muxin Zhang, Qiao Feng, Zhuo Su, Chao Wen, Zhou Xue, and Kun Li. Joint2human: High-quality 3d human generation via compact spherical embedding of 3d joints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[80] Weitian Zhang, Yichao Yan, Yunhui Liu, Xingdong Sheng, and Xiaokang Yang. e^3gen: Efficient, expressive and editable avatars generation. *arXiv preprint arXiv:2405.19203*, 2024. 2, 4, 6, 7

[81] 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. 2

[82] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, BIN FU, Tao Chen, Gang YU, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[83] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. In *International Conference on 3D Vision (3DV)*, 2025. 3