Dream, Lift, Animate:
From Single Images to Animatable Gaussian Avatars

Marcel C. Buehler
ETH Zurich

Ye Yuan
NVIDIA

Xueting Li
NVIDIA

Yangyi Huang
CUHK

Koki Nagano
NVIDIA

Umar Iqbal
NVIDIA

https://research.nvidia.com/labs/dair/dream-lift-animate

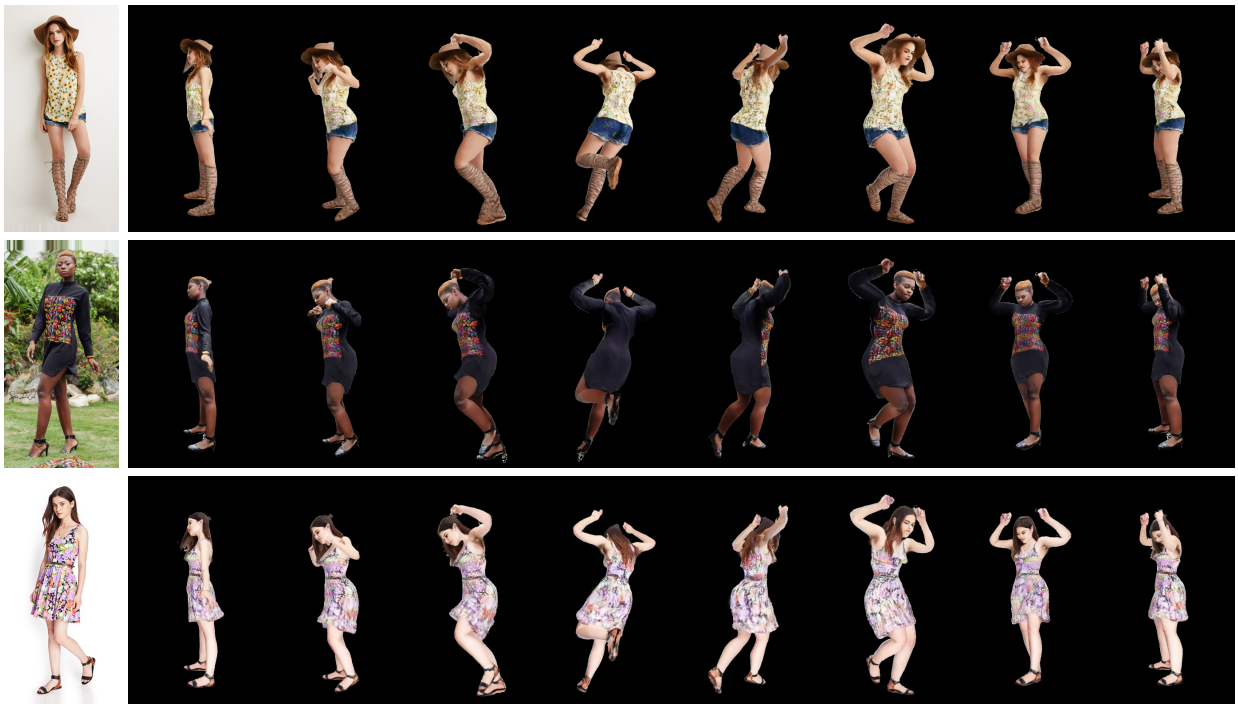

Figure 1. We propose **Dream**, **Lift**, **Animate**, a novel framework to reconstruct high-fidelity, animatable 3D human avatars from a single image by generating multi-view images, lifting them to 3D Gaussians, and mapping them to a pose-aware UV space (Fig. 2). Our approach enables realistic animation and outperforms prior methods in visual quality (Fig. 5 and Tbl. 1).

Abstract

We introduce Dream, Lift, Animate (DLA), a novel framework that reconstructs animatable 3D human avatars from a single image. This is achieved by leveraging multi-view generation, 3D Gaussian lifting, and pose-aware UV-space mapping of 3D Gaussians. Given an image, we first dream plausible multi-views using a video diffusion model, capturing rich geometric and appearance details. These views are then lifted into unstructured 3D Gaussians. To enable animation, we propose a transformer-based encoder that models global spatial relationships and projects these Gaussians into a structured latent representation aligned with the UV space of a parametric body model. This latent code is decoded into UV-space Gaussians that can be animated via body-driven deformation and rendered conditioned on pose and viewpoint. By anchoring Gaussians to the UV manifold, our method ensures consistency during

animation while preserving fine visual details. DLA enables real-time rendering and intuitive editing without requiring post-processing. Our method outperforms state-of-the-art approaches on the ActorsHQ and 4D-Dress datasets in both perceptual quality and photometric accuracy. By combining the generative strengths of video diffusion models with a pose-aware UV-space Gaussian mapping, DLA bridges the gap between unstructured 3D representations and high-fidelity, animation-ready avatars.

1. Introduction

Creating photorealistic 3D human avatars from monocular RGB images remains a fundamental and challenging problem in computer vision and graphics, with wide-ranging applications in gaming, telepresence, virtual try-on, and digital content creation. Achieving high-quality avatar reconstruction from a single image requires addressing three interrelated challenges. First, missing appearance details arising from self-occlusions or limited camera viewpoints must be plausibly and photorealistically hallucinated. Second, sparse 2D information from the input image must be reliably lifted into a geometrically coherent and accurate 3D representation. Third, and most critically, the resulting avatars must be readily animatable, enabling realistic and artifact-free motion synthesis under novel poses and viewpoints while maintaining consistent geometry, texture, and view-dependent effects.

Existing methods typically fall short in simultaneously addressing these challenges. Recent approaches [20, 22, 72] leverage the powerful generative capabilities of multi-view diffusion models [53] to hallucinate missing appearance information; however, these models inherently produce inconsistencies across generated views, leading to avatars that appear overly smoothed, lack critical details, or introduce visually jarring artifacts when animated. Conversely, methods relying on template-based rigging approaches, such as automatically transferring skinning weights from canonical human body models (e.g., SMPL [41]), typically require careful fitting procedures [26, 27, 59, 79]. While powerful, these methods can still encounter challenges when handling non-standard poses, complex clothing, and significant self-occlusions frequently encountered in unconstrained, real-world scenarios, thus affecting their robustness and broader applicability.

To overcome these limitations, we propose **Dream, Lift, Animate (DLA)**, a novel framework designed specifically for reconstructing high-quality, animatable 3D human avatars from a single image. Our method addresses the core reconstruction problem by decomposing it into three complementary stages, each carefully designed to handle limitations inherent in the preceding steps. Specifically, we first leverage a pretrained video diffusion model [66] to **Dream** plausible multi-view observations of the subject, effectively hallucinating realistic geometric and appearance details even in regions unseen from the input viewpoint. Despite the significant progress made by recent diffusion-based approaches, their inherent view-to-view inconsistencies prevent direct avatar reconstruction [72]. To resolve this, we next **Lift** these generated views into an intermediate, unstructured 3D representation based on 3D Gaussian primitives [34] in the input pose space, effectively aggregating appearance and geometry cues from the multi-view observations. Subsequently, to enable animation, we map these unstructured Gaussians into a structured UV-space representation. For this, we use a transformer-based encoder [74] that models global spatial relationships among the unstructured 3D Gaussians and projects them into a structured latent avatar representation, which aligns with the UV space of the SMPL-X body model [48]. This learned mapping also allows our model to reconcile inconsistencies introduced during the initial multi-view generation seamlessly. Finally, in the **Animate** stage, we present a pose- and view-aware Gaussian parameter decoder that converts this latent avatar representation into a UV-space map of Gaussian parameters, enabling spatially coherent and high-fidelity animation through linear blend skinning-based Gaussian deformation.

We demonstrate the advantages of our approach through extensive experiments on the challenging ActorsHQ [30], 4D-Dress [65], and SHHQ [16] datasets. Our method achieves state-of-the-art performance in terms of photometric accuracy, perceptual quality, and articulation realism.

In summary, the core contributions of our work are:

- A novel framework (**DLA**) for reconstructing animatable 3D human avatars from a single monocular RGB image, overcoming the inherent limitations of diffusion-based multi-view generation and template-based rigging approaches.
- A learned transformer-based encoder coupled with a pose- and view-conditioned Gaussian Parameter Decoder that maps unstructured 3D Gaussians into a structured UV-space representation, enabling high-fidelity animation.
- Extensive experiments and ablations demonstrating the effectiveness and versatility of our method, including one-shot reconstruction, realistic animation, and editing.

2. Related Work

One-Shot Human Reconstruction. Reconstructing 3D human avatars from a single image has been explored using both mesh-based and volumetric representations. Mesh-based approaches, including PiFU(-HD) [55, 56] and its extensions [3, 22, 57, 70, 71, 78, 79, 81], typically rely on implicit surfaces or hybrid models [27, 58, 73, 80] to reconstruct geometry and texture. While some decouple

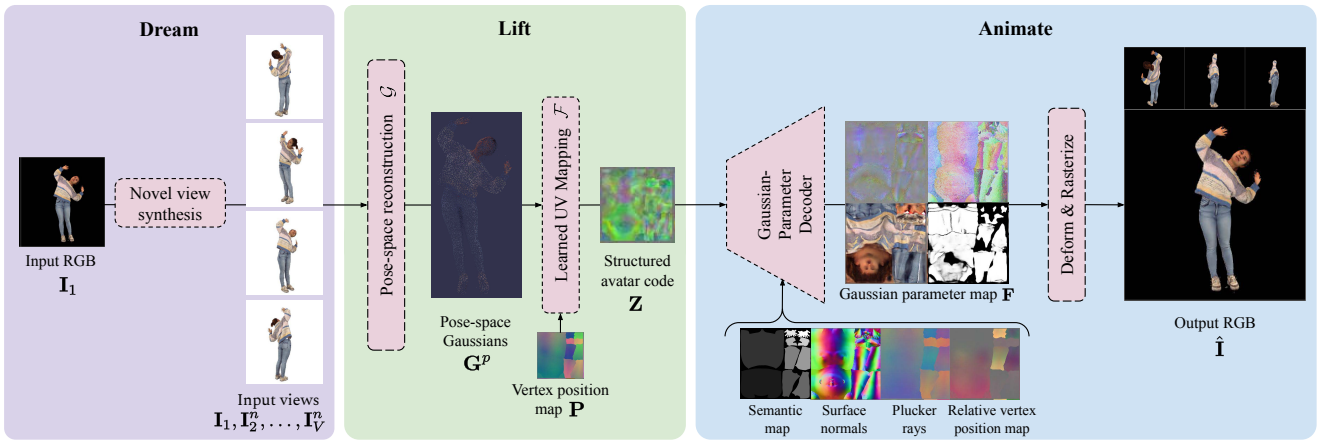

Figure 2. Overview of the proposed **Dream**, **Lift**, **Animate** (**DLA**) framework for reconstructing animatable 3D human avatars from a single image. In the **Dream** stage (Sec. 3.1), we synthesize novel views from the input using a diffusion-based generator. In the **Lift** stage (Sec. 3.2 and Fig. 3), we project the multi-view images into a set of unstructured 3D Gaussians in the *pose space* using a learned Gaussian reconstruction model $\mathcal{G}$. Subsequently, we learn a transformer encoder $\mathcal{F}$ to map 3D Gaussians to a structured latent code $\mathbf{Z}$ in the UV space of a parametric body model. In the **Animate** stage (Sec. 3.3 and Fig. 4), we decode the avatar code into a pose- and view-aware Gaussian parameter map $\mathbf{F}$. This structured representation enables realistic animation and rendering via deformation with a body model.

albedo and shading [3, 13, 57], mesh-based methods often struggle to model realistic view-dependent effects. Volumetric representations address these limitations. Works based on NeRFs [6, 7, 17, 24, 26, 68] achieve high fidelity but remain slow to render. Recent methods adopt 3D Gaussian Splatting (3DGS) [34] for efficient, real-time rendering. Human-focused variants [10, 40, 45, 60, 72] use diffusion-generated multiviews [63] to supervise Gaussian reconstruction. However, these typically model static pose-space geometry and lack support for animation under novel poses.

Animatable Avatars. To enable animation, early works like ARCH [19, 29] use canonicalization with LBS deformation, extended by feedforward models [24, 49]. These methods, however, depend on accurate body registration, which is often challenging for complex poses and clothing. IDOL [83] and LHM [51] (concurrent work) directly predict Gaussian splats on the SMPL-X template to support animation. However, their deformation module does not consider pose-dependent effects. In contrast, our approach conditions the Gaussian parameters on the target pose, enabling pose-dependent effects. In addition, our approach is substantially more lightweight because it decomposes the task into more tractable subproblems of multiview hallucination, 3D lifting, and UV-space mapping. While IDOL and LHM train on multiple nodes (32 NVIDIA A100/H100 GPUs), our model only requires a single node (8 NVIDIA A100 GPUs). Our lightweight design enables modularity and better handling of pose- and view-dependent appearance effects essential for animation.

Generative Priors. Generative models such as GANs [18] and diffusion models [39, 53] have also been used as strong priors for ill-posed problems such as single-image human reconstruction [2, 17, 22, 28, 38, 42, 45, 75, 79]. Recent methods apply diffusion models for view synthesis and geometry prediction [12, 22, 45, 72, 79]. Human3Diffusion [72] jointly trains a 3DGS generator with a multiview diffusion model, but remains limited by low resolution (256×256) and a lack of animation support. SiTH [22] and HumanSplat [45] synthesize unseen views using single- or multi-view diffusion, while SIFU [79] uses a transformer to predict geometry and refines textures with diffusion. Our method builds on both paradigms, i.e., we leverage a pretrained video diffusion model [66] to synthesize unseen views of the person, and we use a PatchGAN [31] to achieve higher realism and high-frequency details.

Generating 3D Avatars as Latents. Learning generative models of humans in a structured latent space enables controllable synthesis and supports downstream tasks such as few-shot reconstruction, inpainting, and editing. Early 3D methods map latent codes to implicit fields [4, 14, 23] or SDFs [44, 46, 77], often using tri-plane features and adversarial supervision. To improve pose control and spatial detail, recent works incorporate body priors [41, 48] with structured latents [1, 11, 23, 25, 35, 43]. However, these methods still fall short in texture realism and often rely on optimization-based inversion [52, 69] for image/text conditioning. In contrast, our method directly maps an input image to a compact, UV-aligned avatar latent that naturally supports animation, editing, and even interpolation.

3. Method

Fig. 2 provides a high-level overview of our proposed framework for reconstructing animatable 3D human avatars

from a single image. Given the input image $\mathbf{I}_1 \in \mathbb{R}^{H_I \times W_I \times C_I}$, our method proceeds in three stages: **Dream**, **Lift**, **Animate**. In the first stage, **Dream**, we employ a video diffusion model [66] to generate plausible multi-view images $\mathbf{I}_2^n, \ldots, \mathbf{I}_V^n$ from the input, addressing self-occlusions and incomplete viewpoints. Although visually compelling, these generated views often exhibit inconsistencies. Thus, in the second stage, **Lift**, we lift these multi-view images into a coherent set of unstructured pose-space 3D Gaussians $\mathbf{G}_1^p, \ldots, \mathbf{G}_K^p$ and use a novel transformer-based encoder to transform them into an animation-friendly structured latent representation $\mathbf{Z}$, which is aligned with the UV space of the SMPL-X body model [48]. Finally, in the third stage, **Animate**, a Gaussian Parameter Decoder predicts a Gaussian parameter map $\mathbf{F}$ in the UV space from the latent code $\mathbf{Z}$, target pose, and viewpoint conditions. We can then sample structured Gaussians $\mathbf{G}_1^s, \ldots, \mathbf{G}_N^s$ from the parameter map. These Gaussians can be readily animated via linear blend skinning (LBS) and rendered in real-time: $\hat{\mathbf{I}} = \mathcal{R}(\{\mathbf{G}_1^s, \ldots, \mathbf{G}_N^s\}, \Theta, \pi)$, where $\mathcal{R}$ is a rendering function that deforms the structured Gaussians $\mathbf{G}_1^s, \ldots, \mathbf{G}_N^s$ according to the target pose Θ using SMPL-X linear blend skinning, and rasterizes them with camera π.

3.1. Dream: Multi-view Generation

As illustrated in Fig. 2, our method begins by generating a set of multi-view images consisting of the original input view $\mathbf{I}_1$ and a set of novel views $\mathbf{I}_2^n, \ldots, \mathbf{I}_V^n$. During training, these views are sourced directly from multi-view datasets with known camera calibrations. At inference time, however, we synthesize novel views using a ControlNet-guided variant [66] of a video diffusion model [64]. Specifically, we first estimate the SMPL-X parameters from the input image [20] and render 2D skeletal poses of the predicted mesh from virtual cameras placed around a 360-degree azimuth. These projected poses serve as control signals to guide the diffusion model in generating photorealistic images from novel viewpoints. While this approach enables hallucination of previously unseen regions and recovers occluded appearance details, the resulting views may still exhibit 3D inconsistencies due to artifacts inherent to the generative diffusion process. Please see the supplementary material for more details.

3.2. Lift: Unstructured Gaussian and Latent Avatar Code Generation

Fig. 3 illustrates how we lift the multiview observations into a unified 3D avatar representation in two sequential substages. We reconstruct per-view 3D Gaussians in the input pose space and map these unstructured 3D Gaussians into a structured latent avatar code in the UV space of the SMPL-X body model, which is more amenable to animation.

Unstructured Gaussians Reconstruction. We employ a pose-space reconstruction model $\mathcal{G}$ to map the multiview images into a set of pixel-aligned 3D Gaussians. The model uses a U-Net-based architecture augmented with cross-view self-attention, following the design of the Large Gaussian Model (LGM) [60]. The resulting Gaussians from all views are then fused into a single set of unstructured 3D Gaussians. Reconstructing Gaussians directly in the input pose space is a tractable and effective strategy, as it avoids the highly non-linear mappings required to directly predict avatar geometry in canonical or UV coordinates. This design choice allows our method to faithfully recover rich appearance and geometric details from the input views while remaining robust to inconsistencies.

Latent Avatar Code Generation. While informative, these unstructured Gaussians are not directly amenable to animation because they lack a consistent topology and structure. To address this, we introduce a transformer-based encoder $\mathcal{F}$ that converts the unstructured Gaussians into a structured latent avatar code $\mathbf{Z}$. The latent code is aligned with the UV space of the SMPL-X model [48], which supports expressive deformation and animation via linear blend skinning (LBS). As illustrated in Fig. 3, we first extract per-Gaussian features by concatenating each Gaussian's raw parameters (e.g., color, position) with intermediate U-Net [54] activations and linearly project them to an embedding space. These U-Net features are pixel-aligned by construction. The per-Gaussian features are then filtered and downsampled with farthest-point sampling [15] into a more compact feature $\mathbf{X} \in \mathbb{R}^{P \times C_p}$ where P is the number of filtered Gaussians and C_p is the combined dimensionality of the Gaussian parameters and U-Net features.

To associate this representation with the SMPL-X UV space, we use a UV-space vertex position map $\mathbf{P} \in \mathbb{R}^{3 \times H_p \times W_p}$ generated by deforming the SMPL-X mesh with the input pose Θ_I and rasterizing the mesh vertex coordinates into UV space. This position map is positionally encoded and used as queries in a cross-attention layer, where the Gaussian features $\mathbf{X}$ serve as keys and values. The result is a compact, spatially-structured latent code $\mathbf{Z}$ aligned with the UV manifold of the SMPL-X body.

This design provides several critical advantages. First, reconstructing Gaussians in the input pose space and lifting them into UV space decouples local appearance estimation from structural reasoning, simplifying both tasks. Second, the feed-forward nature of our lifting pipeline supports fast inference and enables end-to-end differentiable training. Finally, by leveraging a learned reconstruction model and a structured UV mapping, our method effectively absorbs view inconsistencies and artifacts in the generated multiview inputs, producing a coherent and animatable intermediate representation.

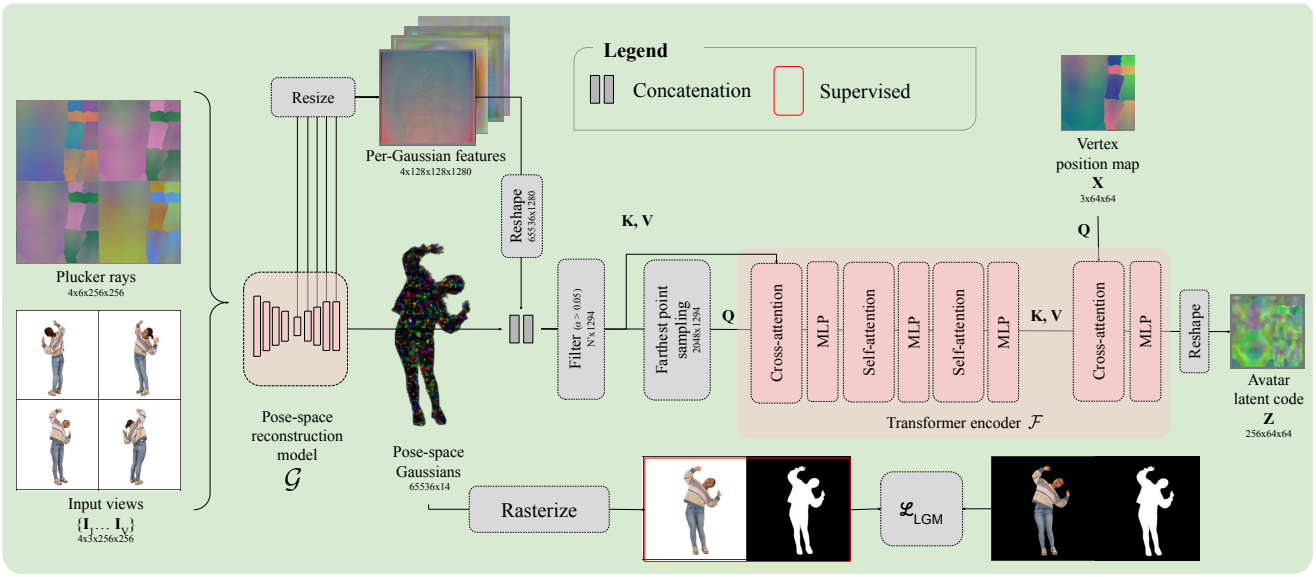

Figure 3. ***Lift*** *from multiview images to an avatar latent code*. The pose-space reconstruction model produces pixel-aligned Gaussian parameters with corresponding feature maps, denoted *pose-space Gaussians* and *per-Gaussian features*. The Gaussians are filtered and subsampled to construct a compact Gaussian feature $\mathbf{X}$ with 2048 Gaussians. These inputs are further processed by a transformer. Specifically, the compact Gaussian feature serves as context (key $\mathbf{K}$ and value $\mathbf{V}$) to a cross-attention layer with queries $\mathbf{Q}$ being the positionally encoded vertex position map $\mathbf{P}$. Finally, the output is reshaped and yields the avatar latent code $\mathbf{Z}$. This figure omits linear projections, skip-connections, and positional encoding for improved readability.

3.3. Animate: Structured Gaussian Generation and Deformation

Gaussian Parameter Decoder. Building upon the UV-aligned latent avatar code $\mathbf{Z}$ produced in the previous stage, we now decode this representation into a fully animatable Gaussian-based avatar. As shown in Fig. 4, the Gaussian Parameter Decoder (GPD) maps the avatar latent code $\mathbf{Z}$ to a UV space Gaussian parameter map, from which structured 3D Gaussians can be sampled. Since convolutional networks greatly benefit from pixel-aligned input features, we design the GPD as a spatially-adaptive ConvNet [47, 82]. The GPD is conditioned on the UV-aligned latent $\mathbf{Z}$ and a one-hot encoded UV segmentation map [48]. We condition with spatially-adaptive normalization [47] because pixelwise normalization has been shown to provide great results for generating images that are aligned to semantic maps [8, 9, 47, 82]. At the highest resolution, the GPD branches out into a *canonical* branch and a *pose- and view-dependent* branch. The pose- and view-dependent branch receives additional inputs with information about the target geometry (surface normals), viewpoint (Plucker rays [32]), and pose (relative vertex position map w.r.t. to the neutral pose). These inputs are created by instantiating the SMPL-X body model with the given target pose and rasterizing its relative vertex location, normals, and plucker rays to the UV space. The inputs are again encoded via spatially adaptive normalization and condition the pose- and view-dependent branch. The two branches are merged with an addition. The output is a UV space Gaussian parameter map $\mathbf{F} \in \mathbb{R}^{H_G \times W_G \times 14}$, which contains the 3D Gaussian parameters, including alphas, colors, as well as positions, rotations and scales defined in the UV tangent space. The two-branch design enables fast inference by caching the canonical branch. Please refer to Sec. 4.2 for more information.

Structured Gaussian Generation and Rendering. We generate the structured 3D Gaussians $\{\mathbf{G}_1^s, \ldots, \mathbf{G}_N^s\}$ by sampling the UV space Gaussian parameter map $\mathbf{F}$ at uniform locations in the UV space and on the mesh surface. Each Gaussian $\mathbf{G}_i^s$ is defined in the UV tangent coordinates and contains the position Δ_i^{xyz}, rotation Δ_i^{R}, scale Δ_i^{s}, alpha α_i, and color $\mathbf{c}_i$. To obtain the Gaussians in the world space, we deform the SMPL-X body mesh using the target pose Θ via linear blend skinning, which yields the UV tangent space coordinates (T_i, R_i, S_i), which consist of position T_i, rotation R_i, and scale S_i. We obtain the final position, rotation, and scale of each Gaussian by

$$\mathbf{G}_i^{\text{xyz}} = T_i + R_i \Delta_i^{\text{xyz}}, \quad \mathbf{G}_i^{\text{R}} = R_i \Delta_i^{\text{R}}, \quad \mathbf{G}_i^{\text{s}} = S_i \Delta_i^{\text{s}} \quad (1)$$

and rasterize the avatar using Gaussian Splatting [34].

3.4. Training

Training presents significant challenges as the model must learn a complex mapping from 2D images through an in-

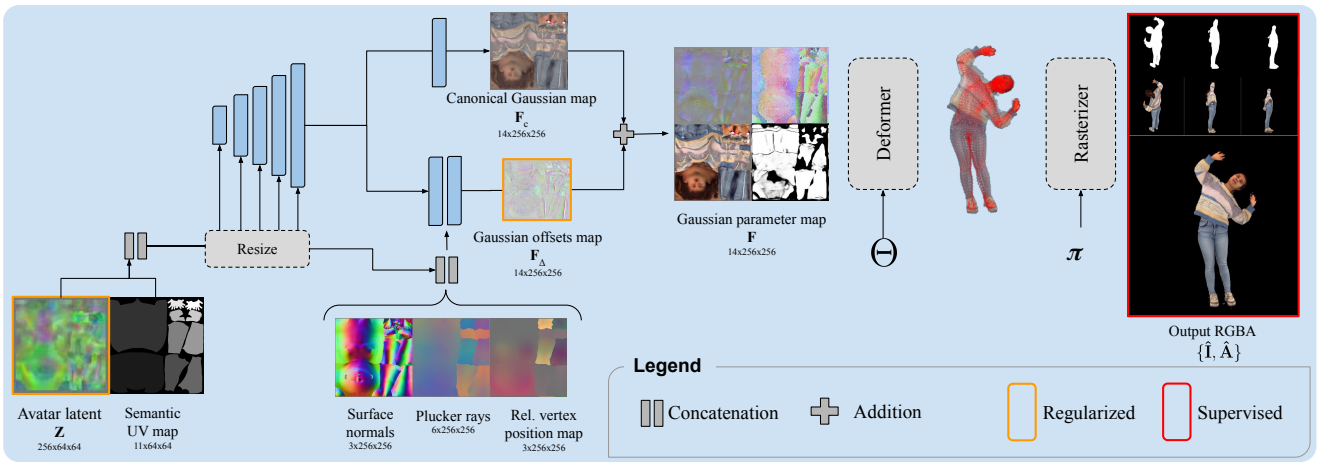

Figure 4. **Animate**. The Gaussian Parameter Decoder (GPD, Sec. 3.3) maps a UV-aligned latent $\mathbf{Z}$ to an animatable 3D Gaussian representation. The GPD upsamples the avatar latent code $\mathbf{Z}$ and produces two output maps: a *canonical* Gaussian map $\mathbf{F}_c$ and an *offset* map $\mathbf{F}_\Delta$. The offset map $\mathbf{F}_\Delta$ adds pose- and view-dependent offsets to the canonical Gaussian $\mathbf{F}_c$, enabling pose- and view-dependent effects. Given a pose Θ and camera π, the Gaussians are deformed with linear blend skinning [48] and rasterized to an RGBA image [34].

termediate UV space to a 3D Gaussian parameter representation. We provide the model with ground truth inputs with 90-degree yaw angle differences and task the model with reconstructing novel viewpoints. Given the Gaussian avatar produced by GPD (Sec. 3.3), we render it using the ground-truth SMPL-X parameters and compute the following losses:

$$\mathcal{L}_{\text{GPD}} = \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} + \lambda_{\text{VGG}}\mathcal{L}_{\text{VGG}} \tag{2}$$

$$+ \lambda_{\text{Mask}}\mathcal{L}_{\text{Mask}} + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}} \tag{3}$$

$$+ \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{C}}\mathcal{L}_{\text{C}}, \tag{4}$$

where the L1 reconstruction loss $\mathcal{L}_{\text{L1}} = \|\mathbf{I}^* - \hat{\mathbf{I}}\|_1$ measures the absolute difference between the rendered output and ground truth image. For mask supervision, we compute $\mathcal{L}_{\text{Mask}} = \|\mathbf{M}^* - \hat{\mathbf{A}}\|_1$, which enforces consistency between the alpha maps from Gaussian splatting $\hat{\mathbf{A}}$ and the ground truth masks $\mathbf{M}^*$. To enhance perceptual quality, we incorporate a GAN loss $\mathcal{L}_{\text{GAN}}$ using a Patch Discriminator [31] with least squares optimization. The perceptual VGG loss $\mathcal{L}_{\text{VGG}}$ leverages an AlexNet [36] backbone with features masked by the ground truth mask $\mathbf{M}^*$. For regularization, we apply a Kullback-Leibler Divergence term $\mathcal{L}_{\text{KL}}$ to constrain the avatar latent maps, and $\mathcal{L}_{\text{C}} = \|\mathbf{F}_\Delta\|_2$ encourages minimal offsets in the Gaussian offsets map.

We also provide intermediate supervision to the pose-space reconstruction model $\mathcal{G}$ (Sec. 3.2). Since the unstructured reconstructed Gaussians are already in pose-space, we splat them and apply the following losses:

$$\mathcal{L}_{\text{UG}} = \lambda^{\text{UG}}_{\text{VGG}}\mathcal{L}^{\text{UG}}_{\text{VGG}} + \lambda^{\text{UG}}_{\text{Mask}}\mathcal{L}^{\text{UG}}_{\text{Mask}}, \tag{5}$$

where $\mathcal{L}^{\text{UG}}_{\text{VGG}}$ and $\mathcal{L}^{\text{UG}}_{\text{Mask}}$ are the VGG loss and mask loss computed on the reconstructed images, respectively. The total loss then becomes: $\mathcal{L} = \mathcal{L}_{\text{GPD}} + \mathcal{L}_{\text{UG}}$. Fig. 3 and Fig. 4 highlight the supervised outputs in red and the regularized features in orange.

4. Experiments

We compare our approach with state-of-the-art methods [22, 61, 79, 83] for human reconstruction from a single image, demonstrating superior performance in both novel view synthesis and animation tasks (Tbl. 1 and Fig. 5). We then showcase the versatility of our framework through various applications like animation (Fig. 1), editing and interpolation (Fig. 7). Finally, we conduct extensive ablation studies (Tbl. 2) to validate our design choices and analyze the impact of different components on the overall performance.

4.1. Comparison with State-of-the-art

We compare with several state-of-the-art methods for one-shot human reconstruction [22, 61, 79, 83]. We measure perceptual (LPIPS) [76] and structural similarity (SSIM) [67], and photometric accuracy (PSNR) for novel view synthesis in Tbl. 1. Fig. 5 shows visual results. Fig. 6 demonstrates animation capabilities.

Method	LPIPS ↓	PSNR ↑	SSIM ↑
DreamGaussian [61]	0.1514	19.48	0.8837
SiTH [22]	0.1577	18.88	0.8764
SIFU [79]	0.1408	19.42	0.8823
IDOL [83]	0.0696	24.48	0.9261
Ours	**0.0580**	**25.58**	**0.9279**

Table 1. Quantitative results on ActorsHQ [30]. We compare novel view synthesis on the input image. Please see Fig. 5 for visuals. The supplementary contains comparisons for novel pose synthesis and comparisons on the 4D-Dress dataset [65].

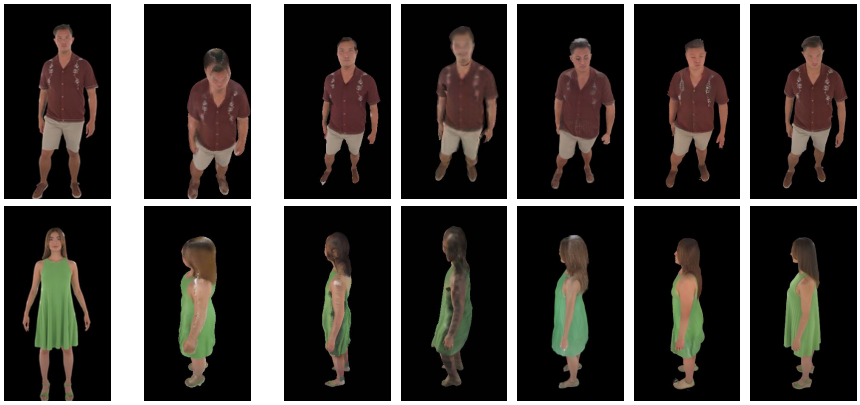

Figure 5. Comparison for novel view synthesis with DreamGaussian [61], SiTH [22], SIFU [79], and IDOL [37]. Tbl. 1 lists metrics.

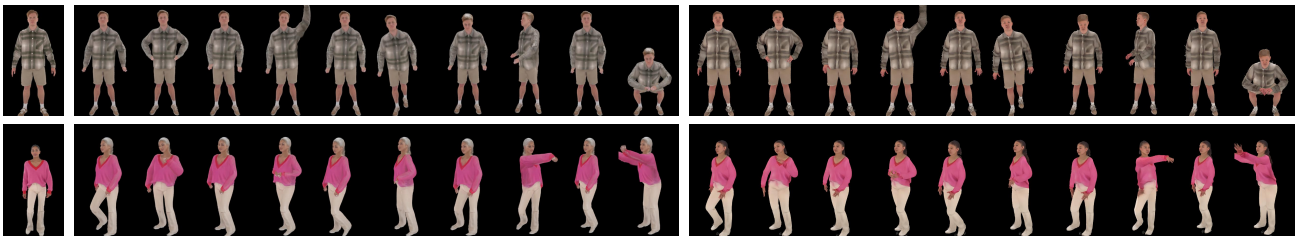

Figure 6. Comparison for animation. Our DLA framework enables detailed renderings for difficult poses, outperforming the state-of-the-art in one-shot animatable avatars in perceptual and photometric metrics. Please see the supp. mat. for more examples and metrics.

Our quantitative evaluation in Tbl. 1 demonstrates superior performance in novel view synthesis. We compare our approach against state-of-the-art methods by taking a single image as input and rendering it from multiple unseen viewpoints, as visualized in Fig. 5. The metrics reported in Tbl. 1 confirm our method's effectiveness across all evaluation criteria. In the following, we discuss the results taking the limitations of the state-of-the-art into consideration.

DreamGaussian [61] employs score distillation sampling [50] to produce static Gaussians that can be converted into a mesh with texture refinement, which does not account for view-dependent effects. SiTH [22] and SiFU [79] also reconstruct textured meshes using implicit functions, but require additional rigging (e.g., with Mixamo) for animation. IDOL [83] offers fast inference but sacrifices quality, as it directly predicts a Gaussian map in the UV space. This presents two drawbacks: first, it attempts to solve two complex steps (Dream and Lift) simultaneously; second, their Gaussians are not conditioned on the target pose, ignoring pose-dependent effects. These limitations result in reduced reconstruction quality, as evidenced in Tbl. 1 and Fig. 5.

Furthermore, we showcase our animation capabilities on ActorsHQ [30] in Fig. 6 and on in-the-wild image [16] in Figures 1. The results demonstrate high-quality pose transfer while preserving the avatar's geometric and appearance. Please see the supplementary material for a comparison for novel pose synthesis, comparisons on the 4D-Dress dataset [65], and experimental details regarding the comparisons.

4.2. Applications

The structured nature of our UV-space Gaussian Parameter Map, decoded from the avatar latent code $\mathbf{Z}$, enables a range of downstream applications. Beyond reconstruction and animation, it supports controllable editing and we observe emergent properties like smooth interpolations.

Editing. Our structured latent representation $\mathbf{Z}$ is spatially aligned with the UV space of the human body, allowing semantic and part-aware manipulations. This alignment enables intuitive avatar editing by performing localized operations on $\mathbf{Z}$. For instance, we can swap specific regions between the latent codes of two different avatars. The decoder then faithfully reconstructs coherent and photorealistic avatars reflecting these edits. This facilitates flexible avatar customization and compositional synthesis (see Fig. 7, bottom).

Emerging Capabilities. Although our framework is not designed as a generative model, we observe that the learned latent space $\mathbf{Z}$ exhibits smooth structure similar to that of generative style spaces [5, 33, 47, 53, 62]. As a result, interpolating between the latent codes of different avatars yields plausible and continuous transitions between identities. Qualitative results are shown in Fig. 7 (top).

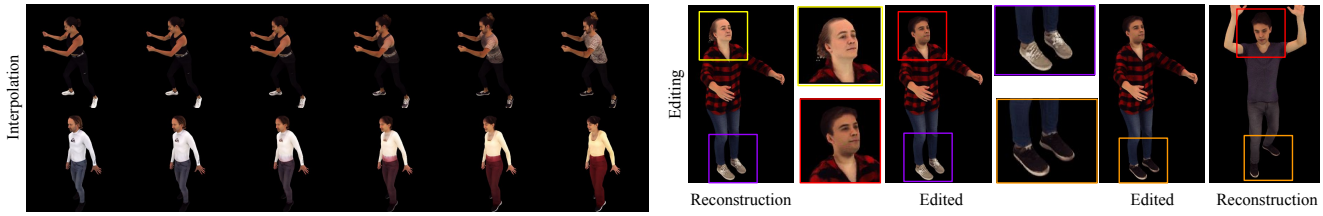

Figure 7. Applications. Our structured latent code affords editing like face swapping and virtual try-on of shoes (right). In addition, we observe emerging capabilities like smooth interpolations between avatar latent codes (left). These examples are reconstructions using multi-view inputs from CustomHumans [21].

Rendering Speed. The design of our Gaussian Parameter Decoder (Fig. 4) allows real-time rendering while considering pose- and view-dependent effects. When rendering novel poses or views for a given subject, we cache the canonical Gaussian map $\mathbf{F}_c$ because it does not depend on the pose and viewpoint—only the offset map $\mathbf{F}_\Delta$ needs to be computed for each frame. This enables rendering frames of resolution 512×512 at 33 FPS (30 ms per frame) on a single NVIDIA RTX 5880.

4.3. Ablations

We report metrics for several ablation studies in Tbl. 2. The first set of ablations examines techniques for mapping pose-space Gaussians from an unstructured cloud to a structured UV space (A). In ablation A.i, we use the SMPL-X body model to directly project pixel-aligned Gaussians to the UV space, similar to RGB texture unprojection when reconstructing textured meshes from multi-view inputs. This approach lacks learning capabilities and fails to account for shapes beyond the body model (like clothing), resulting in a significant drop in perceptual quality (27% decrease in LPIPS). We also experiment with learning the query (A.ii) instead of using the vertex position map ($\mathbf{X}$ in Fig. 3) and found it to yield worse results. We conclude that the vertex position map introduces valuable prior knowledge from the body model, which is missing in the learned query. The second ablation (B) finds a positive impact of view- and pose-conditional inputs in the Gaussian Parameter Decoder (Fig. 4). The third set of ablations (C) considers different inputs to our full model. For inference, our encoder takes the ground-truth frontal input image $\mathbf{I}_1$, three synthesized images $\{\mathbf{I}_2^n, ..., \mathbf{I}_V^n\}$, and estimated SMPL-X parameters as input. We study the effect of noisy SMPL-X parameters in the transformer encoder (C.i), which is particularly relevant for in-the-wild settings where fitting the body model might be challenging. Replacing the ground-truth input image $\mathbf{I}_1$ with the reconstruction from the diffusion model slightly reduces performance (C.ii). Conversely, we simulate the potential of next-generation video diffusion models by feeding only ground-truth novel views as input, which leads to a substantial performance gain (C.iii), indicating room for further improvement by just replacing the video diffusion model with an improved version, thanks to our our modular design. We provide more detailed ablations in the supp. mat.

	LPIPS ↓	PSNR ↑	SSIM ↑
A. UV mapping			
i) Mesh unprojection (no learning)	0.0790	23.66	0.9209
ii) Learned Query	0.0652	24.67	0.9254
B. Animate & render			
No conditioning	0.0644	24.86	0.9248
C. Inputs			
i) Noisy SMPL-X inputs	0.0660	24.77	0.9240
ii) 4 synthetic inputs	0.0657	24.78	0.9237
iii) 4 GT inputs	0.0613	25.71	0.9310
Ours	0.0624	25.09	0.9254

Table 2. Ablation studies (Sec. 4.3).

5. Conclusion

We introduced a novel framework for high-quality 3D human avatar reconstruction from a single image. Our approach, Dream, Lift, Animate, leverages a multiview diffusion model to *dream* plausible unseen viewpoints, which are then *lifted* into an unstructured 3D Gaussian representation via a large feed-forward reconstruction model. A key contribution of our work is a transformer-based encoder that maps these unstructured Gaussians into a structured UV-space representation. This structured form enables realistic animation, fine-grained control, and intuitive editing, while also exhibiting emergent properties such as smooth identity interpolation. Through extensive experiments, we demonstrate that our method achieves state-of-the-art performance in both perceptual quality and photometric accuracy across multiple benchmarks.

Our method has several limitations that warrant discussion. Close body parts may cause color leakage between adjacent regions (e.g., hand skin color affecting nearby clothing). Moreover, while our method can handle slight inconsistencies in the generated multiview images, it cannot handle significant inconsistencies. Finally, reconstructed avatars may exhibit inconsistencies in identity, primarily in facial features. This stems from training dataset biases (THuman2.0, CustomHumans, and 4D-Dress) and the low resolution of face regions in the input. Training our method on in-the-wild monocular images is an interesting future work.

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*, 2023. 3

[2] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia*, 2023. 3

[3] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *CVPR*, 2022. 2, 3

[4] Alexander Bergman, Petr Kellnhofer, Wang Yifan, Eric Chan, David Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. *Advances in Neural Information Processing Systems*, 35:19900–19916, 2022. 3

[5] Marcel C. Buehler, Abhimitra Meka, Gengyan Li, Thabo Beeler, and Otmar Hilliges. Varitex: Variational neural face textures. In *CVPR*, 2021. 7

[6] Marcel C Buehler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, et al. Preface: A data-driven volumetric prior for few-shot ultra high-resolution face synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3402–3413, 2023. 3

[7] Marcel C Buehler, Gengyan Li, Erroll Wood, Leonhard Helminger, Xu Chen, Tanmay Shah, Daoye Wang, Stephan Garbin, Sergio Orts-Escolano, Otmar Hilliges, et al. Cafca: High-quality novel view synthesis of expressive faces from casual few-shot captures. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3

[8] Marcel Buhler, Seonwook Park, Shalini De Mello, Xucong Zhang, and Otmar Hilliges. Content-consistent generation of realistic eyes with style. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1–5. IEEE, 2019. 5

[9] Marcel C Buhler, Andrés Romero, and Radu Timofte. Deepsee: Deep disentangled semantic explorative extreme super-resolution. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 5

[10] Jinnan Chen, Chen Li, Jianfeng Zhang, Lingting Zhu, Buzhen Huang, Hanlin Chen, and Gim Hee Lee. Generalizable human gaussians from single-view image. *arXiv preprint arXiv:2406.06050*, 2024. 3

[11] 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

[12] Armand Comas-Massagué, Di Qiu, Menglei Chai, Marcel Bühler, Amit Raj, Ruiqi Gao, Qiangeng Xu, Mark Matthews, Paulo Gotardo, Sergio Orts-Escolano, et al. Magicmirror: Fast and high-quality avatar generation with a constrained search space. In *European Conference on Computer Vision*, pages 178–196. Springer, 2024. 3

[13] Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Structured 3d features for reconstructing relightable and animatable avatars. In *CVPR*, 2023. 3

[14] 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 *ICCV*, 2023. 3

[15] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y Zeevi. The farthest point strategy for progressive image sampling. *IEEE transactions on image processing*, 6(9):1305–1315, 1997. 4

[16] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu. StyleGAN-Human: A data-centric odyssey of human generation. In *ECCV*, 2022. 2, 7

[17] Xiangjun Gao, Xiaoyu Li, Chaopeng Zhang, Qi Zhang, Yanpei Cao, Ying Shan, and Long Quan. Contex-human: Free-view rendering of human from a single image with texture-consistent synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10084–10094, 2024. 3

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Adv. Neural Inform. Process. Syst.*, 2014. 3

[19] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *CVPR*, 2021. 3

[20] Xu He, Xiaoyu Li, Di Kang, Jiangnan Ye, Chaopeng Zhang, Liyang Chen, Xiangjun Gao, Han Zhang, Zhiyong Wu, and Haolin Zhuang. Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement, 2024. 2, 4

[21] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *CVPR*, 2023. 8

[22] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *CVPR*, 2024. 2, 3, 6, 7

[23] Fangzhou Hong, Zhaoxi Chen, Yushi LAN, Liang Pan, and Ziwei Liu. EVA3d: Compositional 3d human generation from 2d image collections. In *ICLR*, 2023. 3

[24] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *ICCV*, 2023. 3

[25] Tao Hu, Fangzhou Hong, and Ziwei Liu. Structldm: Structured latent diffusion for 3d human generation. In *ECCV*, 2024. 3

[26] Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debing Zhang, and Deng Cai. One-shot implicit animatable avatars with model-based priors. In *ICCV*, 2023. 2, 3

[27] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided reconstruction of lifelike clothed humans. In *3DV*, 2024. 2

[28] Yangyi Huang, Ye Yuan, Xueting Li, Jan Kautz, and Umar Iqbal. Adahuman: Animatable detailed 3d human generation with compositional multiview diffusion, 2025. 3

[29] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable reconstruction of clothed humans. In *CVPR*, 2020. 3

[30] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 2, 6, 7

[31] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 3, 6

[32] Yan-Bin Jia. Plücker coordinates for lines in the space. *Problem Solver Techniques for Applied Computer Science, Com-S-477/577 Course Handout*, 2020. 5

[33] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 7

[34] 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, 3, 5, 6

[35] Nikos Kolotouros, Thiemo Alldieck, Enric Corona, Eduard Gabriel Bazavan, and Cristian Sminchisescu. Instant 3d human avatar generation using image diffusion models. In *ECCV*, 2024. 3

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 2017. 6

[37] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *https://arxiv.org/abs/2311.06214*, 2023. 7

[38] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Mengfei Li, Xiaowei Chi, Siyu Xia, Wei Xue, et al. Pshuman: Photorealistic single-view human reconstruction using cross-scale diffusion. *arXiv preprint arXiv:2409.10141*, 2024. 3

[39] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 3

[40] Zhibin Liu, Haoye Dong, Aviral Chharia, and Hefeng Wu. Human-vdm: Learning single-image 3d human gaussian splatting from video diffusion models, 2024. 3

[41] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 2, 3

[42] Yixing Lu, Junting Dong, Youngjoong Kwon, Qin Zhao, Bo Dai, and Fernando De la Torre. Gas: Generative avatar synthesis from a single image. *arXiv preprint arXiv:2502.06957*, 2025. 3

[43] Yifang Men, Biwen Lei, Yuan Yao, Miaomiao Cui, Zhouhui Lian, and Xuansong Xie. En3d: An enhanced generative model for sculpting 3d humans from 2d synthetic data. In *CVPR*, 2024. 3

[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. 3

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

[46] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3

[47] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 5, 7

[48] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2, 3, 4, 5, 6

[49] Hao-Yang Peng, Jia-Peng Zhang, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. Charactergen: Efficient 3d character generation from single images with multi-view pose canonicalization. *ACRM Trans. Graph.*, 43(4):1–13, 2024. 3

[50] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. In *ICLR*, 2022. 7

[51] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, and Liefeng Bo. Lhm: Large animatable human reconstruction model from a single image in seconds. In *ICCV*, 2025. 3

[52] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 3

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 7

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 4

[55] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 2

[56] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 2

[57] Akash Sengupta, Thiemo Alldieck, Nikos Kolotouros, Enric Corona, Andrei Zanfir, and Cristian Sminchisescu. DiffHuman: Probabilistic Photorealistic 3D Reconstruction of Humans. In *CVPR*, 2024. 2, 3

[58] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid represen-

tation for high-resolution 3d shape synthesis. In *NeurIPS*, 2021. 2

[59] David Svitov, Dmitrii Gudkov, Renat Bashirov, and Victor Lempitsky. Dinar: Diffusion inpainting of neural textures for one-shot human avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7062–7072, 2023. 2

[60] 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. In *ECCV*, 2024. 3, 4

[61] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. DreamGaussian: Generative gaussian splatting for efficient 3d content creation. In *ICLR*, 2024. 6, 7

[62] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Adv. Neural Inform. Process. Syst.*, 2017. 7

[63] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitrii Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[64] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 4

[65] Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4d-dress: A 4d dataset of real-world human clothing with semantic annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 6, 7

[66] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *arXiv preprint arXiv:2406.01188*, 2024. 2, 3, 4

[67] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[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 (CVPR)*, pages 16210–16220, 2022. 3

[69] Yiqian Wu, Hao Xu, Xiangjun Tang, Xien Chen, Siyu Tang, Zhebin Zhang, Chen Li, and Xiaogang Jin. Portrait3d: Text-guided high-quality 3d portrait generation using pyramid representation and gans prior. *ACM Trans. Graph.*, 43(4), 2024. 3

[70] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit clothed humans obtained from normals. In *CVPR*, 2022. 2

[71] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit clothed humans optimized via normal integration. In *CVPR*, 2023. 2

[72] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Human 3diffusion: Realistic avatar creation via explicit 3d consistent diffusion models. In *NeurIPS*, 2024. 2, 3

[73] Yifan Yang, Dong Liu, Shuhai Zhang, Zeshuai Deng, Zixiong Huang, and Mingkui Tan. Hilo: Detailed and robust 3d clothed human reconstruction with high-and low-frequency information of parametric models. In *CVPR*, 2024. 2

[74] Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Trans. Graph.*, 42(4), 2023. 2

[75] Jingbo Zhang, Xiaoyu Li, Qi Zhang, Yanpei Cao, Ying Shan, and Jing Liao. Humanref: Single image to 3d human generation via reference-guided diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1844–1854, 2024. 3

[76] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 6

[77] 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

[78] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[79] Zechuan Zhang, Zongxin Yang, and Yi Yang. SIFU: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *CVPR*, 2024. 2, 3, 6, 7

[80] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13545–13555, 2022. 2

[81] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction, 2021. 2

[82] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *CVPR*, 2020. 5

[83] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image, 2024. 3, 6, 7