# AG3D: Learning to Generate 3D Avatars from 2D Image Collections



**Author(s):**
Dong, Zijian; Chen, Xu; Yang, Jinlong; Black, Michael J.; Hilliges, Otmar; Geiger, Andreas

**Publication date:**
2023

**Permanent link:**
https://doi.org/10.3929/ethz-b-000656339



**Originally published in:**
https://doi.org/10.1109/ICCV51070.2023.01370

# AG3D: Learning to Generate 3D Avatars from 2D Image Collections

Zijian Dong[1,2*] Xu Chen[1,3*] Jinlong Yang[3] Michael J. Black[3] Otmar Hilliges[1] Andreas Geiger[2]
[1]ETH Zürich, Department of Computer Science [2]University of Tübingen
[3]Max Planck Institute for Intelligent Systems, Tübingen

Figure 1: **Sampled 3D Human Appearance and Shape.** Our generative model is learned from an unstructured 2D image collection, yet synthesizes novel 3D humans with high-quality appearance and geometry, different identities and clothing styles including loose clothing such as dresses and skirts. Moreover, our generated 3D humans can be easily animated.

## Abstract

*While progress in 2D generative models of human appearance has been rapid, many applications require 3D avatars that can be animated and rendered. Unfortunately, most existing methods for learning generative models of 3D humans with diverse shape and appearance require 3D training data, which is limited and expensive to acquire. The key to progress is hence to learn generative models of 3D avatars from abundant unstructured 2D image collections. However, learning realistic and complete 3D appearance and geometry in this under-constrained setting remains challenging, especially in the presence of loose clothing such as dresses. In this paper, we propose a new adversarial generative model of realistic 3D people learned from 2D images. Our method captures shape and deformation of the body and loose clothing by adopting a holistic 3D generator and integrating an efficient, flexible, articulation module. To improve realism, we train our model using multiple discriminators while also integrating geometric cues in the form of predicted 2D normal maps. We experimentally find that our method outperforms previous 3D- and articulation-aware methods in terms of geometry and appearance. We validate the effectiveness of our model and the importance of each component via systematic ablation studies.*

[*]Equal contribution

## 1. Introduction

Generative models, like GANs [19], can be trained from large image collections, to produce photo-realistic images of objects [5, 29–31] and even clothed humans [2, 18, 20, 33, 34, 55]. The output, however, is only a 2D image and many applications require diverse, high-quality, virtual 3D avatars, with the ability to control poses and camera viewpoints, while ensuring 3D consistency. To enable the generation of 3D avatars, the research community has been studying generative models that can automatically produce 3D shapes of humans and/or clothing based on input parameters such as body pose and shape [9, 11, 38, 50]. Despite rapid progress, most existing methods do not yet consider texture and require accurate and clean 3D scans of humans for training, which are expensive to acquire and hence limited in quantity and diversity. In this paper, we develop a method that learns a generative model of 3D humans with texture from only a set of unstructured 2D images of various people in different poses wearing diverse clothing; that is, we learn a generative 3D human model from data that is ubiquitous on the Internet. See Fig. 1.

Learning to generate 3D shapes and textures of articulated humans from such unstructured image data is a highly under-constrained problem, as each training instance has a different shape and appearance and is observed only once from a particular viewpoint and in a particular pose. Recent

progress in 3D-aware GANs [6, 22, 48] shows impressive results in learning 3D geometry and appearance of rigid objects from 2D image collections. However, since humans are highly articulated and have more degrees of freedom to model, such methods struggle to generate realistic humans. By modeling articulation, recent work [4, 47] demonstrates the feasibility of learning articulated humans from image collections, allowing the generation of human shapes and images in desired poses, but only in limited quality and resolution. Recently, EVA3D [23] achieves higher resolution by representing humans as a composition of multiple parts, each of which is generated by a small network. However, there is still a noticeable gap between the generated and real humans in terms of appearance and, in particular, geometry. Additionally, the compositional design precludes modeling loose clothing that is not associated with a single body part, such as dresses shown in Fig. 5c.

In this paper, we contribute a new method for learning 3D human generation from 2D image collections, which yields state-of-the-art image and geometric quality and naturally models loose clothing. Instead of representing humans with separate body parts as in EVA3D [23], we adopt a simple monolithic approach that is able to model the human body as well as loose clothing, while adding multiple discriminators that increase the fidelity of perceptually important regions and improve geometric details.

**Holistic 3D Generation and Deformation:** To achieve the goal of high image quality while flexibly handling loose clothing, we propose a novel generator design. We model 3D humans holistically in a canonical space using a monolithic 3D generator and an efficient tri-plane representation [6]. To attain high-quality images it is critically important to enable fast volume rendering. To this end, we adapt the efficient articulation module, Fast-SNARF [8], to our generative setting and further accelerate rendering via empty-space skipping, informed by a coarse human body prior. Our articulation module is more flexible than prior methods that base deformations of the clothed body on SMPL [38], enabling it to faithfully model deformations for points that are far away from the body.

**Modular 2D Discriminators:** We further propose multiple discriminators to improve geometric detail as well as the perceptually-important face region as we found that a single adversarial loss on rendered images is insufficient to recover meaningful 3D geometry in such a highly under-constrained setting. Motivated by the recent success of methods [25, 69] that exploit monocular normal cues [54, 65] for the task of 3D reconstruction, we explore the utility of normal information for guiding 3D geometry in the generative setting. More specifically, we discriminate normal maps rendered from our generative 3D model against 2D normal maps obtained from off-the-shelf monocular estimators [54] applied to 2D images of human subjects. We demonstrate that this additional normal supervision serves as useful and complementary guidance, significantly improving the quality of the generated 3D shapes. Furthermore, we apply separate face discriminators on both the image and normal branch to encourage more realistic face generation.

We experimentally find that our method outperforms previous 3D- and articulation-aware methods by a large margin in terms of both geometry and texture quality, quantitatively (Table 1), qualitatively (Fig. 5) and through a perceptual study (Fig. 4). In summary, we contribute (i) a generative model of articulated 3D humans with SotA appearance and geometry, (ii) a new generator that is efficient and can generate and deform loose clothing, and (iii) several, specialized discriminators that significantly improve visual and geometric fidelity. Code and models are available at https://zj-dong.github.io/AG3D/.

## 2. Related Work

**3D-aware Generative Adversarial Networks:** Generative adversarial networks (GANs) [19] achieve photorealistic image generation [5, 29–31] and show impressive results on the task of 2D human image synthesis [2, 18, 20, 33, 34, 55]. However, these 2D methods cannot guarantee 3D consistency [10, 33, 39] and do not provide 3D geometry. Several methods extend 2D GANs to 3D by combining them with 3D representations, including 3D voxels [44, 64], meshes [36, 58] and point clouds [1, 35]. Recently, many methods represent 3D objects as neural implicit functions [42, 46, 51, 61, 68]. Such representations are also used for 3D-aware generative image synthesis [6, 7, 22, 45, 48, 56, 57]. StyleSDF [48] replaces density with an SDF for better geometry generation and SotA methods like EG3D [6] introduce a tri-plane representation to improve rendering efficiency. Nevertheless, it is not straightforward to extend these methods to non-rigid articulated objects such as humans. In this paper, we propose a 3D- and articulation-aware generative model for clothed humans.

**3D Human Models:** Parametric 3D human body models [3, 28, 38, 49, 66] are able to synthesize minimally clothed human shapes by deforming a template mesh. Extending these mesh models to generate 3D clothing or clothed humans is challenging [40]. In the case of meshes, the geometry is restricted to a fixed mesh topology and large deviations from the template mesh are hard to model. To overcome this limitation, methods such as SMPLicit [11] and gDNA [9] propose to build a 3D generative model of clothed humans based on implicit surface representations, either by adding an implicit garment to the SMPL body or by learning a multi-subject implicit representation with corresponding skinning weights. The main problem of all aforementioned approaches, however, is their reliance on

3D ground truth: their training requires a large number of complete and registered 3D scans of clothed humans in different poses, which are typically acquired using expensive 3D body scanners. Several methods [13, 17, 27, 32, 53, 62, 63, 73] combine NeRF with human priors to enable 3D human reconstruction from multi-view data or even monocular videos. Nevertheless, their proposed human representations can only be utilized to represent human avatars for a single subject, wearing a specific garment.

Recently, some methods have been proposed to learn generative models of 3D human appearance from a collection of 2D images. ENARF-GAN [47] and GNARF [4] leverage 3D human models to learn a 3D human GAN, but they still fail to produce high-quality human images. The concurrent work EVA3D [23] achieves high-resolution human image generation by introducing a compositional part-based human representation. However, none of these methods including other concurrent arXiv papers [26, 67, 72] are able to generate and deform loose clothing, and their geometry typically suffers from noisy artifacts. In contrast, our method generates both high-quality geometry and appearance of diverse 3D-clothed humans even wearing loose clothing, with full control over the pose and appearance. We empirically demonstrate the benefits of our method over the more complex EVA3D model [23] in Section 4.2. A comparison to the recent arXiv papers [26, 67, 72] is not possible since the models and code have not been released.

**3D Shape from 2D Normals:** Several methods predict normals from a single image for general objects [14–16, 24] or clothed humans [54, 65]. These predicted 2D normal cues can be exploited to guide 3D reconstruction using neural field representations. For instance, MonoSDF [69] leverages predicted normals to improve 3D object reconstruction from sparse views. Similarly, SelfRecon [25] uses a normal reconstruction loss to reconstruct a human avatar from a monocular video. PIFuHD [54] and ICON [65] predict normal maps as additional input to support single-view 3D human reconstruction. In this work, we demonstrate that monocular 2D normal cues are useful for learning a generative 3D model of articulated objects.

# 3. Method

Given a large 2D image collection, our goal is to learn a generative model of diverse 3D human avatars with realistic appearance and geometry, while enabling control over pose and identity. An overview of our method is shown in Fig. 2.

In this section, we first introduce an efficient and articulation-aware 3D human generator (Section 3.1) which generates the appearance and shape in canonical space and uses a deformation module to warp into posed space via a learned continuous deformation field. Next, we describe our rendering module, which is accelerated by an empty space skipping strategy that leverages the SMPL body prior. To enable fast training, we use a super-resolution module to lift feature maps to high-resolution images.

We optimize the generator using a combination of adversarial losses (Section 3.2) and an Eikonal loss [21]. While prior work uses a single discriminator formulation, we show that employing several, specialized discriminators improves visual fidelity. To this end, we define discriminators that reason at the level of the whole body and locally at the face region, respectively. We additionally introduce an adversarial normal loss, which significantly improves the quality of the generated geometry.

## 3.1. Holistic 3D Avatar Generator

**Canonical Generator:** Given a latent vector $\mathbf{z} \in \mathbb{R}^{n_z}$ and pose parameters $\mathbf{p} \in \mathbb{R}^{n_p}$, our method first generates 3D human appearance and shape in canonical space (see Fig. 2). Here, we leverage pose conditioning to model local pose-dependent deformations of clothes and bodies. For efficient rendering, the canonical generator builds on the tri-plane representation proposed in ConvONet [52] and EG3D [6] to model 3D features. These are then decoded by an MLP to predict the canonical shape and appearance in 3D space.

We represent geometry using a signed distance field (SDF). Since existing fashion datasets are imbalanced and contain mostly frontal views, learning correct 3D geometry from such datasets is difficult. Following [23], we exploit a human shape prior in the form of a canonical SMPL body [38]. Specifically, for every query point $\mathbf{x}$ in canonical space, we predict an SDF offset $\Delta d(\mathbf{x}, \mathbf{z}, \mathbf{p})$ from the base shape to model details such as hair and clothing. The SDF value $d(\mathbf{x}, \mathbf{z}, \mathbf{p})$ is then calculated as

$$d = d(\mathbf{x}, \mathbf{z}, \mathbf{p}) = d_{\text{SMPL}}(\mathbf{x}) + \Delta d(\mathbf{x}, \mathbf{z}, \mathbf{p}), \tag{1}$$

where $d_{\text{SMPL}}(\mathbf{x})$ is the signed distance to the canonical SMPL surface. Unlike [23], to compute $d_{\text{SMPL}}(\mathbf{x})$ efficiently, we represent the SDF as a low-resolution voxel grid ($128 \times 128 \times 32$), where the value of every grid point is the (pre-computed) distance to the SMPL mesh. We then query $d_{\text{SMPL}}(\mathbf{x})$ by trilinearly interpolating the SDF voxel grid.

We also compute normals in canonical space. The normal $\mathbf{n}$ at a certain canonical point $\mathbf{x}$ is computed as the spatial gradient of the signed distance function at that point:

$$\mathbf{n} = \nabla_{\mathbf{x}} d(\mathbf{x}, \mathbf{z}, \mathbf{p}). \tag{2}$$

The canonical appearance is represented by a 3D texture field $\mathbf{c}$. We also predict features $\mathbf{f}$ that are used to guide the super-resolution module (described later). We denote the entire mapping from 3D point $\mathbf{x}$, latent vector $\mathbf{z}$ and pose condition $\mathbf{p}$ to SDF $d$, normal $\mathbf{n}$, color $\mathbf{c}$ and color features

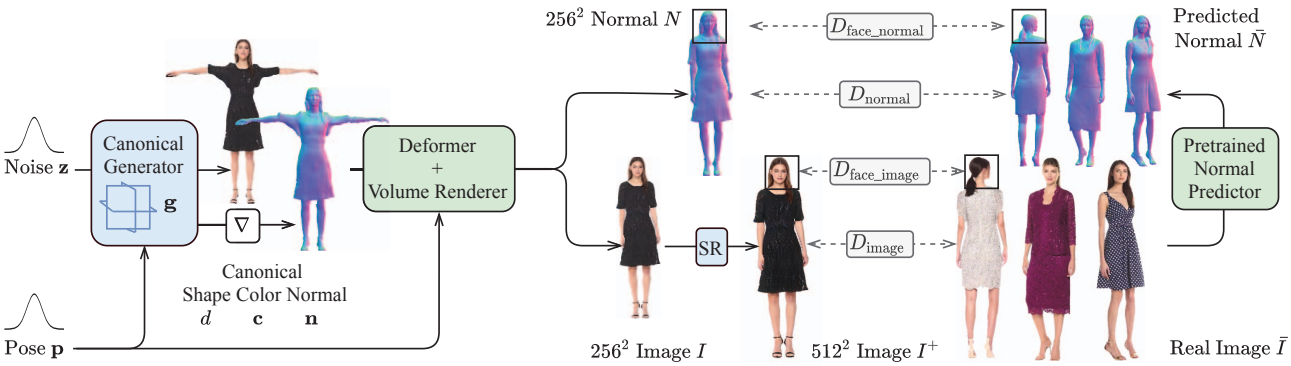



Figure 2: **Method Overview.** *Holistic 3D Human Generation:* Given a latent vector $\mathbf{z}$, our method generates human shape $d$ and appearance $\mathbf{c}$ in canonical space. In addition, we compute surface normals $\mathbf{n}$ via the spatial derivatives of the canonical shape, which is represented as an SDF. These canonical representations are then posed into the target body pose $\mathbf{p}$ via a flexible deformer and then rendered from the target viewpoint. The rendered images are further super-resolved by $2\times$. *Adversarial Training:* We optimize the generator and the super-resolution module using multiple discriminators. In addition to an image discriminator operating on the images, we improve geometry by introducing a normal discriminator that compares our rendered normal maps with the normals of real images predicted by an off-the-shelf normal estimator. To further improve the quality of the perceptually important face region, we add normal and image discriminators for the face region.

$\mathbf{f}$ in *canonical space* as follows:

$$\mathbf{g} : \mathbb{R}^3 \times \mathbb{R}^{n_z} \times \mathbb{R}^{n_p} \rightarrow \mathbb{R} \times \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^{n_f} \qquad (3)$$
$$(\mathbf{x}, \mathbf{z}, \mathbf{p}) \mapsto (d, \mathbf{n}, \mathbf{c}, \mathbf{f}).$$

**Deformer:** To enable animation and to learn from posed images, we require the appearance and 3D shape in *posed space*. In the following we denote quantities in posed space as $(\cdot)'$. Given the bone transformation matrix $\mathbf{B}_i$ for joint $i \in \{1, ..., n_b\}$, a canonical point $\mathbf{x}$ is transformed into its deformed version $\mathbf{x}'$ via

$$\mathbf{x}' = \sum_{i=1}^{n_b} w_i \, \mathbf{B}_i \, \mathbf{x}. \qquad (4)$$

Here, the canonical LBS weight field $\mathbf{w} : \mathbb{R}^3 \rightarrow \mathbb{R}^{n_b}$, with $\mathbf{x} \mapsto (w_1, ..., w_{n_b})$ where $n_b$ is the number of joints, weights the influence of each bone's $\mathbf{B}_i$ transformation on the deformed location $\mathbf{x}'$. This weight field is represented by a low-resolution voxel grid ($64 \times 64 \times 16$). We found that a fixed skinning weights grid, without adaption or varying resolution is sufficient for our task because the variation of the shape parameters estimated from training images is small and our articulation module only requires pose-independent canonical skinning weights. The normal at the deformed point $\mathbf{x}'$ is given by

$$\mathbf{n}' = \frac{(\sum_{i=1}^{n_b} w_i \, \mathbf{R}_i)^{-T} \mathbf{n}}{\|(\sum_{i=1}^{n_b} w_i \, \mathbf{R}_i)^{-T} \mathbf{n}\|} \qquad (5)$$

where $\mathbf{R}_i$ is the rotation component of $\mathbf{B}_i$ [60].

We leverage Fast-SNARF [8] to efficiently warp points *backwards* from posed space $\mathbf{x}'$ to canonical space $\mathbf{x}$ via efficient iterative root finding [8]. The SDF value $d'$, color $\mathbf{c}'$ and feature $\mathbf{f}'$ at the deformed point are obtained by evaluating the generator at the corresponding $\mathbf{x}$. In contrast to [8], which focuses on reconstruction tasks and learns skinning weights on the fly, we constrain the notoriously difficult adversarial training by averaging the skinning weights of the nearest vertices on the canonical SMPL mesh.

**Volume Renderer:** To render a pixel, we follow [43] and cast ray $\mathbf{r}'$ from the camera center $\mathbf{o}'$ along its view direction $\mathbf{v}'$. We use two-pass importance sampling of $M$ points in posed space $\mathbf{x}'_i = \mathbf{o}' + t_i \mathbf{v}'$ and predict their SDF values $d'_i$, colors $\mathbf{c}'_i$, color features $\mathbf{f}'_i$ and normals $\mathbf{n}'_i$. We convert SDF values $d'_i$ to densities $\sigma'_i$ via the method of StyleSDF [48].

The color of each pixel in the rendered image $I$ is computed via numerical integration [43]:

$$I(\mathbf{r}) = \sum_{i=1}^{M} \alpha_i \prod_{i<j} (1 - \alpha_j) \mathbf{c}'_i \quad \alpha_i = 1 - \exp(\sigma'_i \delta_i) \qquad (6)$$

where $\delta_i$ is the distance between samples. 3D normals $N(\mathbf{r})$ and feature vectors $F(\mathbf{r})$ are rendered accordingly.

To accelerate rendering and to reduce memory, we take advantage of the geometric prior of the SMPL model and define the region within a predefined distance threshold to the SMPL surface as the occupied region. For points sampled outside of this region, we set the density to zero.

**Super Resolution:** Although the SMPL-guided volume rendering is more efficient than previous approaches, it is still slow and requires a large amount of memory to render at high resolution. Therefore, we perform volume rendering at a sufficient resolution ($256^2$ pixels) to guarantee good rendering of the normal image $N$ and rely on a super-

(a) View Control (b) Pose Control (c) Interpolation

Figure 3: **Qualitative Results: 3D Human Generation.** We generate 3D human appearance and shape, and render the resulting 3D representations using different body poses and from different viewpoints. In addition, we show virtual people generated by interpolating between latent codes. Overall, our synthesized humans exhibit reasonable appearance and geometric quality, remain consistent across different poses and views, and smoothly interpolate when varying the latent code $\mathbf{z}$.

resolution module [6] to upsample the image feature map $F$ and color $I$ to the final image $I^+$ of size $512^2$ pixels.

### 3.2. Training

We train our model on a large dataset of 2D images using adversarial training, leveraging a combination of multiple discriminators and an Eikonal loss.

**Image Discriminator:** The first discriminator $D_{\text{image}}$ compares full images generated by our method to real images. Following EG3D [6], we apply the discriminator at both resolutions: We upsample our low resolution rendering $I$, concatenate it with the super-resolved image $I^+$, and feed it to a StyleGANv2 [30] discriminator. For real images $\bar{I}$, we downsample and re-upsample them, and concatenate the results with the original image as input to the discriminator.

**Face Discriminator:** We observe that the generated face region suffers from artifacts due to the low resolution of faces within the full-body image. Motivated by 2D human GANs [20], we add a small face discriminator $D_{\text{face\_image}}$. Based on estimated SMPL head keypoints, we crop the head regions of our high resolution output $I^+$ and real data $\bar{I}$ and feed them into the discriminator for comparison.

**Normal Discriminator:** A central goal of our work is to attain geometrically correct 3D avatars. To achieve this, we propose to use geometric cues present in 2D normal maps to guide the adversarial learning towards meaningful 3D geometry. To this end, we use an additional normal discriminator $D_{\text{normal}}$. This normal discriminator compares the predicted 2D normal maps $N$ to 2D normal maps $\bar{N}$ of real images $\bar{I}$ predicted by the 2D normal estimator from PIFuHD [54]. Analogously to the image branch, we use an additional discriminator $D_{\text{face\_normal}}$ to further enhance the geometric fidelity of the generated faces. We refer the reader to the Sup. Mat. for implementation details.

**Eikonal Loss:** To regularize the learned SDFs, we apply an Eikonal loss [21] $\mathcal{L}_{\text{eik}} = \mathbb{E}_{\mathbf{x}_i}\left(\|\nabla(\Delta d(\mathbf{x}_i))\| - 1\right)^2$ to the canonical correspondences $\{\mathbf{x}_i\}$ of sampled points $\{\mathbf{x}'_i\}$.

**Training:** We train our generator and discriminators jointly using the non-saturating GAN objective with R1-regularization [41] and an Eikonal loss. We estimate the 3D human pose for each training image using an off-the-shelf pose detector [71]. The collection of the estimated poses serves as an approximation of the pose distribution. During training, we draw random samples from the pose collection. Please refer to the Sup. Mat. for more training details.

## 4. Experiments

In our experiments, we first demonstrate the quality of generated samples and then compare our method to other SotA baselines. In addition, we provide an ablation study to investigate the importance of each component in our model.

**Datasets:**[*] *DeepFashion* [37] contains an unstructured collection of fashion images of different subjects wearing various types of clothing. We use the curated subset with 8k images from [23] as our training data. *UBCFashion* [70] con-

[*]Disclaimer: Standard fashion datasets lack diversity, see Section 4.4.

| Method | DeepFashion | | | UBCFashion | | |
|---|---|---|---|---|---|---|
| | FID↓ | $FID_{normal}$ ↓ | $FID_{face}$ ↓ | FID↓ | $FID_{normal}$ ↓ | $FID_{face}$ ↓ |
| EG3D | 26.38* | - | - | 23.95* | - | - |
| StyleSDF | 92.40* | - | - | 18.52* | - | - |
| ENARF-GAN | 77.03* | - | - | - | - | - |
| EVA3D | 15.91* | - | - | 12.61* | - | - |
| EVA3D (public) | 20.45 | 30.81 | 17.21 | 19.81 | 49.29 | 54.42 |
| Ours | **10.93** | **20.38** | **14.79** | **11.04** | **18.79** | **15.83** |

Table 1: **Quantitative Comparison with SotA Methods**. We evaluate FID of full images, cropped face images and normal maps generated by our method and the SotA method EVA3D (public) [23] using their released trained models. For reference, we also report quantitative results from the EVA3D paper above the separation line (marked by *).

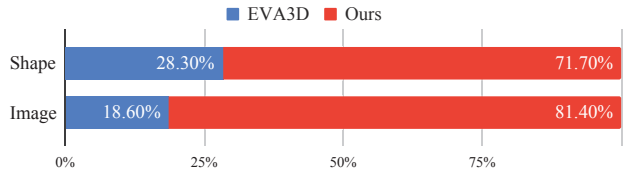



Figure 4: **User Preference**. We conduct a perceptual study with approximately 4000 samples and report how often participants preferred shapes and images generated by our method or those generated by EVA3D [23].

tains 500 sequences of fashion videos with subjects wearing loose clothing such as skirts. Following EVA3D [23], we treat these videos as individual images without assuming temporal information. Pre-processing details can be found in the Sup. Mat.

**Metrics:** We measure the diversity and quality of generated images using the *Fréchet Inception Distance (FID)* between 50k generated images and available real images, denoted by $FID_{image}$. To measure the generated face quality, we report an additional FID specifically for the face region, denoted by $FID_{face}$. Furthermore, we evaluate the quality of the synthesized geometry by computing the FID between our rendered normals and pseudo-GT normal maps predicted by [54] ($FID_{normal}$). We use an inception network [59] pretrained on ImageNet [12] for all FID computation. In addition, we conduct a *Perceptual User Study* among 50 participants with 4000 samples and report how often participants preferred a particular method over ours.

**Baselines:** We compare our method to four baseline methods: EG3D [6], StyleSDF [48], ENARF-GAN [47] and EVA3D [23]. EG3D and StyleSDF are SotA , 3D-aware, generative models of rigid objects. For comparison, these two methods are trained on the aforementioned human datasets. Since these two methods do not model articulation, they have to learn it implicitly. ENARF-GAN and EVA3D additionally model articulation for 3D human generation. The quantitative FID results of all baseline methods are directly taken from the experiment in EVA3D [6]. In addition, we evaluate $FID_{face}$ and $FID_{normal}$ on EVA3D based on their released code and trained model weights.

### 4.1. Quality of 3D Human Generation

We show our qualitative results in Fig. 3. More results can be found in the Sup. Mat. Overall, our method generates realistic human images with faithful details such as clothing patterns, face and hair, and meaningful 3D geometry even with fine structures such as hair and shoe heels. Our method further enables control over the generation as follows.

**View Control:** As shown in Fig. 3a, by learning humans in 3D space, our method can generate 3D-consistent high-quality images and geometry from varied viewpoints.

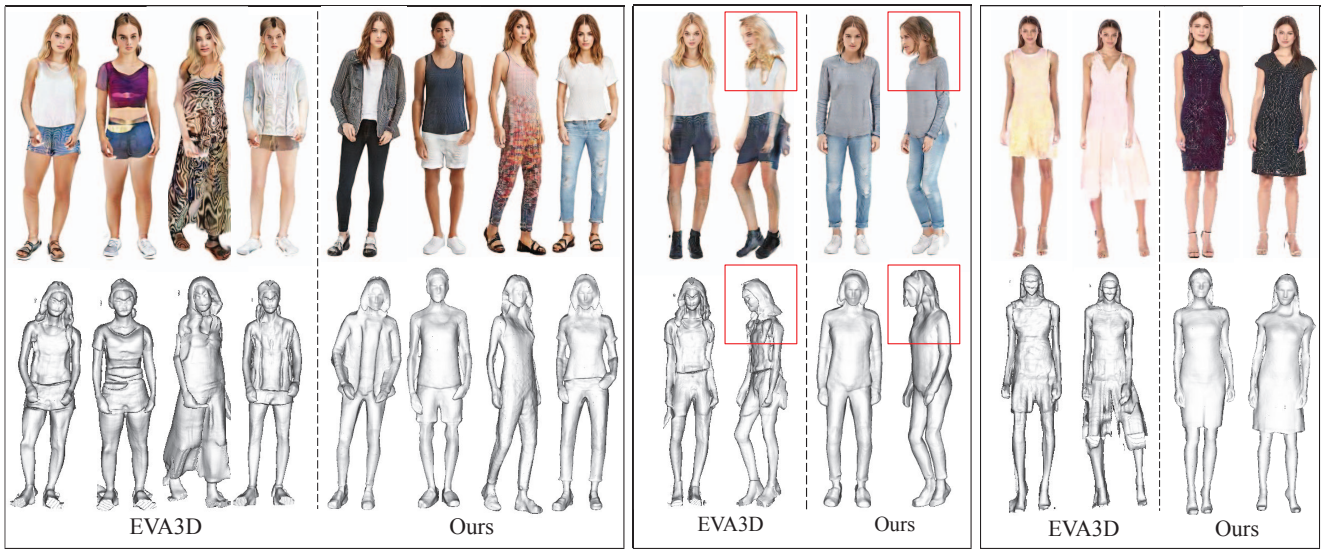
**Pose Control:** The generated 3D humans can also be reposed into unseen poses as shown in Fig. 3b. The images and geometry in different poses are consistent due to the explicit model of human articulation.

**Interpolation:** Our method learns a smooth latent space of 3D human shape and appearance. As shown in Fig. 3c, our method yields smooth transitions of appearance and shape even when interpolating the latent codes of two subjects with different gender and clothing styles.

### 4.2. Comparison to SotA

Table 1 summarizes our quantitative comparisons on both DeepFashion and UBCFashion datasets. Since the SotA method EVA3D [23] outperforms other baselines by a significant margin, we focus our discussion on the comparison with EVA3D only. More comparisons with other baselines can be found in the Sup. Mat.

**Image Quality:** Our method achieves better quantitative results than EVA3D in terms of $FID_{image}$ and $FID_{face}$ on both datasets. This improvement is confirmed by our user study in Fig. 4. Notably, in 81.4% of the cases, participants consider our generated images to be more realistic than EVA3D. A qualitative comparison is shown in Fig. 5a. Our method generates overall sharper images with more de-




(a) Overall Quality (b) Novel Views (c) Loose Clothing

Figure 5: **Qualitative Comparison to EVA3D.** We show random samples of our method and the SotA method EVA3D [23]. Our method achieves better image and shape quality, degrades more gracefully at side views, and better models loose clothing.

| Method | FID ↓ | $FID_{normal}$ ↓ | $FID_{face}$ ↓ |
|---|---|---|---|
| Ours | **10.93** | **20.38** | 14.79 |
| w/o normal GAN | 11.15 | 32.17 | **14.35** |
| w/o face GAN | 11.71 | 23.96 | 20.88 |

Table 2: **Ablation.** We compare our method and ablated baselines in which we remove individual discriminators.

tails due to the use of a holistic 3D generator, and synthesizes more realistic faces due to our face discriminators.

Our improvements are particularly pronounced when considering side views. As shown in Fig. 5b, our method generates sharp and meaningful images also from the side where EVA3D's image quality significantly degrades. This is a consequence of EVA3D's pose-guided sampling strategy. As discussed in their paper, EVA3D had to increase the dataset's frontal bias during training to achieve reasonable geometry and face quality. We hypothesize that this requirement is due to the limited capacity of the lightweight part models. As a consequence, EVA3D overfits more to frontal views and generalizes less well. In contrast, our efficient articulation and rendering modules allow us to exploit a single holistic generator and our face and normal discriminators enable us to directly sample from the data distribution which leads to better generalization.

Interestingly, the (non-animatable) generative model EG3D achieves reasonable FID despite not modeling articulation. This is because the FID evaluation only considers training poses and views, see also Sup. Mat.

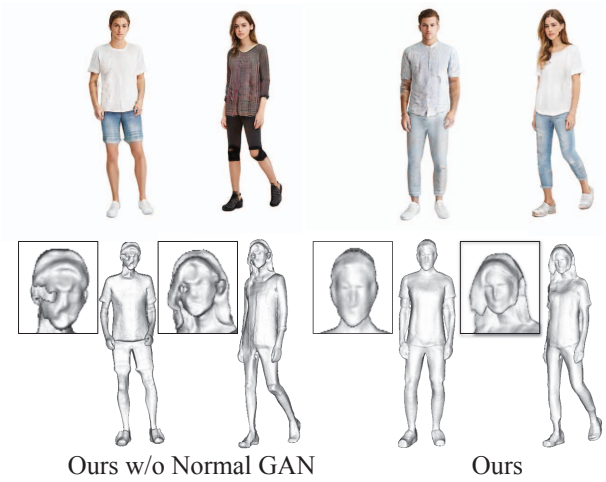



Figure 6: **Ablation of the Normal Discriminator.** Our normal discriminator effectively improves the generated geometry while preserving appearance quality.

**Geometry:** Our method yields significantly better geometry compared to EVA3D, as evidenced by the improvement in $FID_{normal}$ in Table 1 and the perceptual study in Fig. 4. Based on our qualitative results, our geometry is more realistic and detailed, in particular on faces. In contrast, noise and holes can be observed around the shapes generated by EVA3D, despite their surface representation and regularization. We attribute this improvement to our normal discriminators, which provide strong geometric cues.

**Loose Clothing:** As shown in Fig. 5c, our method outperforms EVA3D in modeling loose clothing. Due to its com-

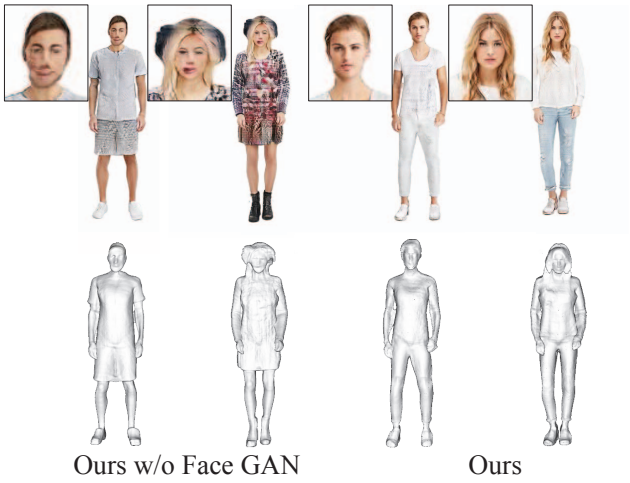



Figure 7: **Ablation of the Face Discriminator.** Our face discriminator effectively improves generated face quality.

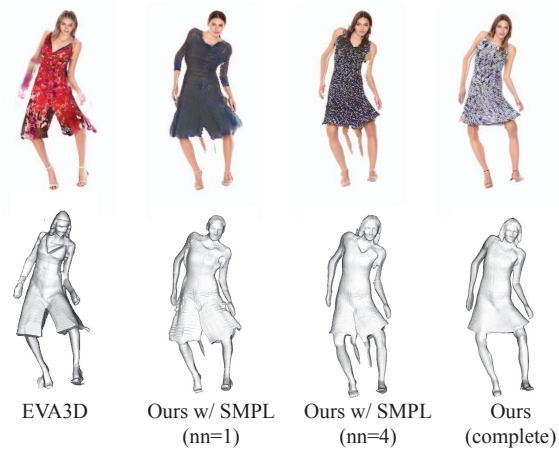



Figure 8: **Ablation Study of the Deformer.** Results with loose clothing in novel poses, generated by EVA3D and our method with different choices for the articulation module.

positional nature, EVA3D is prone to generating artifacts between the legs. In contrast, our holistic representation generates loose clothing without discontinuity artifacts.

**Efficiency:** Our method is more efficient than EVA3D in rendering images and normal maps with the same image and ray sampling resolution. For an image resolution of $256^2$ and with 28 sample points per ray, our method renders normals and images together at 10.5 FPS while EVA3D runs at 5.5 FPS. With a 2D super-resolution module, our method is more than three times faster than EVA3D when rendering images of $512^2$ (9.5FPS vs 3FPS), while achieving better performance in terms of geometry and appearance. More details can be found in Sup. Mat.

### 4.3. Ablation Study

**Normal Discriminator:** Our normal discriminator serves an important role in improving the realism of the generated geometry. Comparing our model to ablated versions where we remove the normal discriminator (w/o normal GAN), we observe a significant $\text{FID}_{\text{normal}}$ improvement (see Table 2). As shown by the qualitative results in Fig. 6, the normal GAN effectively removes holes and noise on the generated surface, especially on faces, while preserving image quality.

**Face Discriminator:** As expected, when training without face discriminators, we observe a large drop in $\text{FID}_{\text{face}}$ (see Table 2). Given that faces are low resolution and hard to generate, our adversarial loss on faces forces the generator to focus on this local region and thus achieves a more realistic generation as shown in Fig. 7.

**Deformer:** To test the importance of our Fast-SNARF based deformation module, we compare our model to a SMPL nearest-neighbor-based deformer (denoted by *Ours w/ SMPL*), where points are deformed based on the skinning weights of their $K$ nearest SMPL vertices in posed space. As shown in Fig. 8, only our method can deform the skirts without splitting them. This is due to our articulation module being able to derive meaningful deformations for points far away from the SMPL surface. In contrast, our ablated baselines, with different choices of $K$, suffer from discontinuity artifacts as they only provide meaningful deformation at regions close to the SMPL surface. Similar artifacts can be observed in EVA3D's results, which we hypothesize stem from their part-based model.

### 4.4. Limitations

Since each training instance is observed only in one pose, the association of pixels to body parts cannot be uniquely determined. Hence, our model sometimes generates wrong clothing patterns under arms, or at hands close to the torso. Future work should investigate techniques to guide association, such as 2D correspondence predictions.

Moreover, samples from generative models reflect the biases present in their training data. The 2D image collections that we use for training focus on fashion images and lack diversity in skin tone, body shape, and age. Our work should be viewed as a methodological proof of concept and contains no mechanisms to combat these biases. To avoid biases, future research and deployable systems should i) be trained on more diverse data or ii) use explicit de-biasing. Further limitations are discussed in the Sup. Mat.

## 5. Conclusion

In this paper we contribute a new controllable generative 3D human model that is learned from unstructured 2D image collections alone and does not leverage any 3D super-

vision. Our model synthesizes high-quality 3D avatars with fine geometric details and models loose clothing more naturally than prior work. We achieve this through a new generator design that combines a holistic 3D generator with an efficient and flexible articulation module. Furthermore, we show that employing several, specialized discriminators that operate on the different branches (RGB and normals) and regions (fully body and facial region), leads to higher visual fidelity. We experimentally demonstrate that our method advances the state-of-the-art in learning a 3D human generator from 2D image collections in terms of both appearance and geometry and that it is the first generative model of 3D humans that can handle the deformations of free-flowing, loose, garments and long hair.

**Acknowledgements:** Zijian Dong was supported by the BMWi in the project KI Delta Learning (project number 19A19013O) and the ERC Starting Grant LEGO-3D (850533). Andreas Geiger is a member of the Machine Learning Cluster of Excellence, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. Xu Chen was supported by the Max Planck ETH Center for Learning Systems. This project was supported by the ERC Starting Grant LEGO-3D (850533), the BMWi project KI Delta Learning (project number 19A19013O) and the DFG EXC number 2064/1 - project number 390727645. We thank Kashyap Chitta, Katja Schwarz, Takeru Miyato and Seyedmorteza Sadat for their feedback, and Tsvetelina Alexiadis for her help with the user study.

**Disclosure:** MJB has received research gift funds from Adobe, Intel, Nvidia, Meta/Facebook, and Amazon. MJB has financial interests in Amazon, Datagen Technologies, and Meshcapade GmbH. MJB's research was performed solely at, and funded solely by, the Max Planck.

## References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3D point clouds. In *International Conference on Machine Learning*, pages 40–49, 2018. 2

[2] 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*, 40(6):1–11, 2021. 1, 2

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 2

[4] Alexander W. Bergman, Petr Kellnhofer, Wang Yifan, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. In *NeurIPS*, 2022. 2, 3

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 1, 2

[6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2, 3, 5, 6

[7] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 2

[8] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J. Black, and Otmar Hilliges. Fast-SNARF: A fast deformer for articulated neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15, 2023. 2, 4

[9] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J. Black, Andreas Geiger, and Otmar Hilliges. gDNA: Towards generative detailed neural avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20427–20437, 2022. 1, 2

[10] Xu Chen, Jie Song, and Otmar Hilliges. Unpaired pose guided human image generation. In *Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 2

[11] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. SMPLicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11875–11885, 2021. 1, 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 6

[13] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20470–20480, 2022. 3

[14] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3D scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 3

[15] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2650–2658, 2015.

[16] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, 27, 2014. 3

[17] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black,

and Timo Bolkart. SCARF: Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, SA'22, page 9, Dec. 2022. 3

[18] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. StyleGAN-Human: A data-centric odyssey of human generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, 2022. 1, 2

[19] 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*, 63(11):139–144, 2020. 1, 2

[20] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. StylePeople: A generative model of fullbody human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5151–5160, 2021. 1, 2, 5

[21] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of the 37th International Conference on Machine Learning,ICML 2020, 13-18 July 2020, Virtual Event*, volume 119, 2020. 3, 5

[22] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. StyleNeRF: A style-based 3D aware generator for high-resolution image synthesis. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022. 2

[23] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. EVA3D: Compositional 3D human generation from 2D image collections. In *ICLR*, 2023. 2, 3, 5, 6, 7

[24] Haiwen Huang, Andreas Geiger, and Dan Zhang. GOOD: Exploring geometric cues for detecting objects in an open world. In *ICLR*, 2023. 3

[25] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. SelfRecon: Self reconstruction your digital avatar from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3

[26] Suyi Jiang, Haoran Jiang, Ziyu Wang, Haimin Luo, Wenzheng Chen, and Lan Xu. HumanGen: Generating human radiance fields with explicit priors. In *CVPR*, 2023. 3

[27] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. NeuMan: Neural human radiance field from a single video. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 402–418. Springer, 2022. 3

[28] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 2

[29] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 1, 2

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 5

[31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8107–8116, 2020. 1, 2

[32] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34:24741–24752, 2021. 3

[33] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 853–862, 2017. 1, 2

[34] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. TryOnGAN: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2021)*, 40(4), 2021. 1, 2

[35] Ruihui Li, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. PU-GAN: A point cloud upsampling adversarial network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7203–7212, 2019. 2

[36] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3D controllable image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2020. 2

[37] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016. 5

[38] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1, 2, 3

[39] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[40] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 2

[41] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International Conference on Machine Learning*, 2018. 5

[42] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2

[43] 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. *Communications of the ACM*, 65(1):99–106, 2021.

4

[44] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3D representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 2

[45] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2

[46] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2

[47] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XVII*, 2022. 2, 3, 6

[48] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-resolution 3D-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 2, 4, 6

[49] Ahmed AA Osman, Timo Bolkart, and Michael J. Black. STAR: Sparse trained articulated human body regressor. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 2020. 2

[50] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. NPMs: Neural parametric models for 3D deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1

[51] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2

[52] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540, 2020. 3

[53] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 3

[54] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2, 3, 5, 6

[55] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and pose control for image synthesis of humans from a single monocular view. *arXiv preprint arXiv:2102.11263*, 2021. 1, 2

[56] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative radiance fields for 3D-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 2

[57] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. VoxGRAF: Fast 3D-aware image synthesis with sparse voxel grids. In *Advances in Neural Information Processing Systems*, 2022. 2

[58] Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3D shape learning from natural images. *arXiv preprint arXiv:1910.00287*, 2019. 2

[59] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 6

[60] Marco Tarini, Daniele Panozzo, and Olga Sorkine-Hornung. Accurate and efficient lighting for skinned models. In *Computer Graphics Forum*, volume 33, 2014. 4

[61] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, pages 27171–27183, 2021. 2

[62] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. ARAH: Animatable volume rendering of articulated human SDFs. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, 2022. 3

[63] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 3

[64] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. *Advances in Neural Information Processing Systems*, 29:82–90, 2016. 2

[65] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13296–13306, 2022. 2, 3

[66] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 2

[67] Zhuoqian Yang, Shikai Li, Wayne Wu, and Bo Dai. 3DHumanGAN: Towards photo-realistic 3D-aware human image generation. In *ICCV*, 2023. 3

[68] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-*

*Fifth Conference on Neural Information Processing Systems*, pages 4805–4815, 2021. 2

[69] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction. 2022. 2, 3

[70] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. DwNet: dense warp-based network for pose-guided human video generation. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 51. BMVA Press, 2019. 5

[71] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 5

[72] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. AvatarGen: A 3D generative model for animatable human avatars. In *Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part III*, 2022. 3

[73] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. HumanNeRF: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7753, 2022. 3