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

Taeksoo Kim¹  Shunsuke Saito²  Hanbyul Joo¹  
¹Seoul National University  ²Meta Reality Labs

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 naïvely 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 point-clouds [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.

3. Preliminaries

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

We collect 3D scans, $S_{sh}$ and $S_{sh+o}$, 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 $z$ 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:

$$o = O(x^c, G(z)),$$

Figure 2. Examples of Our Datasets. Top row: sample scans of $S_{sh}$. Bottom row: sample scans of $S_{sh+o}$. 
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 $z$ via adaptive instance normalization [26].

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

$$x^d = \sum_{i=1}^{n_b} W_i(N(x^c, \beta), z) \cdot B_i(\beta, \theta) \cdot x^c,$$

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 normalizing the body size variation given SMPL shape parameters $\beta \in \mathbb{R}^{10}$. $B_i(\beta, \theta)$ is the transformation of bone $i$ in SMPL model given $\beta$ and pose $\theta \in \mathbb{R}^{24 \times 3}$. To jointly learn the occupancy and deformation networks, we solve for $x^c$ in Eq. 2 given $x^d$ 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 $f_h$, which is the intermediate latent feature before the last layer, in the canonical space:

$$(o_h, d_h, f_h) = O_h(x^c, G_h(z_h)),$$

where $z_h$ 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 $f_o$, which is the intermediate latent feature before the last layer, in the canonical space:

$$(o_o, f_o) = O_o(x^c, G_o(z_o)),$$

where $z_o$ 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 $f_h$ and $f_o$ 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:

$$o_{comp} = O_{comp}(x^c, f_h, f_o)$$
th their SMPL shape and pose parameters. Following the auto-

4.3. Training

loss functions discussed in Sec. 4.3. allowing the model to be interaction-aware.

sively model the residual geometry, which the human mod-

be computed as

human module to account for slight shape variations of the

z

the learnable shape code

end, we first train the human module

M

resent objects as the residual of human geometry. To this

3D scans in an unsupervised manner, our key idea is to rep-

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 rep-

resent objects as the residual of human geometry. To this

end, we first train the human module

M

using

S

sh

with the learnable shape code

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

S

sh+o

, we jointly op-

timize the per-scan latent code

z

sh

for the last stage are re-initialized as the mean of

z

sh

after the second stage, denoted

z

sh.

M

comp

models all training samples using the feature vector from either

M

th

or

M

sh

for the human part, and from

M

o

for the object part. In the case of

S

th

and

S

sh

where scans are with no objects, we introduce a new latent code

z

emp

as an input to

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

L

th

between the predicted occupancy of

M

th

and the ground truth occupancy. Note that

O

th

(·)

and

F

d

(·)

denote the occupancy field and SDF in posed space, respectively. We also use guidance losses

L

bone

, 

L

joint

and

L

warp

to aid training. 

L

bone

encourages the occupancy of

x

bone

to be one, where

x

bone

are randomly

selected points along the SMPL bones in canonical space. 

L

joint

encourages the skinning weights of SMPL joints to be 0.5 for connected two bones and 0 for all other bones. 

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

to zero. Lastly, we use

L

reg,th

to regularize the latent code

z

th

to be close to zero.

\[
L_{th} = BCE((O_{th}(x^c, G_{th}(z_{th})), o_{gt})) \quad (7)
\]

\[
L_{bone} = BCE((O_{th}(x_{bone}, G_{th}(z_{th}))), 1) \quad (8)
\]

\[
L_{joint} = \|W(x_{joint}, z_{th}) - w_{gt}\| \quad (9)
\]

\[
L_{warp} = \|N(v(\beta), \beta) - v(\beta_0)\| \quad (10)
\]

\[
L_{reg,th} = \|z_{th}\| \quad (11)
\]
For training the SDF network, we use L1 loss $L_{\text{sd}f}$ between the predicted and the ground truth signed distance and L2 loss $L_{\text{nm}d}$ between the gradients of SDF and the ground truth normals of points on the surface. We additionally use $L_{\text{igr}}$ for SDF to satisfy the Eikonal equation [23] and $L_{\text{bbox}}$ to prevent SDF values of off-surface points from being the zero-level surface as in [59].

$$L_{\text{sd}f} = |F_{\text{th}}(x^c, G_{\text{th}}(z_{\text{th}})) - d_{\text{gt}}|$$  \hspace{1cm} (12)

$$L_{\text{nm}d} = \|\nabla F_{\text{th}}(x^c, G_{\text{th}}(z_{\text{th}})) - n_{\text{gt}}\|$$  \hspace{1cm} (13)

$$L_{\text{igr}} = (\|\nabla F_{\text{th}}(x^c, G_{\text{th}}(z_{\text{th}}))\| - 1)^2$$  \hspace{1cm} (14)

$$L_{\text{bbox}} = \exp(-\alpha \cdot |F_{\text{th}}(x^c, G_{\text{th}}(z_{\text{th}}))|), \alpha \gg 1$$  \hspace{1cm} (15)

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

$$L_{\text{sh}} = BCE((O_{\text{sh}}^d(x^c, G_{\text{sh}}(z_{\text{sh}})), o_{\text{gt}}))$$  \hspace{1cm} (16)

$$L_{\text{reg}, \text{sh}} = \|z_{\text{sh}}\|$$  \hspace{1cm} (17)

For the last stage, we use the binary cross entropy loss $L_{\text{comp}}$ between the predicted occupancy of $M_{\text{comp}}$ and the ground truth occupancy. We also use $L_{\text{o}}$ between the predicted occupancy of $M_{\text{o}}$ and the residual part of $S_{\text{sh}+o}$ where $M_{\text{h}}$ cannot explain. Moreover, we optimize $z_{\text{sh}}$ by using the binary cross entropy loss $L_{\text{fit}}$ between the output of $M_{\text{sh}}$ and the ground truth occupancy. Finally, we regularize $z_{\text{sh}}$ to be close to $z_{\text{sh}}$ and $z_{\text{o}}$ to be close to zero.

$$L_{\text{comp}} = BCE((O_{\text{comp}}^d(x^c, f_h, f_o), o_{\text{gt}}))$$  \hspace{1cm} (18)

$$L_{\text{o}} = BCE((O_{\text{o}}(x^c, G_{\text{o}}(z_{\text{o}})), (1 - o_{\text{th}}) \cdot o_{\text{comp}})$$  \hspace{1cm} (19)

$$L_{\text{fit}} = BCE((O_{\text{sh}}^d(x^c, G_{\text{sh}}(z_{\text{sh}})), o_{\text{gt}})$$  \hspace{1cm} (20)

$$L_{\text{reg}, \text{sh}} = \|z_{\text{sh}} - z_{\text{sh}}\|$$  \hspace{1cm} (21)

$$L_{\text{reg}, \text{o}} = \|z_{\text{o}}\|$$  \hspace{1cm} (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, $S_{\text{sh}}$ (180 samples) and $S_{\text{sh}+o}$ (342 samples). For $S_{\text{sh}+o}$, 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 $S_{\text{th}}$. We use another set with 300 samples of the source human with backpacks only, denoted as $S_{\text{sh}+bp}$. 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, $S_{\text{unseen}+bp}$.

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

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 $z_{\text{h}}$ and object $z_{\text{o}}$.

Random Generation. The 3D avatars created by attaching specific object latent codes $z_{\text{o}}$ to random sampled human codes $z_{\text{h}}$ are shown in Fig. 5 (bottom). The outputs of the human module $M_{\text{th}}$ 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>FID</th>
<th>User Preference</th>
</tr>
</thead>
<tbody>
<tr>
<td>gDNA (w/ object)</td>
<td>41.71</td>
<td>43.6%</td>
</tr>
<tr>
<td>Arith. gDNA (w/ object)</td>
<td>73.81</td>
<td>13.6%</td>
</tr>
<tr>
<td>Ours (Naive composition)</td>
<td>55.29</td>
<td>22.4%</td>
</tr>
<tr>
<td>Ours</td>
<td>51.03</td>
<td>100% - (above)</td>
</tr>
</tbody>
</table>

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.

**5.3. Comparison with SOTA**

Since our method is the first generative model for composing 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 $S_{th}$, $S_{sh}$, $S_{sh+bp}$ 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 $z_{th}^*$ for the source human without a backpack and another latent code $z_{sh+bp}^*$ for the source human with the backpack. We simply take their subtraction $z_{bp} = z_{sh+bp}^* - z_{sh}^*$, 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 $z_{th} + z_{bp}$. 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 $S_{th}$ 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.

Quantitative 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 $S_{unseen+bp}$ and the generated outputs, follow-

<table>
<thead>
<tr>
<th>Method</th>
<th>Pred-to-Scan↓</th>
<th>Scan-to-Pred↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>gDNA</td>
<td>0.0162</td>
<td>0.0190</td>
</tr>
<tr>
<td>gDNA (w/ object)</td>
<td>0.0218</td>
<td>0.0112</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.0116</strong></td>
<td><strong>0.0099</strong></td>
</tr>
</tbody>
</table>

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 $S_{th}$, we also consider gDNA trained with $S_{th}$, $S_{sh}$ and $S_{sh+bp}$ (‘gDNA (w/ object)’) to enable fitting of the object part. We use scans with backpacks from Renderpeople[^2] and captured dataset $S_{unseen+bp}$ 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Chamfer↓</th>
<th>P2S↓</th>
<th>Normal↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occ</td>
<td>0.0140</td>
<td>0.0169</td>
<td>0.0092</td>
</tr>
<tr>
<td>Occ &amp; SDF</td>
<td>0.0098</td>
<td>0.0128</td>
<td>0.0074</td>
</tr>
</tbody>
</table>

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 $o_b$ 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.

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


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


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


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


