GALA: Generating Animatable Layered Assets from a Single Scan

Taeksoo Kim[1*]   Byungjun Kim[1*]   Shunsuke Saito[2]   Hanbyul Joo[1]

[1]Seoul National University   [2]Codec Avatars Lab, Meta

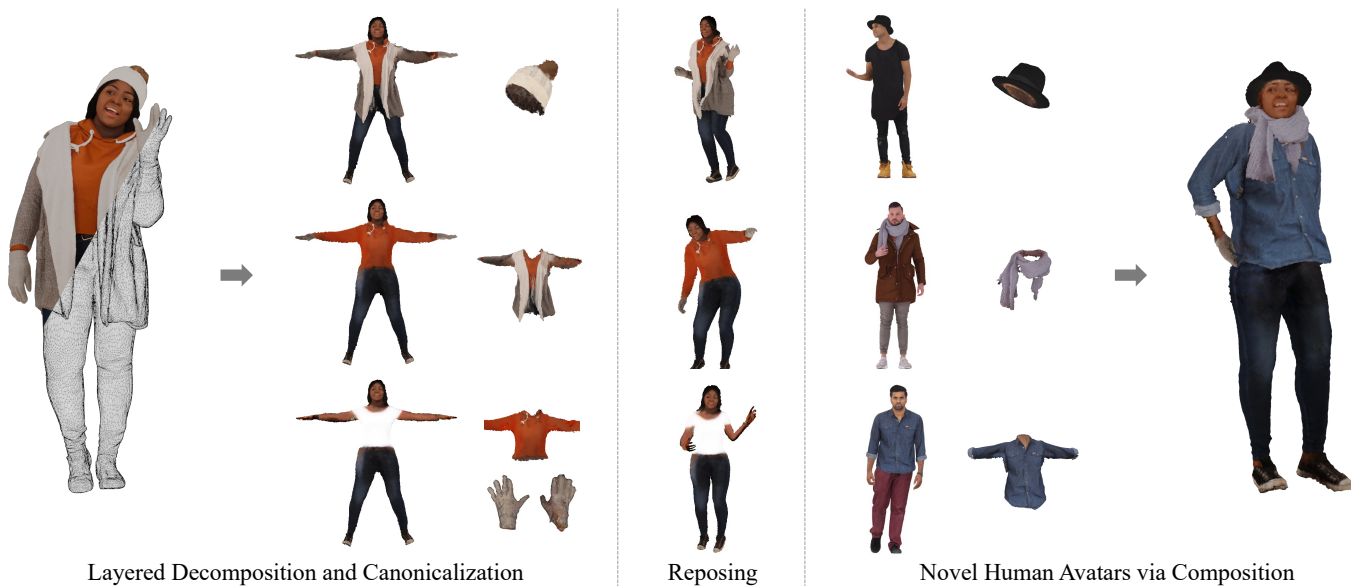{taeksu98, byungjun.kim, hbjoo}@snu.ac.kr   shunsukesaito@meta.com

https://snuvclab.github.io/gala/

Figure 1. **GALA.** Given a single-layer 3D mesh of a clothed human (left), our approach enables **G**eneration of **A**nimatable **L**ayered **A**ssets for 3D garment transfer and avatar customization in any poses by decomposing and inpainting the geometry and texture of each layer with a pretrained 2D diffusion model in a canonical space.

Layered Decomposition and Canonicalization   Reposing   Novel Human Avatars via Composition

Abstract

*We present **GALA**, a framework that takes as input a single-layer clothed 3D human mesh and decomposes it into complete multi-layered 3D assets. The outputs can then be combined with other assets to create novel clothed human avatars with any pose. Existing reconstruction approaches often treat clothed humans as a single-layer of geometry and overlook the inherent compositionality of humans with hairstyles, clothing, and accessories, thereby limiting the utility of the meshes for down-stream applications. Decomposing a single-layer mesh into separate layers is a challenging task because it requires the synthesis of plausible geometry and texture for the severely occluded regions. Moreover, even with successful decomposition, meshes are not normalized in terms of poses and body shapes, failing coherent composition with novel identities and poses. To address these challenges,*

we propose to leverage the general knowledge of a pretrained 2D diffusion model as geometry and appearance prior for humans and other assets. We first separate the input mesh using the 3D surface segmentation extracted from multi-view 2D segmentations. Then we synthesize the missing geometry of different layers in both posed and canonical spaces using a novel pose-guided Score Distillation Sampling (SDS) loss. Once we complete inpainting high-fidelity 3D geometry, we also apply the same SDS loss to its texture to obtain the complete appearance including the initially occluded regions. Through a series of decomposition steps, we obtain multiple layers of 3D assets in a shared canonical space normalized in terms of poses and human shapes, hence supporting effortless composition to novel identities and reanimation with novel poses. Our experiments demonstrate the effectiveness of our approach for decomposition, canonicalization, and composition tasks compared to existing solutions.

*Equal contribution

1. Introduction

As social interactions become increasingly online, the ability to customize digital representations of oneself is more important than ever. This is particularly critical in the domain of virtual try-on and photorealistic avatar customization. However, creating assets that can be easily layered on top of any avatars typically requires substantial manual efforts by artists. Our goal is to enable the automatic creation of reusable 3D layered assets that can be effortlessly composed to any human with any pose.

Unlike artist-created 3D assets, reconstruction-based 3D assets are becoming widely accessible. In addition to high-quality 3D scans [4, 62], single-view reconstruction methods [3, 66, 67] or text-to-3d generation techniques [12, 44, 60] further simplify the creation of 3D models. Despite these advancements, using these 3D models for virtual try-on or avatar customization remains an open challenge because these models are typically single-layered. Different attributes such as hair, clothing, and accessories are glued into a single triangle mesh, and anything beneath the outermost layer is fully occluded. Moreover, self-contact regions are connected, making re-animation challenging.

To address this, we propose a fully automatic framework for creating compositional layered 3D assets from a single-layer scan. Unlike the existing text-based 3D generation methods [44, 60] that only support the generation of each asset in isolation, our approach learns to decompose a mesh into multiple layers and inpaint missing geometry and appearance for compositing the decomposed assets into novel identities. Our key idea is to complement missing geometric and appearance information by leveraging a strong image prior built from a large-scale image collections. In particular, we leverage a latent diffusion model [63] that is trained on an extremely large corpus of images. Using a score distillation sampling (SDS), we inpaint the occluded regions while retaining the originally visible regions.

For reposing, simply inpainting the geometry and appearance in an input posed space is not sufficient. For garment transfer across different identities with various poses, we need to represent the target asset and the remaining human layer in individual canonical spaces. However, we observe that the vanilla SDS loss often provides poor guidance by ignoring the target pose information. We address the lack of pose-sensitivity in the SDS loss by introducing a pose-guided SDS loss. Specifically, we derive the SDS loss with a pose-conditioned diffusion model [88]. This allows us to supervise the shape and appearance jointly in both posed and canonical spaces. Once we obtain the canonicalized object and human layers, we can mix and match with other assets to create virtual try-on as shown in Fig. 1. The composite results are further refined with penetration handling.

As there is no established benchmark for decomposition, canonicalization, and composition from a single scan, we establish a new evaluation protocol to quantitatively assess our approach. For decomposition, our approach significantly outperforms recent text-driven 3D editing methods. We also show that the proposed pose-guided SDS enables robust canonicalization even for challenging cases, outperforming existing methods. Lastly, we show garment transfer to create novel avatars only from a collection of single-layer clothed humans. Our contributions can be summarized as follows:

- We propose a new task of multi-layer decomposition and composition from a single-layer scan, which offers a practical compositional asset creation pipeline.
- We present a pose-guided SDS loss, enabling the robust modeling of layered clothed humans in a canonical space for garment transfer and reposing from a single scan.
- We provide a comprehensive analysis of generating animatable layered assets from a single scan with a newly established evaluation protocol. We will release code for benchmarking future research on this novel task.

2. Related Work

Clothed Human Modeling. 3D parametric human models [35, 46, 58, 82] have been proposed to model diverse poses and shapes of humans, allowing us to reconstruct minimally clothed 3D humans [8, 36, 58, 64, 87]. To represent clothed humans, follow-up work leverages 3D displacements on top of the template body model [1, 2, 47], or separate mesh layers [6, 59]. Yet, the topological constraints and the resolution of the template model limit their ability to model clothing with complex shapes and high-frequency details. In recent years, deep implicit shape representations [17, 48, 53, 56, 79] have emerged as a significant breakthrough in modeling 3D humans, demonstrating their efficacy in reconstructing detailed clothed humans from images, scans, depth maps, or pointclouds [22, 52, 66–68, 74, 76, 80, 81, 85, 89]. Extending work enables the animations of these reconstructions [14, 16, 22, 51, 52, 68, 74, 76] by learning a canonical 3D shape in a space normalized in terms of human poses and shapes. Since these approaches treat the clothed human as a single-layer mesh, several work [5, 7, 13, 57, 59, 75] attempts to model the clothing of humans as a separate layer. SMPLicit [19] models clothing with implicit shape representation on top of the parametric mesh model. There are a few attempts to enable compositional and animatable modeling of avatars. SCARF [23] and MEGAN separately models humans and clothing from video observations using a hybrid representation of mesh and NeRF [53]. MEGANE [42] models high-fidelity compositional heads and eyeglasses from multi-view videos. NCHO [40] learns a compositional generative model of humans and objects from multiple scans with and without objects. Unlike existing approaches, our approach enables the modeling of animatable *multi-layer* assets from a *single scan*. To enable this, we exploit an image prior from a pretrained diffusion model [63].

3D Content Generation. Recent advancements in 3D representations [53, 79] and generative modeling [26, 29] have spurred active research for 3D content generation. Generation from text, in particular, has gained popularity due to its intuitive interface. Early work like Text2Shape [11] trains text and shape encoders to learn joint embeddings to generate text-consistent 3D shapes. Due to the challenges of collecting large-scale paired text-3D datasets, several approaches [30, 34, 50, 54] utilize pretrained CLIP model [61] for text-guided 3D content generation. With the recent rise of diffusion models [29, 73] for high-quality image generation [21, 63], DreamFusion [60] proposes score distillation sampling (SDS) loss for optimizing 3D scenes represented as NeRF [53] by leveraging the 2D diffusion prior. Various 3D representations such as point clouds [55, 86], meshes [12, 44, 45], and neural fields [49, 70] have also been utilized. Some approaches [31, 32, 39, 72, 77] incorporates additional 3D datasets with diffusion model to enable high-quality 3D generation. MVdream [72] generates multi-view images by finetuning the diffusion model with multi-view rendering of Objaverse [20]. Chupa [39] and HumanNorm [32] finetune the diffusion model to generate normal or depth maps of 3D humans. However, current 3D content generation methods generate 3D assets as a *single-layer mesh*, limiting their utility for composition with other assets. In contrast, our approach leverages the 2D diffusion prior to create decomposed layers of attributes in a canonical space, facilitating garment transfer and reposing.

3D Editing and Completion. Editing or completing 3D scenes has traditionally been a task for experienced artists, but recent work [18, 25, 27, 38, 43, 69, 90] shows the great potential of text-based automatic 3D content manipulation. Instruct-NeRF2NeRF [27] edits the pretrained NeRF using prompt by iteratively updating training images. DreamEditor [90] exploits mesh-based neural fields [83] to enable local and flexible editing via SDS loss [60] using a diffusion model finetuned with DreamBooth [65]. Vox-E [69] similarly utilizes SDS loss but enables local editings using the 3D attention map aggregated from the 2D attention maps of a diffusion model. FocalDreamer [43] employs an additive approach to edit the geometry of input 3D scans, creating reusable independent assets. While our approach shares the motivation of FocalDreamer [43] in the sense of generating reusable 3D assets, our method does not require the designated editing region as an additional input and focuses on the *decomposition* of the input 3D scan into multiple reusable layers instead of the addition of new components.

3. Preliminaries

3.1. Score Distillation Sampling

To synthesize 3D scenes without requiring large-scale 3D datasets, DreamFusion [60] introduces Score Distillation

Sampling (SDS) loss. SDS loss leverages the knowledge of a pretrained 2D diffusion model. Given the target prompt, the loss optimizes over the 3D volume parameterized with θ using the differentiable renderer g, such that the generated image $\boldsymbol{x} = g(\theta)$ closely resembles samples from the frozen diffusion model, ϕ. The gradient of the loss is calculated as,

$$\nabla_\theta \mathcal{L}^{SDS}(\boldsymbol{x}, \phi) = \mathbb{E}\left[\omega(t)(\hat{\epsilon}_\phi(\boldsymbol{x}_t; \boldsymbol{y}, t) - \epsilon)\frac{\partial \boldsymbol{x}}{\partial \theta}\right], \quad (1)$$

where $\boldsymbol{y}$ denotes text condition and t is the noise level. $\boldsymbol{x}_t$ denotes the noised image, $\hat{\epsilon}_\phi(\boldsymbol{x}_t; \boldsymbol{y}, t)$ represents the noise prediction for the sampled noise ϵ, and $\omega(t)$ is the weighting function defined by the scheduler of the diffusion model.

3.2. Deep Marching Tetrahedra

We adopt Deep Marching Tetrahedra [71] (DMTet) as our geometric representation. DMTet employs a deformable tetrahedral grid denoted as (X_T, T), where X_T represents the grid's 3D vertices and T defines the tetrahedral structure, where each tetrahedron contains four vertices in X_T. For each vertex $\boldsymbol{x}_i \in X_T$, DMTet predicts the signed distance value $s(\boldsymbol{x}_i)$ from the surface and the position offset $\Delta\boldsymbol{x}_i$ of each vertex and extracts a triangular mesh from the implicit field using the differentiable Marching Tetrahedral (MT) layer. Since the pipeline is fully differentiable, losses defined explicitly on the surface mesh can be used for optimizing the surface geometry represented by DMTet.

4. Method

Our method decomposes a single-layer 3D human scan into two complete layers of the target object and the rest of the scan in separate canonical spaces. Following the previous work [12], we first model the geometry and subsequently model the appearance, and adopt DMTet [71] as our geometric representation (Sec. 4.1). To reconstruct visible parts of each layer, we lift multi-view 2D segmentations of the target object onto the input 3D scan. Using forward linear blend skinning (LBS), we transform the canonical geometry of each layer to the pose of the input scan and reconstruct visible parts based on the acquired segmentation. We further leverage a 2D diffusion prior via our pose-guided SDS loss applied in canonical space to enable canonicalization of a *single* scan and complete the geometry of the occluded regions (Sec. 4.2). Once we optimize the geometry of each layer, we model the appearance using similar SDS losses in the canonical space (Sec. 4.3). Lastly, we refine the composition of the decomposed layers by reducing self-penetration (Sec. 4.4). Fig. 2 shows an overview of our pipeline.

4.1. Representation and Initialization

We model the geometry of the human and an object in separate canonical spaces using DMTet [71]. For a given
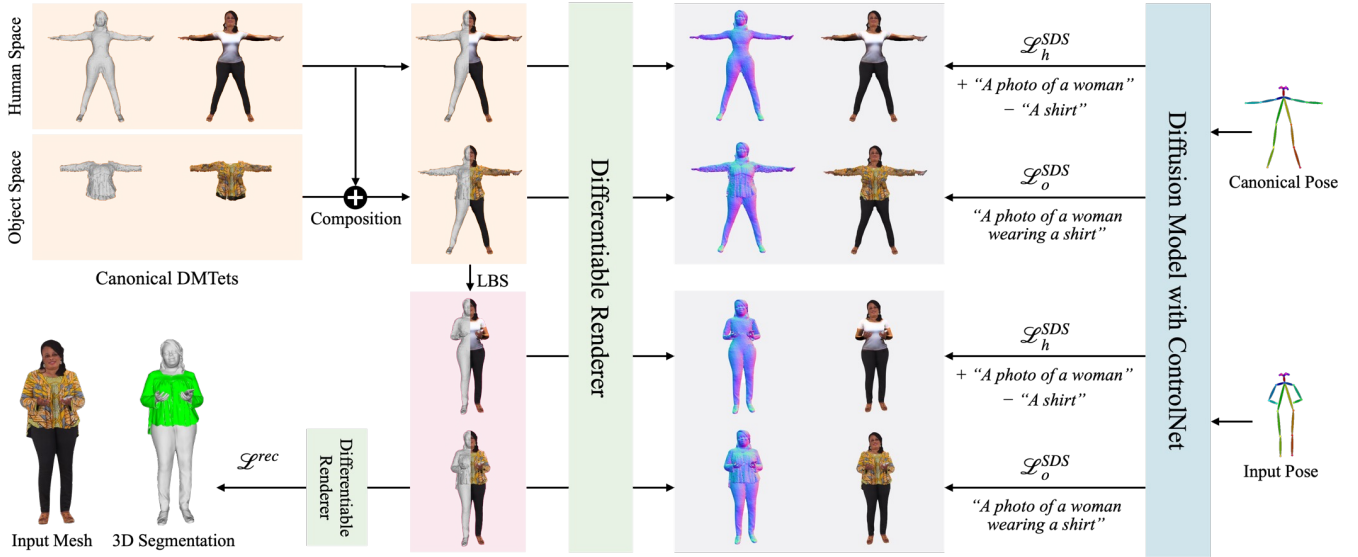
Figure 2. **Overview.** GALA learns an object and the remaining human layers in a canonical space using DMTet [71]. The canonical space colored orange and the original posed space colored purple are differentiably associated with linear blend skinning (LBS). Our novel pose-guided SDS loss (right) guides the decomposition and inpainting in both the canonical and posed space. We also retain the fidelity of visible regions via a reconstruction and segmentation loss (left-bottom).

tetrahedral grid for the human (X_{T_h}, T_h) and for the object (X_{T_o}, T_o), we utilize MLP networks $\boldsymbol{\Psi}_h$ and $\boldsymbol{\Psi}_o$ to predict the signed distance and the deformation offset of every vertex of the grids. Using the predicted signed distance and offset, the canonical human mesh, $\mathcal{M}_h^c = (\mathcal{V}_h^c, \mathcal{F}_h)$, and the canonical object mesh, $\mathcal{M}_o^c = (\mathcal{V}_o^c, \mathcal{F}_o)$, can be extracted from each grid via a differentiable MT layer, where $\mathcal{V}_h^c$ and $\mathcal{V}_h^c$ denotes the vertices, and $\mathcal{F}_h$ and $\mathcal{F}_o$ denotes the faces of each mesh. To obtain a posed mesh, we transform every vertex of the reconstructed mesh via forward linear blend skinning (LBS) [14, 68], utilizing the skinning weights of the nearest neighbor vertex of the canonical SMPL-X mesh [58]. Formally, a vertex $\boldsymbol{v}^c \in \mathcal{V}_h^c \cup \mathcal{V}_o^c$ in canonical space is transformed into a posed space with,

$$\bar{\boldsymbol{v}}^p = (\sum_{i=1}^{n_b} w_i \cdot \mathbf{T}_i(\boldsymbol{\beta}, \boldsymbol{\theta})) \cdot \begin{bmatrix} \mathbf{I} & \mathcal{B}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) \\ 0 & 1 \end{bmatrix} \cdot \bar{\boldsymbol{v}}^c, \quad (2)$$

where $\bar{\boldsymbol{v}}^p, \bar{\boldsymbol{v}}^c$ are homogeneous coordinates of $\boldsymbol{v}^p, \boldsymbol{v}^c$ respectively, n_b is the number of bones, w_i is the blend skinning weight of the bone i, and $\mathbf{T}_i(\boldsymbol{\beta}, \boldsymbol{\theta}) \in \mathbb{R}^{4 \times 4}$ is the transformation of the bone i in SMPL-X model given shape parameter $\boldsymbol{\beta} \in \mathbb{R}^{10}$ and pose parameter $\boldsymbol{\theta} \in \mathbb{R}^{55 \times 3}$. Blend shapes $\mathcal{B}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi})$ are the summation of identity blend shapes, pose blend shapes, and the expression blend shapes, where $\boldsymbol{\psi} \in \mathbb{R}^{10}$ is the expression parameter. By transforming all vertices, we get the posed human mesh, $\mathcal{M}_h^p = (\mathcal{V}_h^p, \mathcal{F}_h)$, and the posed object mesh, $\mathcal{M}_o^p = (\mathcal{V}_o^p, \mathcal{F}_o)$. For ease of notation, we use $LBS(\cdot)$ to specify the relationship between the canonical mesh and posed mesh as follows:

$$\mathcal{M}_{\{h,o\}}^p = LBS(\mathcal{M}_{\{h,o\}}^c). \quad (3)$$

We initialize our DMTets using SMPL-X mesh in canonical pose. We sample points $\boldsymbol{q} \in \mathbb{R}^3$ in each space, compute

the signed distance $SDF(\boldsymbol{q})$ from each point to the SMPL-X mesh, and optimize the following loss functions.

$$\mathcal{L}_h^{init} = \|s(\boldsymbol{q}; \boldsymbol{\Psi}_h) - SDF(\boldsymbol{q})\|_2^2 \quad (4)$$

$$\mathcal{L}_o^{init} = \|s(\boldsymbol{q}; \boldsymbol{\Psi}_o) - SDF(\boldsymbol{q})\|_2^2. \quad (5)$$

4.2. Geometry Decomposition and Canonicalization

Given an input scan, we decompose and canonicalize the scan into two separate geometries of human and object, $\mathcal{M}_h^{c*}, \mathcal{M}_o^{c*}$, which minimizes the following total loss:

$$\mathcal{L}_{geo} = \lambda_{h_{geo}}^{rec} \mathcal{L}_{h_{geo}}^{rec} + \lambda_{o_{geo}}^{rec} \mathcal{L}_{o_{geo}}^{rec} \quad (6)$$
$$+ \lambda_{h_{geo}}^{SDS} \mathcal{L}_{h_{geo}}^{SDS} + \lambda_{o_{geo}}^{SDS} \mathcal{L}_{o_{geo}}^{SDS} + \lambda_{comp}^{seg} \mathcal{L}_{comp}^{seg},$$

$$\mathcal{M}_h^{c*}, \mathcal{M}_o^{c*} = \underset{\mathcal{M}_h^c, \mathcal{M}_o^c}{\arg\min} \mathcal{L}_{geo}. \quad (7)$$

We describe each loss in the following.

Reconstruction Loss. To decouple the geometry of the human and object, we employ the 3D surface segmentation of the target object. Specifically, we rasterize the scan from multiple viewpoints and perform binary segmentation in 2D, distinguishing the target object from other parts using an off-the-shelf open-vocabulary segmentation tool [41]. Utilizing the aggregated pixel-to-face correspondence established during the rasterization process, we cast votes for each face of the mesh to determine whether it belongs to the specified object or not. Consequently, the given input scan in posed space, denoted as $\mathcal{M}^{scan}$, is partitioned into two incomplete surface meshes: the object mesh, $\mathcal{M}_o^{scan}$, and the remaining human figure mesh, $\mathcal{M}_h^{scan}$, as shown in Fig. 3.

To preserve the identity of visible regions of the input scan, we employ rendering-based reconstruction losses in the posed space. Using a differentiable rasterizer $\mathcal{R}$ and a sampled camera $\boldsymbol{k}$, we render masks $\mathbf{A} \in \{0,1\}^{H \times W}$ and normal maps $\mathbf{N} \in \mathbb{R}^{H \times W}$ of the generated posed meshes $\mathcal{M}_h^p$ and $\mathcal{M}_o^p$, where H, W are the height and width of the rendered masks and normal maps.

$$\mathbf{A}_h^p, \mathbf{N}_h^p = \mathcal{R}(\mathcal{M}_h^p, \boldsymbol{k}) = \mathcal{R}(LBS(\mathcal{M}_h^c), \boldsymbol{k}) \qquad (8)$$

$$\mathbf{A}_o^p, \mathbf{N}_o^p = \mathcal{R}(\mathcal{M}_o^p, \boldsymbol{k}) = \mathcal{R}(LBS(\mathcal{M}_o^c), \boldsymbol{k}) \qquad (9)$$

Together with the mask and normal map of the input mesh, we additionally render segmentation masks $\mathbf{S}_h^{scan}, \mathbf{S}_o^{scan} \in \{0,1\}^{H \times W}$ for the human and the object using the 3D surface segmentation:

$$\mathbf{A}^{scan}, \mathbf{N}^{scan}, \mathbf{S}_h^{scan}, \mathbf{S}_o^{scan} = \mathcal{R}(\mathcal{M}^{scan}, \boldsymbol{k}). \qquad (10)$$

Finally, the losses for reconstruction are defined as follows:

$$\mathcal{L}_{h_{geo}}^{rec} = \|\mathbf{N}_h^p \odot \mathbf{S}_h^{scan} - \mathbf{N}^{scan} \odot \mathbf{S}_h^{scan}\|_2^2, \qquad (11)$$

$$\mathcal{L}_{o_{geo}}^{rec} = \|\mathbf{N}_o^p \odot \mathbf{S}_o^{scan} - \mathbf{N}^{scan} \odot \mathbf{S}_o^{scan}\|_2^2 \qquad (12)$$
$$+ \|\mathbf{A}_o^p - \mathbf{A}^{scan} \odot \mathbf{S}_o^{scan}\|_2^2,$$

where $\odot$ is the Hadamard product. We employ extra mask loss to regularize the shape of the object in posed space, assuming that the object is layered on top of the human. Furthermore, to capture the intricate details of human faces and hands, we render close-up views of these regions by zooming in on the corresponding joints of the posed SMPL-X mesh and apply the same reconstruction losses.

Pose-guided SDS Loss. Our goal is to obtain complete 3D assets in a neutral pose from a *single* posed scan, which can be animated into arbitrary poses without artifacts. The core challenges lie in the difficulty of (1) completing the occluded regions of both assets and (2) modeling canonical shape of each asset from a single scan. To overcome both challenges, we propose a pose-guided SDS loss that leverages the prior of the pretrained diffusion model equipped with ControlNet [88] conditioned with OpenPose poses [10]. The gradient of our pose-guided SDS loss is defined as:

$$\nabla_{\boldsymbol{\Psi}} \mathcal{L}_{pose}^{SDS}(\boldsymbol{z}_t(\mathbf{X}), \boldsymbol{y}, \boldsymbol{p}, \phi)$$
$$= \mathbb{E}[\omega(t)(\hat{\epsilon}_\phi(\boldsymbol{z}_t(\mathbf{X}); \boldsymbol{y}, \boldsymbol{p}, t) - \epsilon) \frac{\partial \mathbf{X}}{\partial \boldsymbol{\Psi}} \frac{\partial \boldsymbol{z}_t(\mathbf{X})}{\partial \mathbf{X}}], \qquad (13)$$

where $\mathbf{X}$ is the rendered normal or texture of the mesh $\mathcal{M}$, $\boldsymbol{z}_t(\mathbf{X})$ is the latent embedding with noise from the forward process. $\boldsymbol{y}$ represents the positive and negative text prompts where positive prompts describe the underlying human and negative prompts describe the target object to remove. $\boldsymbol{p}$

is the pose condition for ControlNet [88] converted by the mapping from SMPL-X joints to OpenPose joints.

However, when the pose-guided SDS loss and reconstruction loss are applied in the posed space through the forward transformation of Eq. (2), the output canonical shape suffers from undesired artifacts due to many-to-one mapping from the canonical space to the posed space (see Fig. 9 (b)). While previous approaches [14, 78] address this ambiguity by jointly learning from multiple scans or images with various poses, we observe that these approaches perform poorly when given only a single scan.

To enable plausible canonicalization from a single scan, we apply our pose-guided SDS loss (Eq. (13)) in the canonical space. The gradients are derived as follows:

$$\nabla_{\boldsymbol{\Psi}_h} \mathcal{L}_{h_{geo}}^{SDS} = \nabla_{\boldsymbol{\Psi}_h} \mathcal{L}_{pose}^{SDS}(\boldsymbol{z}_t(\tilde{\mathbf{N}}_h^c), \boldsymbol{y}_h, \boldsymbol{p}^c, \phi), \qquad (14)$$

$$\nabla_{\boldsymbol{\Psi}_o} \mathcal{L}_{o_{geo}}^{SDS} = \nabla_{\boldsymbol{\Psi}_o} \mathcal{L}_{pose}^{SDS}(\boldsymbol{z}_t(\tilde{\mathbf{N}}_{comp}^c), \boldsymbol{y}_{comp}, \boldsymbol{p}^c, \phi), \qquad (15)$$

where $\mathcal{L}_{h_{geo}}^{SDS}, \mathcal{L}_{o_{geo}}^{SDS}$ are the loss for the human and object space, respectively. $\tilde{\mathbf{N}}_h^c, \tilde{\mathbf{N}}_{comp}^c$ are the rendered normal map concatenated with the mask of the human mesh and the composite mesh in the canonical space, and $\boldsymbol{z}_t(\tilde{\mathbf{N}}_h^c), \boldsymbol{z}_t(\tilde{\mathbf{N}}_{comp}^c)$ are the downsampled version of them with noise produced by the forward diffusion process as in Fantasia3D [12]. $\boldsymbol{y}_h$, $\boldsymbol{y}_{comp}$ are the text prompts for the human and object space, and $\boldsymbol{p}^c$ is the neutral pose condition. Remarkably, our pose-guided SDS loss in the canonical space along with the reconstruction loss in the posed space, effectively inpaints the occluded regions and eliminates artifacts caused by the many-to-one mapping between the canonical space and posed space. To further remove artifacts tightly attached to the human torso and assure the quality of decomposition in the input pose, we additionally apply our pose-guided SDS loss in a set of pre-defined poses including the input pose.

For the object space, we apply our pose-guided SDS loss to the canonical composite mesh $\mathcal{M}_{comp}^c$ (Eq. (15)) with the gradient of the human mesh detached. Since the OpenPose ControlNet [88] is trained to generate pose-consistent human images, we obtain better guidance for the object space through pose-guided SDS loss with the rendering of the composite mesh than the object mesh. Please refer to the supplementary material for details.

Segmentation Loss The aforementioned reconstruction loss constrains each layer in isolation. However, we observe that this alone is not sufficient to prevent penetration of the layer beneath when incomplete regions are synthesized via pose-guided SDS loss. Thus, we additionally incorporate a segmentation loss to further regularize the geometry after composition. Specifically, we assign one-hot encoded vector attributes $[1, 0]$ and $[0, 1]$, respectively to every face of $\mathcal{M}_h^p$ and $\mathcal{M}_o^p$, and rasterize both meshes together to get the
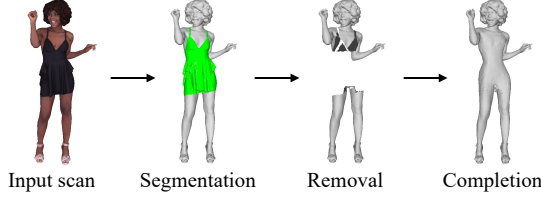
Figure 3. **Decomposition and Synthesis.** We decompose humans and objects using 3D segmentation lifted from 2D and synthesize plausible geometry of the missing regions using pose-guided SDS.

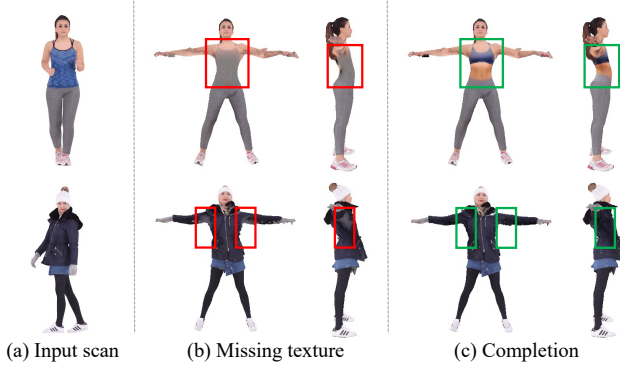(a) Input scan   (b) Missing texture   (c) Completion

Figure 4. **Texture Generation.** Applying SDS loss in canonical space generates texture for regions occluded by objects along with self-occluded regions.

segmentation masks for the human and the object, $\mathbf{S}_h^p$ and $\mathbf{S}_o^p$. We minimize the difference between $\mathbf{S}_h^p$ and $\mathbf{S}_o^p$, and the rendered segmentation masks of the input scan, $\mathbf{S}_h^{scan}$ and $\mathbf{S}_o^{scan}$, with the following loss:

$$\mathcal{L}_{comp}^{seg} = \|\mathbf{S}_h^p - \mathbf{S}_h^{scan}\|_2^2 + \|\mathbf{S}_o^p - \mathbf{S}_o^{scan}\|_2^2. \quad (16)$$

4.3. Appearance Completion

Given the inpainted canonical human mesh $\mathcal{M}_h^c$, and object mesh $\mathcal{M}_o^c$, we model the appearance of each mesh represented as vertex colors. We employ MLP networks $\mathbf{\Gamma}_h$ and $\mathbf{\Gamma}_o$ to predict the albedo of every vertex.

The total loss for optimizing the texture is defined as,

$$\mathcal{L}_{tex} = \lambda_{h_{tex}}^{rec}\mathcal{L}_{h_{tex}}^{rec} + \lambda_{o_{tex}}^{rec}\mathcal{L}_{o_{tex}}^{rec} \\ + \lambda_{h_{tex}}^{SDS}\mathcal{L}_{h_{tex}}^{SDS} + \lambda_{o_{tex}}^{SDS}\mathcal{L}_{o_{tex}}^{SDS}. \quad (17)$$

Similar to Sec. 4.2, we utilize the 3D surface segmentation to initialize the color of the visible regions in the input mesh. Specifically, we differentiably render RGB images, $\mathbf{I}_h^p$, $\mathbf{I}_o^p$, and $\mathbf{I}^{scan}$ of the posed meshes, $\mathcal{M}_h^p$ and $\mathcal{M}_o^p$, and the input scan, $\mathcal{M}^{scan}$, and optimize the following losses:

$$\mathcal{L}_{h_{tex}}^{rec} = \|\mathbf{I}_h^p \odot \mathbf{S}_h^{scan} - \mathbf{I}^{scan} \odot \mathbf{S}_h^{scan}\|_2^2, \quad (18)$$

$$\mathcal{L}_{o_{tex}}^{rec} = \|\mathbf{I}_o^p \odot \mathbf{S}_o^{scan} - \mathbf{I}^{scan} \odot \mathbf{S}_o^{scan}\|_2^2. \quad (19)$$

To generate textures for the fully occluded regions, we utilize the pose-guided SDS loss as shown in Fig. 4. We use the vertex colors of $\mathcal{M}_h^c$ and $\mathcal{M}_{comp}^c$ to render the RGB images, $\mathbf{I}_h^c$ and $\mathbf{I}_{comp}^c$, and optimize our texture MLPs, $\mathbf{\Gamma}_h$ and $\mathbf{\Gamma}_o$, by computing the gradients of following pose-guided SDS losses:

$$\nabla_{\mathbf{\Gamma}_h}\mathcal{L}_{h_{tex}}^{SDS} = \nabla_{\mathbf{\Gamma}_h}\mathcal{L}_{pose}^{SDS}(\boldsymbol{z}_t(\mathbf{I}_h^c), \boldsymbol{y}_h, \boldsymbol{p}^c, \boldsymbol{\phi}), \quad (20)$$

$$\nabla_{\mathbf{\Gamma}_o}\mathcal{L}_{o_{tex}}^{SDS} = \nabla_{\mathbf{\Gamma}_o}\mathcal{L}_{pose}^{SDS}(\boldsymbol{z}_t(\mathbf{I}_{comp}^c), \boldsymbol{y}_{comp}, \boldsymbol{p}^c, \boldsymbol{\phi}), \quad (21)$$

where $\boldsymbol{z}_t(\mathbf{I}_h^c)$ and $\boldsymbol{z}_t(\mathbf{I}_{comp}^c)$ represent the latent embeddings of $\mathbf{I}_h^c$ and $\mathbf{I}_{comp}^c$, achieved using the pretrained image encoder of the diffusion model [63]. All other notations remain consistent with those used in Eq. (14) and Eq. (15).

4.4. Composition

When composing the generated assets to novel identities, penetration of the human layer beneath could happen. To resolve this, we also introduce a refinement step. Given a canonical human mesh $\mathcal{M}_h^c$, and a canonical object mesh $\mathcal{M}_o^c$, we optimize the vertex positions of $\mathcal{M}_h^c$ along their normal directions, $\boldsymbol{n}_h$. For each vertex $\boldsymbol{v}_h \in \mathcal{V}_h^c$ of $\mathcal{M}_h^c$, we find its nearest neighbor vertex $\boldsymbol{v}_h^{nn} \in \mathcal{V}_o^{c'} \in \mathcal{V}_o^c$ of $\mathcal{M}_o^c$, where $\mathcal{V}_o^{c'}$ denotes the visible vertices among $\mathcal{V}_o^c$. We introduce a penalty when $\overrightarrow{\boldsymbol{v}_h\boldsymbol{v}_h^{nn}}$ and $\boldsymbol{n}_h$ are oriented in opposite directions. Similarly, for each vertex $\boldsymbol{v}_o \in \mathcal{V}_o^{c'}$, we find its nearest neighbor vertex $\boldsymbol{v}_o^{nn} \in \mathcal{V}_h^c$, and penalize when $\overrightarrow{\boldsymbol{v}_o\boldsymbol{v}_o^{nn}}$ and $\boldsymbol{n}_o$ have the same direction, where $\boldsymbol{n}_o$ denotes the normals of $\boldsymbol{v}_o$. Formally, we minimize the following loss,

$$\mathcal{L}_{ref} = -\frac{\overrightarrow{\boldsymbol{v}_h\boldsymbol{v}_h^{nn}}}{\|\overrightarrow{\boldsymbol{v}_h\boldsymbol{v}_h^{nn}}\|} \cdot \boldsymbol{n}_h + \frac{\overrightarrow{\boldsymbol{v}_o\boldsymbol{v}_o^{nn}}}{\|\overrightarrow{\boldsymbol{v}_o\boldsymbol{v}_o^{nn}}\|} \cdot \boldsymbol{n}_o + \lambda_{dis}\|\Delta\boldsymbol{v}_h\|_2^2, \quad (22)$$

where the last term regularizes the displacements of $\boldsymbol{v}_h$.

5. Experiments

5.1. Datasets and Metrics

RenderPeople [62]: RenderPeople provides high-quality sigle-layer 3D human scans, and we select 30 scans to cover diverse categories of target objects to decompose. We evaluate the quality of the decomposition against state-of-the-art (SOTA) methods [9, 69]. Following the evaluation protocol of previous editing work [9, 24, 27], we utilize the CLIP text-image direction similarity which measures the alignment of the performed edit with the text instruction. We also present a novel metric named, pixel-wise object removal score (POR Score), which measures the ratio of the number of pixels of the target object, before and after the edit. During evaluation, we render both the input and the edited output from 30 evenly distributed viewpoints and measure each metric.
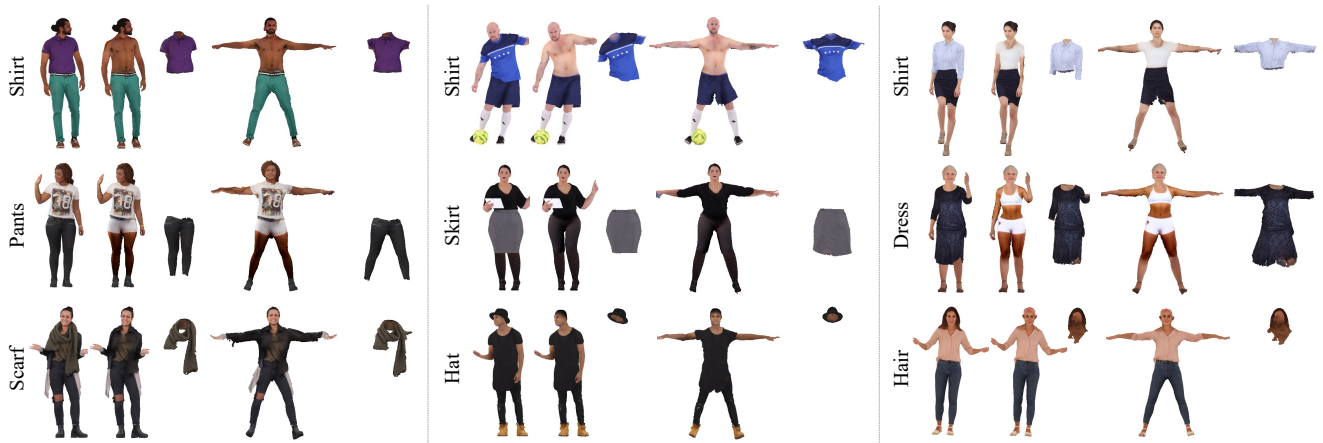
Figure 5. **Decomposition and Canonicalization.** In each set, we show the decomposition and canonicalization results of the leftmost sample.
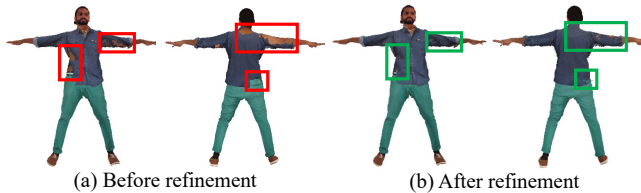

(a) Before refinement   (b) After refinement

Figure 6. **Refinement.** Our refinement stage successfully reduces the misalignment between humans and objects.
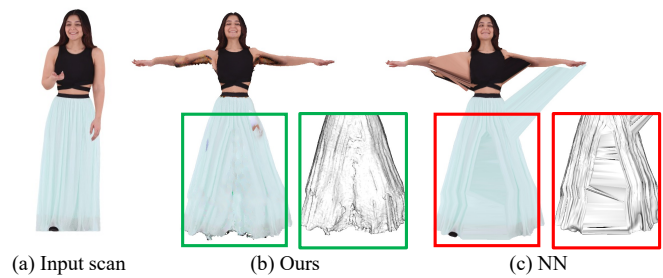

(a) Input scan   (b) Ours   (c) NN

Figure 7. **Loose Clothing.** Our method excels in modeling the canonical geometry of loose clothing such as dresses or skirts compared to existing canonicalization methods.

	CLIP TI Direction Similarity ↑	POR Score ↓
Ours	**0.1117**	**0.1144**
I-N2N [27]	0.0621	0.4871
Vox-E [69]	0.0374	0.5583

Table 1. **Quantitative Comparison on Decomposition.** We report CLIP similarity and pixel-wise object removal score to provide quantitative metrics for the subjective editing task.

CAPE Dataset [47] CAPE dataset contains the 3D sequences of clothed humans along with the corresponding SMPL parameters. We utilize CAPE dataset to evaluate the quality of canonicalization in comparison to existing methods and conduct ablation studies. For evaluation, we use 18 subjects, each wearing diverse clothing types that include both long and short upper and lower garments. For each subject, we select 100 scans with equal intervals in the sequence, and perform canonicalization using the last scan. We then pose the modeled canonical shape into poses of the preceding 99 scans, and calculate Intersection over Union (IoU) and Chamfer distance (Chamf) to measure the alignment. Since the dataset provides parameters of SMPL, we adapt our pipeline to use SMPL instead of SMPL-X.[1]

5.2. Qualitative Evaluation

Decomposition and Canonicalization. Fig. 5 shows that our method synthesizes realistic geometry and texture for the occluded regions, and enables robust canonicalization of both humans and objects, even in challenging poses.

Layered Decomposition. In Fig. 1, we highlight the key advantage of our method by applying a series of decompositions to the input scan. By recomposing the decomposed assets, our method enables the decomposition of specific layers of clothing which was previously not feasible.

Composition and Refinement. Fig. 1 shows generated avatars via combinations of the decomposed assets. The composition outputs can be further refined as shown in Fig. 6.

Loose Clothing. As shown in Fig. 7, our approach enables the successful canonicalization and modeling of loose clothing, where a simple canonicalization method based on nearest neighbor [28, 33] struggles.

5.3. Quantitative Evaluation

Decomposition. We evaluate the quality of decomposition against the SOTA text-guided 3D editing methods [27, 69], which we believe is the closest to our task. Instruct-NeRF2NeRF [27] is a text-guided NeRF [53] editing method based on the Instruct-Pix2Pix [9]. Vox-E [69] edits a 3D scene by first fitting a ReLU field [37] with multi-view images and then editing the learned ReLU field using SDS loss. We provide prompts for each method to remove the target object and compare the decomposition results. Tab. 1 shows that our method outperforms SOTA baselines, achieving the highest CLIP text-to-image similarity and the lowest POR Score. We also provide qualitative comparison in Fig. 8.

---

[1]All datasets used in this research were exclusively downloaded, accessed, and utilized at SNU.

(-) pants   (-) hat   (-) hair

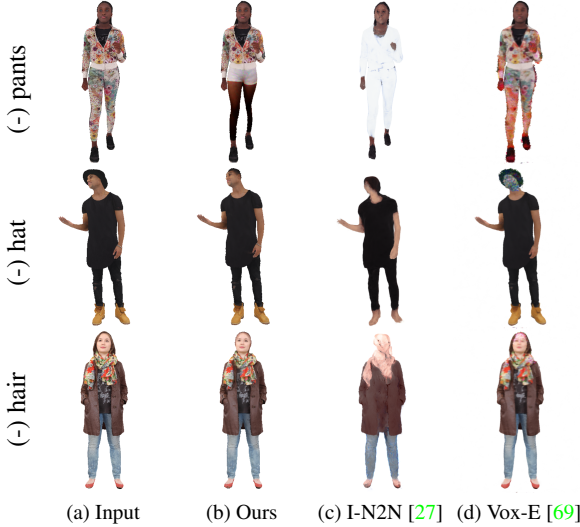(a) Input   (b) Ours   (c) I-N2N [27]   (d) Vox-E [69]

Figure 8. **Qualitative Comparison.** In contrast to our approach, other methods often face challenges in effectively removing the intended object or resulting in deterioration in unrelated areas.

Method	IoU↑	Chamf↓
Ours	84.70%	0.821
NN	83.93%	0.845
KNN	83.93%	0.846
Fast-SNARF (w/ 1 scan) [15]	27.90%	11.729
Fast-SNARF (w/ 10 scans)	41.56%	10.647
Fast-SNARF (w/ 50 scans)	86.87%	1.000
Fast-SNARF (w/ 100 scans)	91.35%	0.675
Fast-SNARF (w/ 200 scans)	94.47%	0.473

Table 2. **Quantitative comparison of canonicalization.** Chamfer distances are in centimeters. We use $K = 6$ for KNN.

Canonicalization. We compare our canonicalization results with baseline methods. To solely assess the quality of canonicalization, we exclude the decomposition process by modeling the whole scan in a single space. We employ three baselines for comparison. Nearest Neighbor (NN), transforms each vertex to its canonical position based on the skinning weights of the nearest neighbor SMPL vertex [33]. K-Nearest Neighbor (KNN) uses the weighted average of skinning weights of k-nearest neighbor SMPL vertices [84]. Tab. 2 demonstrates that our method outperforms the baselines, reporting the highest IoU and the lowest Chamfer distance when transformed into various poses. We also compare our results with Fast-SNARF [15], the current SOTA for canonicalization from multiple scans. However, we observed severe instability in the learning of MLP-based skinning fields with a small number of scans. Thus, we discard the skinning field in Fast-SNARF, and use the nearest neighbor skinning weights instead for comparison. Tab. 2 shows that our method achieves comparable performance with Fast-SNARF trained with up to 50 scans. Note that the original Fast-SNARF is trained with a significantly larger dataset

Cano. SDS Loss	Pose-Guided SDS	IoU↑	Chamf↓
✗	✗	79.97%	1.384
✓	✗	82.89%	1.227
✓	✓	**83.59%**	**1.184**

Table 3. **Ablation study.** We ablate the SDS loss in the canonical space and the pose-guided SDS loss.

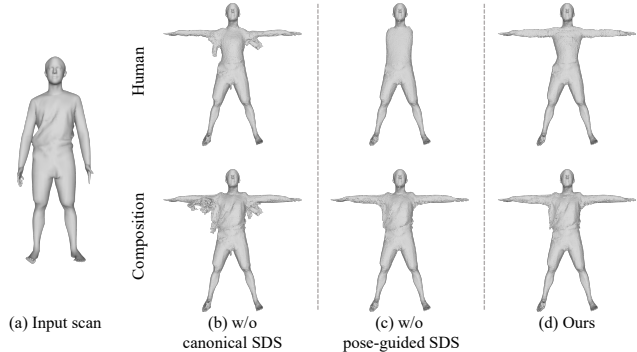(a) Input scan   (b) w/o canonical SDS   (c) w/o pose-guided SDS   (d) Ours

Figure 9. **Ablation study.** We show the effect of applying SDS loss in canonical space and the importance of the pose-guided SDS loss for robust canonicalization.

of around 3000 scans. Please refer to the supplementary material for qualitative comparison.

Ablation Study. Tab. 3 and Fig. 9 summarize an ablation study to evaluate our design choices. First, we validate the importance of the SDS loss in the canonical space. Without the SDS loss in the canonical space, we observe artifacts in the canonical shape as shown in Fig. 9 (b), leading to implausible reposing results. We further validate our pose-guided SDS by using the vanilla SDS loss without a pose condition. As illustrated in Fig. 9 (c), the use of the vanilla SDS loss leads to noticeable artifacts near the armpits and often lack large body parts. In contrast, using the proposed pose-guided SDS loss achieves more plausible canonicalization without artifacts as shown in Fig. 9 (d) and Tab. 3.

6. Conclusion

We presented GALA, a framework that turns a single static scan into reusable and animatable layered assets. Our experiments show that decomposing and inpainting separate layers in 3D is now possible with the 2D diffusion prior. Our pose-guided SDS loss allows us to jointly optimize each component in both posed and canonical space to produce clean textured 3D geometry. The resulting layered assets can be plausibly composed with novel identities and be further reposed to a target pose. Our method demonstrates significant improvements compared to existing editing or canonicalization methods, both qualitatively and quantitatively.

References

[1] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *Proc. CVPR*, 2018. 2

[2] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proc. CVPR*, 2019. 2

[3] T. Alldieck, M. Zanfir, and C. Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proc. CVPR*, 2022. 2

[4] AXYZ DESIGN. https://secure.axyz-design.com. 2

[5] H. Bertiche, M. Madadi, and S. Escalera. Cloth3d: clothed 3d humans. In *Proc. ECCV*, 2020. 2

[6] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proc. ICCV*, 2019. 2

[7] B. L. Bhatnagar, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *Proc. ECCV*, 2020. 2

[8] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Proc. ECCV*, 2016. 2

[9] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proc. CVPR*, 2023. 6, 7

[10] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE TPAMI*, 2019. 5

[11] K. Chen, C. B. Choy, M. Savva, A. X. Chang, T. Funkhouser, and S. Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Proc. ACCV*, 2018. 3

[12] R. Chen, Y. Chen, N. Jiao, and K. Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proc. ICCV*, 2023. 2, 3, 5

[13] X. Chen, A. Pang, W. Yang, P. Wang, L. Xu, and J. Yu. Tightcap: 3d human shape capture with clothing tightness field. *ACM TOG*, 2021. 2

[14] X. Chen, Y. Zheng, M. J. Black, O. Hilliges, and A. Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proc. ICCV*, 2021. 2, 4, 5

[15] X. Chen, T. Jiang, J. Song, M. Rietmann, A. Geiger, M. J. Black, and O. Hilliges. Fast-snarf: A fast deformer for articulated neural fields. *IEEE TPAMI*, 2022. 8

[16] X. Chen, T. Jiang, J. Song, J. Yang, M. J. Black, A. Geiger, and O. Hilliges. gdna: Towards generative detailed neural avatars. In *Proc. CVPR*, 2022. 2

[17] Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. In *Proc. CVPR*, 2019. 2

[18] J. Chibane, T. Alldieck, and G. Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proc. CVPR*, 2020. 3

[19] E. Corona, A. Pumarola, G. Alenya, G. Pons-Moll, and F. Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Proc. CVPR*, 2021. 2

[20] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi. Objaverse: A universe of annotated 3d objects. In *Proc. CVPR*, 2023. 3

[21] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3

[22] Z. Dong, C. Guo, J. Song, X. Chen, A. Geiger, and O. Hilliges. Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In *Proc. CVPR*, 2022. 2

[23] Y. Feng, J. Yang, M. Pollefeys, M. J. Black, and T. Bolkart. Capturing and animation of body and clothing from monocular video. In *Proc. ACM SIGGRAPH Asia*, 2022. 2

[24] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM TOG*, 2022. 6

[25] W. Gao, N. Aigerman, G. Thibault, V. Kim, and R. Hanocka. Textdeformer: Geometry manipulation using text guidance. In *ACM TOG*, 2023. 3

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3

[27] A. Haque, M. Tancik, A. Efros, A. Holynski, and A. Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proc. ICCV*, 2023. 3, 6, 7, 8

[28] T. He, Y. Xu, S. Saito, S. Soatto, and T. Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proc. ICCV*, 2021. 7

[29] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[30] F. Hong, M. Zhang, L. Pan, Z. Cai, L. Yang, and Z. Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *TOG*, 2022. 3

[31] S. Hu, F. Hong, T. Hu, L. Pan, H. Mei, W. Xiao, L. Yang, and Z. Liu. Humanliff: Layer-wise 3d human generation with diffusion model. *arXiv preprint*, 2023. 3

[32] X. Huang, R. Shao, Q. Zhang, H. Zhang, Y. Feng, Y. Liu, and Q. Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. *arXiv preprint arXiv:2310.01406*, 2023. 3

[33] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung. Arch: Animatable reconstruction of clothed humans. In *Proc. CVPR*, 2020. 7, 8

[34] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole. Zero-shot text-guided object generation with dream fields. In *Proc. CVPR*, 2022. 3

[35] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proc. CVPR*, 2018. 2

[36] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proc. CVPR*, 2018. 2

[37] A. Karnewar, T. Ritschel, O. Wang, and N. Mitra. Relu fields: The little non-linearity that could. In *Proc. ACM SIGGRAPH*, 2022. 7

[38] Y. Kasten, O. Rahamim, and G. Chechik. Point-cloud completion with pretrained text-to-image diffusion models. In *NeurIPS*, 2023. 3

[39] B. Kim, P. Kwon, K. Lee, M. Lee, S. Han, D. Kim, and H. Joo. Chupa: Carving 3d clothed humans from skinned shape priors

using 2d diffusion probabilistic models. *Proc. ICCV*, 2023. 3

[40] T. Kim, S. Saito, and H. Joo. Ncho: Unsupervised learning for neural 3d composition of humans and objects. In *Proc. ICCV*, 2023. 2

[41] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. In *Proc. ICCV*, 2023. 4

[42] J. Li, S. Saito, T. Simon, S. Lombardi, H. Li, and J. Saragih. Megane: Morphable eyeglass and avatar network. In *CVPR*, 2023. 2

[43] Y. Li, Y. Dou, Y. Shi, Y. Lei, X. Chen, Y. Zhang, P. Zhou, and B. Ni. Focaldreamer: Text-driven 3d editing via focal-fusion assembly. *arXiv preprint arXiv:2308.10608*, 2023. 3

[44] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3d: High-resolution text-to-3d content creation. In *Proc. CVPR*, 2023. 2, 3

[45] Z. Liu, Y. Feng, M. J. Black, D. Nowrouzezahrai, L. Paull, and W. Liu. Meshdiffusion: Score-based generative 3d mesh modeling. In *Proc. ICLR*, 2023. 3

[46] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015. 2

[47] Q. Ma, J. Yang, A. Ranjan, S. Pujades, G. Pons-Moll, S. Tang, and M. J. Black. Learning to dress 3d people in generative clothing. In *Proc. CVPR*, 2020. 2, 7

[48] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proc. CVPR*, 2019. 2

[49] G. Metzer, E. Richardson, O. Patashnik, R. Giryes, and D. Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proc. CVPR*, 2023. 3

[50] O. Michel, R. Bar-On, R. Liu, S. Benaim, and R. Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proc. CVPR*, 2022. 3

[51] M. Mihajlovic, Y. Zhang, M. J. Black, and S. Tang. Leap: Learning articulated occupancy of people. In *CVPR*, 2021. 2

[52] M. Mihajlovic, S. Saito, A. Bansal, M. Zollhoefer, and S. Tang. Coap: Compositional articulated occupancy of people. In *Proc. CVPR*, 2022. 2

[53] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proc. ECCV*, 2020. 2, 3, 7

[54] N. Mohammad Khalid, T. Xie, E. Belilovsky, and T. Popa. Clip-mesh: Generating textured meshes from text using pre-trained image-text models. In *Proc. ACM SIGGRAPH Asia*, 2022. 3

[55] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 3

[56] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. CVPR*, 2019. 2

[57] C. Patel, Z. Liao, and G. Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proc. CVPR*, 2020. 2

[58] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proc. CVPR*, 2019. 2, 4

[59] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM TOG*, 2017. 2

[60] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Proc. ICLR*, 2023. 2, 3

[61] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021. 3

[62] Renderpeople, 2018. `https://renderpeople.com/3d-people`. 2, 6

[63] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022. 2, 3, 6

[64] Y. Rong, T. Shiratori, and H. Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proc. ICCV*, 2021. 2

[65] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. CVPR*, 2023. 3

[66] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. ICCV*, 2019. 2

[67] S. Saito, T. Simon, J. Saragih, and H. Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proc. CVPR*, 2020. 2

[68] S. Saito, J. Yang, Q. Ma, and M. J. Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Proc. CVPR*, 2021. 2, 4

[69] E. Sella, G. Fiebelman, P. Hedman, and H. Averbuch-Elor. Vox-e: Text-guided voxel editing of 3d objects. In *Proc. ICCV*, 2023. 3, 6, 7, 8

[70] J. Seo, W. Jang, M.-S. Kwak, J. Ko, H. Kim, J. Kim, J.-H. Kim, J. Lee, and S. Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. In *Proc. ICLR*, 2024. 3

[71] T. Shen, J. Gao, K. Yin, M.-Y. Liu, and S. Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *NeurIPS*, 2021. 3, 4

[72] Y. Shi, P. Wang, J. Ye, L. Mai, K. Li, and X. Yang. Mvdream: Multi-view diffusion for 3d generation. In *Proc. ICLR*, 2024. 3

[73] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *Proc. ICLR*, 2021. 3

[74] G. Tiwari, N. Sarafianos, T. Tung, and G. Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *Proc. ICCV*, 2021. 2

[75] R. Vidaurre, I. Santesteban, E. Garces, and D. Casas. Fully convolutional graph neural networks for parametric virtual

try-on. *Comput. Graph. Forum*, 39(8):145–156, 2020. 2

[76] S. Wang, M. Mihajlovic, Q. Ma, A. Geiger, and S. Tang. Metaavatar: Learning animatable clothed human models from few depth images. In *NeurIPS*, 2021. 2

[77] T. Wang, B. Zhang, T. Zhang, S. Gu, J. Bao, T. Baltrusaitis, J. Shen, D. Chen, F. Wen, Q. Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proc. CVPR*, 2023. 3

[78] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proc. CVPR*, 2022. 5

[79] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar. Neural fields in visual computing and beyond. *Comput. Graph. Forum*, 2022. 2, 3

[80] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black. Icon: implicit clothed humans obtained from normals. In *Proc. CVPR*, 2022. 2

[81] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proc. CVPR*, 2023. 2

[82] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proc. CVPR*, 2020. 2

[83] B. Yang, C. Bao, J. Zeng, H. Bao, Y. Zhang, Z. Cui, and G. Zhang. Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *Proc. ECCV*, 2022. 3

[84] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Analyzing clothing layer deformation statistics of 3d human motions. In *Proc. ECCV*, 2018. 8

[85] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proc. CVPR*, 2021. 2

[86] X. Zeng, A. Vahdat, F. Williams, Z. Gojcic, O. Litany, S. Fidler, and K. Kreis. Lion: Latent point diffusion models for 3d shape generation. In *NeurIPS*, 2022. 3

[87] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, and Z. Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proc. ICCV*, 2021. 2

[88] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. ICCV*, 2023. 2, 5

[89] Z. Zheng, T. Yu, Y. Liu, and Q. Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE TPAMI*, 2021. 2

[90] J. Zhuang, C. Wang, L. Liu, L. Lin, and G. Li. Dreameditor: Text-driven 3d scene editing with neural fields. *Proc. ACM SIGGRAPH Asia*, 2023. 3