gDNA: Towards Generative Detailed Neural Avatars

Xu Chen[1,3]   Tianjian Jiang[1]   Jie Song[1]   Jinlong Yang[3]
Michael J. Black[3]   Andreas Geiger[2,3]   Otmar Hilliges[1]

[1]ETH Zürich, Department of Computer Science   [2]University of Tübingen
[3]Max Planck Institute for Intelligent Systems, Tübingen
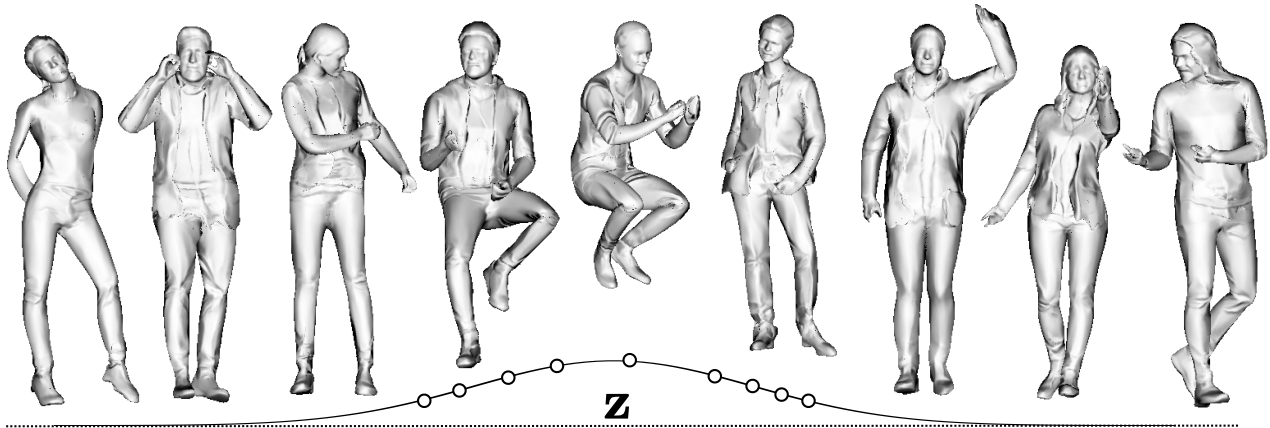https://xuchen-ethz.github.io/gdna

Figure 1. **Generative Detailed Neural Avatars.** We propose a method to *generate* 1) a diverse set of 3D virtual humans of 2) varied identity, gender and shape, appearing in 3) different clothing styles and poses, with 4) realistic and stochastic details such as wrinkles in garments. Our multi-subject method learns shape, articulation and clothing details from few posed scans without requiring skinning weight supervision. The method is able to synthesize novel identities that are not in the training set and generalizes to unseen poses.

Abstract

To make 3D human avatars widely available, we must be able to generate a variety of 3D virtual humans with varied identities and shapes in arbitrary poses. This task is challenging due to the diversity of clothed body shapes, their complex articulations, and the resulting rich, yet stochastic geometric detail in clothing. Hence, current methods that represent 3D people do not provide a full generative model of people in clothing. In this paper, we propose a novel method that learns to generate detailed 3D shapes of people in a variety of garments with corresponding skinning weights. Specifically, we devise a multi-subject forward skinning module that is learned from only a few posed, un-rigged scans per subject. To capture the stochastic nature of high-frequency details in garments, we leverage an adversarial loss formulation that encourages the model to capture the underlying statistics. We provide empirical evidence that this leads to realistic generation of local details such as wrinkles. We show that our model is able to generate natural human avatars wearing diverse and detailed clothing. Furthermore, we show that our method can be used on the task of fitting human models to raw scans, outperforming the previous state-of-the-art.

1. Introduction

The ability to easily create diverse high-quality virtual humans with full control over their pose has many applications in movie production, games, VR/AR, architecture, and computer vision. While modern computer graphics techniques achieve photorealism, they typically require a lot of expertise and extensive manual effort. Our goal is to make 3D human avatars widely accessible by learning a *generative model of people*. Towards this goal, we propose the first method that can *generate* 1) diverse 3D virtual humans with 2) various identities and shapes, appearing in 3) different clothing styles and poses, with 4) realistic and stochastic high-frequency details such as wrinkles in garments.

Generative modeling of 3D rigid objects has recently seen rapid progress, fueled by continuous and resolution-independent neural 3D representations [13, 40, 43, 48, 54]. However, modeling clothed humans and their articulation is more difficult due to the complex interaction of garments, their topology, and pose-driven deformations. Recent work leverages neural implicit surfaces to learn high-quality articulated avatars for a *single* subject [12, 15, 52, 58] but these methods are not generative, i.e. they cannot synthesize novel human identities and shapes. Generative models of *clothing* exist that augment SMPL by predicting displacements from the body mesh (CAPE [36]), or by drap-

ing an implicit garment representation on a T-posed body (SMPLicit [14]), and relying on SMPL's learned skinning for reposing. We show empirically that holistic modeling of identity, shape, articulation and clothing leads to higher fidelity generation and animation of virtual humans and to higher accuracy in fitting to 3D scans.

Taking a step towards fully **g**enerative modeling of **d**etailed **n**eural **a**vatars, we propose gDNA, a method that synthesizes 3D surfaces of *novel* human shapes, with control over the clothing style and pose, and that produces realistic high-frequency details of the garments (Fig. 1). To leverage raw (*posed*) 3D scans, we build a multi-subject implicit generative representation. We build upon SNARF [12], a recent method for learning *single-subject* articulation-dependent effects that has been shown to generalize well to unseen poses. SNARF [12] requires *many* poses of a *single* subject for training. In contrast, our *multi-subject* method can be learned from *very few* posed scans (1-3) of many different subjects. This is achieved via the addition of a latent space for the conditional generation of shape and skinning weights for clothed humans. Furthermore, a learned warping field yields accurate deformations, using the same skinning field, independent of body size.

Clothing wrinkles are produced by an underlying stochastic process. To capture these effects, we propose a method that learns the underlying statistics of 3D clothing details via an adversarial loss. Previous mesh-based approaches formulate this in UV-space [28], which is not directly applicable to implicit surfaces due to the lack of mesh connectivity. To learn high-frequency details, we first predict a 3D normal field, conditioned on the coarse shape features. To backpropagate the adversarial loss to the 3D normal field we establish 3D-2D correspondences by augmenting forward skinning with an implicit surface renderer. We show that adversarial training leads to significantly improved fidelity of 3D geometric details, see Fig. 9.

Trained from posed scans only, we demonstrate the first method that can generate a large variety of 3D clothed human shapes with detailed wrinkles under pose control. The generated samples can be reposed via the learned skinning weights. We evaluate gDNA quantitatively, qualitatively, and through a perceptual study; gDNA strongly outperforms baselines. Furthermore, we show that gDNA can be used for fitting and re-animation of 3D scans, outperforming the state of the art (SOTA). In summary, we contribute:

- The first method to generate a large variety of animatable 3D human shapes in detailed garments; that
- learns from raw posed 3D scans without requiring canonical shapes, detailed surface registration, or manually defined skinning weights, and
- a technique to significantly improve the geometric detail in clothing deformation, based on recovering the underlying statistics of cloth deformation.

2. Related Work

2D and 3D Generative Models: Most modern methods for synthesizing natural images leverage generative adversarial networks (GANs) [18] or variational auto-encoders (VAEs) [27]. These methods have achieved a high level of photorealism [24–26] and can yield impressive results on the task of synthesizing 2D images of humans [4, 11, 19, 29, 34, 53]. However, such methods reason in 2D and hence 3D consistency cannot be guaranteed [11, 29, 34] nor is extracting 3D geometry from such approaches straightforward.

Several methods for the task of learning *rigid* 3D shapes exist. Early methods rely on voxel [61] or point cloud [3] representations. More recently, several methods represent object shapes by learning an implicit function using neural networks [13, 40, 48]. Such representations have also been proposed for the task of *generative* modeling of 3D shapes [10, 13, 16, 40, 42, 43, 48, 54]. However, these methods are typically not easily extended to non-rigid clothed humans. In this paper, we study the problem of 3D implicit *generative* modeling of *non-rigid* human shape.

3D Human Models: Parametric 3D human body models [7, 23, 33, 46, 64] can synthesize 3D human shapes from a set of low-dimensional control parameters by deforming a template mesh. This idea has also been extended to model clothed humans [5, 36]. However, geometric expressivity is limited due to the fixed mesh topology and the bounded resolution of the template mesh.

To overcome the topology and resolution limitations of meshes, other representations, including point clouds [35, 37, 65], implicit surfaces [12, 47, 52, 55, 58, 60], and radiance fields [32, 45, 49, 56, 63], have been explored. In particular, neural implicit surface representations have emerged as a powerful tool to model 3D (clothed) human shapes [6, 15, 17, 20, 21, 30, 41, 50, 51, 62, 66, 67] due to their topological flexibility and resolution independence. Recent work [12, 52, 58] uses implicit surfaces to learn human avatars for a *single* subject, wearing a *specific* garment. These methods model clothing details such as wrinkles as a deterministic function of the body poses. However, due to hysteresis and complex material properties, garment folds and wrinkles are stochastic and existing methods struggle to capture these effects. In contrast, we propose a *multi-subject* generative model of 3D humans that provides separate control over poses, garments and can synthesize realistic geometric details.

CAPE [36] and SMPLicit [14] are generative models of *clothing* only, based on meshes and implicit surfaces respectively. Both methods are purely additive, that is they drape an implicit garment over the SMPL body [14] or predict the displacement parameters of a SMPL+D template mesh [36]. We experimentally show that this leads to lower fidelity in generated samples and higher error when fitting
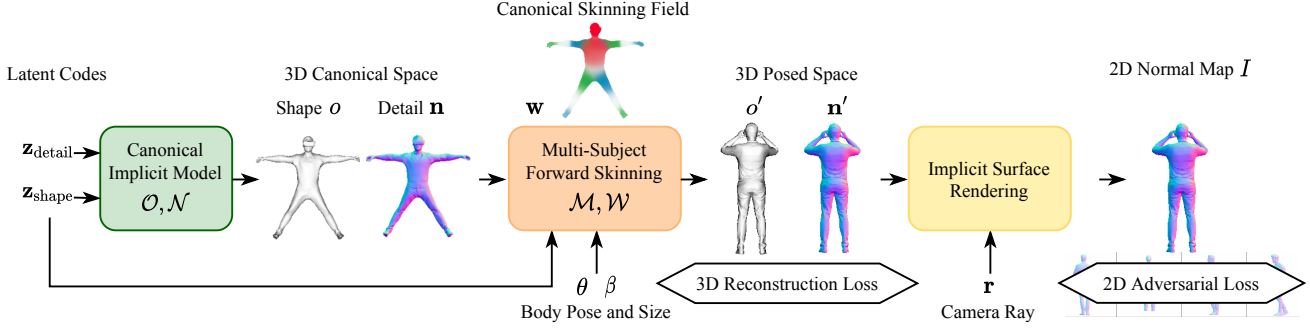
Figure 2. **Method Overview.** We represent clothed humans using coarse shapes and detailed normals in a pose- and body-size-independent canonical space. This canonical representation can then be deformed to target body poses and sizes via a multi-subject forward skinning module. The deformed shapes are compared to raw posed scans via a 3D reconstruction loss to learn canonical shapes and skinning. To improve surface details, we augment the forward skinning module with an implicit surface renderer to generate 2D normal maps and learn detailed 3D normal fields by applying a 2D adversarial loss formulation.

to 3D scans. NPMs [47] provide a latent space of multiple subjects for fitting to RGB-D depth maps or 3D scans.

A common problem of all aforementioned approaches is the specific training data requirements these models impose: They either require synthetic data in canonical space [14, 47], or precise registration of a template mesh to posed scans [36, 47]. The former are rare and suffer from a domain gap, while the latter is challenging to attain. Our method overcomes this issue by requiring only a *few* training samples of each subject in *posed* space. We show that our method learns complex shape and clothing details and models realistic deformation even from such limited data.

Adversarial Training of Clothing Details: Adversarial loss formulations have been used to learn detailed cloth wrinkles by optimizing 2D representations such as UV normal maps [28] or depth images [59]. It is noteworthy that implicit surfaces lack a notion of connectivity and therefore, incorporating 2D representations that have been designed to augment explicitly parametrized meshes is not straightforward. In contrast, we propose a formulation that leverages a 2D adversarial loss computed with posed images to optimize a 3D implicit representation in canonical space. Finally, our focus is the generation of human shapes appearing in varied clothing styles and diverse identities while previous methods focus on reconstruction [59] or single garment pose-dependent wrinkle enhancement [28].

3. Method

Our goal is to build a model that generates diverse 3D clothed humans with varying identities and fine-grained geometric details in arbitrary poses. Our model is learned from a sparse set of static scans without assuming surface correspondences. Our method is summarized Fig. 2.

First, we formulate a pose- and body-size-independent canonical representation of clothed human shapes (Section 3.1). Second, to learn the canonical shape and defor-

mation properties from very few posed scans of each of the subjects, we extend a single-subject differentiable forward skinning method [12] to multiple subjects via a latent space of shape, articulation and garment (Section 3.2). Finally, to learn rich yet stochastic geometric details, we learn a detailed 3D normal field via a 2D adversarial loss formulation. To achieve this, we augment the forward skinning module with an implicit surface renderer (Section 3.3). Training details are discussed in Section 3.4.

3.1. Canonical Representation

Our method is based on neural implicit representations, leveraging their topological flexibility and resolution independence. We model the clothed human shape and geometric clothing details jointly.

Coarse Shape: We model the shape in canonical space as the $\tau = 0.5$ level set of a neural occupancy function:

$$\mathcal{S}(\mathbf{z}_{\text{shape}}) = \{\mathbf{x} \mid \mathcal{O}(\mathbf{x}, \mathbf{z}_{\text{shape}}) = \tau\}, \quad (1)$$

where $\mathcal{O}$ is a neural network that predicts the occupancy probability o for any 3D point $\mathbf{x}$ in canonical space. The prediction is conditioned on a *shape* code $\mathbf{z}_{\text{shape}} \in \mathbb{R}^{L_{\text{shape}}}$:

$$\mathcal{O} : \mathbb{R}^3 \times \mathbb{R}^{L_{\text{shape}}} \to [0, 1] \times \mathbb{R}^{L_{\mathbf{f}}} \quad (2)$$
$$(\mathbf{x}, \mathbf{z}_{\text{shape}}) \mapsto (o, \mathbf{f})$$

This occupancy network also outputs a feature vector $\mathbf{f}$ of dimension $L_{\mathbf{f}}$ for each surface point. This feature carries coarse shape information and is used to predict fine details.

We combine a 3D CNN-based feature generator and a locally conditioned MLP to model $\mathcal{O}$. A 3D style-based [25, 42] generator, illustrated in Fig. 3 first produces a 3D feature volume conditioned on $\mathbf{z}_{\text{shape}}$ via adaptive instance normalization [22]. The final occupancy is obtained via trilinear sampling of the feature volume and by feeding the feature and the 3D coordinate into an MLP.
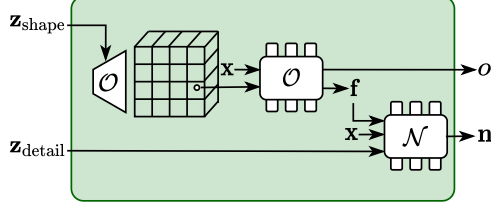
Figure 3. **Canonical Implicit Model.** Given latent codes $\mathbf{z}_{\text{shape}}$ and $\mathbf{z}_{\text{detail}}$, this module predicts the occupancy probability o and the normals $\mathbf{n}$ for 3D points in canonical space. To predict o, a 3D style-based generator first produces a 3D feature volume conditioned on $\mathbf{z}_{\text{shape}}$. The final occupancy of a 3D point is obtained via trilinear sampling of the feature volume and by feeding the feature and the 3D coordinate into an MLP. To predict normals $\mathbf{n}$, we use an MLP conditioned on feature $\mathbf{f}$ and latent code $\mathbf{z}_{\text{detail}}$.

Detailed Surface Normals: Learning an occupancy field for multiple subjects and garment types with accurate and detailed normals is challenging and we empirically show that a naïve implementation leads to artifacts on the surface (cf. Fig. 9). Analogous to normal mapping for polygon meshes [8, 28], we model surface details via normals in canonical 3D space. Such surface normals can be represented by the gradient of the implicit function, but this results in considerable computational complexity. Therefore, we use an MLP to predict surface normals similar to [58]. However, because implicit surfaces have no notion of connectivity, we propose a geometry-aware approach to link coarse geometry and the detailed normal field. More specifically, we condition the surface normal prediction on the underlying shape, leveraging the feature $\mathbf{f}$ from the occupancy network. We further condition the field on a latent $\mathbf{z}_{\text{detail}} \in \mathbb{R}^{L_{\text{detail}}}$ to enable generation of controllable details for the same coarse shape:

$$\mathcal{N} : \mathbb{R}^3 \times \mathbb{R}^{L_{\text{detail}}} \times \mathbb{R}^{L_{\mathbf{f}}} \to \mathbb{R}^3 \tag{3}$$
$$(\mathbf{x}, \mathbf{z}_{\text{detail}}, \mathbf{f}) \mapsto \mathbf{n}$$

3.2. Multi-Subject Forward Skinning

We additionally model the deformation properties and define the body size ($\boldsymbol{\beta}$) and pose ($\boldsymbol{\theta}$) parameters to be consistent with SMPL, enabling use of existing datasets (e.g. AMASS [38]) for animation. The body size parameter $\boldsymbol{\beta}$ is a 10-dimensional vector, and the body pose parameter $\boldsymbol{\theta}$ represents the joint angles of SMPL's skeleton.

Single-Subject Skinned Representation: To animate implicit human shapes in controllable body poses $\boldsymbol{\theta}$, recent work [12, 41, 52, 58] generalizes mesh-based linear blend skinning algorithms to neural implicit surfaces. The skeletal deformation of each 3D point is modeled as the weighted average of a set of bone transformations, with weights at each point predicted by an MLP. A key difference is whether this skinning weight field is defined in canonical space or in
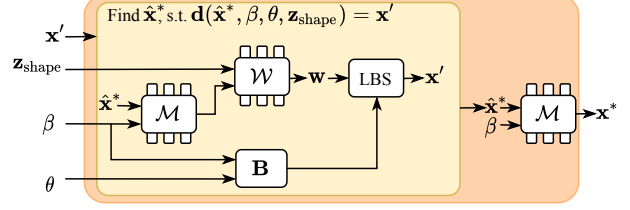
posed space. We follow Chen et al. [12] who define the skinning field in canonical space:

$$\mathcal{W} : \mathbb{R}^3 \to \mathbb{R}^{n_b} \tag{4}$$
$$\mathbf{x} \mapsto \mathbf{w},$$

where n_b denotes the number of bones and the weights $\mathbf{w} = \{w_1, \ldots, w_{n_b}\}$ of each point $\mathbf{x}$ are enforced to satisfy $w_i \geq 0$ and $\sum_i w_i = 1$ by a softmax activation function. As shown in [12], defining the skinning weights field in canonical space is desirable because the skinning weights are then pose-independent, thus easier to learn and enabling generalization to out-of-distribution poses.

Multi-Subject Skinned Representation: We extend this forward skinning idea to multiple subjects. Since the skinning weight field is defined in canonical space, the model can aggregate information over multiple training instances. Importantly, this enables us to learn skinning from *one or a few* poses of *multiple* subjects, instead of requiring *many* poses of the *same* subject.

To achieve this, we decouple the effects originating from the body size variation $\boldsymbol{\beta}$ and the clothed human shape $\mathbf{z}_{\text{shape}}$. We model the skinning field in a body-size-neutral space, analogously to the canonical surface representations. To capture diverse clothed human shapes, we condition the field on the latent shape code $\mathbf{z}_{\text{shape}}$:

$$\mathcal{W} : \mathbb{R}^3 \times \mathbb{R}^{L_{\text{shape}}} \to \mathbb{R}^{n_b} \tag{5}$$
$$(\mathbf{x}, \mathbf{z}_{\text{shape}}) \mapsto \mathbf{w}$$

We then model body size change with an additional warping field. Given a point $\hat{\mathbf{x}}$ in $\boldsymbol{\beta}$-size space, the warping field maps it back to the mean size by predicting its canonical correspondence $\mathbf{x}$ (see Fig. 4):

$$\mathcal{M} : \mathbb{R}^3 \times \mathbb{R}^{L_{\boldsymbol{\beta}}} \to \mathbb{R}^3 \tag{6}$$
$$(\hat{\mathbf{x}}, \boldsymbol{\beta}) \mapsto \mathbf{x}$$

Figure 4. **Multi-subject Forward Skinning.** This module deforms the canonical occupancy and normal field to the target pose $\boldsymbol{\theta}$ and size $\boldsymbol{\beta}$ by establishing correspondences from posed space $\mathbf{x}'$ to canonical space $\mathbf{x}^*$. First, given a deformed point $\mathbf{x}'$, we find its correspondence in resized canonical space $\hat{\mathbf{x}}^*$ by iteratively finding the root of (9). Subsequently, we map $\hat{\mathbf{x}}^*$ to $\mathbf{x}^*$ in size-neutral canonical space using the warping field.

In this formulation, $\boldsymbol{\beta}$ captures body shape variations analogously to SMPL, e.g. body height. Therefore, the canonical shape network only needs to model the remaining shape variations beyond SMPL, e.g. clothing and hair, controlled by $\mathbf{z}_{\text{shape}}$. The final resized canonical surface is defined by:

$$\hat{\mathcal{S}}(\mathbf{z}_{\text{shape}}, \boldsymbol{\beta}) = \{\hat{\mathbf{x}} \mid \mathcal{O}(\mathcal{M}(\hat{\mathbf{x}}, \boldsymbol{\beta}), \mathbf{z}_{\text{shape}}) = \tau\} \quad (7)$$

Given the target body pose $\boldsymbol{\theta}$, a point $\hat{\mathbf{x}}$ in $\boldsymbol{\beta}$-size space is transformed to posed space $\mathbf{x}'$ via

$$\begin{aligned} \mathbf{x}' &= \mathbf{d}(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}_{\text{shape}}) \\ &= \sum_{i=1}^{n_b} \mathcal{W}_i(\mathcal{M}(\hat{\mathbf{x}}, \boldsymbol{\beta}), \mathbf{z}_{\text{shape}}) \cdot \boldsymbol{B}_i(\boldsymbol{\beta}, \boldsymbol{\theta}) \cdot \hat{\mathbf{x}}, \end{aligned} \quad (8)$$

where $\boldsymbol{B}_i(\boldsymbol{\beta}, \boldsymbol{\theta})$ are the bone transformation matrices obtained from the parametric skeleton of SMPL.

Implicit Differentiable Forward Skinning: While our model learns a canonical representation, its supervision is provided in posed space. Given a point $\mathbf{x}'$ in posed space we need to determine its correspondence in canonical space $\mathbf{x}$ to compare the predicted occupancy and normals to ground-truth. We first find the correspondence $\hat{\mathbf{x}}^*$ of $\mathbf{x}'$ in resized canonical space and then map $\hat{\mathbf{x}}^*$ to canonical space $\mathbf{x}^*$. An overview is provided in Fig. 4. While the goal is to determine $\mathbf{x}' \mapsto \hat{\mathbf{x}}$, we only have direct access to the inverse mapping defined by forward skinning Eq. (8), which is not invertible. Following [12], we determine the correspondence numerically by finding the root of the equation:

$$\mathbf{d}(\hat{\mathbf{x}}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}_{\text{shape}}) - \mathbf{x}' = \mathbf{0}, \quad (9)$$

using Broyden's method [9]. Subsequently, the canonical correspondence $\mathbf{x}^*$ is given by:

$$\mathbf{x}^* = \mathcal{M}(\hat{\mathbf{x}}^*, \boldsymbol{\beta}) \quad (10)$$

We can now determine the occupancy at $\mathbf{x}'$ as $o' = \mathcal{O}(\mathbf{x}^*, \mathbf{z}_{\text{shape}})$ and the normal $\mathbf{n}'$ as

$$\mathbf{n}' = (\sum_{i=1}^{n_b} \mathcal{W}_i(\mathbf{x}^*, \mathbf{z}_{\text{shape}}) \cdot \mathbf{R}_i)^{-T} \mathcal{N}(\mathbf{x}^*, \mathbf{f}, \mathbf{z}_{\text{detail}}) \quad (11)$$

where $\mathbf{R}_i$ denotes the rotational component of $\boldsymbol{B}_i$.

For convenient future reference, we define the occupancy field $\mathcal{O}'$ and normal function $\mathcal{N}'$ in posed space as:

$$\mathcal{O}' : (\mathbf{x}', \mathbf{z}_{\text{shape}}, \boldsymbol{\beta}, \boldsymbol{\theta}) \mapsto o', \mathbf{f} \quad (12)$$
$$\mathcal{N}' : (\mathbf{x}', \mathbf{z}_{\text{detail}}, \mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}) \mapsto \mathbf{n}' \quad (13)$$

3.3. Implicit Surface Rendering

Geometric clothing details are challenging to learn due to their stochastic nature. In 2D image generation tasks, GANs have achieved impressive results on learning high fidelity local textures. We propose to learn better geometric details $\mathcal{N}$ using an adversarial loss. Towards this goal, we augment the forward skinning module with an implicit renderer to establish direct correspondences between 2D projections of 3D points in posed space and corresponding 3D points in canonical space, enabling end-to-end training.

Implicit Rendering with Skinning: Given a pixel $\mathbf{p}$ in the 2D posed normal map, its correspondence in deformed 3D space $\mathbf{x}'$ can be determined by the intersection between the ray through $\mathbf{p}$ and the forward skinned surface:

$$\mathcal{O}'(\mathbf{x}', \mathbf{z}_{\text{shape}}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \tau, \text{with } \mathbf{x}' = \mathbf{r}_c + t \cdot \mathbf{r}_d \quad (14)$$

where $\mathbf{r}_d$ and $\mathbf{r}_c$ denote the ray direction and origin, and t is the scalar distance along the ray. Following [44], we determine the intersection point $\mathbf{x}'$ by finding the first change of occupancy $\mathcal{O}'$ along the ray using the Secant method. We also obtain the canonical correspondence point $\mathbf{x}$ of $\mathbf{p}$ via forward skinning. Solving the 3D canonical correspondence for each pixel, yields the 2D normal map I:

$$I_{\mathbf{p}} = \mathcal{N}'(\mathbf{x}', \mathbf{z}_{\text{detail}}, \mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}) \quad (15)$$

3.4. Training

We train our method via a set of posed scans and their corresponding SMPL parameters $\boldsymbol{\theta}, \boldsymbol{\beta}$. We follow the auto-decoding framework of [48], and assign one shape code $\mathbf{z}_{\text{shape}}$ and one detail code $\mathbf{z}_{\text{detail}}$ to each training sample. These are initialized to be zero and optimized jointly with the network weights. To enable sampling, we fit a Gaussian distribution to the latent codes after training.

We split training into two stages: We first train the coarse shape, skinning, and warping networks and then train the normal network. This two-stage training is essential. Otherwise, the normal supervision will be back-propagated to wrong locations in canonical space due to wrong correspondences before training of shape and skinning converges.

For the first stage, we use the binary cross entropy loss $\mathcal{L}_{\text{BCE}}$ between predicted occupancy $\mathcal{O}'(\mathbf{x}', \mathbf{z}_{\text{shape}}, \boldsymbol{\beta}, \boldsymbol{\theta})$ and ground-truth o_{gt}. Following [12], we add auxiliary losses $\mathcal{L}_{\text{bone}}$ and $\mathcal{L}_{\text{joint}}$ to guide learning during early iterations:

$$\mathcal{L}_{\text{bone}} = BCE(\mathcal{O}(\mathbf{x}_{\text{bone}}, \mathbf{z}_{\text{shape}}), 1) \quad (16)$$
$$\mathcal{L}_{\text{joint}} = \|\mathbf{w}(\mathbf{x}_{\text{joint}}, \mathbf{z}_{\text{shape}}) - \mathbf{w}_{\text{joint, target}}\|_2^2 \quad (17)$$

where $\mathbf{x}_{\text{bone}}$ are randomly sampled points on canonical bones, $\mathbf{x}_{\text{joint}}$ are randomly sampled canonical joints, and $\mathbf{w}_{\text{joint, target}}$ is a vector that is 0.5 for the neighboring bones and 0 elsewhere (for details see Sup. Mat.). To ensure that the warping field changes body size consistently, we enforce the warping field to warp SMPL vertices $\mathbf{v}(\boldsymbol{\beta})$ to the corresponding location in the neutral shape $\mathbf{v}(\boldsymbol{\beta}_0)$:

$$\mathcal{L}_{\text{warp}} = \|\mathcal{M}(\mathbf{v}(\boldsymbol{\beta}), \boldsymbol{\beta}) - \mathbf{v}(\boldsymbol{\beta}_0)\|_2^2 \quad (18)$$

Finally, we regularize the latent code to be close to the origin of the latent space via $\mathcal{L}_{\text{reg,shape}} = \|\mathbf{z}_{\text{shape}}\|_2^2$.

Figure 5. **Clothed Human Generation.** We show randomly sampled clothed human shapes generated by our method. *Top:* generated canonical shapes with detailed normals and skinning weights. *Bottom:* generated shapes reposed (2 poses each). Shapes are visualized as normal maps to highlight the detailed geometry synthesized by our method. For shaded results, see other figures, e.g. Fig. 1.

The normal prediction network is trained subsequently. Here we penalize differences between the predicted and GT normal $\mathbf{n}'_{gt}$ for randomly sampled surface points:

$$\mathcal{L}_{norm} = 1 - {\mathbf{n}'_{gt}}^T \cdot \mathcal{N}'(\mathbf{x}', \mathbf{z}_{detail}, \mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}) \qquad (19)$$

In addition, we apply non-saturating adversarial losses [18] $\mathcal{L}_{adv} = -\log(1 + \exp(D(I)))$ with R_1 gradient penalty [39] on the predicted 2D normal maps I and the real normal maps rendered from the posed scans I_{real}. D is a jointly trained discriminator (see Sup. Mat. for details). We further regularize $\mathbf{z}_{detail}$ with $\mathcal{L}_{reg,detail} = \|\mathbf{z}_{shape}\|_2^2$.

3.5. Inference

We generate human avatars by randomly sampling $\mathbf{z}_{shape}$ and $\mathbf{z}_{detail}$ from the estimated Gaussian distribution. We then extract meshes in resized canonical space using MISE [40] from the implicit representation $\hat{\mathcal{S}}(\mathbf{z}_{shape}, \boldsymbol{\beta})$ and predict the vertex normal with our normal field. Finally, we pose the meshes to desired poses $\boldsymbol{\theta}$ following Eq. (8).

4. Experiments

Our main goal is to generate 3D human avatars. Since we are the first to tackle this problem setting, we compare our method to carefully designed ablative baselines, enabling analysis of each component of our method. We also evaluate the expressiveness of our model by fitting it to unseen scans and compare the accuracy to SOTA 3D human shape modeling methods. We outline the evaluation protocols in the following and refer the readers to Sup. Mat. for details.

Datasets:

3D Scans: We train our model on commercial scans [1, 2].

SIZER: Following [14], we use the SIZER dataset [57] to evaluate fitting. This dataset contains 3D scans of humans in 21 garments, including shirts, T-shirts, coats and pants.

Metrics:

Fréchet Inception Distance (FID): To evaluate generation quality, we compute FID between 2D normal maps of training scans and those of randomly generated 3D shapes.

User Preference: We conduct a perceptual study among 44 subjects and report how often participants preferred a particular method over ours.

Surface Distance: To evaluate fitting accuracy, we measure the one-directional Chamfer distance between predicted surfaces and the target scans, following SMPLicit [14].

Baselines:

NPMs [47]: NPMs learn the latent space of human shapes and deformation from ground-truth canonical shapes and vertex displacements, obtained from synthetic 3D animations [31] and real scans with registered surfaces [36].

SMPLicit [14]: SMPLicit learns a generative model of 3D garments, and drapes these over the SMPL T-pose. This model is trained with a collection of 3D synthetic garments.

4.1. Quality of Generated Samples

Random Generation of Canonical Shapes: We show random samples generated by our method in Fig. 5 (top). While trained with posed scans only, our method learns plausible canonical shapes with surface details.

Disentangled Pose and Shape: The generated shapes can be reposed as desired, even to poses far beyond the training pose distribution (cf. Fig. 5 bottom and Fig. 1).

Interpolation: Interpolating the shape and details codes, yields smooth transitions of shapes and details between two very different samples, as shown in Fig. 6.
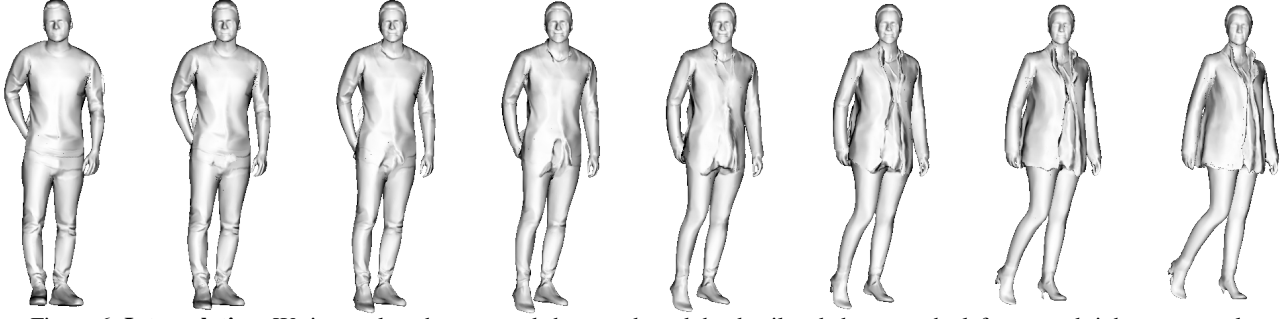
Figure 6. **Interpolation.** We interpolate the pose and shape code and the detail code between the leftmost and rightmost sample.

Figure 7. **Disentangled Generation of Shape and Details.** We generate samples with the same coarse shape and different detail codes. Note that the details appear noticeably different from each other while all match the underlying coarse shape.
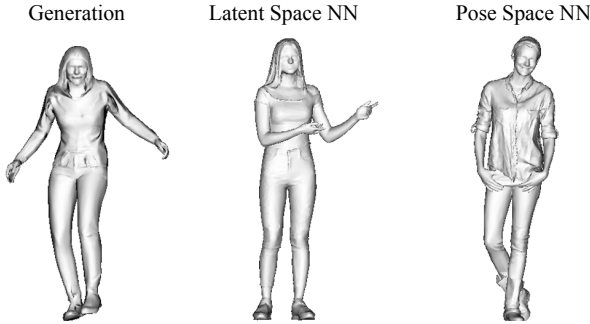

Figure 8. **Nearest Training Samples.** Note that the pose and clothing of the generated shape is noticeably different from that of the nearest training sample.

Disentangled Shape and Details: Our disentangled formulation allows us to generate diverse clothing details for the same coarse shape. Fig. 7 shows results with the same coarse shape $\mathbf{z}_{\text{shape}}$ but different details codes $\mathbf{z}_{\text{detail}}$. While the coarse shape remains the same, gDNA generates varied plausible wrinkles that match the underlying coarse shape.

Extrapolation Beyond Training Distribution: To further illustrate generalization, we show the training samples with the most similar pose and latent code to the generated sample in Fig. 8. The nearest neighbors are noticeably different from our generation, demonstrating that our method generalizes and is able to generate novel shapes in novel poses.

4.2. Ablation Study

We now ablate our design choices. The results are summarized in Tab. 1 and Fig. 9.

Canonical Space Modeling: We verify the necessity to model shapes in canonical space and joint learning of skinning weights. Towards this goal, we implement a baseline that generates posed shapes directly given the latent code and the body pose as input. As shown in Fig. 9 (first row), the individual samples lack details as the baseline must capture a large shape space caused by the pose change. Since the method does not reason about articulation, the sampled shapes suffer from invalid pose configurations, leading to high FID values as shown in Tab. 1 (*Pose ONet*).

Adversarial Learning: The adversarial loss plays an important role in improving the perceptual realism of the generated samples, as evidenced by the FID improvement from *Detailed Normal (w/o Adversarial)* to *Ours* in Tab. 1. The normals estimated directly from the occupancy field suffer from artifacts on the surface (Fig. 9 (second row)). Training without adversary leads to overly smooth geometry (Fig. 9 (third row)), as the reconstruction loss induces a bias that averages out details. In contrast, our method produces realistic high-frequency details (Fig. 9 (bottom)). Notably, in 21.3% of the cases, users consider our generated shapes to be even more realistic than real scans.

4.3. Comparison with SOTA on Model Fitting

While our main goal is to generate clothed human shapes, our model can be fit to raw observations, just like existing 3D parametric human or clothing models. We consider two recent SOTA methods, i.e. NPMs [47] and SMPLicit [14]. We follow SMPLicit [14] and fit ours and the baselines to scans from the SIZER dataset.

Accuracy: While not designed for fitting, our method achieves better accuracy than previous special purpose methods, as demonstrated in Tab. 2. Our method captures the person identity and clothing shapes more faithfully than NPMs and SMPLicit, and our results exhibit more details such as wrinkles (Fig. 10 top). Since the model is trained directly from posed scans, disentangling pose and shape, it learns about real clothing details and can reproduce them.

Method	FID↓	User Preference↑
Pose ONet	43.80	8.11%
Coarse Shape	29.34	26.1%
Detailed Normal (w/o Adv.)	42.18	15.4%
Ours	**11.54**	–
Ground-truth Scans	N/A	78.7%

Table 1. **Ablation Study.** We report FID and user preference. The user preference score indicates how often participants of our perceptual study preferred a particular method over ours.
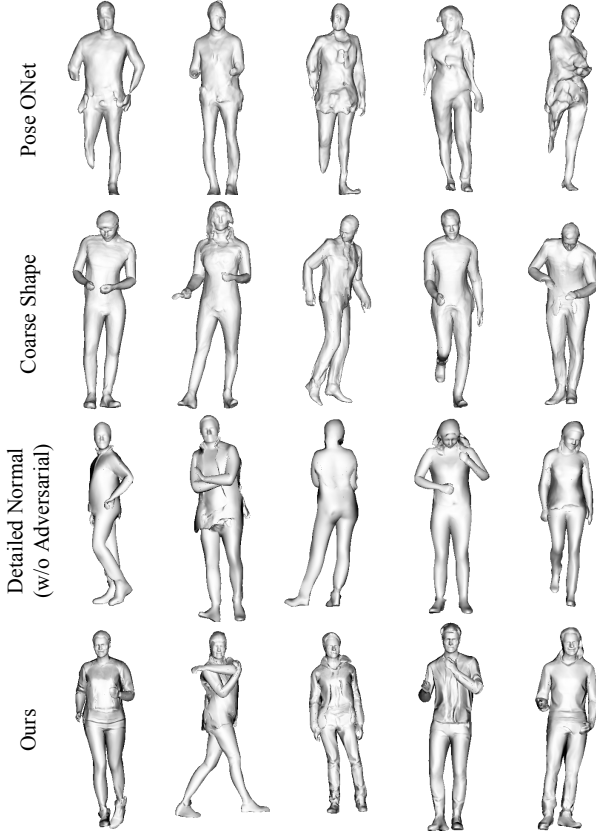
Method	Pred-to-Scan ↓	Scan-to-Pred ↓
SMPLicit [14]	N/A	0.0240
NPMs [47]	0.0156	0.0215
Ours	**0.0134**	**0.0123**

Table 2. **Fitting Comparison.** We report the distance between the target scan and 3D shapes fit by SOTA methods and ours. Pred-to-Scan metric does not apply to multi-layer surfaces from SMPLicit.

Figure 9. **Generation Comparison.** We show random samples from ablative baselines and our method. Without the adversarial loss, the generated shapes appear either bumpy (*Coarse Shape*) or over-smooth (*Detailed Normal w/o Adv.*).
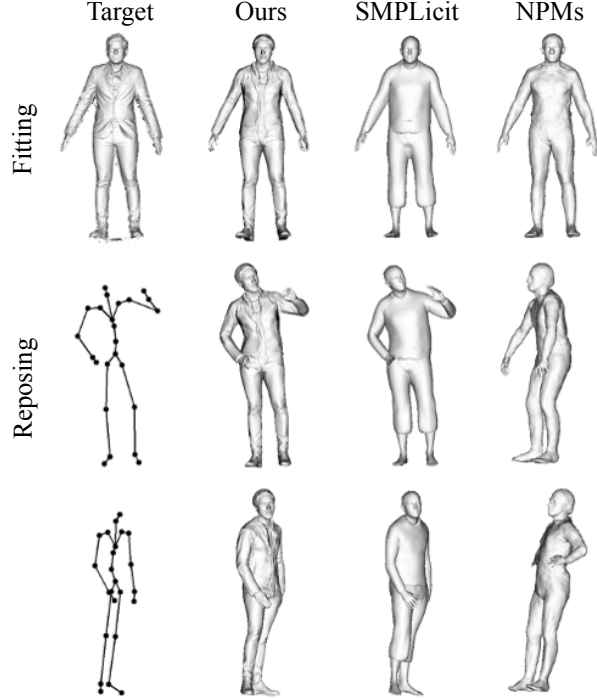

Figure 10. **Fitting and Reposing.** We compare model fitting results on the SIZER dataset with SMPLicit [14] and NPMs [47]. We also show the fitted shapes reposed into target poses. As NPMs do not allow specifying target pose, random poses are shown.

Reposing Scans: During fitting we also recover skinning weights. This enables reposing of the shape as demonstrated in Fig. 10 bottom.

5. Conclusion

We propose gDNA, a generative model of 3D clothed humans that can produce a large variety of clothed people with detailed wrinkles and explicit pose control. Using implicit multi-subject forward skinning enables learning from only a few posed scans per subject. To model the stochastic details of garments, we exploit a 2D adversarial loss to update a 3D normal field. We demonstrate that gDNA can be used in various applications such as animation and 3D fitting, outperforming state-of-the-art methods.

Limitations: Learning loose clothing (e.g. skirts) from deformed observations remains challenging due to the topology ambiguity and the large pose-dependent non-linear cloth deformation. Please refer to Sup. Mat. for more discussions about limitations and societal impact.

References

[1] https://3dpeople.com/. 6

[2] https://renderpeople.com/. 6

[3] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3D point clouds. In *Proc. of the International Conf. on Machine learning (ICML)*, 2018. 2

[4] Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with Style: Detail-preserving pose-guided image synthesis with conditional StyleGAN. *ACM Transactions on Graphics*, 2021. 2

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

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

[7] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM Trans. on Graphics*, 2005. 2

[8] James F Blinn. Simulation of wrinkled surfaces. *ACM Trans. on Graphics*, 1978. 4

[9] Charles G Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation*, 1965. 5

[10] Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges. Category level object pose estimation via neural analysis-by-synthesis. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 2

[11] Xu Chen, Jie Song, and Otmar Hilliges. Unpaired pose guided human image generation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 2

[12] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 1, 2, 3, 4, 5

[13] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[14] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. SMPLicit: Topology-aware generative model for clothed people. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 6, 7, 8

[15] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[16] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W Taylor, and Joshua M Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

[17] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 2, 6

[19] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. StylePeople: A generative model of fullbody human avatars. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[20] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[21] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. ARCH++: Animation-ready clothed human reconstruction revisited. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[22] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2017. 3

[23] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total Capture: A 3D deformation model for tracking faces, hands, and bodies. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[24] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[26] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2014. 2

[28] Z. Lähner, D. Cremers, and T. Tung. DeepWrinkles: Accurate and realistic clothing modeling. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 3, 4

[29] Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler. A generative model for people in clothing. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2017. 2

[30] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 2

[31] Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, and Matthias Nießner. 4DComplete: Non-rigid motion estimation beyond the observable surface. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 6

[32] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural Actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. on Graphics*, 2021. 2

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. on Graphics*, 2015. 2

[34] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[35] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J. Black. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[36] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3, 6

[37] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J Black. The power of points for modeling humans in clothing. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

[38] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 4

[39] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *Proc. of the International Conf. on Machine learning (ICML)*, 2018. 6

[40] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 6

[41] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 4

[42] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3D representations from natural images. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2, 3

[43] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[44] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[45] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

[46] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. STAR: Sparse trained articulated human body regressor. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 2

[47] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Neural parametric models for 3D deformable shapes. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2, 3, 6, 7, 8

[48] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 5

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

[50] 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 *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[51] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[52] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 4

[53] Kripasindhu Sarkar, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. HumanGAN: A generative model of humans images. In *Proc. of the International Conf. on 3D Vision (3DV)*, 2021. 2

[54] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative radiance fields for 3D-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

[55] Ruizhi Shao, Hongwen Zhang, He Zhang, Yanpei Cao, Tao Yu, and Yebin Liu. DoubleField: Bridging the neural surface and radiance fields for high-fidelity human rendering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[56] Shih-Yang Su, Frank Yu, Michael Zollhoefer, and Helge Rhodin. A-NeRF: Surface-free human 3D pose refinement

via neural rendering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[57] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 6

[58] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-GIF: Neural generalized implicit functions for animating people in clothing. In *International Conference on Computer Vision (ICCV)*, October 2021. 1, 2, 4

[59] Lizhen Wang, Xiaochen Zhao, Tao Yu, Songtao Wang, and Yebin Liu. NormalGAN: Learning detailed 3D human from a single RGB-D image. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[60] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. MetaAvatar: Learning animatable clothed human models from few depth images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[61] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2

[62] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[63] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-NeRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[64] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[65] Ilya Zakharkin, Kirill Mazur, Artur Grigorev, and Victor Lempitsky. Point-based modeling of human clothing. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

[66] Yufeng Zheng, Victoria Fernández Abrevaya, Xu Chen, Marcel C Bühler, Michael J Black, and Otmar Hilliges. IMavatar: Implicit morphable head avatars from videos. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[67] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2021. 2