NCHO: Unsupervised Learning for Neural 3D Composition of Humans and Objects

Taeksoo Kim[1] Shunsuke Saito[2] Hanbyul Joo[1]
[1]Seoul National University [2]Meta Reality Labs
taeksu98@snu.ac.kr shunsuke.saito16@gmail.com hbjoo@snu.ac.kr

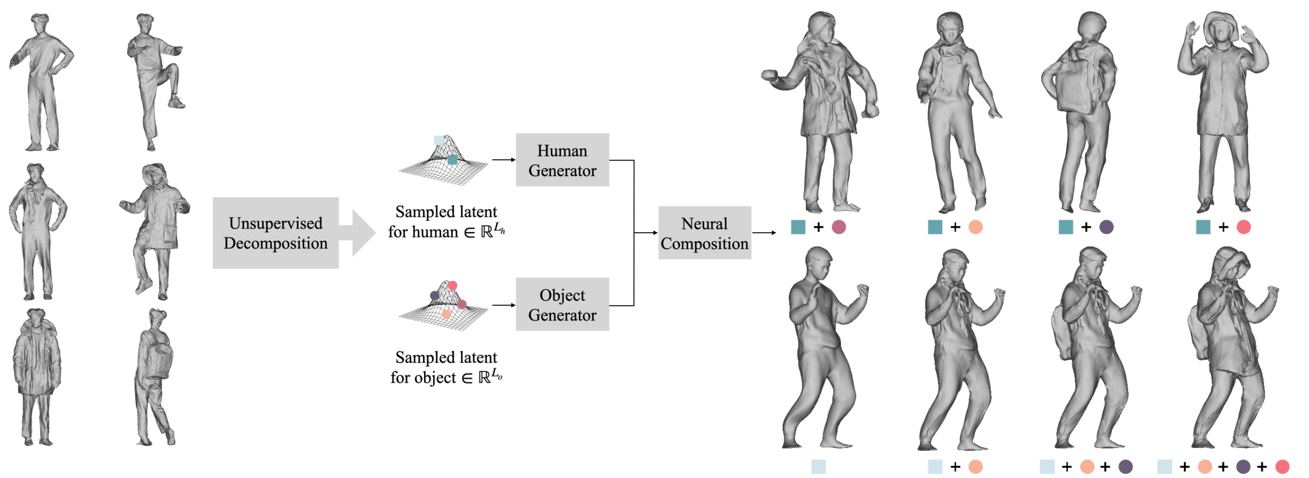

Figure 1. From raw 3D scans of the "source human" in casual clothing with additional outwear or objects (left), our method automatically decomposes objects from the source human and builds a compositional generative model that enables 3D avatar creations of novel human identities with variety of outwear and objects (right) in an unsupervised manner.

Abstract

Deep generative models have been recently extended to synthesizing 3D digital humans. However, previous approaches treat clothed humans as a single chunk of geometry without considering the compositionality of clothing and accessories. As a result, individual items cannot be naturally composed into novel identities, leading to limited expressiveness and controllability of generative 3D avatars. While several methods attempt to address this by leveraging synthetic data, the interaction between humans and objects is not authentic due to the domain gap, and manual asset creation is difficult to scale for a wide variety of objects. In this work, we present a novel framework for learning a compositional generative model of humans and objects (backpacks, coats, scarves, and more) from real-world 3D scans. Our compositional model is interaction-aware, meaning the spatial relationship between humans and objects, and the mutual shape change by physical contact is fully incorporated. The key challenge is that, since humans and objects are in contact, their 3D scans are merged into a single piece. To decompose them without manual annotations, we propose to leverage two sets of 3D scans of a single person with and without objects. Our approach learns to decompose objects and naturally compose them back into a generative human model in an unsupervised manner. Despite our simple setup requiring only the capture of a single subject with objects, our experiments demonstrate the strong generalization of our model by enabling the natural composition of objects to diverse identities in various poses and the composition of multiple objects, which is unseen in training data. The project page is available at https://taeksuu.github.io/ncho.

1. Introduction

Generative modeling of 3D humans from real-world data has shown promise to represent and synthesize diverse human shapes, poses, and motions. Especially, the ability to create realistic humans with diverse clothing and accessories (e.g., backpacks, scarves, and hats) is indispensable for a myriad of applications including AR/VR, entertainment, and virtual try-on. The early work [4, 28, 36, 50, 71] has demonstrated success in modeling "undressed" human bodies from real-world scans. More recently, the research community has been increasingly focused on the generative modeling of clothed humans [13, 16, 38], to better represent humans in everyday life.

Recent advancements in shape representations such as Neural Fields [69] mitigate the need for pre-defining topology or template of clothing, enabling to build animatable clothed humans from raw 3D scans [14, 56]. Along with its advantage in strong expressive power for avatar modeling, this approach also allows the models to learn faithful physical interactions between objects and humans. However, since raw 3D scans do not provide a clear separation of different components, existing approaches typically treat human bodies, clothing, and accessories as an entangled block of geometry [13]. In this paper, we argue that this leads to suboptimal expressiveness and composability of the generative avatars. Many applications require more intuitive control to add, replace, or modify objects while maintaining human identity. To make avatars explicitly compositable with objects, several works propose to leverage synthetic data [6, 16, 27]. However, the manual creation of 3D assets remains a challenge and is extremely difficult to scale. Moreover, the physical interaction of bodies, clothing, and accessories in synthetic data tends to be less faithful due to the domain gap.

In contrast to prior methods, our goal is to build a compositional generative model of objects and humans from real-world observations. The core challenge lies in the difficulty of learning the composition and decomposition of objects in contact from raw 3D scans. Capturing objects in isolation does not lead to faithful composition due to the lack of realistic deformations induced by physical contact. Thus, while it is essential to collect 3D scan data on objects and humans in contact, the joint scanning of humans with objects only provides an entangled block of 3D geometry as mentioned, and accurately segmenting different components requires non-trivial 3D annotation efforts.

Upon these challenges, our contributions are: scalable data capture protocol, unsupervised decomposition of objects and humans, and generalizable neural object composition.

Scalable Data Capture. Capturing multiple identities with various poses and objects requires prohibitively large time and storage. To overcome this issue, we propose to collect human-object interactions with diverse poses only from a single subject, referred to as the "source human". To enable decomposition of objects, we also capture the same person without any objects, where the deviation between two sets defines "objects" in our setup. Examples are shown in Fig. 2. This capture protocol offers sufficient diversity in poses and object types within a reasonable capture time.

Unsupervised Decomposition of Objects. To separate objects from the source human, we leverage the expressiveness of the generative human model based on implicit surface representation [13]. We train a human module without objects, and then jointly optimize the latent codes of the avatar and a generative model for objects to best explain the 3D scans of the person with objects. While the human module accounts for state differences in pose and clothing, the object-only module learns to synthesize the residual geometry as an object layer in an unsupervised manner. Notably, objects in our work are defined as residual geometry that cannot be explained by the trained human-only module.

Neural Object Composition. While the unsupervised decomposition successfully separates objects from the source human, we observe that naively composing it to novel identities from other datasets [52, 75] leads to undesired artifacts and misalignment in the contact regions. To address this, we propose a neural composition method by introducing another composition MLP that takes latent features from both human and object modules to make a final shape prediction. Due to the local nature of MLPs, our approach plausibly composes objects to novel identities without retraining as in Fig. 1.

Our experiments show that our compositional generative model is superior to existing approaches without explicit disentanglement of objects and humans [13]. In addition, we show that our model can be used for fine-grained controls including object removal from 3D scans and multiple object compositions on a human, demonstrating the utility and expressiveness of our approach beyond our training data.

2. Related Work

3D Human Models. Representing plausible 3D human bodies while handling diverse variations in shapes and poses is a long-standing problem. Due to the challenge in modeling diverse shape variation, the early work [4,28,36,50,71] mainly focuses on the undressed 3D human body by learning mesh-based statistical models deformed from a template mesh. To model dressed 3D humans, follow-up work [1, 2, 37] adds 3D offsets on top of the parametric undressed human body models to represent clothing. Yet, the topological constraints and the resolution of the template model restrict these methods from modeling arbitrary shapes of clothing with high-frequency details. Recently emerging deep implicit shape representation [15, 39, 42, 49] provides a breakthrough in expressing 3D humans by leveraging neural networks for representing continuous 3D shape space, where its efficacy is demonstrated in reconstructing clothed humans with high-fidelity from images [54, 55, 70]. There also has been an actively growing field to represent animatable 3D human avatars using 3D scans, depth maps, or pointclouds [13, 14, 17, 19, 25, 40, 41, 56, 62, 65]. However, prior 3D human models have paid little attention to the joint modeling of humans and objects in close contact.

2D/3D Generative Models Generative models intend to express the plausible variations over the latent space, which can be used to create diverse realistic samples. There have been extensive studies in 2D generative modeling to create realistic photos [29–31] via generative adversarial networks (GANs) [21, 22], variational autoencoders (VAEs) [33], and more recently, diffusion models [18, 24, 53, 60]. Generative 3D modeling has also been actively explored. By leveraging the availability of a large-scale 3D object scans [12], many approaches present generative models for 3D objects [11, 15, 39, 44, 45, 49, 57]. Relatively few approaches have been presented for generative 3D human modeling, due to the lack of available diverse 3D datasets for people [3, 13, 16, 37, 71]. We show that our scalable data capture protocol and compositional generative model enable the synthesis of 3D humans with diverse objects in novel poses.

Compositional Models. Compositional generative models via neural networks have been explored to represent different components as independent models, representing a whole scene by compositing them together. These approaches pursue controlling or sampling one component without affecting the rest. The early approaches focus on building such models in 2D for creating realistic 2D images via generative models [5, 35, 63, 77]. More recent approaches explore the compositional reasoning for 3D [9, 34, 45, 47, 66, 67, 73, 74]. Most approaches in this direction aim at synthesizing realistic novel views by compositing NeRFs [42] for 3D objects and scenes [45, 67, 73] and for human faces [9, 45, 72]. However, these approaches do not consider mutual shape deformations between objects. Human bodies are also treated as a composition of multiple body parts. These approaches attain final composition output by either max-pooling the outputs of individual components [17, 40] or by using another neural network [3, 7, 46, 61]. While a recent work shows interaction-aware 3D composition reasoning is possible for faces and eyeglasses with extensive annotations and data preprocessing [34], our approach supports diverse object categories without requiring any manual annotations.

Garment Modeling. Due to the deformable nature of garments, capturing and modeling 3D clothing is challenging. Only a few 3D garment datasets have been presented [6, 76], where laborious segmentation and post-processing are required to separate the garments from dummies or human bodies. While most methods reconstruct a clothed 3D human as a single chunk of geometry [54, 55, 70], there exist methods reconstructing the 3D clothing as a separate layer on top of parametric mesh model (e.g., SMPL) using segmentation [20] or synthetic 3D assets [16, 27]. Virtual try-on has also been actively explored in graphics via physics simulation [68] and or synthetic data [64]. In contrast, our approach learns a generative clothing and accessory model from real-world observations in an unsupervised fashion.

Figure 2. **Examples of Our Datasets.** Top row: sample scans of $\mathcal{S}_{sh}$. Bottom row: sample scans of $\mathcal{S}_{sh+o}$.

3. Preliminaries

Data Acquisition. To model humans and objects in contact, we capture two sets of datasets, $\mathbf{S}_{sh}$ and $\mathbf{S}_{sh+o}$. $\mathbf{S}_{sh}$ consists of 3D scans of a single identity, denoted as "source human" with various poses. $\mathbf{S}_{sh+o}$ consists of 3D scans of the source human with a variety of objects or additional outwear as shown in Fig. 2. In this work, we choose coats, vests, backpacks, scarves, and hats to demonstrate the generality of our approach for outwear and everyday accessories. To support the generative modeling of objects, we capture multiple objects in each category. In addition to $\mathbf{S}_{sh}$ and $\mathbf{S}_{sh+o}$, we also use another 3D human dataset [75] to train a target generative human model for composition, denoted $\mathbf{S}_{th}$.

We collect 3D scans, $\mathbf{S}_{sh}$ and $\mathbf{S}_{sh+o}$, with a capture system of 8 synchronized and calibrated Azure Kinects (see supp. mat. for details). We apply KinectFusion [43] to fuse the depth maps, and then reconstruct watertight meshes with screened-poisson surface reconstruction [32]. We also detect 2D keypoints using OpenPose [10] and apply the multi-view extension of SMPLify [8] to obtain SMPL parameters [36] for each scan.

Generative Articulated Models. We adopt the generative human model [13] which extends forward skinning with root finding [14] for cross-identity modeling. We briefly discuss the framework and highlight our key modifications. The key idea in gDNA [13] is to represent occupancy fields conditioned by identity-specific latent codes $\mathbf{z}$ in a canonical space, and transform them into a posed space using forward linear blend skinning (LBS). The occupancy fields of a person in the canonical space can be represented as follows:

$$o = O(\mathbf{x}^c, G(\mathbf{z})), \tag{1}$$

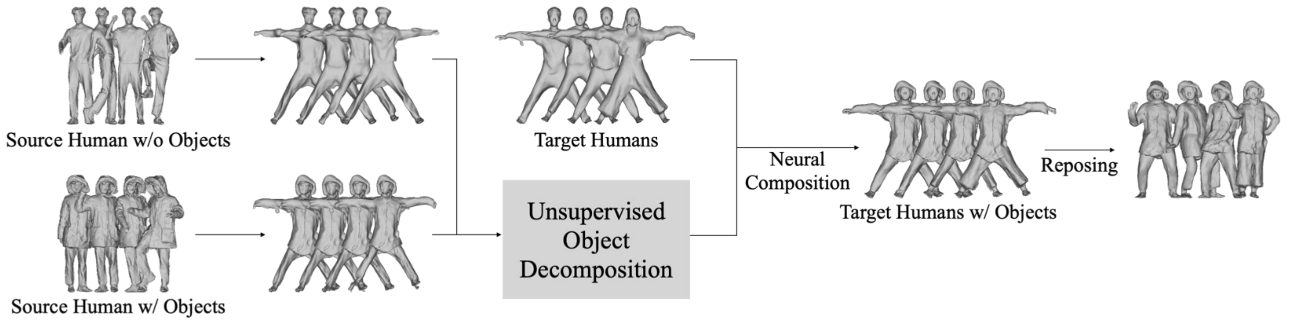

Figure 3. **Overview.** From captured scans of the source human with and without objects, our method successfully decomposes objects from humans without any supervision, allowing a generative model to learn the shapes of various objects. These objects are then added to novel identities via neural composition, resulting in the creation of diverse human avatars with controllable objects.

where $G(\cdot)$ is a spatially varying feature generator taking the latent code. While the original work [13] uses 3D feature voxels for the output of G, we use a tri-plane feature representation [11], which achieves better performance with higher memory efficiency. The generated feature map is conditioned on the latent code $\mathbf{z}$ via adaptive instance normalization [26].

To query the occupancy of a point $\mathbf{x}^d$ in a posed space, we transform the canonical coordinate $\mathbf{x}^c$ as follows:

$$\mathbf{x}^d = \sum_{i=1}^{n_b} W_i(N(\mathbf{x}^c, \beta), \mathbf{z}) \cdot \mathbf{B}_i(\beta, \theta) \cdot \mathbf{x}^c, \tag{2}$$

where n_b is the number of bones, W_i is the skinning weight of bone i, predicted from the identity conditioned skinning network W, and N is the warping network for neutralizing the body size variation given SMPL shape parameters $\beta \in \mathbb{R}^{10}$. $\mathbf{B}_i(\beta, \theta)$ is the transformation of bone i in SMPL model given β and pose $\theta \in \mathbb{R}^{24\times3}$. To jointly learn the occupancy and deformation networks, we solve for $\mathbf{x}^c$ in Eq. 2 given $\mathbf{x}^d$ using iterative root finding [14]. We also discard the surface normal prediction networks used in [13]. Instead of hallucinating details via adversarial learning, we model detailed geometry by jointly representing shapes as SDF together with the occupancy fields as in [58]. As we can directly supervise SDF on surface normals [23], we model detailed geometry as true surface. However, we find that directly replacing the occupancy with SDF leads to unstable training. To mitigate instability, we propose a hybrid modeling of occupancy and SDF, and disable the backpropagation of gradients from SDF to the deformation networks so that it is only supervised by the occupancy head. See supp. mat. for details.

4. Method

Fig. 3 shows an overview of our pipeline. Our goal is to decompose and compose generative objects on target humans from raw 3D scans. To this end, we introduce a generative human module and an object module.

The Human Module $\mathcal{M}_h = (G_h, O_h, D_h)$ represents the human part and it is composed of a feature generator G_h, a decoder O_h, and deformation networks $D_h = (W_h, N_h)$. As an output, $\mathcal{M}_h$ produces an occupancy value o_h, SDF d_h, and a feature vector $\mathbf{f}_h$, which is the intermediate latent feature before the last layer, in the canonical space:

$$(o_h, d_h, \mathbf{f}_h) = O_h(\mathbf{x}^c, G_h(\mathbf{z}_h)), \tag{3}$$

where $\mathbf{z}_h$ is a learnable latent code for the human part. Note that the hybrid modeling of occupancy and SDF is applied only to the human module as our unsupervised decomposition losses require occupancy. We leave extending the hybrid model to the remaining modules for future work.

The Object Module $\mathcal{M}_o = (G_o, O_o)$ is responsible for modeling the geometry of the object part. Because the object module and the human module share the same canonical space, the object module does not require separate deformation networks. $\mathcal{M}_o$ returns an occupancy value o_o, and a feature vector $\mathbf{f}_o$, which is the intermediate latent feature before the last layer, in the canonical space:

$$(o_o, \mathbf{f}_o) = O_o(\mathbf{x}^c, G_o(\mathbf{z}_o)), \tag{4}$$

where $\mathbf{z}_o$ is a learnable latent code for the object part.

4.1. Neural Object Composition

While compositing the occupancy of the human and object modules in a closed-form [17,40] is possible, we observe that this leads to misalignment in the contact regions and floating artifacts. To address these issues, we introduce a neural composition module parameterized by MLPs.

The composition module $\mathcal{M}_{comp} = (O_{comp}, D_{comp})$ is used to integrate humans and objects in the canonical space. We directly feed the feature vectors $\mathbf{f}_h$ and $\mathbf{f}_o$ from the human module and object module respectively as inputs. $\mathcal{M}_{comp}$ outputs the final occupancy value o_{comp}, after composition in the canonical space:

$$o_{comp} = O_{comp}(\mathbf{x}^c, \mathbf{f}_h, \mathbf{f}_o) \tag{5}$$

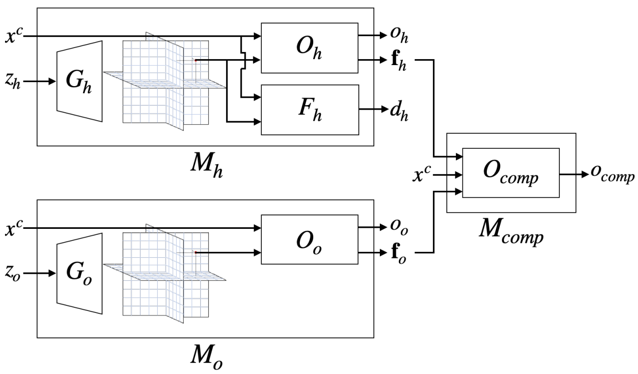

Figure 4. **Model.** Given latent code $\mathbf{z}_h$, $\mathcal{M}_h$ predicts the occupancy fields and SDFs for humans in canonical space. Similarly, with latent code $\mathbf{z}_o$, $\mathcal{M}_o$ predicts the occupancy fields for objects. The features $\mathbf{f}_h$ and $\mathbf{f}_o$ from each network are passed to $\mathcal{M}_{comp}$ to predict the occupancy fields for final compositional outputs of humans and objects in the same canonical space.

Similar to the human module, the deformation networks $D_{comp} = (W_{comp}, N_{comp})$ provide the mapping from the canonical space to the posed space. The entire model is illustrated in Fig. 4.

4.2. Unsupervised Object Decomposition

To achieve the decomposition of object layers from raw 3D scans in an unsupervised manner, our key idea is to represent objects as the residual of human geometry. To this end, we first train the human module $\mathcal{M}_h$ using $\mathbf{S}_{sh}$ with the learnable shape code $\mathbf{z}_{sh}$ for each scan. This allows the human module to account for slight shape variations of the source human. In the next step, using $\mathbf{S}_{sh+o}$, we jointly optimize the per-scan latent code $\mathbf{z}_{sh}$ together with the object module $\mathcal{M}_o$ while freezing $\mathcal{M}_h$. Given the composed occupancy o_{comp} in Eq. 5 and the predicted occupancy of the human module o_h, the target occupancy of the object module can be computed as

$$(1 - o_h) \cdot o_{comp}. \tag{6}$$

This formulation encourages the object module to exclusively model the residual geometry, which the human module is not able to represent, enabling disentangled control over humans and objects. Moreover, it simultaneously encapsulates the spatial information between humans and objects, allowing the model to be interaction-aware.

We jointly optimize the neural composition module and the object module $\mathcal{M}_o$ in an end-to-end manner using the loss functions discussed in Sec. 4.3.

4.3. Training

Our system is trained using $\mathbf{S}_{th}$, $\mathbf{S}_{sh}$ and $\mathbf{S}_{sh+o}$ with their SMPL shape and pose parameters. Following the auto-decoding framework of [49], we jointly optimize the latent code $\mathbf{z}$ assigned for each scan along with the network weights during training. Every scan in each dataset is assigned its own latent code, denoted $\mathbf{z}_{th} \in \mathbb{R}^{L_{th}}$ for scans in $\mathbf{S}_{th}$, $\mathbf{z}_{sh} \in \mathbb{R}^{L_{sh}}$ for scans in $\mathbf{S}_{sh}$ and $\mathbf{z}_o \in \mathbb{R}^{L_o}$ for scans in $\mathbf{S}_{sh+o}$. Note that all latent codes except $\mathbf{z}_o$ are initialized with zero. For $\mathbf{z}_o$, we use one-hot encoding for each object category using the first 5 bits to enable random sampling from a specific category and initialize the remaining bits with zero. We denote the human module for generating target human shapes, $\mathcal{M}_{th}$, and the human module for modeling shapes of the source human, $\mathcal{M}_{sh}$.

Training consists of three stages: We first train $\mathcal{M}_{th}$ and $\mathbf{z}_{th}$ with $\mathbf{S}_{th}$ which consists 3D scans of general humans. This is to leverage the wide variation of shapes and poses of samples in $\mathbf{S}_{th}$ for training the multi-subject forward skinning module, D_{th}. For later stages, D_{th} is used to initialize other deformation networks with its warping network N_{th} frozen, to let all samples share the same canonical space. Next, we train $\mathcal{M}_{sh}$ and $\mathbf{z}_{sh}$ with $\mathbf{S}_{sh}$ which consists 3D scans of the source human without any object. For the last stage, using all the samples, we train $\mathcal{M}_o$, $\mathcal{M}_{comp}$, $\mathbf{z}_{sh}$, and $\mathbf{z}_o$ with the pre-trained $\mathcal{M}_{th}$, $\mathcal{M}_{sh}$ and $\mathbf{z}_{th}$ frozen. Note that $\mathbf{z}_{sh}$ for the last stage are re-initialized as the mean of $\mathbf{z}_{sh}$ after the second stage, denoted $\bar{\mathbf{z}}_{sh}$.

$\mathcal{M}_{comp}$ models all training samples using the feature vector from either $\mathcal{M}_{th}$ or $\mathcal{M}_{sh}$ for the human part, and from $\mathcal{M}_o$ for the object part. In the case of $\mathbf{S}_{th}$ and $\mathbf{S}_{sh}$ where scans are with no objects, we introduce a new latent code $\mathbf{z}_{emp}$ as an input to $\mathcal{M}_o$ for no objects. See supp. mat. for the detailed training procedure.

Losses: For the first stage, we use losses following [13]. We use the binary cross entropy loss $\mathcal{L}_{th}$ between the predicted occupancy of $\mathcal{M}_{th}$ and the ground truth occupancy. Note that $O^d(\cdot)$ and $F^d(\cdot)$ denote the occupancy field and SDF in posed space, respectively. We also use guidance losses $\mathcal{L}_{bone}$, $\mathcal{L}_{joint}$ and $\mathcal{L}_{warp}$ to aid training. $\mathcal{L}_{bone}$ encourages the occupancy of $\mathbf{x}_{bone}$ to be one, where $\mathbf{x}_{bone}$ are randomly selected points along the SMPL bones in canonical space. $\mathcal{L}_{joint}$ encourages the skinning weights of SMPL joints to be 0.5 for connected two bones and 0 for all other bones. $\mathcal{L}_{warp}$ encourages deformation network N to change body size consistently, by enforcing vertices of a fitted SMPL to warp to vertices of the mean SMPL shape, achieved by having shape parameter β as zero. Lastly, we use $\mathcal{L}_{reg\_th}$ to regularize the latent code $\mathbf{z}_{th}$ to be close to zero.

$$\mathcal{L}_{th} = BCE((O^d_{th}(\mathbf{x}^c, G_{th}(\mathbf{z}_{th})), o_{gt}) \tag{7}$$

$$\mathcal{L}_{bone} = BCE((O_{th}(\mathbf{x}_{bone}, G_{th}(\mathbf{z}_{th})), 1) \tag{8}$$

$$\mathcal{L}_{joint} = \|W(\mathbf{x}_{joint}, \mathbf{z}_{th}) - \mathbf{w}_{gt}\| \tag{9}$$

$$\mathcal{L}_{warp} = \|N(\mathbf{v}(\beta), \beta) - \mathbf{v}(\beta_0)\| \tag{10}$$

$$\mathcal{L}_{reg\_th} = \|\mathbf{z}_{th}\| \tag{11}$$

For training the SDF network, we use L1 loss $\mathcal{L}_{sdf}$ between the predicted and the ground truth signed distance and L2 loss $\mathcal{L}_{nml}$ between the gradients of SDF and the ground truth normals of points on the surface. We additionally use $\mathcal{L}_{igr}$ for SDF to satisfy the Eikonal equation [23] and $\mathcal{L}_{bbox}$ to prevent SDF values of off-surface points from being the zero-level surface as in [59].

$$\mathcal{L}_{sdf} = |F^d_{th}(\mathbf{x^c}, G_{th}(\mathbf{z}_{th})) - d_{gt}| \tag{12}$$

$$\mathcal{L}_{nml} = \|\nabla F^d_{th}(\mathbf{x^c}, G_{th}(\mathbf{z}_{th})) - n_{gt}\| \tag{13}$$

$$\mathcal{L}_{igr} = (\|\nabla F_{th}(\mathbf{x^c}, G_{th}(\mathbf{z}_{th}))\| - 1)^2 \tag{14}$$

$$\mathcal{L}_{bbox} = exp(-\alpha \cdot |F_{th}(\mathbf{x^c}, G_{th}(\mathbf{z}_{th}))|), \alpha \gg 1 \tag{15}$$

For the second stage, we use the binary cross entropy loss $\mathcal{L}_{sh}$ between the predicted occupancy of $\mathcal{M}_{sh}$ and the ground truth occupancy, and $\mathcal{L}_{reg\_sh}$ to regularize the latent code $\mathbf{z}_{sh}$ to be close to zero. Since we initialize D_{sh} with pre-trained D_{th}, additional guidance losses are not required.

$$\mathcal{L}_{sh} = BCE((O^d_{sh}(\mathbf{x}^c, G_{sh}(\mathbf{z}_{sh})), o_{gt}) \tag{16}$$

$$\mathcal{L}_{reg\_sh} = \|\mathbf{z}_{sh}\| \tag{17}$$

For the last stage, we use the binary cross entropy loss $\mathcal{L}_{comp}$ between the predicted occupancy of $\mathcal{M}_{comp}$ and the ground truth occupancy. We also use $\mathcal{L}_o$ between the predicted occupancy of $\mathcal{M}_o$ and the residual part of $\mathbf{S}_{sh+o}$ where $\mathcal{M}_h$ cannot explain. Moreover, we optimize $\mathbf{z}_{sh}$ by using the binary cross entropy loss $\mathcal{L}_{fit}$ between the output of $\mathcal{M}_{sh}$ and the ground truth occupancy. Finally, we regularize $\mathbf{z}_{sh}$ to be close to $\overline{\mathbf{z}}_{sh}$ and $\mathbf{z}_o$ to be close to zero.

$$\mathcal{L}_{comp} = BCE((O^d_{comp}(\mathbf{x}^c, \mathbf{f}_h, \mathbf{f}_o), o_{gt}) \tag{18}$$

$$\mathcal{L}_o = BCE((O_o(\mathbf{x}^c, G_o(\mathbf{z}_o)), (1 - o_h) \cdot o_{comp}) \tag{19}$$

$$\mathcal{L}_{fit} = BCE((O^d_{sh}(\mathbf{x}^c, G_{sh}(\mathbf{z}_{sh})), o_{gt}) \tag{20}$$

$$\mathcal{L}_{reg\_sh} = \|\mathbf{z}_{sh} - \overline{\mathbf{z}}_{sh}\| \tag{21}$$

$$\mathcal{L}_{reg\_o} = \|\mathbf{z}_o\| \tag{22}$$

5. Experiments

We evaluate our generative composition model across various scenarios. We first demonstrate the quality of the random 3D avatar creations from our model and the disentangled natures of human and object controls. Quantitative and qualitative comparisons against the previous SOTA [13] are performed, incorporating a user study via CloudResearch Connect. We also conduct ablation studies to validate our design choices.

5.1. Dataset

Our 3D Scans: As described in Sec. 3, we use our multi-Kinect system to capture the source human with and without objects, $\mathbf{S}_{sh}$ (180 samples) and $\mathbf{S}_{sh+o}$ (342 samples). For $\mathbf{S}_{sh+o}$, we consider 4 categories of objects: 5 backpacks (77 samples in total), 6 outwear (94 samples), 8 scarves (89 samples), and 6 hats (82 samples). We run quantitative evaluation by focusing on backpacks as other objects such as outwear are already incorporated in $\mathbf{S}_{th}$. We use another set with 300 samples of the source human with backpacks only, denoted as $\mathbf{S}_{sh+bp}$. To build a testing set for FID computation in this quantitative evaluation, we further capture 343 samples of 3 different unseen identities who wear unseen backpacks. We denote this test dataset, $\mathbf{S}_{unseen+bp}$.

THuman2.0 [75]: THuman2.0[1] provides high-quality 3D dataset for dressed humans. We use 526 samples for $\mathbf{S}_{th}$.

Figure 5. **Random Generation.** Top row: randomly sampled outputs of the human module before composition. Bottom row: composition outputs of target humans on top with specific objects.

5.2. Qualitative Evaluation

We demonstrate the expressive power and controllability of our composition model via inferences in various scenarios by controlling latent codes for humans $\mathbf{z}_h$ and object $\mathbf{z}_o$.

Random Generation. The 3D avatars created by attaching specific object latent codes $\mathbf{z}_o$ to random sampled human codes $\mathbf{z}_h$ are shown in Fig. 5 (bottom). The outputs of the human module $\mathcal{M}_h$ are also shown on the top of Fig. 5 for reference. Our model enables the creation of diverse 3D avatars with controllable objects.

Disentangled Controls over Human and Objects. To further test the disentangled nature of our composition model, we create 3D humans with objects by changing either human latent code or object latent code, as shown in Fig. 6. The examples on the top vary the human part by keeping the same object code that represents a scarf. On the bottom examples, we vary object codes for a fixed identity shown on the leftmost side. These results show the core advantage of our composition model in individual controls.

[1]The THuman2.0 dataset was downloaded, accessed, and used in this research exclusively at SNU.

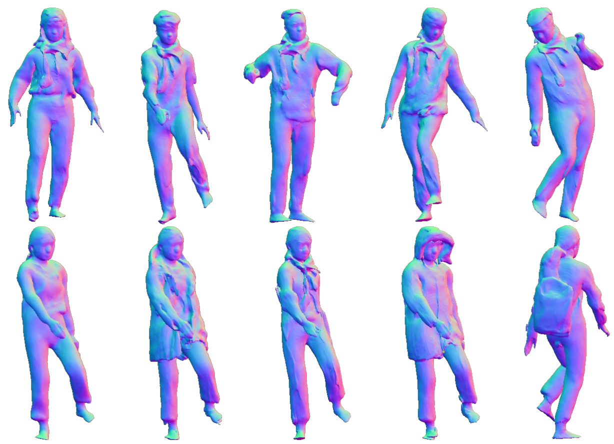

Figure 6. **Disentangled Human and Object.** Top row: composition outputs of the same object (a scarf), added to different human identities. Bottom row: composition outputs of different objects added to the single human identity shown in the leftmost column.

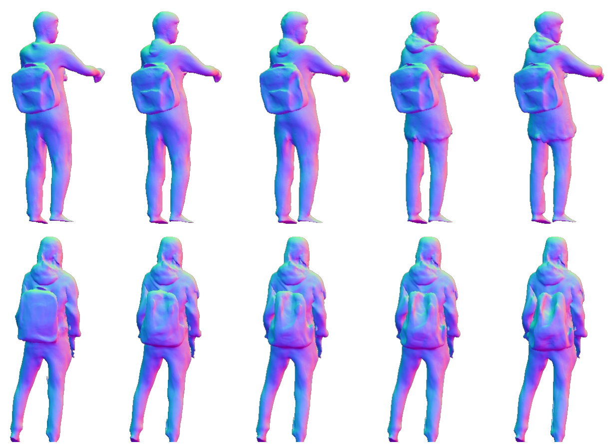

Figure 7. **Interpolation.** Top row: human module interpolation. Bottom row: object module interpolation. Notice that interpolating one module doesn't deteriorate the geometry of the other.

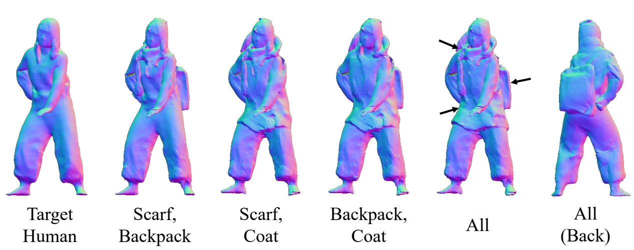

Figure 8. **Composition of Multiple Objects.** Two or more objects are added to the leftmost human. Note that our train data contain no scans of the source human with multiple objects.

Interpolation. Fig. 7 demonstrates smooth interpolation of each module without deteriorating the other module.

Composition of Multiple Objects. Fig. 8 shows that our system allows the composition of multiple objects. To add multiple objects, we use the latent code of each object and get the occupancy and the feature vector of objects. Using the normalized occupancy of multiple objects as weights, we calculate the weighted sum of feature vectors. The aggregated feature is then fed to the composition module along with the human feature to get the final composition output. Note that our dataset has no such sample with multiple objects.

Method	FID	User Preference
gDNA (w/ object)	41.71	43.6%
Arith. gDNA (w/ object)	73.81	13.6%
Ours (Naive composition)	55.29	22.4%
Ours	51.03	100% - (above)

Table 1. Quantitative evaluation of the importance of compositional modeling. User preference score reflects the frequency with which participants of our perceptual study favored each method over ours.

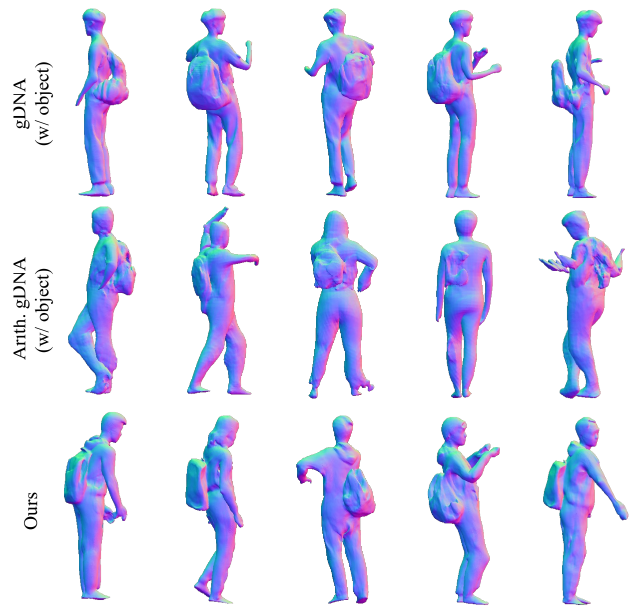

Figure 9. **Qualitative Comparison on Compositional Modeling.** Compared to our method, baselines suffer from generating outputs of diverse humans with complete objects.

5.3. Comparison with SOTA

Since our method is the first generative model for compositing humans and objects, there is no direct competitor, and comparison with the previous non-compositional model such as gDNA is non-trivial. To make the assessment possible at our best, we consider a specific scenario where a user wants to create samples with a specific object category, being the backpack here. To provide such controllability on gDNA, we first extend the gDNA model with our dataset. Note that, in this evaluation, we use the same dataset $\mathbf{S}_{th}$, $\mathbf{S}_{sh}$, $\mathbf{S}_{sh+bp}$ for training both our model and gDNA.

Extending gDNA for Composition. We train gDNA model using the public code with our datasets. Both human-only outputs and the ones with a backpack can be sampled from the trained model. To intentionally generate outputs with a backpack, we search the latent codes associated with the training samples with backpacks and fit a gaussian from which we can perform a sampling. We denote this baseline method as *'gDNA (w/ object)'*.

The second possible extension of gDNA is based on the arithmetic operation among gDNA's latent codes, which is widely used for GAN-based image manipulation [51]. We found that gDNA's original framework allows some level of composition by adding or subtracting the latent codes. Specifically, we choose a latent code $\mathbf{z}^*_{sh}$ for the source human without a backpack and another latent code $\mathbf{z}^*_{sh+bp}$ for the source human with the backpack. We simply take their subtraction $\mathbf{z}_{bp} = \mathbf{z}^*_{sh+bp} - \mathbf{z}^*_{sh}$, which can be considered as a residual for the backpack. We found that composition can be performed by adding this residual to another human's latent code, that is $\mathbf{z}_{bp} + \mathbf{z}_{th}$. We denote this baseline method as *'Arith. gDNA (w/ object)'*.

Qualitative Comparison with User Study. The visual comparison between ours and the extended gDNAs is shown in Fig. 9. In the first row, we show random samples generated from *'gDNA (w/ object)'*. Since the human scans with the backpack are only of the source human's (other samples from $\mathbf{S}_{th}$ do not have any backpack), the generated outputs lack shape variety for the human part, producing always the source human's identity. In the second row of Fig. 9, backpacks are added to novel identities; however, the method suffers from lack of details on both humans and objects. In contrast, the outputs of our method shown in the last row show strong generalization by creating diverse human identities with naturally attached detailed objects.

To further validate this comparison, we perform a user test (A/B test) on CloudResearch Connect. We render samples from three viewpoints (same views for all) and show ours with each baseline (A/B examples) in a random order to each subject. Each subject answers 5 questions per baseline by choosing more authentic 3D human samples. The data was collected from 50 subjects. The results are shown in the "User Preference" column in Tab. 1. As shown, our methods are preferred over extended gDNA baselines. Moreover, to confirm the diversity of identities in our method and *'gDNA (w/ object)'*, 50 subjects were shown the rendering of the source human and were asked to choose samples that don't resemble the source human. Samples of our method were chosen by 92.4%, indicating that *'gDNA (w/ object)'* suffers to generate novel identities with a backpack.

Quantiative Evaluation via FID. To evaluate the generation quality of our method, we compare Fréchet Inception Distance (FID) between the 2D normal renderings of the test dataset $\mathbf{S}_{unseen+bp}$ and the generated outputs, following [13]. The result is shown in Tab. 1. *'gDNA (w/ object)'* has a relatively better score than ours, due to the fact that it only samples 3D humans around $\mathbf{S}_{sh+bp}$, which are always close to the GT samples. A more fair comparison is between ours and *'Arith. gDNA (w/ object)'*, where both approaches try to attach the backpack to novel identities. Our method significantly outperforms this baseline.

Method	Pred-to-Scan↓	Scan-to-Pred↓
gDNA	0.0162	0.0190
gDNA(w/ object)	0.0218	0.0112
Ours	**0.0116**	**0.0099**

Table 2. Fitting accuracy comparison with the SOTA method [13].

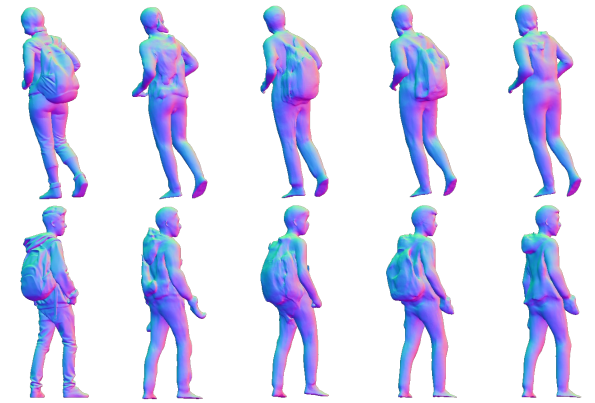

Figure 10. **Fitting and Object Removal.** Compared to baselines, our method successfully explains both human shapes and object shapes, enabling the natural removal of objects after fitting.

Performance on Fitting. We evaluate the expressiveness of our model by fitting it to unseen scans with objects. As a baseline, we consider gDNA [13] as it demonstrates better fitting results on 3D clothed human scans over other SOTA methods [16, 48]. Besides the original gDNA trained with $\mathbf{S}_{th}$, we also consider gDNA trained with $\mathbf{S}_{th}$, $\mathbf{S}_{sh}$ and $\mathbf{S}_{sh+bp}$ (*'gDNA (w/ object)'*) to enable fitting of the object part. We use scans with backpacks from Renderpeople[2] [52] and captured dataset $\mathbf{S}_{unseen+bp}$ for fitting comparison.

As shown in Tab. 2, our method reports better fitting accuracy than the baselines. Our method effectively fits the geometry of both humans and objects while baselines only reconstruct either the human part or the object part as shown in Fig. 10. Moreover, since our method separately models humans and objects, it enables the high-quality removal of objects after fitting.

[2]RenderPeople was downloaded, accessed, and used in this research exclusively at SNU.

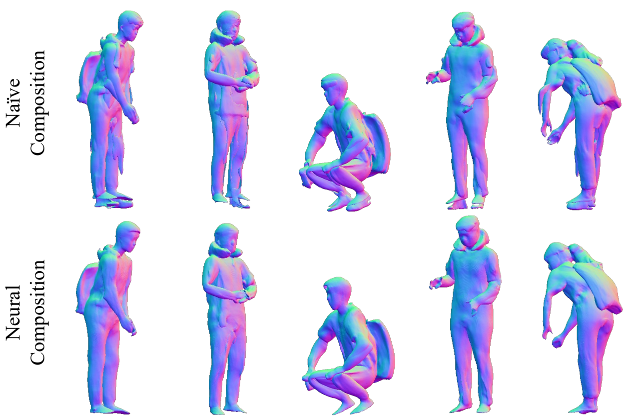

Figure 11. **Composition Comparison.** While naive composition suffers from severe artifacts, neural composition reduces these artifacts and produces high-quality outputs.

Method	Chamfer↓	P2S↓	Normal↓
Occ	0.0140	0.0169	0.0092
Occ & SDF	**0.0098**	**0.0128**	**0.0074**

Table 3. Reconstruction accuracy comparison between the original gDNA and gDNA trained with additional SDF loss terms.

5.4. Ablation Study

Neural Composition. Our system provides two ways of extracting the final composition output. One is by using o_{comp}: neural composition, and the other is by using the maximum value between o_h and o_o of queried points: naive composition. We verify the necessity of using neural composition in order to generate high-quality outputs of humans with objects. Compared to naive composition, neural composition remarkably reduces the artifacts induced by the imperfect fitting of the source human, resulting in lower FID values (Tab. 1). Qualitative comparison is presented in Fig. 11.

Hybrid Modeling of Occupancy and SDF. We verify the significance of predicting both occupancy and SDF over predicting only occupancy to generate outputs with higher frequency details. For each method, we reconstruct the ground truth data used for training with assigned latent codes. We then compute the Chamfer distance and point-to-surface distance (P2S) between the ground truth and the reconstruction output. We also render 2D normal maps from fixed views and compute the L2 error (Normal). As demonstrated in Tab. 3, reconstruction outputs are improved when both occupancy and SDF are predicted. In Fig. 12 we show the qualitative comparison between samples generated via each method. See supp. mat. for further experimental results.

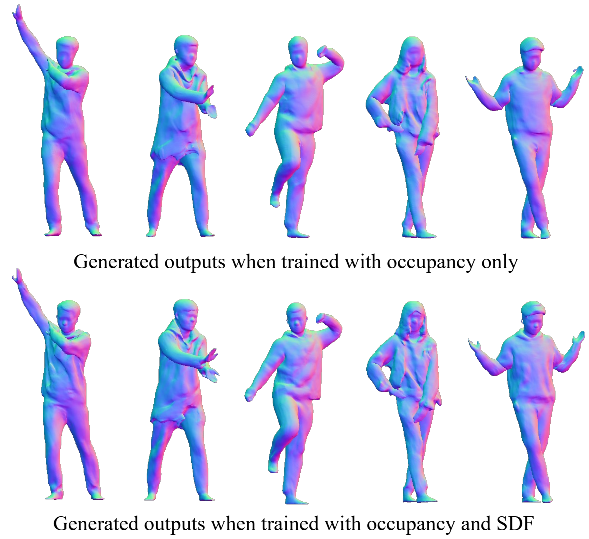

Figure 12. **Qualitative Comparison on Introducing SDF Network in the Human Module.** Top row: Generated outputs when trained with occupancy only. Bottom row: Generated outputs when trained with the hybrid modeling of occupancy and SDF. Additionally predicting the SDF improves the details of generated outputs.

6. Discussion

We present a novel framework for learning a compositional generative model of humans and objects (backpacks, coats, scarves, and more). Our compositional generative model provides separate controllability for the human part and object part. To train our compositional model without manual annotation for the object geometries, we propose to leverage 3D scans of a single person with and without objects. Our results show that the learned generative model for the object part can be authentically transferred to novel human identities.

Limitations and Future Work. While our approach is general and supports diverse objects, decomposing thin layers of clothing in an unsupervised manner remains a challenge due to the limited precision of 3D scans and inherent limitations of the SDF representation. Furthermore, it is important to note that our method is also subject to the quality of gDNA, where the upper limit of the quality achievable in our composition output is constrained by the quality limitations of gDNA. Extending our approach towards better quality outputs with textures, or to model compositions of humans and objects from RGB images would be an exciting research direction for future work.

Acknowledgements: The work of H. Joo and T. Kim was supported by SNU Creative-Pioneering Researchers Program, NRF grant funded by the Korea government (MSIT) (No. 2022R1A2C2092724), and IITP grant funded by the Korea government(MSIT) (No.2022-0-00156 and No.2021-0-01343). H. Joo is the corresponding author.

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *CVPR*, 2019. 2

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, 2018. 2

[3] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *ICCV*, 2021. 3

[4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *TOG*. 2005. 2

[5] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. Compositional gan: Learning image-conditional binary composition. *IJCV*, 2020. 3

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

[7] Sourav Biswas, Kangxue Yin, Maria Shugrina, Sanja Fidler, and Sameh Khamis. Hierarchical neural implicit pose network for animation and motion retargeting. *arXiv preprint arXiv:2112.00958*, 2021. 3

[8] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 3

[9] Mallikarjun BR, Ayush Tewari, Xingang Pan, Mohamed Elgharib, and Christian Theobalt. gcorf: Generative compositional radiance fields. *3DV*, 2022. 3

[10] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3

[11] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 3, 4

[12] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[13] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars. In *CVPR*, 2022. 2, 3, 4, 5, 6, 8

[14] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *ICCV*, 2021. 2, 3, 4

[15] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 2, 3

[16] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *CVPR*, 2021. 2, 3, 8

[17] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *ECCV*, 2020. 2, 3, 4

[18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3

[19] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In *CVPR*, 2022. 2

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

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 3

[23] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *ICML*. 2020. 4, 6

[24] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 2022. 3

[25] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *TOG*, 2022. 2

[26] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 4

[27] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *ECCV*, 2020. 2, 3

[28] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 2

[29] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ICLR*, 2017. 3

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3

[31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 3

[32] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *TOG*, 2013. 3

[33] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2013. 3

[34] Junxuan Li, Shunsuke Saito, Tomas Simon, Stephen Lombardi, Hongdong Li, and Jason Saragih. Megane: Morphable eyeglass and avatar network. In *CVPR*, 2023. 3

[35] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *CVPR*, 2018. 3

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

[37] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *CVPR*, 2020. 2, 3

[38] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *ICCV*, 2021. 2

[39] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 2, 3

[40] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. Coap: Compositional articulated occupancy of people. In *CVPR*, 2022. 2, 3, 4

[41] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. Leap: Learning articulated occupancy of people. In *CVPR*, 2021. 2

[42] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3

[43] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011. 3

[44] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *ICCV*, 2019. 3

[45] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 3

[46] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, 2021. 3

[47] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *CVPR*, 2021. 3

[48] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *ICCV*, 2021. 8

[49] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 2, 3, 5

[50] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 2

[51] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016. 8

[52] Renderpeople, 2018. https://renderpeople.com/3d-people. 2, 8

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

[54] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 2, 3

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

[56] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*, 2021. 2

[57] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *NeurIPS*, 2020. 3

[58] Vincent Sitzmann, Eric R. Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions. In *NeurIPS*, 2020. 4

[59] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *NeurIPS*, 2020. 6

[60] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3

[61] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *NeurIPS*, 2021. 3

[62] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *ICCV*, 2021. 2

[63] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, 2017. 3

[64] Raquel Vidaurre, Igor Santesteban, Elena Garces, and Dan Casas. Fully convolutional graph neural networks for parametric virtual try-on. In *CGF*, 2020. 3

[65] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. In *NeurIPS*, 2021. 2

[66] Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. Learning compositional radiance fields of dynamic human heads. In *CVPR*, 2021. 3

[67] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *ECCV*, 2022. 3

[68] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, et al. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *TOG*, 2022. 3

[69] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *CGF*, 2022. 2

[70] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: implicit clothed humans obtained from normals. In *CVPR*, 2022. 2, 3

[71] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *CVPR*, 2020. 2, 3

[72] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe hd: A high-resolution 3d-aware generative model. In *CVPR*, 2022. 3

[73] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *ICCV*, 2021. 3

[74] Hong-Xing Yu, Leonidas J. Guibas, and Jiajun Wu. Unsupervised discovery of object radiance fields. In *ICLR*, 2022. 3

[75] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021. 2, 3, 6

[76] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *ECCV*, 2020. 3

[77] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A Efros. Learning a discriminative model for the perception of realism in composite images. In *ICCV*, 2015. 3