Gaussian Wardrobe: Compositional 3D Gaussian Avatars for Free-Form Virtual Try-on

Zhiyi Chen*[1] Hsuan-I Ho*[1] Tianjian Jiang[1]
Jie Song[3,4] Manuel Kaufmann[1,2] Chen Guo[1]

[1]Department of Computer Science, ETH Zürich
[2]ETH AI Center, ETH Zürich
[3]HKUST(GZ) [4]HKUST

https://ait.ethz.ch/gaussianwardrobe

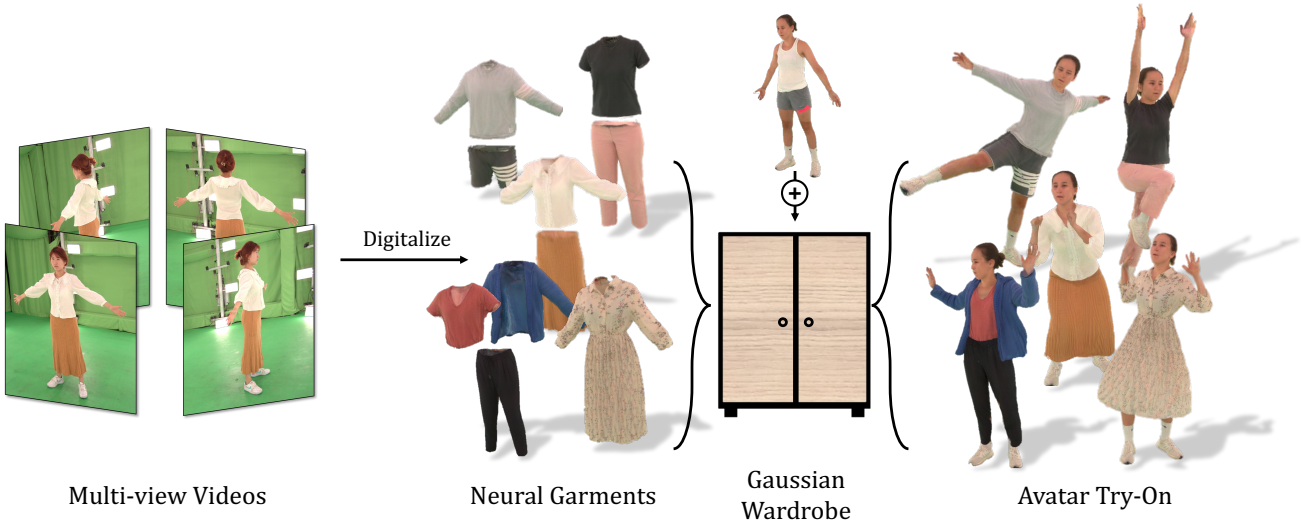

Figure 1. **Gaussian Wardrobe** is a novel approach to digitalize compositional 3D Gaussian avatars from multi-view videos. The learned neural garments are subject-agnostic. Therefore, they can be stored, reused, and seamlessly recombined with new subjects. Leveraging Gaussian Wardrobe, we realize a practical 3D avatar virtual try-on application. Our method demonstrates the capabilities of modeling the dynamics of challenging free-form clothing such as skirts, dresses, and open jackets.

Abstract

We introduce Gaussian Wardrobe, a novel framework to digitalize compositional 3D neural avatars from multi-view videos. Existing methods for 3D neural avatars typically treat the human body and clothing as an inseparable entity. However, this paradigm fails to capture the dynamics of complex free-form garments and limits the reuse of clothing across different individuals. To overcome these problems, we develop a novel, compositional 3D Gaussian representation to build avatars from multiple layers of free-form garments. The core of our method is decomposing neural avatars into bodies and layers of shape-agnostic neural garments. To achieve this, our framework learns to disentangle each garment layer from multi-view videos and canonicalizes it into a shape-independent space. In experiments, our method models photorealistic avatars with high-fidelity dynamics, achieving new state-of-the-art performance on novel pose synthesis benchmarks. In addition, we demonstrate that the learned compositional garments contribute to a versatile digital wardrobe, enabling a practical virtual try-on application where clothing can be freely transferred to new subjects.

*Equal contribution

1. Introduction

Recent advancements in neural representations have enabled the creation of photorealistic human avatars with remarkable fidelity [12, 32]. These avatars are key across various XR applications, including telepresence, digital fashion, and entertainment. However, these methods predominantly train a dedicated neural network to model a clothed human as an indivisible whole. This one-to-one paradigm limits scalability and prevents avatars for broad user bases. To overcome this bottleneck, a fundamental property is required for a scalable avatar system: *Compositionality*.

Current neural avatar methods typically represent a clothed human as a single entity [10, 25, 32, 42, 47]. While this formulation can efficiently drive the avatars with skeletal deformations from a parametric body model (*e.g.*, SMPL-X [46]), it strongly assumes that clothing is a simple extension of the body and directly deforms with it. This strategy, therefore, fails to capture the dynamics of common garments like open jackets or skirts, which have topologies distinct from the body and do not conform to underlying bone movements. Some recent attempts [7, 8, 11, 58] have begun to incorporate explicit clothing modeling as a distinct entity. Nevertheless, the underlying deformations of the avatar and its garments remain entangled. This contradicts the real-world nature of clothing, which can be reused, swapped across individuals, and realistically animated on novel human subjects. The deformation aspect has been largely overlooked and represents the critical factor of *compositionality*. Ideally, neural avatars should be decomposed into compositional neural garments, defined both in terms of geometry and garment-specific deformation. These garments ought to be subject-agnostic and designed for free recombination across avatars, thereby enabling a versatile digital wardrobe.

In this work, we present **Gaussian Wardrobe**, a novel framework for digitalizing compositional neural avatars from multi-view videos. Our method encodes garments of arbitrary topology as animatable 3D Gaussian representations, enabling realistic and dynamic deformations. The central innovation lies in decomposing a complete human avatar into a collection of subject-agnostic neural garments. As illustrated in Fig. 1, the garments together form a digital wardrobe that can be recombined and transferred across subjects, thereby facilitating applications like 3D virtual try-on. The resulting avatars exhibit high-fidelity rendering while faithfully capturing realistic motion dynamics across diverse clothing categories.

At the core of our method is a novel compositional representation designed for garments with arbitrary topologies. We build upon the recent template-based method of Animatable Gaussians [32], but introduce three critical modifications to achieve our goals of compositionality. i) We segment the input template mesh into distinct body and garment components to enable a compositional representation. This decomposition facilitates more accurate modeling of clothing geometry, appearance, and non-linear motion using 3D Gaussians. ii) All garments are canonicalized into a zero-shaped space to guarantee cross-subject compositionality. In this way, the learned appearance and deformation models become inherently shape-agnostic and thus transferable across subjects. iii) A dedicated learning framework with carefully designed loss functions is developed to disentangle each garment layer from the input videos. Once trained, these layers are represented as reusable neural assets, collectively forming a digital wardrobe for diverse downstream tasks.

Complementing this reconstruction scheme, we design a practical 3D virtual try-on application with Gaussian Wardrobe. Our system leverages the created collection of reusable subject-agnostic neural garments and transfers them to a new user's body shape. During a try-on session, the target user's identity is defined by their body Gaussians and shape parameters. We then combine this identity with the 3D Gaussians of the selected garments and animate the avatar according to a sequence of driving poses. To mitigate potential surface penetrations under drastically out-of-distribution pose animations, we propose an online penetration detection mechanism integrated into the rendering pipeline. This strategy enables the detection of interpenetrations between layers and the correction of resulting visual artifacts during rendering.

In our experiments, we demonstrate that the proposed framework creates photorealistic and compositional avatars across diverse garments with complex topologies. We further establish state-of-the-art performance on the 4D-DRESS [62] and Actor-HQ [23] benchmarks for novel pose synthesis. In summary, our contributions are threefold:

- A novel compositional 3D Gaussian-based approach capable of modeling 3D human avatars in complex, free-form garments.
- A reconstruction scheme that learns to decompose an avatar into distinct neural garments and models their dynamic deformations from multi-view video.
- A novel framework to produce subject-agnostic neural garments that enable compositional 3D virtual try-on across subjects with diverse body shapes.

2. Related Work

Animatable 3D human avatar. Traditionally, 3D human avatars were created using pre-scanned human meshes [13–15, 66] or extensions of parametric body models [1, 2, 28, 59, 67]. However, these approaches often suffer from limited representational capability. To address this issue, the research community introduced neural representations, such as implicit neural fields [39, 44], Neural Radiance Fields (NeRF) [40], and 3D Gaussians [29], to model clothed hu-

mans as neural avatars. Numerous works have since been proposed to digitize these avatars from 3D scans [5, 32, 53], monocular videos [10, 21, 22, 25–27, 42, 47], and even single images [19, 51, 64, 65, 71]. While these methods produce high-quality appearances, they represent the body and clothing as a single entity, which limits their expressiveness for modeling free-flowing garments.

Recently, another stream of work has emerged that explicitly models garments as a separate entity. These approaches extend the underlying parametric body models with additional implicit [6–8, 11, 24, 41, 60] or explicit representations [33, 36–38, 49, 58, 68, 74] to capture the complex surface deformations of clothing. However, a fundamental limitation is that the deformations of the avatar and its garments remain entangled, thereby restricting their applicability in realistic virtual try-on settings. While some attempts incorporate physically-simulated training data or physics-informed objectives [9, 31, 50, 52, 55, 72], they typically focus on modeling the geometric deformation of a single garment and do not support holistic animation of the human body together with garments comprising multiple layers. The work most closely related to ours is LayGA [34], which models multi-layer garments using 3D Gaussians derived from an underlying SMPL-X body mesh. Although this approach enables compositional virtual try-on, its reliance on parametric body meshes restricts its ability to represent free-form loose clothing, such as open jackets and skirts, whose dynamics are only weakly correlated with skeletal deformations. In this work, we address this limitation by integrating mesh-based garment modeling with state-of-the-art animatable 3D Gaussian representations. This integration enables our method not only to support novel clothing transfer for tight-fitting garments but also to achieve compelling results when dressing subjects in loose outfits.

Avatar virtual try-on and editing. The literature on virtual try-on has mostly focused on composing outfits within 2D images [16, 54, 56]. With recent developments in XR technology, 3D virtual try-on has emerged as a significant research area. Conventionally, 3D avatar customization was achieved by combining extensive collections of human scans with artist-designed 3D assets [3, 75]. To avoid this time-consuming process, recent approaches [4, 17, 43, 73] have leveraged Score Distillation Sampling (SDS) [48] to edit avatar appearance and clothing using text prompts. While these techniques offer flexibility, their lengthy optimization times limit their practicality.

More closely related to our work are methods that reconstruct interchangeable 3D avatars for virtual try-on and editing [18, 34, 70]. For instance, CustomHumans [18] learns a generative model over SMPL-X body meshes, enabling the swapping of local appearances at each mesh face. Similarly, LayGA [34] uses SMPL-X meshes to predict 3D Gaussians for garments, while LayerAvatar [70] employs diffusion models to generate clothing textures and displacements from the UV maps of parametric bodies. A significant limitation of these methods is that their compositional ability stems from the shared topology of the underlying parametric mesh. This dependence restricts them to modeling only a few garment layers and fails to capture the free-form dynamics of real-world clothing. In contrast, Gaussian Wardrobe is inspired by mesh-based neural avatars [32, 50], which offer greater capacity to model diverse garments and achieve a more realistic virtual try-on experience.

3. Gaussian Wardrobe

Preliminary. Our compositional Gaussian Avatar is built upon the framework of Animatable Gaussians [32]. Given a multi-view video $\{I_{m,n}\}$ of a human subject, where m is the frame index and n is the camera index, we register the subject's shape and per-frame body pose parameters. We use the SMPL-X [46] parametric body model, which defines shape with $\beta \in \mathbb{R}^{10}$ PCA coefficients and pose with parameters $\theta \in \mathbb{R}^{21+30+3+1}$ (comprising 21 body joint rotations, 30 hand parameters, 3 facial joint rotations, and a global orientation).

The Animatable Gaussians method utilizes a mesh template and a pose-conditioned U-Net to represent the avatar. To create the template, we first select the first video frame $\{I_{1,n}\}$, which should feature a pose close to the standard A-pose. We then reconstruct a mesh from this frame using the multi-view data and repose it into a pre-defined canonical pose. This template, denoted as $\mathcal{M}$ in Fig. 2 (left), is used to render pose-dependent positional maps for any given body pose θ_m. The U-Net is trained to predict the parameters of 3D Gaussian primitives for various input poses, which enables the avatar to be driven by novel movements.

Method overview. The overall framework of our method, Gaussian Wardrobe, is summarized in Fig. 2. In Sec. 3.1, we present our approach for augmenting the Animatable Gaussians representation to support compositionality and model free-form neural garments. Next, in Sec. 3.2, we introduce the learning framework used to decompose individual garment layers from multi-view videos. Finally, in Sec. 3.3, we demonstrate a practical application of 3D virtual try-on using a collection of trained neural garments that form a digital wardrobe.

3.1. Compositional Gaussian Representation

A single template mesh is insufficient for modeling the complex dynamics of free-form clothing. Moreover, because the template encodes subject-specific shape information (β), it is not suitable for generalization across differ-

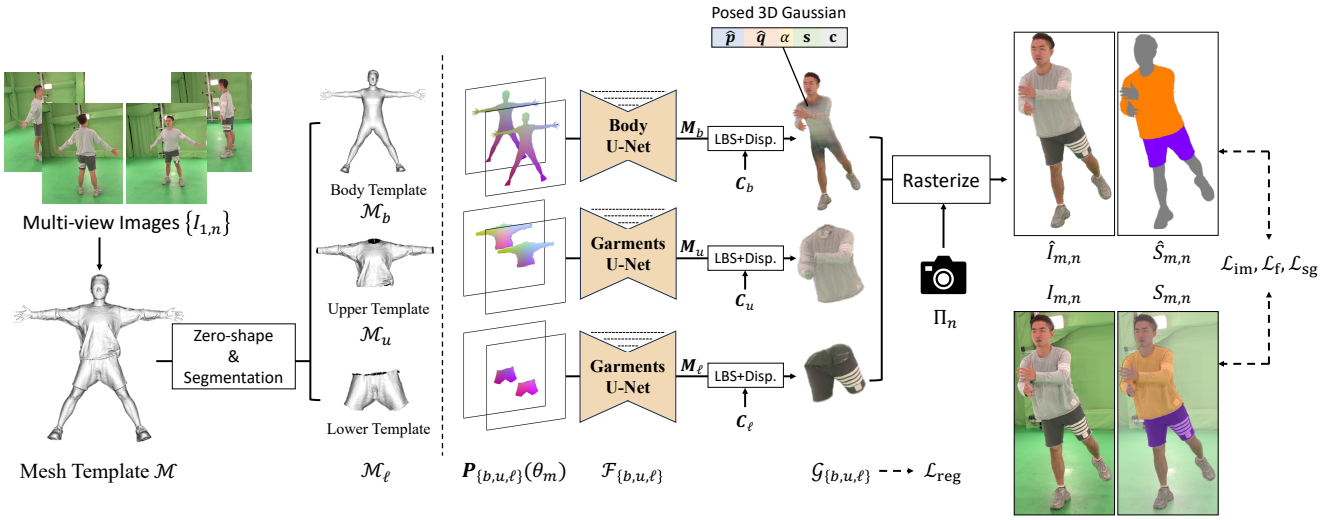

Figure 2. **Gaussian Wardrobe** digitalizes compositional neural avatars from multi-view videos. Our pipeline consists of two major components: (left) a compositional Gaussian representation and (right) a framework for learning neural garments. We first reconstruct a mesh template $\mathcal{M}$ from the first video frame and segment it into body $\mathcal{M}_b$, garment templates $\mathcal{M}_{\{u,\ell\}}$ in the zero-shaped canonical space. During training, each layer learns a separate U-Net $\mathcal{F}$ to predict the parameters of 3D Gaussian primitives $\mathbf{M}$ from pose-conditioned positional maps $\mathbf{P}$. We composite the 3D Gaussians $\mathcal{G}$ from all layers and render RGB images $\hat{I}$ and segmentation masks $\hat{S}$ to compute the training loss $\mathcal{L}$. The learned neural garments are shape-agnostic and can seamlessly transfer to other subjects for avatar virtual try-on.

ent subjects. To address these limitations, we design a new compositional Gaussian representation.

First, to make the template shape-agnostic, we deform it into a canonical, zero-shape space by removing the subject-specific shape blendshapes β. Specifically, each vertex $\mathbf{v}_i^t$ on the original template is displaced to a new position $\mathbf{v}_i^c$ following: $\mathbf{v}_i^c = \mathbf{v}_i^t - \sum_{b=1}^{10} \beta_b \mathbf{o}_{bi}$. Here, β_b is the b-th PCA shape coefficient, and $\mathbf{o}_{bi}$ is the corresponding blendshape offset for the i-th vertex. To obtain these offsets for our custom template $\mathcal{M}$, we adopt the strategy from Fast-SNARF [5], which involves voxelizing and diffusing the SMPL-X blendshape offsets into a 64^3 voxel grid. The value of $\mathbf{o}_{bi}$ for each vertex is then queried by trilinearly interpolating the values in this grid at the position of $\mathbf{v}_i^t$.

After obtaining the shape-agnostic template, we perform 3D segmentation [62] to separate the body from the different garment layers. Specifically, we segment the template into upper ($\mathcal{M}_u$), lower ($\mathcal{M}_\ell$), and optional outer ($\mathcal{M}_o$) garment layers to model their dynamic deformations. To capture the underlying skeletal motion, we use a separate SMPL-X body mesh ($\mathcal{M}_b$). The meshes for regions like hair and shoes are merged with this body mesh to better preserve their details. We visualize these multi-layer, shape-agnostic templates in Fig. 2 (left).

For each layer $L \in \{b, u, \ell\}$, we follow the approach of Animatable Gaussians [32] to generate two feature maps from its corresponding template mesh $\mathcal{M}_L$. First, we create front-and-back coordinate maps $\mathbf{C}_L \in \mathbb{R}^{2\times H\times W\times 3}$ by rasterizing the template from orthographic front and back views. Each pixel in these maps stores the 3D position of a point on the template, defining the initial state of the 3D Gaussians. Second, we deform the template mesh using Linear Blend Skinning (LBS) with the registered pose parameters θ_m to generate posed positional maps $\mathbf{P}_L(\theta_m) \in \mathbb{R}^{2\times H\times W\times 3}$. These maps have the same dimensions as the coordinate maps, but each pixel stores the new 3D position of the corresponding point after deformation.

These positional maps serve as input to a set of layer-specific neural networks, $\mathcal{F}_L$. Each network is trained to predict a corresponding front-and-back Gaussian map $\mathbf{M}_L \in \mathbb{R}^{2\times 2H\times 2H\times (14)}$ which represents the clothing or body part in the target pose θ_m:

$$\mathcal{M}_L \mapsto \big(\mathbf{C}_L, \mathbf{P}_L(\theta_m)\big) \xmapsto{\mathcal{F}_L} \mathbf{M}_L, \quad L \in \{b, u, \ell\}. \quad (1)$$

Each pixel in the output Gaussian map $\mathbf{M}_L$ defines a 3D Gaussian primitive with parameters: positional offset, rotation, opacity, scale, and color $[\Delta\mathbf{p}_i^c, \mathbf{q}_i, \alpha_i, \mathbf{s}_i, \mathbf{c}_i] \in \mathbb{R}^{14}$. The associated covariance matrix is calculated as $\mathbf{\Sigma}_i = \mathbf{RSS}^\intercal\mathbf{R}^\intercal$, where $\mathbf{R}$ is the rotation matrix derived from $\mathbf{q}_i$, and $\mathbf{S}$ is the diagonal scaling matrix derived from $\mathbf{s}_i$.

3.2. Learning Compositional Neural Garments

Given multi-view video frames $\{I_{m,n}\}$ and their corresponding body poses θ_m, our goal is to train a separate neural network for each garment layer to model its dynamics. To do so, the predicted shape-agnostic Gaussian primitives must be deformed back into the target posed space. This allows them to be rendered and compared against the ground-truth 2D images for computing training losses.

For a Gaussian primitive from any layer $L \in \{b, u, \ell\}$, we start with its canonical position $\mathbf{p}_i^c$ from the coordinate map $\mathbf{C}_L$. We first apply the network's predicted positional offset $\Delta\mathbf{p}_i^c$. Next, we transform this updated point from the canonical space into the final posed space by reapplying the subject-specific shape blendshapes β and performing LBS with respect to the pose θ_m. This process updates the position and rotation of each Gaussian primitive as follows:

$$\begin{aligned} \hat{\mathbf{p}}_i &= \sum_{j=1}^{55} w_{ij}\big(\mathbf{R}_j(\mathbf{p}_i^c + \Delta\mathbf{p}_i^c + \sum_{b=1}^{10} \beta_b \mathbf{o}_{bi}) + \mathbf{t}_j\big), \\ \hat{\mathbf{\Sigma}}_i &= \sum_{j=1}^{55} w_{ij}(\mathbf{R}_j \mathbf{\Sigma}_i \mathbf{R}_j^\intercal). \end{aligned} \tag{2}$$

Here, $(\mathbf{R}_j, \mathbf{t}_j)$ is the j-th bone transformation mapping from the canonical pose to the target pose θ_m, and w_{ij} are the skinning weights obtained via the voxelized field strategy discussed in Sec. 3.1.

After this deformation, we obtain a set of renderable 3D Gaussians for each layer $\mathcal{G}_L$ correctly positioned in the posed space. As shown in Fig. 2 (right), we composite the Gaussians from all layers and use a splatting-based rasterizer $\mathcal{R}_{splat}(\cdot)$ to render RGB images ($\hat{I}_{m,n}$) and segmentation masks $\hat{S}_{m,n}$ from the n-th camera viewpoint $\mathbf{\Pi}_n$:

$$\big(\mathcal{G}_b,\, \mathcal{G}_u,\, \mathcal{G}_\ell\big) \xmapsto{\mathcal{R}_{\text{splat}}(\mathbf{\Pi}_n)} \big(\hat{I}_{m,n},\, \hat{S}_{m,n}\big). \tag{3}$$

Finally, we jointly optimize all layer-specific U-Nets using the objective functions described next.

Photometric losses. Similar to standard 3D Gaussian Splatting, we enforce photometric loss between the rendered images ($\hat{I}_{m,n}$) and the ground-truth video frames ($I_{m,n}$). The loss is a combination of L_1 loss, SSIM [63] loss, and LPIPS [69] loss terms:

$$\begin{aligned} \mathcal{L}_{\text{im}} = \sum_{m,n} \big(&\|\hat{I}_{m,n} - I_{m,n}\|_1 + \\ &\lambda_1 \mathcal{L}_{\text{SSIM}}(\hat{I}_{m,n}, I_{m,n}) + \lambda_2 \mathcal{L}_{\text{Lpips}}(\hat{I}_{m,n}, I_{m,n})\big). \end{aligned} \tag{4}$$

To improve facial detail, we extract the face regions (F) from both the rendered and ground-truth images and compute an additional face-specific LPIPS loss:

$$\mathcal{L}_{\text{f}} = \sum_{m,n} \lambda_f \mathcal{L}_{\text{Lpips}}(\hat{F}_{m,n}, F_{m,n}). \tag{5}$$

Segmentation losses. A primary objective of our method is to ensure that each garment layer is correctly decomposed and separated. To enforce this, we employ a segmentation loss inspired by D3GA [74]. We first assign a distinct color

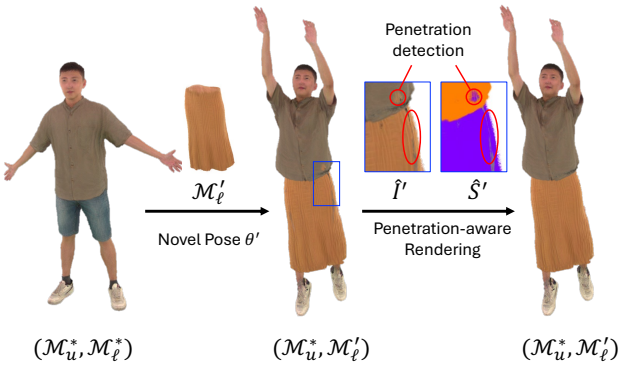

Figure 3. **Exemplar of avatar virtual try-on.** Given a reconstructed avatar (left), we replace its lower garment with a new skirt, $\mathcal{M}'_\ell$. The combined avatar can be animated to a novel pose θ', which may introduce minor penetration artifacts (middle). We resolve these artifacts on-the-fly during rendering with our online penetration detection algorithm (right).

to the Gaussians from each layer and render multi-class segmentation masks $\hat{S}_{m,n}$ (visualized in Fig. 2 (right)). These are compared against ground-truth masks $S_{m,n}$ obtained from an off-the-shelf segmentation model [62].

Furthermore, since the underlying body has minimal pose-dependent dynamics, its shape should remain close to the registered template. We enforce this by rendering body-only masks, $\hat{S}^b_{m,n} = \mathcal{R}_{\text{splat}}(\mathcal{G}_b; \mathbf{\Pi}_n)$, and comparing them to masks rendered from the input body template $S^b_{m,n}$. The total segmentation loss is:

$$\mathcal{L}_{\text{sg}} = \sum_{m,n} \big(\lambda_{\text{sg}} \|\hat{S}_{m,n} - S_{m,n}\|_2 + \lambda_{\text{bs}} \mathcal{L} \|\hat{S}^b_{m,n} - S^b_{m,n}\|_1\big) \tag{6}$$

Regularization. We include two types of regularization to ensure the reconstructed neural body and garments are physically plausible.

First, to prevent inner layers from penetrating outer layers, we introduce a penetration loss. For a Gaussian at position $\mathbf{p}_u$ on an outer layer (e.g., a shirt), we find its nearest neighbor at position $\mathbf{p}_b$ on the adjacent inner layer (e.g., the body). We then enforce that the signed distance between them along the inner layer's normal vector $\mathbf{n}_b$ should be larger than a minimum threshold ϵ. This is formulated as a squared hinge loss:

$$\mathcal{L}_{\text{pe}} = \max(\epsilon - (\mathbf{p}_u - \mathbf{p}_b) \cdot \mathbf{n}_b, 0)^2. \tag{7}$$

Note that this regularization is applied between all adjacent layers. Please see the supplementary material for details on estimating the normal vectors $\mathbf{n}_b$.

Second, we apply several geometric regularizations for stable convergence. These include an offset term $\mathcal{L}_{\text{o}} = ||\Delta\mathbf{p}||_2^2$ to penalize large Gaussian displacements; a

smoothing term $\mathcal{L}_{\text{sm}} = ||\Delta\mathbf{p} - \Delta\mathbf{p}_{\mathcal{N}}||_2^2$ to ensure coherent offsets among neighboring Gaussians; and a body opacity term $\mathcal{L}_{\text{bo}} = -\log(\alpha_b)$ to encourage the body layer to be fully opaque. The regularization terms are combined into a single loss:

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{pe}}\mathcal{L}_{\text{pe}} + \lambda_{\text{o}}\mathcal{L}_{\text{o}} + \lambda_{\text{sm}}\mathcal{L}_{\text{sm}} + \lambda_{\text{bo}}\mathcal{L}_{\text{bo}}. \tag{8}$$

Full Objective. Finally, we jointly optimize all U-Nets by minimizing the full loss function, which is a weighted sum of all components:

$$\mathcal{L}_{\text{full}} = \mathcal{L}_{\text{im}} + \mathcal{L}_{\text{f}} + \mathcal{L}_{\text{sg}} + \mathcal{L}_{\text{reg}}. \tag{9}$$

The $\lambda_{(\cdot)}$ terms are hyperparameters used to balance each loss component. Their specific values are listed in the supplementary material.

3.3. Avatar Virtual Try-on

Once trained, each shape-agnostic garment layer can be stored, reused, and seamlessly swapped across multiple human subjects. We exploit this compositional nature to develop a practical virtual try-on application.

In this application, the user's identity is defined by their body shape parameters β^* and the corresponding body template and U-Net $(\mathcal{M}_b^*, \mathcal{F}_b^*)$. This ensures that their unique skin tone, hairstyle, and face remain consistent during a try-on session. With the body fixed, we can replace any garment layer's template and network, such as $(\mathcal{M}_{\{u,\ell,o\}}^*, \mathcal{F}_{\{u,\ell,o\}}^*)$ with other clothing items from our digital wardrobe. For example, a user can swap shorts for a skirt by replacing $(\mathcal{M}_\ell^*, \mathcal{F}_\ell^*)$ with a new pair $(\mathcal{M}_\ell', \mathcal{F}_\ell')$ as shown in Fig. 3 (left). Finally, we composite a new set of 3D Gaussians for the complete outfit and render a final image $\hat{I}'$. This step follows the deformation process in Eq. (2) but uses the user's shape β^*, a novel driving poseθ', and a target camera matrix $\mathbf{\Pi}'$:

$$(\mathcal{M}_b^*,\, \mathcal{M}_u^*,\, \mathcal{M}_\ell') \xmapsto{(\theta',\beta^*)} (\mathcal{G}_b^*,\, \mathcal{G}_u^*,\, \mathcal{G}_\ell') \xmapsto{\mathcal{R}_{\text{splat}}(\mathbf{\Pi}')} \hat{I}'. \tag{10}$$

Penetration-aware Rendering. While our penetration regularizer effectively suppresses penetration during training, minor visual artifacts may still persist during animation, particularly with out-of-distribution poses (see Fig. 3 (middle)). To resolve this, we developed a simple yet efficient online correction algorithm that integrates directly into the rendering pipeline. During inference, alongside the RGB image, we render multi-class segmentation masks $\hat{S}'$ for the composed outfit. We then apply a contour-finding algorithm [57] to these masks to detect discontinuous regions, which indicate potential inter-garment penetration. By analyzing the depth values of the Gaussians within these detected regions, we can confirm if a pixel from an inner layer was rendered on top of an outer layer. If penetration is confirmed, we correct the erroneously rendered pixel by replacing its color with that of the correct outermost garment. As shown in Fig. 3 (right), this post-processing step effectively removes visual artifacts. This algorithm is efficient for correcting penetrations on the fly.

It is worth noting that we do not explicitly handle inter-layer penetration like previous methods [34, 60], since our goal is to produce realistic 2D rendering rather than to solve the correct 3D geometry. With our online correction strategy, we can mitigate the expensive geometry optimization process and simplify the problem to selecting among the rendering layers more efficiently.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
4D-DRESS [62]			
Animatable Gaussians [32]	27.75	0.9577	0.0531
LayGA [34]	27.58	0.9574	0.0543
Gaussian Wardrobe (Ours)	**28.06**	**0.9579**	**0.0527**
ActorsHQ [23]			
Animatable Gaussians [32]	27.92	0.9411	0.0418
LayGA [34]	27.80	0.9413	0.0421
Gaussian Wardrobe (Ours)	**28.38**	**0.9436**	**0.0397**

Table 1. **Evaluation of novel pose synthesis on 4D-DRESS and ActorsHQ datasets.** Best results are highlighted in **bold**. Our method consistently outperforms all baselines on all evaluation metrics (*cf*. Fig. 4).

4. Experiments

4.1. Experiment Setup

Datasets. We evaluate our method on two public datasets: 4D-DRESS [62] and ActorsHQ [23].

- 4D-DRESS contains multi-view videos of 32 subjects in diverse outfits and challenging poses. For our method, we select 7 subjects (16 distinct garments in total) and use 700-1100 frames from 12 camera views for training, holding out 300 frames per subject for evaluation. Due to the significant training time of baseline methods, we retrain them on a representative subset of 3 of these subjects for comparison.
- ActorsHQ features multi-view videos of 8 subjects performing simpler body poses. From this dataset, we utilize 3 subjects, with approximately 1000 training frames and 350 test frames captured from 14 views. All methods were trained on the same data for this dataset.

Evaluation Protocol. For the quantitative evaluation of novel pose synthesis, we report three standard metrics: PSNR, SSIM [63], and LPIPS [69]. Additionally, to specifically assess the quality of our garment decomposition in the ablation studies, we employ segmentation metrics including mean mIoU, Recall, and F1-Score.

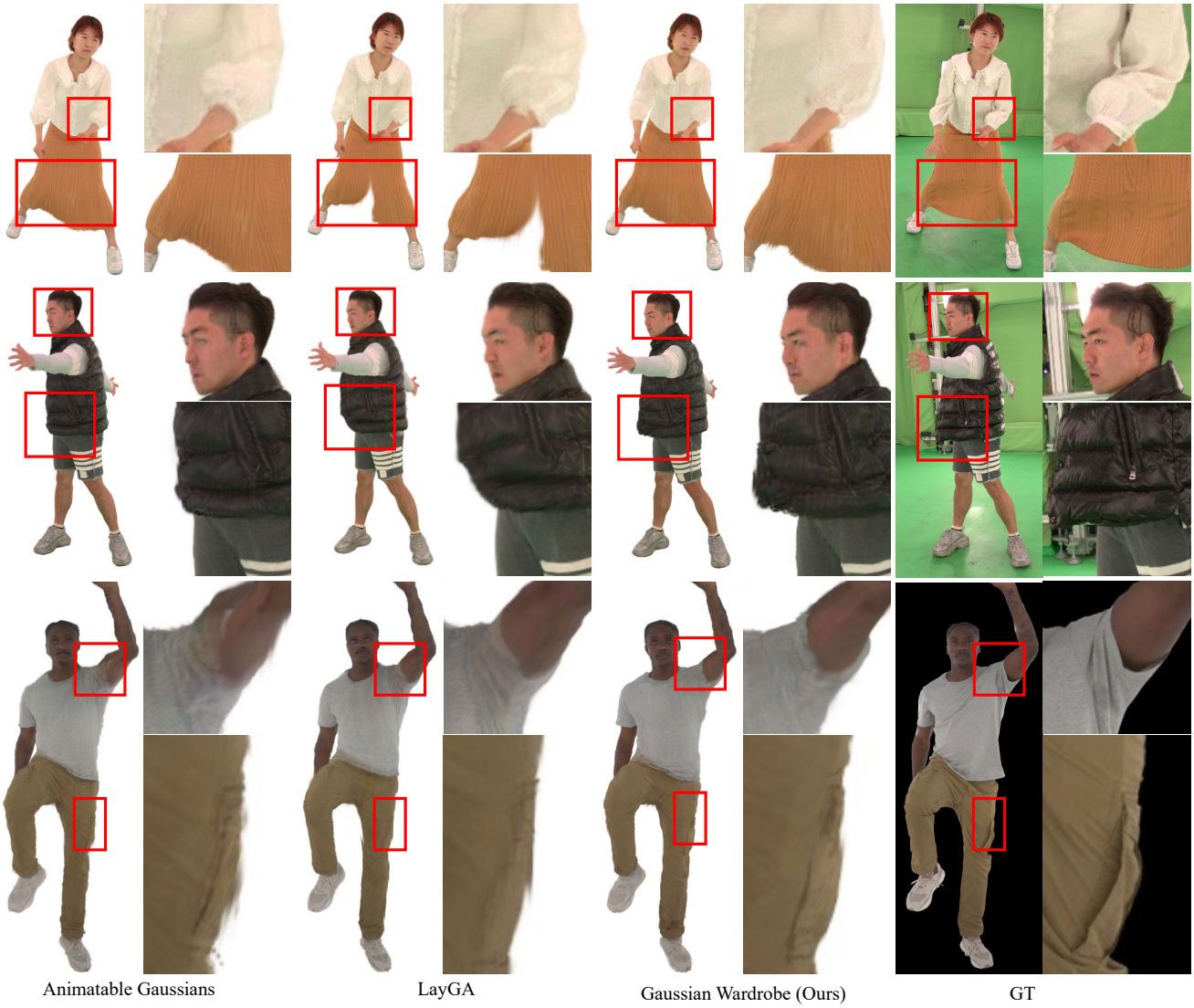

Figure 4. **Qualitative comparisons of novel pose synthesis on 4D-Dress and ActorsHQ datasets.** Our method can better model the dynamics of free-form garments, such as skirts (top) and vests (middle), and generate realistic renderings with sharper facial and garment details. In contrast, the baseline methods suffer from artifacts, such as blurry faces and semi-transparent clothing, and fail to reproduce fine details like wrinkles or pockets.

4.2. Novel Pose Synthesis Comparisons

We compare our method against two state-of-the-art baselines: Animatable Gaussians [32] and LayGA [34]. For Animatable Gaussians, we use the official public implementation and adhere to the default training configurations. As the source code for LayGA is not public, we re-implemented the method based on direct guidance from the original authors to ensure fidelity. To ensure a fair comparison, all methods are evaluated on the identical, held-out test sets corresponding to the subjects on which they were trained.

The quantitative results in Tab. 1, demonstrate that our method outperforms both baselines across all metrics on both datasets. Qualitative comparisons, shown in Fig. 4, further highlight the strengths of our approach. Our method better models the complex dynamics of free-form garments, such as the flowing motion of a skirt or the flaps of a vest, where SMPL-X mesh-based methods (i.e., LayGA) often fail. Furthermore, our approach captures finer details, particularly in facial regions and along clothing boundaries. We attribute this improvement to our explicit decomposition of the avatar, which allows the model to learn dedicated representations for the body and each garment layer. This addresses the blurred or entangled artifacts common in unified models like Animatable Gaussians. For more qualitative comparisons on free-form garments, please refer to the supplementary video.

Method	mIoU ↑	Recall ↑	F1-Score ↑
w/o $\mathcal{L}_{sg}$	0.848	0.898	0.924
w/o $\mathcal{L}_{reg}$	0.879	0.922	0.940
w/o $\mathcal{L}_{pe}$	0.883	0.927	0.942
Ours Full	**0.893**	**0.936**	**0.947**

Table 2. **Ablation study of the loss terms.** We assess the effectiveness of each loss term from the full objective. We rendered segmentation masks of garment layers and compared them against the ground-truth masks (*cf*. Fig. 5). The results show that the full model achieves the best performance.

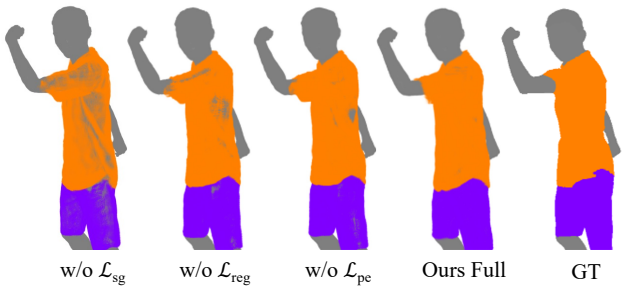

Figure 5. **Visualization on the effectiveness of loss terms.** We rendered segmentation masks of garment layers to visualize the impact of each regularization term. The evaluation shows that our full model produces cleanest results, while the absence of $\mathcal{L}_{\text{pe}}$, $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{sg}$ leads to self-penetration or irregular segmentation.

4.3. Ablation Studies

We conduct a series of ablation studies to validate the loss functions designed for our learning framework. In each experiment, we remove a specific loss term during training and evaluate its impact on garment decomposition using segmentation metrics. The quantitative and qualitative results are summarized in Tab. 2 and Fig. 5, respectively.

First, we examine the effect of the segmentation loss $\mathcal{L}_{sg}$. Without this term, the body and clothing representations become significantly entangled. As visualized during virtual try-on in Fig. 6, this entanglement prevents successful garment transfer, making the garment's appearance corrupted when applied to a new subject. Furthermore, the segmentation loss alone is insufficient to prevent physical artifacts like inter-layer penetration between the body and clothing. This is evident in the variants that lack our regularization $\mathcal{L}_{reg}$ and penetration $\mathcal{L}_{pe}$ losses, where visible penetration artifacts occur. By combining all proposed loss terms, our full model achieves a clean, penetration-free decomposition. These results validated the necessity of each component. Please refer to the supplementary material for more results.

4.4. Virtual Try-On Applications

A key application of our framework is a flexible 3D virtual try-on system. Because our learned garments are subject-agnostic, they can be reused and transferred across different subjects. As demonstrated in Fig. 7, these garments seamlessly adapt to new subjects with varying body shapes and identities. It is worth noting that the resulting avatars are not static; they can be animated by novel pose sequences and exhibit realistic, free-flowing dynamics in unseen poses. Please refer to our supplementary video for more demonstrations of these dynamic try-on results.

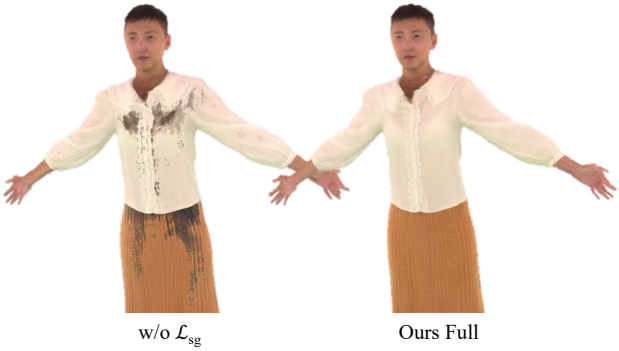

Figure 6. **Importance of segmentation loss in virtual try-on.** We visualize the results of virtual try-on using the variants our method trained with and without $\mathcal{L}_{sg}$. The garment's appearance becomes corrupted due to the entanglement of body and clothing.

Figure 7. **Virtual try-on application.** Our method enables flexible 3D virtual try-on across different subjects. The resulting avatars can be animated to a novel pose.

5. Conclusion

Conclusion. We introduced Gaussian Wardrobe, a novel framework for creating compositional 3D neural avatars. By representing garments as distinct, shape-agnostic 3D Gaussian models, our method enables flexible transfer of clothing across different subjects. Through extensive quantitative and qualitative experiments, we demonstrated state-of-the-art performance in both reconstruction and animation. We then showcased a virtual try-on application to validate practical relevance. In summary, Gaussian Wardrobe offers a promising paradigm for scalable, personalized, and dynamic digital wardrobes. We believe this work paves the way for the future of XR and virtual interaction.

Acknowledgements. This work was partially supported by the Swiss SERI Consolidation Grant "AI-PERCEIVE".

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[2] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[3] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[4] Yukang Cao, Masoud Hadi, Liang Pan, and Ziwei Liu. Gs-vton: Controllable 3d virtual try-on with gaussian splatting. *arXiv preprint arXiv:2410.05259*, 2024. 3

[5] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J. Black, and Otmar Hilliges. Fast-snarf: A fast deformer for articulated neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 3, 4, 2

[6] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[7] 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 Conference Paper*, 2022. 2

[8] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J Black. Learning disentangled avatars with hybrid 3d representations. *arXiv preprint arXiv:2309.06441*, 2023. 2, 3

[9] Artur Grigorev, Michael J Black, and Otmar Hilliges. Hood: Hierarchical graphs for generalized modelling of clothing dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[10] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3D avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[11] Chen Guo, Tianjian Jiang, Manuel Kaufmann, Chengwei Zheng, Julien Valentin, Jie Song, and Otmar Hilliges. Reloo: Reconstructing humans dressed in loose garments from monocular video in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2, 3

[12] Chen Guo, Junxuan Li, Yash Kant, Yaser Sheikh, Shunsuke Saito, and Chen Cao. Vid2avatar-pro: Authentic avatar from videos in the wild via universal prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[13] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 2019. 2

[14] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics (TOG)*, 2021. 2

[16] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[17] Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

[18] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[19] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[20] Hsuan-I Ho, Chen Guo, Po-Chen Wu, Ivan Shugurov, Chengcheng Tang, Abhay Mittal, Sizhe An, Manuel Kaufmann, and Linguang Zhang. Phd: Personalized 3d human body fitting with point diffusion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2025. 2

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

[22] Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[23] 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)*, 2023. 2, 6

[24] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[25] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[26] Tianjian Jiang, Hsuan-I Ho, Manuel Kaufmann, and Jie Song. Prioravatar: Efficient and robust avatar creation from monocular video using learned priors. In *SIGGRAPH Asia Conference Paper*, 2025.

[27] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[28] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 2023. 2

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 2

[31] Yifei Li, Hsiao-yu Chen, Egor Larionov, Nikolaos Sarafianos, Wojciech Matusik, and Tuur Stuyck. Diffavatar: Simulation-ready garment optimization with differentiable simulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[32] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4, 6, 7, 1

[33] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[34] Siyou Lin, Zhe Li, Zhaoqi Su, Zerong Zheng, Hongwen Zhang, and Yebin Liu. Layga: Layered gaussian avatars for animatable clothing transfer. In *SIGGRAPH Conferenc Paper*, 2024. 3, 6, 7, 1, 2

[35] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 2

[36] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J. Black. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[37] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[38] Qianli Ma, Jinlong Yang, Michael J. Black, and Siyu Tang. Neural point-based shape modeling of humans in challenging clothing. In *International Conference on 3D Vision (3DV)*, 2022. 3

[39] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[40] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[41] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 3d clothed human reconstruction in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[42] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2, 3

[43] Hui En Pang, Shuai Liu, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Disco4d: Disentangled 4d human generation and animation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[44] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[45] Pramish Paudel, Anubhav Khanal, Danda Pani Paudel, Jyoti Tandukar, and Ajad Chhatkuli. ihuman: Instant animatable digital humans from monocular videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 304–323, 2024. 2

[46] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[47] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[48] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 3

[49] Sergey Prokudin, Qianli Ma, Maxime Raafat, Julien Valentin, and Siyu Tang. Dynamic point fields. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

[50] Boxiang Rong, Artur Grigorev, Wenbo Wang, Michael J. Black, Bernhard Thomaszewski, Christina Tsalicoglou, and Otmar Hilliges. Gaussian Garments: Reconstructing simulation-ready clothing with photorealistic appearance from multi-view video. In *International Conference on 3D Vision (3DV)*, 2025. 3

[51] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[52] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Snug: Self-supervised neural dynamic garments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[53] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[54] Dan Song, Xuanpu Zhang, Juan Zhou, Weizhi Nie, Ruofeng Tong, Mohan Kankanhalli, and An-An Liu. Image-based virtual try-on: A survey. *International Journal of Computer Vision (IJCV)*, 2025. 3

[55] Zhaoqi Su, Liangxiao Hu, Siyou Lin, Hongwen Zhang, Shengping Zhang, Justus Thies, and Yebin Liu. Caphy: Capturing physical properties for animatable human avatars. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

[56] Ke Sun, Jian Cao, Qi Wang, Linrui Tian, Xindi Zhang, Lian Zhuo, Bang Zhang, Liefeng Bo, Wenbo Zhou, Weiming Zhang, and Daiheng Gao. Outfitanyone: Ultra-high quality virtual try-on for any clothing and any person. *arXiv preprint arXiv:2407.16224*, 2024. 3

[57] Satoshi Suzuki and Keiichi Abe. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 1985. 6, 1

[58] Jeff Tan, Donglai Xiang, Shubham Tulsiani, Deva Ramanan, and Gengshan Yang. Dressrecon: Freeform 4d human reconstruction from monocular video. In *International Conference on 3D Vision (3DV)*, 2025. 2, 3

[59] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[60] Onat Vuran and Hsuan-I Ho. Remu: Reconstructing multi-layer 3d clothed human from images. In *British Machine Vision Conference (BMVC)*, 2025. 3, 6

[61] Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. Styleavatar: Real-time photo-realistic portrait avatar from a single video. In *SIGGRAPH Conferenc Paper*, 2023. 1

[62] 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, 4, 5, 6

[63] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003. 5, 6

[64] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[65] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[66] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 2018. 2

[67] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[68] Hongwen Zhang, Siyou Lin, Ruizhi Shao, Yuxiang Zhang, Zerong Zheng, Han Huang, Yandong Guo, and Yebin Liu. Closet: Modeling clothed humans on continuous surface with explicit template decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[69] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5, 6

[70] Weitian Zhang, Sijing Wu, Manwen Liao, and Yichao Yan. Disentangled clothed avatar generation via layered representation. *arXiv preprint arXiv:2501.04631*, 2025. 3

[71] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[72] Yang Zheng, Qingqing Zhao, Guandao Yang, Wang Yifan, Donglai Xiang, Florian Dubost, Dmitry Lagun, Thabo Beeler, Federico Tombari, Leonidas Guibas, et al. Physavatar: Learning the physics of dressed 3d avatars from visual observations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 3

[73] Jiayin Zhu, Linlin Yang, and Angela Yao. Instructhumans: Editing animated 3d human textures with instructions. *arXiv preprint arXiv:2404.04037*, 2024. 3

[74] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. In *International Conference on 3D Vision (3DV)*, 2025. 3, 5

[75] Xingxing Zou, Xintong Han, and Waikeung Wong. Cloth4d: A dataset for clothed human reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3