Disco4D: Disentangled 4D Human Generation and Animation from a Single Image

Hui En Pang[1], Shuai Liu[3], Zhongang Cai[1,2,3], Lei Yang[2,3],
Tianwei Zhang[1], Ziwei Liu[1]
[1]S-Lab, Nanyang Technological University [2]SenseTime Research [3]Shanghai AI Laboratory

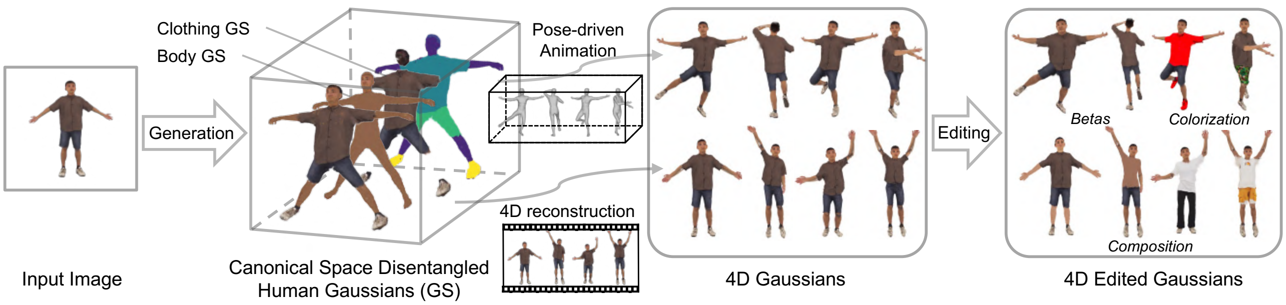

Figure 1. **Disco4D is a novel Gaussian Splatting framework for 4D disentangled human generation from a single image. Clothing and assets are modeled as Gaussians on top of the SMPL-X body, enabling dynamic animation and flexible editing.**

Abstract

*We present **Disco4D**, a novel Gaussian Splatting framework for 4D human generation and animation from a single image. Different from existing methods, Disco4D distinctively disentangles clothings (with Gaussian models) from the human body (with SMPL-X model), significantly enhancing the generation details and flexibility. Specifically, **1)** Disco4D learns to efficiently fit the clothing Gaussians over the SMPL-X Gaussians. **2)** Next, Disco4D adopts diffusion models to enhance the 3D generation process, e.g., modeling occluded parts not visible in the input image. **3)** Finally, Disco4D learns an identity encoding for each clothing Gaussian to facilitate the separation and extraction of clothing assets. Furthermore, Disco4D naturally supports 4D human animation with vivid dynamics. Extensive experiments demonstrate the superiority of Disco4D on 4D human generation and animation tasks. Our code is available at https://github.com/disco-4d/Disco4D.*

1. Introduction

The development of high-fidelity 3D digital humans is increasingly important across a variety of augmented and virtual reality applications. To streamline the creation of these digital avatars from easily accessible in-the-wild images, a multitude of research efforts have been made on reconstructing 3D clothed human models from a single image [2, 33, 40–42, 81, 82, 103, 104, 104, 117]. These works predominantly focus on the simultaneous reconstruction of the human body and clothing. Unfortunately, these works have inherent limitations, and integrating them into applications that require virtual try-on or avatar customization poses significant challenges. This is primarily because the models are rendered as single-layer, non-animatable meshes where distinct attributes (e.g., hair, clothing, accessories) are merged into one continuous surface, with underlying layers completely obscured and self-contact areas inseparably connected. Such limitation complicates the re-animation and dynamic customization tasks. Existing works that perform layered reconstruction [25, 26] rely on self-rotating video inputs with extensive frames and viewpoints and involve substantial processing times.

To address these issues, we propose Disco4D, a novel 4D clothed human reconstruction method that *distinctly separates the human body from clothing elements* from a single image. It supports human animation as the 4th dimension, which cannot be realized by prior static 3D reconstruction works [2, 33, 42, 81, 82, 103, 104, 117]. To achieve this, it employs the SMPL-X [71] parametric model to represent the human body, capitalizing on its efficacy in capturing body structure and kinematics. Conversely, clothing, along with dynamic and variable elements such as hair and accessories, is represented using Gaussian models, which are able to model the large variability in clothing. By binding Gaussians to a SMPL-X model and fixing it during the training phase, Disco4D ensures the integrity of the body while focusing the learning process on the appearance aspects. To model occluded portions not visible in the input image, diffusion models are used to enhance the 3D generation process. Moreover, Disco4D includes an identity grouping mechanism for the Gaussians, which is instrumental in maintaining the separability and individuality of each clothing asset.

The independent reconstruction of clothing and body offers several advantages. (1) *Enhanced reconstruction fidelity*. The SMPL-X body serves as a stable anchor for the clothing to conform to. By isolating the focus to learn clothing Gaussians, we achieve a more refined geometry and intricate detailing in the clothed model. (2) *Fine-grained categorization and extraction of clothing items*. Disco4D is able to separate clothing Gaussians into their respective categories, which is crucial for the recovery and utilization of individual clothing assets. (3) *Extensive editing capabilities*. Disco4D supports different editing functions, including the removal of specific items, inpainting (altering color or material), and other modifications. Such rich editing options allow for precise adjustments to individual assets without inadvertently affecting adjacent elements. This level of control is particularly beneficial in applications requiring detailed customization, such as virtual fashion design and digital content creation. (4) *Improved animation capabilities*. The body Gaussians adhere to the deformations dictated by the SMPL-X model, while clothing Gaussians conform to the underlying body movements but also exhibit behaviors true to their material characteristics. The disentangled deformation allows for nuanced adjustments to clothing behavior in response to complex body movements, thereby elevating the quality of clothed human animation.

2. Related Works

Table 1 summarizes the relevant 3D/4D generation methods. We describe their details below.

2.1. 3D Generation

Single-image 3D Generation. Single-image reconstruction leverages advanced methods [45, 67] to generate 3D assets in the form of 3D point clouds or NeRF [65] from one image. While earlier efforts using auto-encoders focused on synthetic objects [12, 14, 21, 88, 93, 105], newer approaches treat the task as conditional generation, employing diffusion models [35] for 3D generation from both image and text [19, 35, 60, 64, 74, 76, 79, 90]. One-2-3-45 [59] uses 2D diffusion models [60, 87] to generate multi-view images for reconstruction, while LRM [36] adopts transformer-based architecture to scale up the task on large datasets [19, 112]. Gaussian-based methods [47], particularly DreamGaussian [89] and LGM [91], offer efficient, high-resolution 3D model generation from text or images. Recently, video diffusion models have attracted significant attention due to their remarkable ability to generate intricate scenes and complex dynamics with great spatio-temporal consistency [4, 7–9, 31, 56, 116]. They are employed to generate consistent multi-view images, and then reconstruct underlying 3D assets with high quality [15].

Table 1. **3D/4D generation methods from a single image.**

Method	Type	Layered	Animatable
LGM [91]	General	✗	✗
PiFU [81]	Human-centric	✗	✗
DreamFusion [74]	General	✗	✗
DreamGaussian [89]	General	✗	✗
PiFU [81]	Human-centric	✗	✗
D-IF [107]	Human-centric	✗	✗
HiLo [108]	Human-centric	✗	✗
ECON [104]	Human-centric	✗	✗
SHERF [40]	Human-centric	✗	✓
Disco4D	Human-centric	✓	✓

Single-image human-centric 3D Generation. Significant research efforts have been made for 3D human reconstruction, which can be classified into the following categories. (1) *Explicit-shape-based methods* rely on Human Mesh Recovery (HMR) using parametric models like SMPL [62] and SMPL-X [71] to generate 3D body meshes [16, 17, 22, 24, 44, 46, 48, 49, 51, 66, 80, 118]. To account for 3D garments, several approaches incorporate offsets [99, 119] or templates, utilize deformable garment templates [6, 43], or employ non-parametric forms for clothed figures [27, 104, 115]. Despite their advancements, they face limitations in handling complex outfit variations and loose clothing due to inherent topological constraints. (2) *Implicit-function-based methods* utilize implicit representations like occupancy or distance fields for modeling clothed humans with complex geometries, such as loose garments. Techniques range from end-to-end regression of free-form implicit surfaces [2, 81, 82] to use of geometric priors [33, 42, 103, 104, 117] and implicit shape completion [104]. Notable works such as PIFu [81], ARCH(++) [33, 42], and PaMIR [117] can extract textured models from images, but struggle with depth ambiguities and texture inconsistencies. (3) *NeRF-based methods* incorporate model-based priors (i.e., SMPL-X) for accurate human reconstruction. Efforts like SHERF [40] and ELICIT [41] improve the reconstruction coherence by addressing 2D observation incompleteness leveraging appearance priors. Most of these 3D clothed human reconstruction and animation works [2, 33, 42, 81, 82, 103, 104, 117] require training on human-specific datasets, which brings another limitation on the availability of such datasets.

3D Clothing Modeling. Reconstructing clothing from images and videos as a separate layer over the human body poses significant challenges due to the diversity of clothing topologies. Previous efforts relied on either template meshes or implicit surface models, and required extensive, high-quality 3D data from simulations [5, 70, 83, 95] or tailored template meshes [13, 32, 73, 100]. New methods were developed [34, 43] for multi-clothing models and versatile template meshes, respectively, facilitating diverse clothing topology encoding. However, these techniques typically fall short in capturing the clothing texture and appearance. The reliance on predefined clothing style templates further constrains their ability to handle real-world clothing variations. Corona et al. [18] addressed these shortcomings by representing clothing layers with deep unsigned distance functions and an auto-decoder for style and cut differentiation, though this often produces overly-smooth reconstructions [18]. On the other hand, SCARF [25] and DELTA [26] significantly enhance the visual fidelity by applying NeRF to clothing layers, but require self-rotating video inputs and considerable processing times.

2.2. 4D Animation

4D Animation. This task aims at capturing dynamic 3D scenes over time. Two primary approaches have emerged: modeling 4D scenes by adding time dimension t or latent codes to spatial coordinates [28, 57, 98]; combining deformation fields with static 3D scenes [20, 58, 68, 69, 75, 92, 114]. Recent efforts in explicit or hybrid representations, like planar decomposition [11, 84, 85], hash representations [94], and other innovative methods [1, 23, 29], have improved reconstruction speed and quality. Gaussian Splatting, especially, stands out for balancing efficiency with quality, with dynamic 3D Gaussians [63] and 4D Gaussian Splatting [97, 110] techniques introducing time-dependent deformations to enhance reconstructions. Notably, DreamGaussian4D [78] stands out by minimizing the optimization time while achieving high-quality 4D reconstructions.

Human-centric 4D Animation. Recent works leverage Gaussian-based methods [38, 50, 54, 55, 61, 77, 113] for 4D human reconstruction, requiring extensive frame sequences (50-100 frames) and/or multiple viewpoints. Currently there has not been any work on 4D layered human generation and animation from a single image or a video with few images, which will be achieved in this paper.

3. Methodology

3.1. Preliminaries

3D Gaussian Splatting employs explicit 3D Gaussian points as its primary rendering entities. A 3D Gaussian point is defined as a function $G(x) = e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$, where μ and Σ are the spatial mean and covariance matrix, respectively. Each Gaussian is also associated with its own rotation r, scaling s, opacity α, a view-dependent color c represented by spherical harmonic coefficients f.

SMPL-X parameterization [71] is an extension of the SMPL body model [62] with face and hand, designed to capture a more accurate representation of intricate body movements. **SMPL-X** is defined as a function $M(\beta, \theta, \psi) : \mathbb{R}^{|\beta|\times|\theta|\times|\psi|} \rightarrow \mathbb{R}^{3N}$, parametrized by the pose $\theta \in \mathbb{R}^{3J}$ (where J denotes the number of full body joints), body shape $\beta \in \mathbb{R}^{|\beta|}$ and facial expression $\psi \in \mathbb{R}^{|\psi|}$.

3.2. Overview

Given a single image, Disco4D generates animatable 3D clothed human avatars in a bottom-up manner, facilitating natural separability. Our generated 3D clothed avatars, denoted as S_{human}, are represented as the concatenation of S_{body} and S_{cloth}. Inspired by prior works [89, 91], S capitalizes on Gaussian representations:

$$S = G(\mu, r, s, \alpha, c, e), \tag{1}$$

where μ, r, s, α, c and e denote *positions*, *rotation*, *scaling*, *opacity*, *spherical harmonics coefficients* and *identity encoding*, respectively. Different from traditional Gaussian representations, we add identity encoding e to associate each Gaussian with its clothing category.

Figure 2 depicts our framework. We start by generating colored SMPL-X Gaussians representing the body beneath clothing (Sec. 3.3). We obtain a visual hull for canonicalization and refine Gaussian predictions to align and envelop the SMPL-X mesh (Sec. 3.4). Next, we iteratively optimize canonical clothing Gaussians external to the SMPL-X mesh (Sec. 3.5). Lastly, we showcase the animation and editing of generated clothed avatars (Sec. 3.6). Notably, we leverage diffusion models to refine textures during 3D generation (Sec. 3.5) and extrapolate unseen views during 4D animation (Sec. 3.6).

3.3. SMPL-X Gaussians

Given an image, we first estimate coarse SMPL-X parameters with an off-the-shelf model [10], and then refine coarse predictions by fitting on 2D keypoints and clothing segmentation masks [72], obtaining pixel-aligned SMPL-X parameters (β, θ, ψ).

Mesh Binding. To convert the SMPL-X [71] mesh $M(\beta, \theta, \psi)$ into Gaussians S_{body} for rendering, flat 3D Gaussians are bound to each mesh triangle, similar to SuGaR [30]. Gaussian means μ_{body} are computed using predefined barycentric coordinates, while Gaussian rotations r_{body} derive from surface normals. The initial scaling s_{body} ensures dense mesh coverage, with the last axis set to 0.1 for a uniformly thin surface. For color representation beneath clothing, opacity α_{body} is set to 1.0, with spherical harmonics c_{body} optimized for each Gaussian. Visible skin color is supervised, while occluded skin color aligns with visible regions. A fixed label e_{body} is assigned for rendering, remaining unchanged during training. When optimizing clothing Gaussians S_{cloth}, SMPL-X Gaussians S_{body} parameters stay fixed, preserving the body structure while allowing flexible learning for clothing.

3.4. Initialization of Clothing Gaussians

Cloth styles are diverse, making proper initialization crucial for effective clothing modeling. In synchronization with

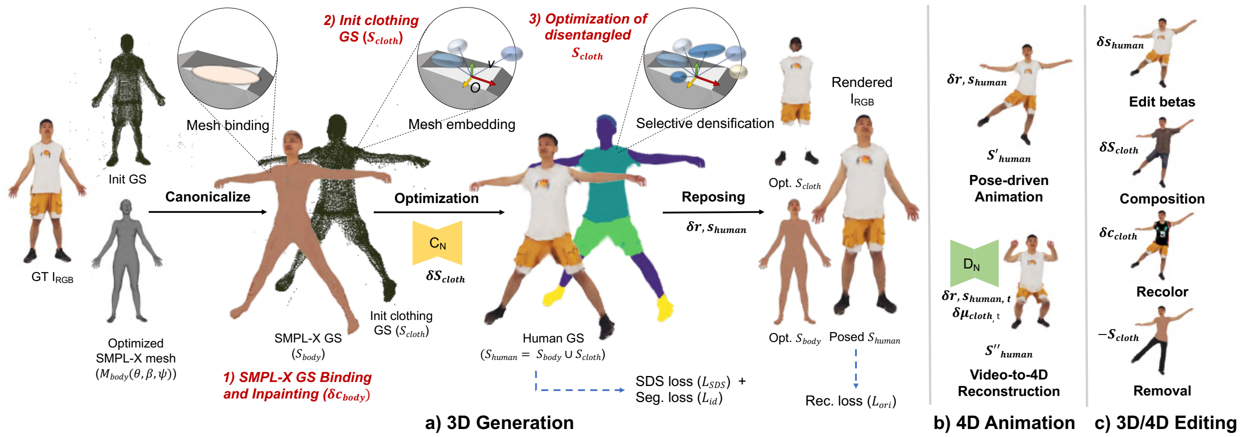

Figure 2. **Framework Overview of Disco4D. (a) 3D Generation** utilizes a single image to obtain disentangled body and clothing Gaussians. Body, face and hand poses are refined to be pixel-aligned. For faster initialization, clothing Gaussians and visual hull are obtained with Gaussian Reconstruction Models. These clothing Gaussians are embedded to SMPL-X mesh and adopt the local coordinate system of the triangle. Subsequently, the iterative optimization process (pruning, identity encoding and densifying) separates the body and garments. The learned identity encodings guide the densification of the clothing Gaussians. **(b) 4D Animation** is achieved by either direct driving of SMPL-X poses or leveraging video to learn extra clothing deformation (refer to Figure 3 for more details). Various **(c) 3D/4D Editing** operations can be performed with our disentangled representation.

estimating SMPL-X, we first employ the Video Diffusion Model [7] to estimate multi-view images. Subsequently, we leverage Gaussian Reconstruction Models [91] to obtain initial 3D Gaussians and their corresponding visual hull. Yet, the reconstructed 3D outputs often suffer from geometric inaccuracies, such as incorrect poses due to pose ambiguity or missing limbs. To address this, we refine the coarse visual hull to ensure it accurately aligns with and overlays the SMPL-X mesh and encapsulates a good geometry for the clothed figure. With SMPL-X aligned visual hull, we derive the refined Gaussians by adopting properties from their nearest neighbors. The refined visual hull and Gaussians are then canonicalized for the optimization phase.

Mesh embedding. Each 3D clothing Gaussian is embedded on a triangle of the canonical mesh, defining its position in both canonical and posed spaces. The mean vertex position $\mathbf{O}$ serves as the origin of the local coordinate system, with the Gaussian positioned by an offset vector $\mathbf{v} = \sigma\mathbf{i}+\beta\mathbf{j}+\gamma\mathbf{k}$, where σ, β, and γ are the components of the displacement vector along the tangent $\mathbf{i}$, bitangent $\mathbf{j}$, and normal $\mathbf{k}$. Unlike SplattingAvatar [86], which displaces Gaussians along the normal, our approach allows embedding to the most suitable triangle rather than the nearest one. For example, hair Gaussians are tagged to head faces instead of the nearest face for reposing [39, 86] (Figure 9 in Appendix). In animation, the Gaussian rotates with its embedded triangle face (δr), while scaling (δs) is adjusted dynamically based on changes in edge lengths. During optimization, Gaussian and embedding parameters ($\mathbf{O}$, $\mathbf{v}$, δr, and δs) are jointly updated.

3.5. Optimization of Separable Gaussians

With the SMPL-X Gaussian and initialized clothing Gaussian, we aim to optimize canonical clothing Gaussians S_{cloth} outside the SMPL-X mesh. This involves three steps: **1)** we use Signed Distance Function (SDF) loss and pruning to discourage and remove Gaussians that reside within the body; **2)** we introduce *identity encoding* e to attach a clothing label for each clothing Gaussian, by lifting multi-view 2D segmentations of the target object onto the 3D Gaussians; and **3)** guided by e_{body} and e_{cloth}, we selectively densify only the relevant clothing points while ignoring body points. Once the disentangled clothing is obtained, we use SDS loss to in-paint high-resolution texture from the reference image to individual clothing Gaussians, thereby enriching the details of unseen regions.

SDF Loss and Pruning. In reality, the clothing is always external to the body. During refinement, we ensure that the clothing Gaussians are positioned externally to the SMPL-X mesh by applying the SDF loss and a pruning strategy. Specifically, the SDF loss $\mathcal{L}_{sdf}$ penalizes any new densified Gaussians that intrude into the space of the SMPL-X mesh, ensuring that the clothing Gaussians consistently remain outside the body's surface. Pruning is applied at fixed intervals to reinforce this separation, and systematically remove any Gaussians located within the SDF of the SMPL-X mesh.

Identity encoding. To associate each Gaussian to its clothing category, we introduce *Identity Encoding* (e), a learnable and compact vector of length 15, representing clothing categories from SegFormer [101] segmentation masks[1]. During training, the encodings are rendered into 2D segmentation masks in a differentiable manner following [111]. For classification, we apply a softmax to the rendered features E_{id} and use cross-entropy loss $\mathcal{L}_{2d}$ for (K+1)-category classifica-

[1] Categories: 0: "Background", 1: "Hat", 2: "Hair", 3: "Sunglasses", 4: "Upper-clothes", 5: "Skirt", 6: "Pants", 7: "Dress", 8: "Belt", 9: "Left-shoe", 10: "Right-shoe", 11: "Face", 12: "Skin", 13: "Bag", 14: "Scarf"

tion. An unsupervised 3D regularization loss $\mathcal{L}_{3d}$ promotes spatial consistency among the top k-nearest 3D Gaussians' Identity Encodings. Consequently, the overall identity loss is $\mathcal{L}_{id} = \mathcal{L}_{2d} + \mathcal{L}_{3d}$. Refer to Appendix 7.2 for more details.

Densification of clothing Gaussians. To learn clothing more efficiently, we perform sampling for categorical Gaussians that belong to the same clothing category and embedding. We find the k-nearest Gaussian points for the resampled points and inherit their Gaussian properties (scaling, rotation, opacity, SH properties). By selectively densifying clothing Gaussians, we only add necessary Gaussians while ignoring body Gaussians.

Anisotropy. To prevent overly-skinny kernels that point outward from the object surface under large deformations, we enforce the anisotropy of Gaussian kernels following [102]. During optimization, we employ $\mathcal{L}_{\text{ani}} = \frac{1}{|P|}\sum_{p\in P}\max\left(\frac{\max(s_p)}{\min(s_p)}, \tau\right) - \tau$, where s_p is the scalings of 3D Gaussians. This loss constrains the ratio between the major and minor axis lengths below threshold τ.

Total loss. To inpaint occluded textures, we use the $\mathcal{L}_{SDS}$ loss on the Gaussians in the canonical pose after optimizing the front view for 500 steps. Combined with the conventional 3D Gaussian Loss $\mathcal{L}_{ori}$ on image rendering, the total loss L for end-to-end optimization of clothing Gaussians via network C_N is:

$$\mathcal{L} = \lambda_{ori}\mathcal{L}_{ori} + \lambda_{id}\mathcal{L}_{id} + \lambda_{ani}\mathcal{L}_{ani} + \lambda_{sdf}\mathcal{L}_{sdf} + \lambda_{SDS}\mathcal{L}_{SDS} \quad (2)$$

3.6. 4D Human Animation and Editing

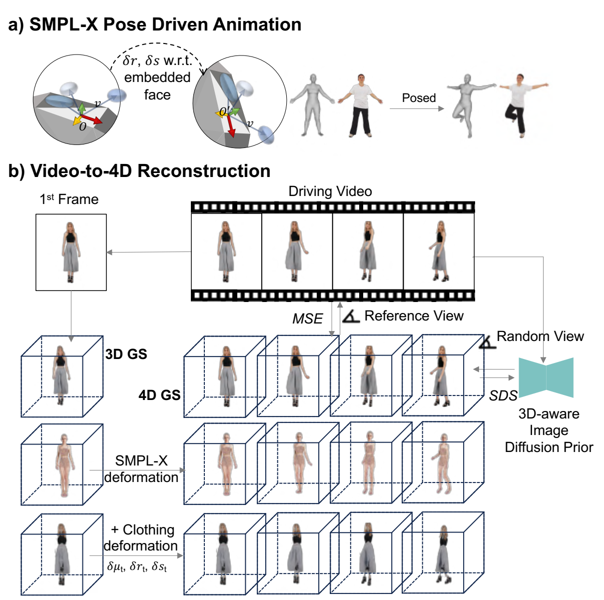

Figure 3. **4D animation** is achieved by (a) driving SMPL-X poses or (b) using video to learn additional clothing deformations. From the first frame, a static 3D disentangled GS model is generated. Pose transformations deform body and clothing Gaussians, and a deformation network is optimized to capture additional clothing deformations over time.

Disco4D's disentangled representation naturally supports animation and editing. The canonical Gaussians S_{body} and S_{cloth} enable separate deformations for clothing and body, ensuring realistic animation. Besides, individual clothing categories can be easily edited using image or text prompts. The learned clothing can be transferred to different body shapes and poses, for versatile customization.

Animating Gaussians. As shown in Figure 2, Disco4D enables animation of the canonical human Gaussian via two methods. Firstly, Gaussians can be directly driven using 3D SMPL-X sequences obtained from a motion database or estimated from 2D videos. Secondly, Disco4D enhances the model by learning detailed clothing dynamics from monocular videos. This disentanglement enables the focused modeling of clothing dynamics without altering the underlying human representation.

To extend static 3D Gaussians into dynamic 4D Gaussians, a deformation network is trained to predict changes in position, rotation, and scale of the reposed clothing Gaussians based on a timestamp, as described in DreamGaussian4D [78]. Unlike [78], which learns deformations for all Gaussians, Disco4D models body Gaussians using the SMPL-X mesh, while clothing Gaussians employ posed transformations and learned deformations. The transformation is defined as $S'' = D_N(S', t)$ where D_N is the deformation network, S' is the spatial descriptions of the reposed 3D clothing Gaussian, t is the timestamp, and S'' is the spatial descriptions of the deformed and reposed 3D clothing Gaussians. Following [78], the deformation model is initialized to predict zero deformation at the start of training to avoid divergence between dynamic and static models. The weights and biases of the final prediction heads are initialized to zero, and skip connections are introduced to enable gradient backpropagation.

To optimize the deformation field using the reference view video, we minimize the reconstruction loss $\mathcal{L}_{Ref}$ between the rendered image and video frame at each timestep. To propagate the motion from the reference view to the entire 3D model, we leverage Zero-1-to-3-XL [19] to predict the deformation of the unseen part to calculate $\mathcal{L}_{SDS}$. Despite per-frame predictions of image diffusion models, the fixed color and opacity of static 3D Gaussians help preserve temporal consistency.

Editing Clothing Gaussians. We extract the Gaussians corresponding to the specific category and edit them. This allows fine-grained editing and ensures that other Gaussians are not affected. Instead of fine-tuning all 3D Gaussians, we freeze the properties for most of the well-trained Gaussians and only adjust a small part of 3D Gaussians relevant to the target categories. For 3D object removal, we simply delete the 3D Gaussians of the editing target. For 3D object colorization by in-painting or text guidance, we reinitialise the color and tune the color (SH) parameters of the corresponding Gaussian group, while fixing the 3D positions and other properties to preserve the learned 3D geometry.

4. Experiments

Our detailed implementation and experiment setup can be found in Appendix 7.3.

4.1. 3D Generation

Generation and Disentanglement. Our generation and disentanglement results are presented in Figure 4 and Table 2. We assessed the disentanglement quality using the Synbody [109] and CloSe [3] datasets, rendering 30 and 110 clothed human meshes respectively from four angles and evaluating CLIP-similarity, PSNR, SSIM, and LPIPS for various poses and views within the CloSe dataset. Disco4D leverages diffusion models without requiring training on human specific datasets. Therefore, we compare it with DreamGaussian [89] and LGM [91] which reconstruct 3D objects from diffusion models. Additionally, we conducted comparisons with SHERF, a human-centric baseline for evaluating novel poses and views. Figure 4 shows Disco4D has higher fidelity and better geometry for body parts such as face and limbs due to the representation using SMPL-X Gaussians. It outperforms DreamGaussian and SHERF on SynBody and CloSe benchmarks. Disco4D performs worse than LGM on novel views, likely due to its optimization of Gaussians in canonical space for pose generalization, compromising view-specific detail.

Editing. We can edit specific clothing appearance given an image or text prompt, repose the person and transfer person characteristics. The disentanglement allows fine-grained editing and modification of individual assets without affecting other assets, and stacking multiple edits (Figure 4).

User study. We conducted a user study to evaluate the generative quality of our image-to-3D Gaussians reconstruction on random in-the-wild images from SHHQ, detailed in Table 3. This study focuses on reference view consistency and overall generation quality, crucial aspects in image reconstruction tasks. We rendered 360-degree rotation videos for 25 images generated by DreamGaussian, LGM, and Disco4D. We invited 43 volunteers to rate 24∼27 mixed samples from these methods on image consistency and overall model quality, yielding 1080 valid scores. As shown in Table 3, Disco4D was preferred, demonstrating better alignment with the original image content and superior overall quality.

4.2. 4D Animation

Pose-Driven Animation. Disco4D generates canonical Gaussians that can be animated with any pose sequence. Figure 12 in the Appendix demonstrates our animation capabilities and compares them with current SOTA 2D animation methods. Using identical inputs—a single frame and pose sequence—our approach more effectively preserves the body shape and fine details such as facial features and clothing. It surpasses Animate-Anyone [37] and Magic-Animate [106] in accurately modeling fine-grained body parts like hands and faces, and exhibits greater consistency compared to CHAMP [120]. The disentanglement feature of Disco4D further allows for direct manipulation of Clothing Gaussians, as shown in Figure 6.

4D Reconstruction. For the 4D-Dress Dataset [96], we evaluated 8 sequences, assessing CLIP similarity scores against ground-truth meshes and disentangled assets, along with novel view performance (PSNR, SSIM, LPIPS) from four viewpoints. Table 4 summarizes our quantitative results, benchmarking Disco4D against existing video-to-4D general GS approaches, such as DreamGaussian4D [78], as well as human-centric GS methods, including MonoHuman [113], GART [54], and GaussianAvatar [38]. We evaluate on monocular videos comprising 14 frames, captured from a limited front-view perspective, without full-body visibility across frames.

Disco4D outperforms MonoHuman [113], GART [54], and GaussianAvatar [38] (Table 4) as these methods reconstruct using known video information, unable to model unseen regions. Consequently, these methods cannot accurately model back views from front-facing videos, leading to artifacts in other perspectives and canonical space (see Figure 5). In contrast, Disco4D first performs reconstruction and subsequently incorporates details, such as clothing deformation, from the input frames, enabling consistent reconstruction even in unseen viewpoints.

While DreamGaussian4D [78] is capable of modeling back-view information, the details remain coarse. Our results demonstrate that initializing with our model from the first frame (DreamGaussian4D Disco4D-init) significantly outperforms other initialization methods (DreamGaussian4D-LGM init, DreamGaussian init) in both fidelity and geometry (Table 4). Nevertheless, without incorporating human priors, DreamGaussian4D [78] still faces challenges, such as missing limbs and difficulty modeling fine details like facial features (see Figure 13 in Appendix).

Reposing our canonical avatar enables us to align the body and assets accurately with the inferred postures from the source video, yielding high-quality reconstruction of faces, hands, and garments. Our reposed method surpasses DreamGaussian4D in geometry and fidelity by incorporating human priors. However, reposing alone cannot capture clothing dynamics. To address this, our disentangled approach models clothing deformations on the reposed Gaussians, guided by a diffusion model. As demonstrated in Figure 13 and Table 4, this process enhances the accuracy of clothing resemblance to the ground truth. The combination of asset repositioning and learned deformations improves modeling quality, with repositioning handling pose-driven changes and learned deformations simulating dynamic asset movements as observed in the driving video.

Table 2. **CLIP-embedding loss for generated humans and segmented assets, and performance (PSNR, SSIM, LPIPS) comparisons for novel poses and views on the Synbody and CloSe datasets across DreamGaussian, LGM, SHERF, and Disco4D.**

Method	SynBody							CloSe									
	CLIP				Novel View			CLIP				Novel View			Novel Pose		
	All ↑	Pants ↑	Shirt ↑	Shoes ↑	PSNR ↑	SSIM ↑	LPIPS ↓	All ↑	Pants ↑	Shirt ↑	Shoes ↑	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
DreamGaussian	0.751	0.715	0.710	0.749	13.118	0.883	0.229	0.734	0.693	0.674	0.767	20.08	0.939	0.089	-	-	-
LGM	0.807	0.724	0.747	0.760	12.884	**0.876**	0.228	0.829	0.727	0.712	0.778	**20.50**	**0.939**	**0.077**	-	-	-
SHERF	0.766	0.649	0.636	0.714	15.189	0.852	0.189	0.777	0.785	0.729	0.801	18.96	0.912	0.083	15.54	0.844	0.165
Disco4D	**0.851**	**0.784**	**0.753**	**0.801**	**15.691**	0.848	**0.185**	**0.856**	**0.858**	**0.810**	**0.842**	20.10	0.918	0.081	**17.96**	**0.851**	**0.136**

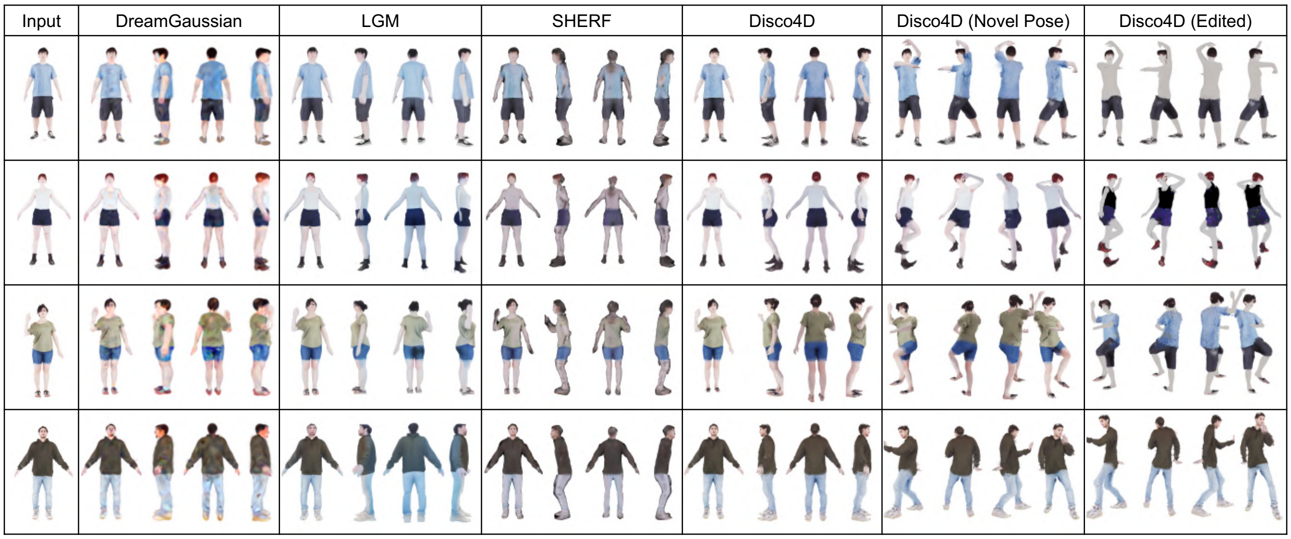

Figure 4. **Qualitative comparison of image generation across DreamGaussian, LGM, SHERF, and Disco4D.**

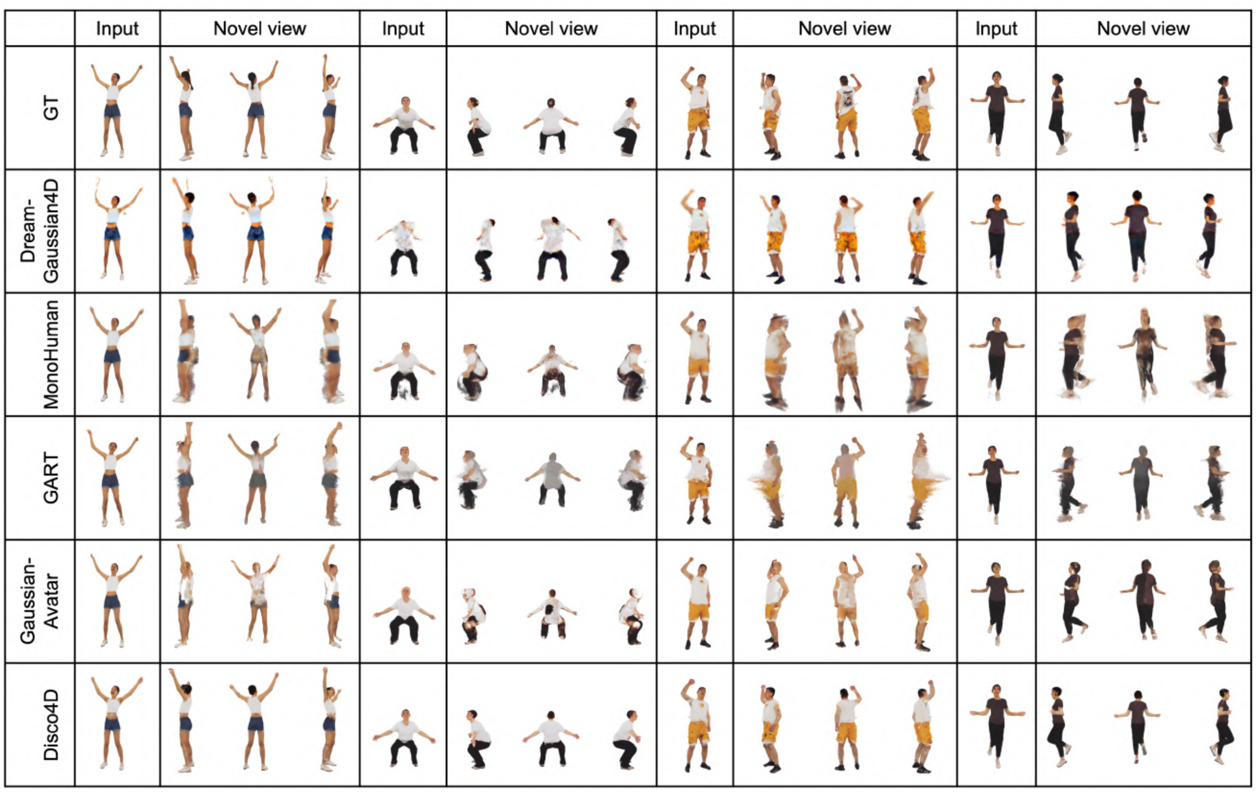

Figure 5. **Qualitative comparison of 4D generation between DreamGaussian4D, MonoHuman, GART, GaussianAvatar, and Disco4D.**

Table 3. **User study rates quality of generated 3D Gaussians from 1-5. The higher the better.**

Metric	Image Consistency ↑	Overall Quality ↑
DreamGaussian	2.017	1.852
LGM	2.338	2.017
Disco4D	**3.142**	**3.037**

Table 4. **CLIP-embedding loss for generated humans and segmented assets, and performance (PSNR, SSIM, LPIPS) comparison on 4D-Dress across various video-to-4D methods.**

	All ↑	Assets ↑	PSNR ↑	SSIM ↑	LPIPS ↓
DreamGaussian4D	0.784	0.769	20.54	0.93	0.080
MonoHuman	0.762	0.743	20.22	0.92	0.086
GART	0.800	0.772	18.81	0.92	0.086
GaussianAvatar	0.822	0.768	20.01	0.93	0.069
DreamGaussian4D (LGM init)	0.809	0.795	19.16	0.93	0.086
DreamGaussian4D (Disco4D init)	0.870	0.849	21.02	0.93	0.065
Disco4D (reposed)	0.853	0.774	23.94	0.95	0.049
Disco4D (reposed)+learned deformations	**0.900**	**0.865**	**25.46**	**0.96**	**0.035**

Figure 6. **First frame Editing and Animation**. Betas Editing, Recoloring (Text/Image-guided), Composition (Removal, Swap).

4D Editing. For a normal pipeline in character animation, editing the person in the video requires high consistency throughout all frames. For pose-driven animation methods, first frame editing and generation is required. Our method directly edits the Gaussians, which is more straightforward, fine-grained and consistent. This is seen from Figure 6.

4.3. Ablation Studies

Initialization of Clothing Gaussians. This process is crucial for high fidelity reconstruction. As shown in Figure 8 in the Appendix, we evaluate different strategies, including random, surface, and hull-based initialization. Hull-based initialization significantly enhances the model accuracy and realism over other methods. Initialization directly on the SMPL-X surface often leads to inaccurate geometries, particularly with complex or loose garments, creating elongated, thin Gaussians and visual artifacts. In contrast, hull-based initialization captures garment details more effectively and maintains pose consistency, closely aligning with the true geometry of the clothed body.

Geometry of Clothing Gaussians. Figure 14 in the Appendix highlights the differences in clothing geometry between DreamGaussian [89], LGM [91] and Disco4D. In DreamGaussian, all points are confined within the body geometry, whereas in LGM, about half of the points extend beyond the SMPL-X body. Removing internal points leaves sparse, translucent representations for clothing. This sparsity suggests reliance on internal points for visual representation, failing to accurately depict the object's geometry where appearance should primarily originate from surface points. Often, clothing Gaussian points are incorrectly positioned inside the body's hull rather than on the surface. To better represent clothing geometry, Disco4D positions all clothing Gaussians externally to the SMPL-X body mesh, accurately reflecting the garment's actual physical characteristics.

Clothing editing. Figure 14 shows our editing results with the prompt "Color the top pink". Disco4D allows for precise editing of the targeted clothing without affecting other areas.

5. Discussion

Despite achieving impressive results, some failure cases still exist, as shown in Figure 7 in the Appendix. Disco4D relies on robust and pixel-aligned SMPL-X estimation, which is still an unsolved problem. It occasionally fails for poor visual hull initialization. The extraction of mesh assets from clothing Gaussians using Local Density Query, as per DreamGaussian [89], currently loses fine-grained details. Enhancing the detail level of geometry derived from clothing Gaussians could bolster the utility of reconstructed assets in animation and simulation applications. Furthermore, the initialized visual hulls obtained from multi-view SMPL-X guided images are often of suboptimal quality and suffer from poor side and back views, necessitating refinement. Improving pose guidance models to achieve more accurate visual hulls could alleviate the need for extensive refinement. In addition, future works could look into modeling multi-layered clothing and reconstructing the occluded clothing. Disco4D has many positive applications, but it also has the potential to facilitate deepfake avatars and raise IP concerns. Regulations should be built to address these issues alongside its benefits in the entertainment industry.

6. Conclusion

We propose Disco4D, a novel approach for the generation of 3D animatable clothed human Gaussians from a single image, emphasizing high-fidelity detail and separation of assets. We manage to compositionally generate separate components, such as haircut, accessories, and decoupled outfits. Our core insight is the fixing of SMPL-X Gaussians, fitting segmented Gaussians over SMPL-X Gaussians, and application of diffusion models to enhance 3D reconstruction, including modeling occluded parts not visible in the input image. Its capability to separate assets offers significant advantages, including localized, fine-grained editing of individual assets and enhanced animatability.

Acknowledgements

We sincerely thank the anonymous reviewers for their valuable comments on this paper. This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme(AISG Award No: AISG3-PhD-2023-08-049T). This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20221-0012, MOE-T2EP20223-0002), and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Jad Abou-Chakra, Feras Dayoub, and Niko Sünderhauf. Particlenerf: Particle based encoding for online neural radiance fields. *arXiv preprint arXiv:2211.04041*, 2022. 3

[2] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic Monocular 3D Reconstruction of Humans Wearing Clothing. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:1496–1505, 2022. 1, 2

[3] Dimitrije Antić, Garvita Tiwari, Batuhan Ozcomlekci, Riccardo Marin, and Gerard Pons-Moll. CloSe: A 3D clothing segmentation dataset and model. In *International Conference on 3D Vision (3DV)*, 2024. 6

[4] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation, 2024. 2

[5] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: clothed 3d humans. In *European Conference on Computer Vision*, pages 344–359. Springer, 2020. 2

[6] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019. 2

[7] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 2, 4

[8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models, 2023.

[9] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2

[10] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[11] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. *CVPR*, 2023. 3

[12] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 2

[13] 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

[14] Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. Bsp-net: Generating compact meshes via binary space partitioning. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[15] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators, 2024. 2

[16] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12352 LNCS: 769–787, 2020. 2

[17] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[18] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *CVPR*, 2021. 3

[19] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan

Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 2, 5

[20] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[21] Shivam Duggal and Deepak Pathak. Topologically-aware deformation fields for single-view 3d reconstruction. *CVPR*, 2022. 2

[22] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J. Black. Learning to Regress Bodies from Images using Differentiable Semantic Rendering. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11230–11239, 2021. 2

[23] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*. ACM, 2022. 3

[24] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021. 2

[25] 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. 1, 3

[26] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J. Black. Learning disentangled avatars with hybrid 3d representations. *arXiv*, 2023. 1, 3

[27] Valentin Gabeur, Jean-Sebastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images, 2019. 2

[28] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 3

[29] Shanyan Guan, Huayu Deng, Yunbo Wang, and Xiaokang Yang. Neurofluid: Fluid dynamics grounding with particle-driven neural radiance fields, 2022. 3

[30] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *arXiv preprint arXiv:2311.12775*, 2023. 3

[31] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning, 2023. 2

[32] Oshri Halimi, Fabian Prada, Tuur Stuyck, Donglai Xiang, Timur Bagautdinov, He Wen, Ron Kimmel, Takaaki Shiratori, Chenglei Wu, and Yaser Sheikh. Garment avatars: Realistic cloth driving using pattern registration, 2022. 2

[33] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. ARCH++: Animation-Ready Clothed Human Reconstruction Revisited. *Proceedings of the IEEE International Conference on Computer Vision*, pages 11026–11036, 2021. 1, 2

[34] Zhu Heming, Cao Yu, Jin Hang, Chen Weikai, Du Dong, Wang Zhangye, Cui Shuguang, and Han Xiaoguang. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *Computer Vision – ECCV 2020*, pages 512–530. Springer International Publishing, 2020. 2

[35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 2

[36] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2

[37] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 6

[38] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 6

[39] Shoukang Hu and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. *arXiv preprint arXiv:*, 2023. 4

[40] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. *arXiv preprint arXiv:2303.12791*, 2023. 1, 2

[41] 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 *IEEE Conference on Computer Vision (ICCV)*, 2023. 2

[42] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable Reconstruction of Clothed Humans. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3090–3099, 2020. 1, 2

[43] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*. Springer, 2020. 2

[44] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar Fine-Tuning for 3D Human Model Fitting Towards In-the-Wild 3D Human Pose Estimation. *Proceedings - 2021 International Conference on 3D Vision, 3DV 2021*, pages 42–52, 2021. 2

[45] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions, 2023. 2

[46] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-End Recovery of Human Shape and Pose. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 2

[47] 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, 15

[48] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 2

[49] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11035–11045, 2021. 2

[50] Muhammed Kocabas, Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats, 2023. 3

[51] Nikos Kolotouros, Georgios Pavlakos, Michael Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:2252–2261, 2019. 2

[52] Georgios Kopanas, Julien Philip, Thomas Leimkühler, and George Drettakis. Point-based neural rendering with per-view optimization. *Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering)*, 40(4), 2021. 15

[53] Georgios Kopanas, Thomas Leimkühler, Gilles Rainer, Clément Jambon, and George Drettakis. Neural point catacaustics for novel-view synthesis of reflections. *ACM Transactions on Graphics*, 41(6): Article–201, 2022. 15

[54] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models, 2023. 3, 6

[55] Mengtian Li, Shengxiang Yao, Zhifeng Xie, and Keyu Chen. Gaussianbody: Clothed human reconstruction via 3d gaussian splatting, 2024. 3

[56] Xuanyi Li, Daquan Zhou, Chenxu Zhang, Shaodong Wei, Qibin Hou, and Ming-Ming Cheng. Sora generates videos with stunning geometrical consistency. *arXiv preprint arXiv: 2402.17403*, 2024. 2

[57] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[58] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[59] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 2

[60] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 2

[61] Yang Liu, Xiang Huang, Minghan Qin, Qinwei Lin, and Haoqian Wang. Animatable 3d gaussian: Fast and high-quality reconstruction of multiple human avatars. *arXiv preprint arXiv:2311.16482*, 2023. 3

[62] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2, 3, 15

[63] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024. 3

[64] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8446–8455, 2023. 2

[65] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[66] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022. 2

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

[68] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 3

[69] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), 2021. 3

[70] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 2

[71] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, DImitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 10967–10977, 2019. 1, 2, 3, 15

[72] 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. 3

[73] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017. Two first authors contributed equally. 2

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

[75] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[76] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 2

[77] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. 2024. 3

[78] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 3, 5, 6

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

[80] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1749–1759, 2021. 2

[81] 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 *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[82] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 1, 2

[83] Igor Santesteban, Miguel A. Otaduy, and Dan Casas. Learning-Based Animation of Clothing for Virtual Try-On. *Computer Graphics Forum (Proc. Eurographics)*, 2019. 2

[84] Sara Fridovich-Keil and Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023. 3

[85] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[86] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 4

[87] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023. 2

[88] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction, 2023. 2

[89] 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, 3, 6, 8

[90] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22819–22829, 2023. 2

[91] 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. 2, 3, 4, 6, 8

[92] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2021. 3

[93] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d scene representation and rendering. In *arXiv:2010.04595*, 2020. 2

[94] Haithem Turki, Jason Y. Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes, 2023. 3

[95] Raquel Vidaurre, Igor Santesteban, Elena Garces, and Dan Casas. Fully Convolutional Graph Neural Networks for Parametric Virtual Try-On. *Computer Graphics Forum (Proc. SCA)*, 2020. 2

[96] 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. 6, 16

[97] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Wang Xinggang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023. 3

[98] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9421–9431, 2021. 3

[99] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *Proceedings of International Conference on 3D Vision (3DV '20)*, pages 322 – 332, 2020. 2

[100] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM Transactions on Graphics*, 40(6):1–15, 2021. 2

[101] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *CoRR*, abs/2105.15203, 2021. 4

[102] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. *arXiv preprint arXiv:2311.12198*, 2023. 5

[103] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, 2022. 1, 2

[104] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[105] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *Advances in Neural Information Processing Systems 32*, pages 492–502. 2019. 2

[106] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. 2023. 6

[107] Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. D-IF: Uncertainty-aware Human Digitization via Implicit Distribution Field. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 2

[108] 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 *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 10671–10681, 2024. 2

[109] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20282–20292, 2023. 6

[110] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023. 3

[111] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. *arXiv preprint arXiv:2312.00732*, 2023. 4, 15

[112] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, 2023. 2

[113] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023. 3, 6

[114] Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove. Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13144–13152, 2021. 3

[115] Ilya Zakharkin, Kirill Mazur, Artur Grigorev, and Victor Lempitsky. Point-based modeling of human clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14718–14727, 2021. 2

[116] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. *arXiv:2311.10982*, 2023. 2

[117] Zheng Zerong, Yu Tao, Liu Yebin, and Dai Qionghai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction, 2021. 1, 2

[118] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[119] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4491–4500, 2019. 2

[120] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance, 2024. 6

7. Appendix

Due to space constraints in the main paper, we provide additional information and details in this supplementary material. These include:

- A more detailed discussion of the preliminaries in Section 7.1.
- Comprehensive training details of the Identity Encoding Loss in Section 7.2.
- Implementation details outlined in Section 7.3.
- Analysis of failure cases in Section 7.4.
- Ablation of initialization strategies in Section 7.5
- Examples of hair tagging in Section 7.6
- Additional visualizations presented in Section 7.9.
- Detailed information on the demo video [Disco4D-demo.mp4] in Section 7.10.

7.1. Preliminary

3D Gaussian Splatting utilizes explicit 3D Gaussian points as the core elements for rendering. Each 3D Gaussian point is defined by the function:

$$G(x) = e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)},$$

where μ represents the spatial mean, and Σ denotes the covariance matrix. Additionally, each Gaussian is assigned an opacity value α and a view-dependent color c, parameterized by spherical harmonic coefficients f. During rendering, these 3D Gaussians are projected onto the 2D view plane via a splatting technique. The 2D projection is computed using the projection matrix, while the 2D covariance matrices are approximated as: $\Sigma' = J_g W_g \Sigma W_g^T J_g^T$, where W_g is the viewing transformation, and J_g is the Jacobian of the affine approximation for perspective projection. The final pixel color is obtained through alpha-blending of N layered 2D Gaussians from front to back $C = \sum_{i\in N} T_i \alpha_i c_i$, with $T_i = \prod_{j=1}^{i}(1-\alpha_j)$.

The opacity α is determined by multiplying γ with the contribution of the 2D covariance, derived from Σ' and the pixel coordinate in image space. The covariance matrix Σ is parameterized using a quaternion q and a 3D scaling vector v to aid in optimization.

SMPL-X parameterization [71] extends the original SMPL body model [62] by incorporating detailed face and hand deformations to capture more expressive human movements. **SMPL-X** expands **SMPL** joint set by including additional joints for facial features, toes and fingers, enabling a more accurate representation of complex body movements. **SMPL-X** is defined by a function $M(\beta,\theta,\psi) : \mathbb{R}^{|\beta|\times|\theta|\times|\psi|} \rightarrow \mathbb{R}^{3N}$, where $\theta \in \mathbb{R}^{3K}$ represents the pose (with K being the number of body joints), $\beta \in \mathbb{R}^{|\beta|}$ represents body shape, and $\psi \in \mathbb{R}^{|\psi|}$ captures facial expressions. Further details can be found in [71].

7.2. Training details of Identity Encoding loss

To optimize the introduced Identity Encoding of each Gaussian, we render these encoded identity vectors into 2D images in a differentiable manner following [111]. We adapt the differentiable 3D Gaussian renderer from [47], approaching the rendering process similarly to the color optimization using spherical harmonic (SH) coefficients, as described in [47]. In this method, 3D Gaussian splatting utilizes neural point-based α'-rendering [52, 53], where the influence weight α' is calculated in 2D for each Gaussian and pixel. Following the approach in [47], the influence of all Gaussians on a pixel is computed by sorting them based on depth and blending the N ordered Gaussians that overlap with that pixel:

$$E_{id} = \sum_{i\in\mathcal{N}} e_i \alpha_i \prod_{j=1}^{i-1}(1-\alpha'_j) \tag{3}$$

Here, the rendered 2D mask identity feature E_{id} is the sum of the Identity Encoding e_i (of length 15) for each Gaussian, weighted by the Gaussian's influence factor α'_i on that pixel. The value of α'_i is determined by evaluating a 2D Gaussian with covariance $\Sigma 2D$, which is scaled by a learned per-point opacity α_i:

$$\Sigma 2D = JW\Sigma 2D^{3D} W^T J^T \tag{4}$$

where Σ^{3D} is the 3D covariance matrix, $\Sigma 2D$ represents the splatted 2D counterpart, J is the Jacobian of the affine approximation for the 3D-to-2D projection, and W is the world-to-camera transformation matrix.

To ensure consistency in the Identity Encoding e_i during training, we apply an unsupervised 3D regularization loss. This loss encourages the Identity Encodings of the top k-nearest 3D Gaussians to remain close in feature space, promoting spatial consistency. Using the softmax function F, we define the KL divergence loss with m sampled points as follows:

$$\mathcal{L}_{3d} = \frac{1}{m}\sum_{j=1}^{m} D_{KL}(P||Q) = \frac{1}{mk}\sum_{j=1}^{m}\sum_{i=1}^{k} F(e_j)\log\left(\frac{F(e_j)}{F'(e_j)}\right) \tag{5}$$

Here, P is the sampled Identity Encoding e of a 3D Gaussian, and Q consists of the k-nearest neighbors in 3D space, represented as $e'_1, e'_2, ..., e'_k$. The total identity encoding loss is then defined as:

$$\mathcal{L}_{id} = \mathcal{L}_{2d} + \mathcal{L}_{3d} \tag{6}$$

7.3. Implementation details

The 3D generation experiments were conducted using a single 24GB RTX3090 GPU, while the 4D generation experiments utilized a single 48GB RTX6000 GPU. For the 3D generation process, the SMPL-X fitting was performed with 3000 iterations in 3 minutes, followed by skin color inpainting on SMPL-X Gaussians for 100 iterations in 30 seconds. Reconstruction and disentanglement optimization required 3000 iterations, completed in 12 minutes. In video reconstruction, SMPL-X fitting aligned 14 frames in 6 minutes for in-the-wild videos. The 4D-Dress [96] experiments involved 1000 iterations for clothing deformation over 18 minutes.

7.4. Failure cases

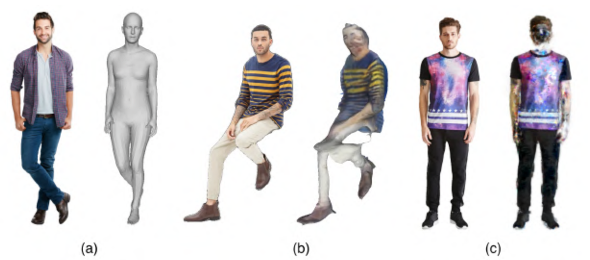

Figure 7. **Failure cases of Disco4D.** (a) Poor SMPL-X estimation (b) Poor visual hull initialization (c) Misclassification of clothing categories.

Disco4D relies on robust and pixel-aligned SMPL-X estimation, which is still an unsolved problem, especially for challenging poses. In Figure 7a, it is difficult to correct the pose with keypoints and segmentation mask due to depth ambiguity. Disco4D occasionally fails for poor visual hull initialization (7b), which is common for difficult poses. Lastly, poor disentanglement is a common problem due to misclassification of clothing category by the segmentation model. This is seen in Figure 7c where the arms are wrongly classified under the "top" category.

7.5. Initialization

We evaluate random, surface, and hull-based initialization strategies. Surface initialization on SMPL-X often produces inaccurate geometries for complex or loose garments, leading to elongated Gaussians and artifacts. Hull-based initialization better captures garment details, preserves pose consistency, and aligns closely with the true clothed body geometry, as seen in Figure 8.

7.6. Hair tagging

In our approach, hair Gaussians are tagged to head faces rather than the nearest face during reposing. Reposing hair Gaussians according to the nearest face, as commonly done in previous works, often results in artifacts such as disjointed hair (Figure 9). By leveraging the learned identity encoding, we assign a unified identity to hair Gaussians, enabling them to be reposed cohesively as a single entity, thereby preserving the structural integrity of the hair during transformations.

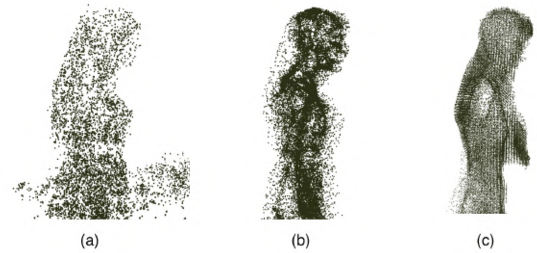

Figure 8. **Ablation of initialization.** (a) Random Initialization (b) SMPL-X Initialization (c) Visual Hull Initialization.

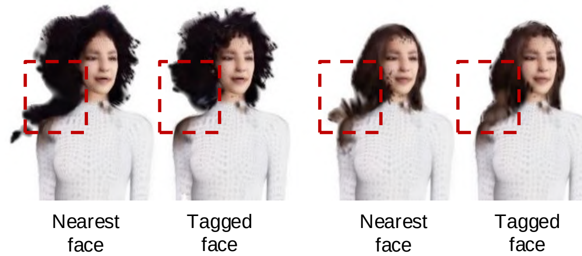

Figure 9. **Visualization of hair tagging.**

7.7. In-the-wild evaluation

Figure 10. **Qualitative evaluation on ITW images.**

Figure 11. **Qualitative evaluation on avatars clothed in dress.**

Our focus on studio and synthetic datasets (e.g., Synbody, CloSe, and 4DDress) was due to the availability of ground-truth data from multiple views, enabling rigorous quantitative evaluation. ITW images lack such ground-truth data, making comparisons challenging. Nevertheless, our solution applies to ITW images, with some examples shown in Fig. 10. Examples of avatars clothed in dress are added in Fig 11, driven with poses from subjects in Fig. 10.

7.8. Facial detail

Additional visualizations of well known individuals are provided in Fig. 10 and Fig. 11.

7.9. Extra visualizations

Figure 12 presents visual comparisons with 2D animation methods. Figure 13 illustrates ablation results for 4D reconstruction. Finally, ablation studies on point geometry and editing are provided in Figure 14.

7.10. Demo video

Extended visualizations and results showcasing 3D generation and disentanglement, pose-driven animation, video-to-4D reconstruction, and fine-grained editing of animated outputs are demonstrated in the accompanying demo video [Disco4D-demo.mp4]. A sample of the video is shown in Figure 15.

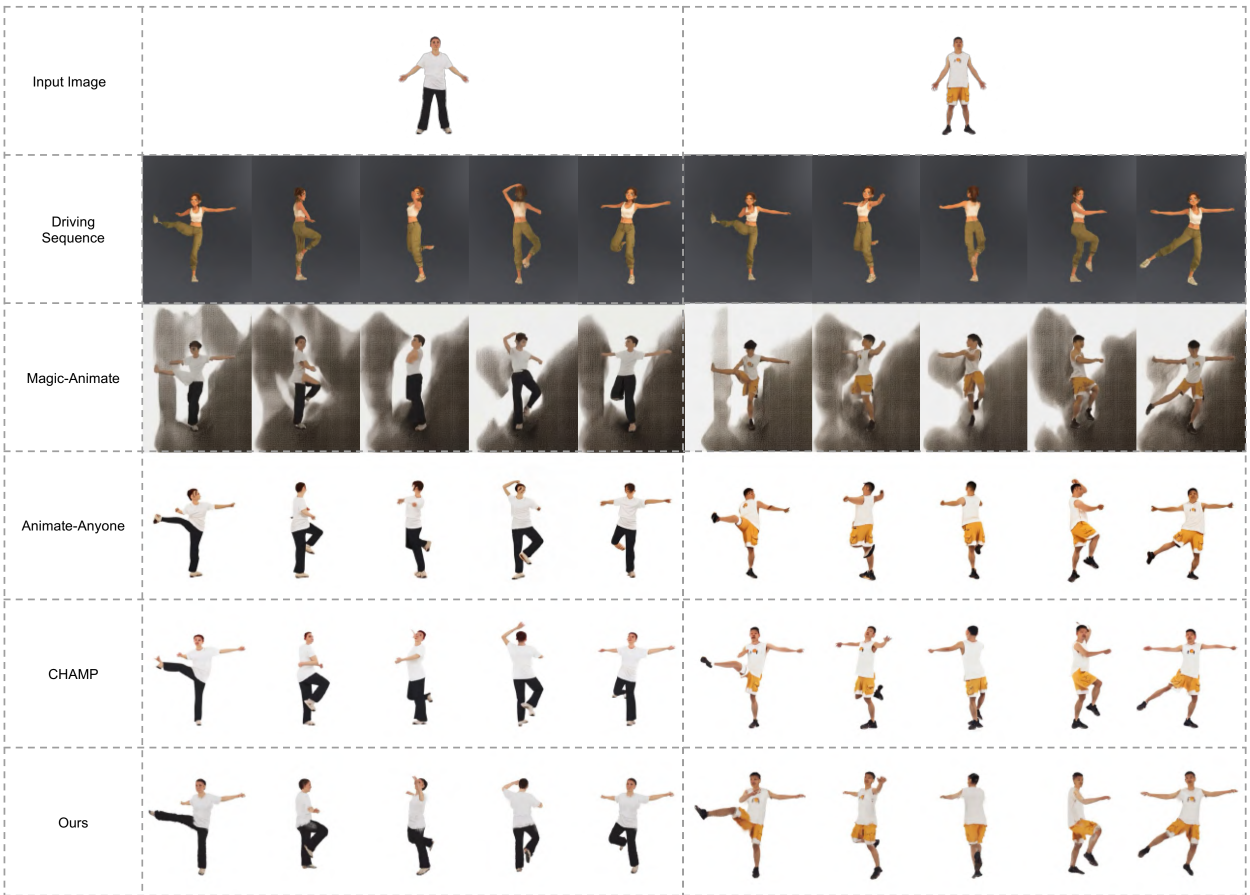

Figure 12. **Comparison to 2D animation methods.** Compared to Magic-Animate and Animate-Anyone, we have better preservation of body shape and details. Compared to CHAMP, we have better geometry and consistency.

Figure 13. **4D reconstruction results on 4D-Dress Dataset.**

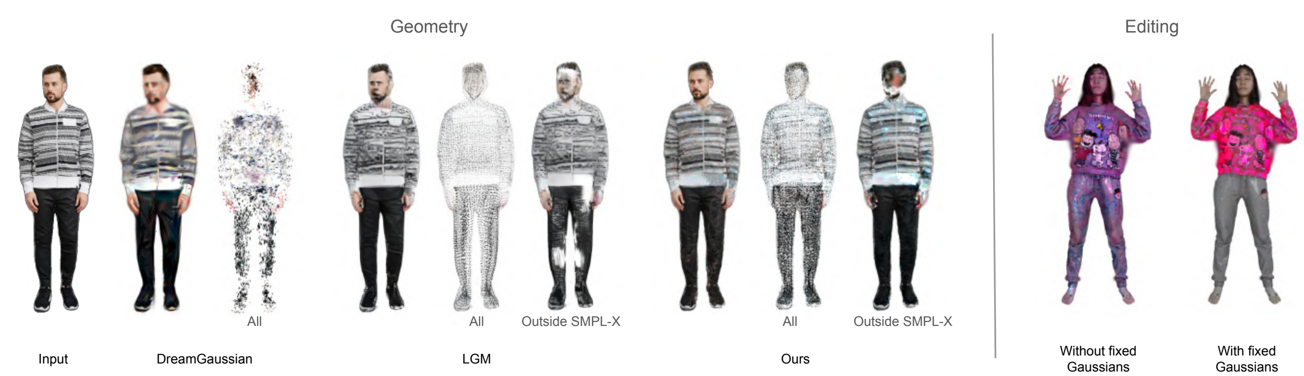

Figure 14. **Ablation of points geometry (left) and editing results (right). Points ("All") are visualised with a Gaussian Scale of 0.1.**

Figure 15. **Additional visualizations showcasing generation, disentanglement, animation, and editing. Full demonstrations are available in the accompanying demo video [Disco4D-demo.mp4]. This figure provides a sample from the demo.**