En3D: An Enhanced Generative Model for Sculpting 3D Humans from 2D Synthetic Data

Yifang Men[1], Biwen Lei[1], Yuan Yao[1], Miaomiao Cui[1], Zhouhui Lian[2], Xuansong Xie[1]
[1]Institute for Intelligent Computing, Alibaba Group
[2]Wangxuan Institute of Computer Technology, Peking University, China
https://menyifang.github.io/projects/En3D/index.html

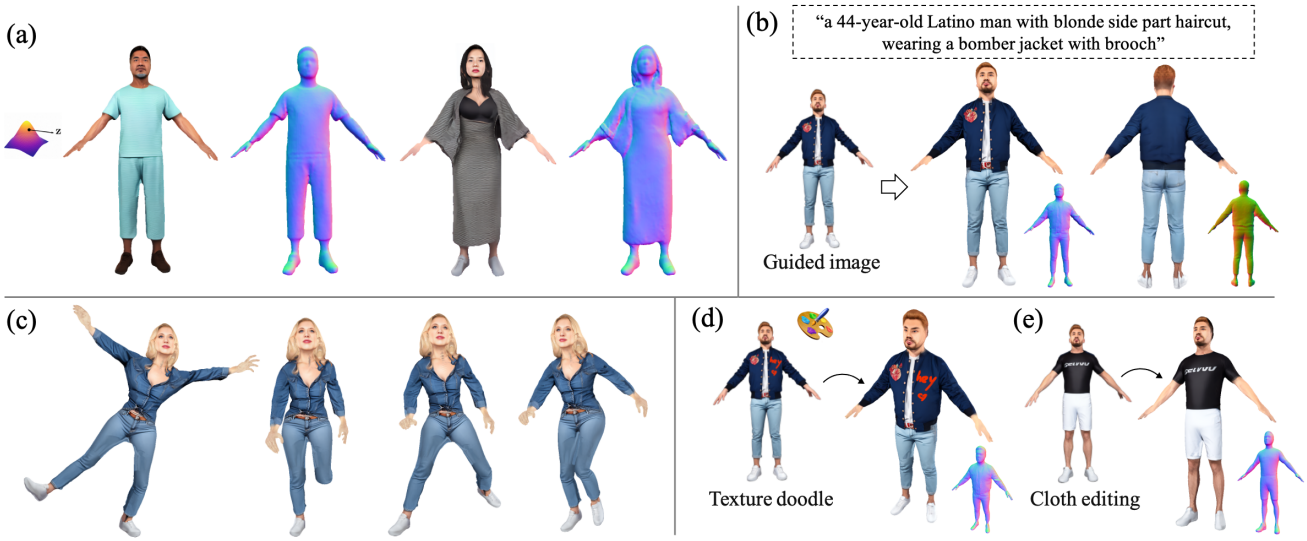


Figure 1. Given random noises or guided texts, our generative scheme can synthesize high-fidelity 3D human avatars that are visually realistic and geometrically accurate. These avatars can be seamlessly animated and easily edited. Our model is trained on 2D synthetic data without directly relying on any pre-existing 3D or 2D collections.

Abstract

We present En3D, an enhanced generative scheme for sculpting high-quality 3D human avatars. Unlike previous works that rely on scarce 3D datasets or limited 2D collections with imbalanced viewing angles and imprecise pose priors, our approach aims to develop a zero-shot 3D generative scheme capable of producing visually realistic, geometrically accurate and content-wise diverse 3D humans without directly relying on pre-existing 3D or 2D assets. To address this challenge, we introduce a meticulously crafted workflow that implements accurate physical modeling to learn the enhanced 3D generative model from synthetic 2D data. During inference, we integrate optimization modules to bridge the gap between realistic appearances and coarse 3D shapes. Specifically, En3D comprises three modules: a 3D generator that accurately models generalizable 3D humans with realistic appearance from synthesized balanced, diverse, and structured human images; a geometry sculptor that enhances shape quality using multi-view normal constraints for intricate human structure; and a texturing module that disentangles explicit texture maps with fidelity and editability, leveraging semantical UV partitioning and a differentiable rasterizer. Experimental results show that our approach significantly outperforms prior works in terms of image quality, geometry accuracy and content diversity. We also showcase the applicability of our generated avatars for animation and editing, as well as the scalability of our approach for content-style free adaptation.

1. Introduction

3D human avatars play an important role in various applications of AR/VR such as video games, telepresence and virtual try-on. Realistic human modeling is an essential task, and many valuable efforts have been made by leveraging neural implicit fields to learn high-quality articulated avatars [8, 10, 41, 48]. However, these methods are directly

learned from monocular videos or image sequences, where subjects are single individuals wearing specific garments, thus limiting their scalability.

Generative models learn a shared 3D representation to synthesize clothed humans with varying identities, clothing and poses. Traditional methods are typically trained on 3D datasets, which are limited and expensive to acquire. This data scarcity limits the model's generalization ability and may lead to overfitting on small datasets. Recently, 3D-aware image synthesis methods [5, 18, 37] have demonstrated great potential in learning 3D generative models of rigid objects from 2D image collections. Follow-up works show the feasibility of learning articulated humans from image collections driven by SMPL-based deformations, but only in limited quality and resolution. EVA3D [15] represents humans as a composition of multiple parts with NeRF representations. AG3D [9] incorporates an efficient articulation module to capture both body shape and cloth deformation. Nevertheless, there remains a noticeable gap between generated and real humans in terms of appearance and geometry. Moreover, their results are limited to specific views (i.e., frontal angles) and lack diversity (i.e., fashion images in similar skin tone, body shape, and age).

The aim of this paper is to propose a zero-shot 3D generative scheme that does not directly rely on any pre-existing 3D or 2D datasets, yet is capable of producing high-quality 3D humans that are visually realistic, geometrically accurate, and content-wise diverse. The generated avatars can be seamlessly animated and easily edited. An illustration is provided in Figure 1. To address this challenging task, our proposed method inherits from 3D-aware human image synthesis and exhibits substantial distinctions based on several key insights. Rethinking the nature of 3D-aware generative methods from 2D collections [5, 9, 15], they actually try to learn a generalizable and deformable 3D representation, whose 2D projections can meet the distribution of human images in corresponding views. Thereby, it is crucial for accurate physical modeling between 3D objects and 2D projections. However, previous works typically leverage pre-existing 2D human images to estimate physical parameters (i.e., camera and body poses), which are inaccurate because of imprecise SMPL priors for highly-articulated humans. This inaccuracy limits the synthesis ability for realistic multi-view renderings. Second, these methods solely rely on discriminating 2D renderings, which is ambiguous and loose to capture inherent 3D shapes in detail, especially for intricate human appearance.

To address these limitations, we propose a novel generative scheme with two core designs. Firstly, we introduce a meticulously-crafted workflow that implements accurate physical modeling to learn an enhanced 3D generative model from synthetic data. This is achieved by instantiating a 3D body scene and projecting the underlying 3D skeletons into 2D pose images using explicit camera parameters. These 2D pose images act as conditions to control a 2D diffusion model, synthesizing realistic human images from specific viewpoints. By leveraging synthetic view-balanced, diverse and structured human images, along with known physical parameters, we employ a 3D generator equipped with an enhanced renderer and discriminator to learn realistic appearance modeling. Secondly, we improve the 3D shape quality by leveraging the gap between high-quality multi-view renderings and the coarse mesh produced by the 3D generative module. Specifically, we integrate an optimization module that utilizes multi-view normal constraints to rapidly refine geometry details under supervision. Additionally, we incorporate an explicit texturing module to ensure faithful UV texture maps. In contrast to previous works that rely on inaccurate physical settings and inadequate shape supervision, we rebuild the generative scheme from the ground up, resulting in comprehensive improvements in image quality, geometry accuracy, and content diversity. In summary, our contributions are threefold:

- We present a zero-shot generative scheme that efficiently synthesizes high-quality 3D human avatars with visual realism, geometric accuracy and content diversity. These avatars can be seamlessly animated and easily edited, offering greater flexibility in their applications.
- We develop a meticulously-crafted workflow to learn an enhanced generative model from synthesized human images that are balanced, diverse, and also possess known physical parameters. This leads to diverse 3D-aware human image synthesis with realistic appearance.
- We propose to integrate optimization modules into the 3D generator, leveraging multi-view guidance to enhance both shape quality and texture fidelity, thus achieving realistic 3D human assets.

2. Related work

3D Human Modeling. Parametric models [3, 19, 20, 29, 38] serve as a common representation for 3D human modeling, they allows for robust control by deforming a template mesh with a series of low-dimensional parameters, but can only generate naked 3D humans. Similar ideas have been extended to model clothed humans [2, 31], but geometric expressivity is restricted due to the fixed mesh topology. Subsequent works [6, 39, 39] further introduce implicit surfaces to produce complex non-linear deformations of 3D bodies. Unfortunately, the aforementioned approaches all require 3D scans of various human poses for model fitting, which are difficult to acquire. With the explosion of NeRF, valuable efforts have been made towards combining NeRF models with explicit human models [8, 10, 28, 41, 48]. Neural body [41] anchors a set of latent codes to the vertices of the SMPL model [29] and transforms the spatial locations of the codes to the volume in the observation space. Human-

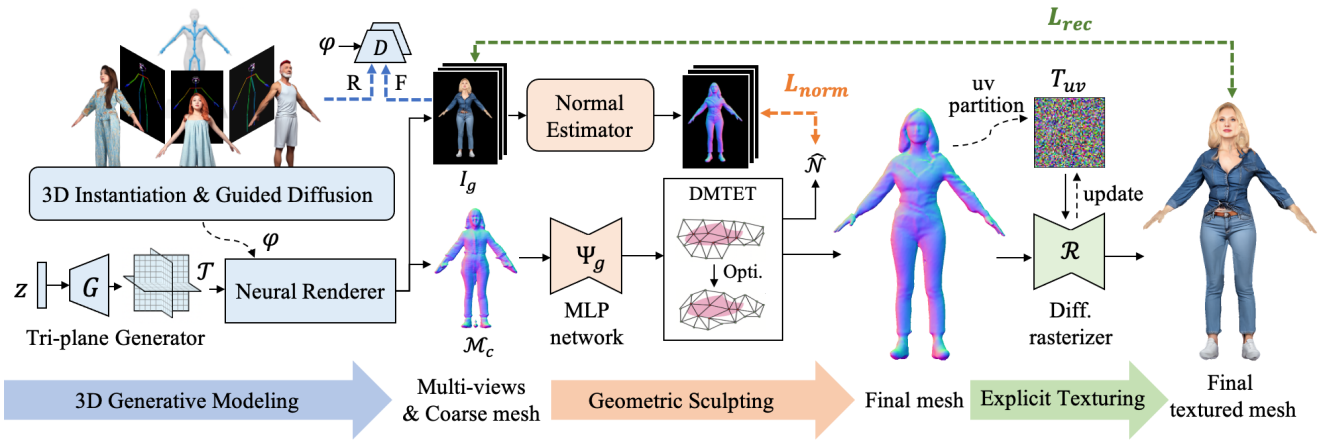


Figure 2. An overview of the proposed scheme, which consists of three modules: 3D generative modeling (3DGM), geometric sculpting (GS), and explicit texturing (ET). 3DGM employs synthesized diverse, balanced and structured human images with accurate camera parameters φ to learn generalizable 3D humans with the triplane-based architecture. GS is integrated as an optimization module by utilizing multi-view normal constraints to refine and carve geometry details. ET utilizes UV partitioning and a differentiable rasterizer to disentangle explicit UV texture maps. Not only multi-view renderings but also realistic 3D models can be acquired finally by our method.

NeRF [48] optimizes for a canonical, volumetric T-pose of the human with a motion field to map the non-rigid transformations. Text to 3D humans [16, 24] are introduced by combining SDS loss [42] with diffusion guidance. Nevertheless, these methods are learned directly from image sequences or guided by single text prompts, where subjects are fitting to single scenes, thus limiting their scalability.

Generative 3D-aware Image Synthesis. Recently, 3D-aware image synthesis methods have lifted image generation with explicit view control by integrating the 2D generative models [21–23] with 3D representations, such as voxels [13, 34, 35, 49], meshes [27, 47] and points clouds [1, 26]. GRAF [46] and π-GAN[4] firstly integrate the implicit representation networks, i.e., NeRF [33], with differentiable volumetric rendering for 3D scene generation. However, they have difficulties in training on high-resolution images due to the costly rendering process. Subsequent works have sought to improve the efficiency and quality of such NeRF-based GANs, either by adopting a two-stage rendering process [5, 12, 36, 37, 51] or a smart sampling strategy [7, 55]. StyleSDF [37] combines a SDF-based volume renderer and a 2D StyleGAN network [22] for photorealistic image generation. EG3D [5] introduces a superior triplane representation to leverage 2D CNN-based feature generators for efficient generalization over 3D spaces. Although these methods demonstrate impressive quality in view-consistent image synthesis, they are limited to simplified rigid objects such as faces, cats and cars.

To learn highly articulated humans from unstructured 2D images, recent works [9, 11, 15, 17, 52, 53] integrate the deformation field to learn non-rigid deformations based on the body prior of estimated SMPL parameters. EVA3D [15] represents humans as a composition of multiple parts with NeRF representations. Instead of directly rendering the image from a 3D representation, 3DHumanGAN [52] uses an equivariant 2D generator modulated by 3D human body prior, which enables to establish one-to-many mapping from 3D geometry to synthesized textures from 2D images. AG3D [9] combines the 3D generator with an efficient articulation module to warp 3D objects from canonical space to posed space via a learned continuous deformation field. However, a gap still exists between the generated and real humans in terms of appearance, due to the imprecise priors from complex poses as well as the data biases from limited human poses and imbalanced viewing angles in the dataset.

3. Method Description

Our goal is to develop a zero-shot 3D generative scheme that does not directly rely on any pre-existing 3D or 2D collections, yet is capable of producing high-quality 3D humans that are visually realistic, geometrically accurate and content-wise diverse to generalize to arbitrary humans.

An overview of the proposed scheme is illustrated in Figure 2. We build a sequential pipeline with the following three modules: 3D generative modeling (3DGM), geometric sculpting (GS), and explicit texturing (ET). The first module synthesizes view-balanced, structured and diverse human images with known camera parameters. Subsequently, it learns a 3D generative model from these synthetic data, focusing on realistic appearance modeling (Section 3.1). To overcome the inaccuracy of the 3D shape, the GS module is incorporated during the inference process. It optimizes a hybrid representation with multi-view normal constraints to carve intricate mesh details (Section 3.2).

Additionally, the ET module is employed to disentangle explicit texture by utilizing semantical UV partitioning and a differentiable rasterizer (Section 3.3). By combining these modules, we are able to synthesize high-quality and faithful 3D human avatars by incorporating random noises or guided texts/images (Section 3.4).

3.1. 3D generative modeling

Without any 3D or 2D collections, we develop a synthesis-based flow to learn a 3D generative module from 2D synthetic data. We start by instantiating a 3D scene through the projection of underlying 3D skeletons onto 2D pose images, utilizing accurate physical parameters (i.e., camera parameters). Subsequently, the projected 2D pose images serve as conditions to control the 2D diffusion model [54] for synthesizing view-balanced, diverse, and lifelike human images. Finally, we employ a triplane-based generator with enhanced designs to learn a generalizable 3D representation from the synthetic data. Details are described as follows.

3D instantiation. Starting with a template body mesh (e.g., SMPL-X [40]) positioned and posed in canonical space, we estimate the 3D joint locations $\mathcal{P}_{3d}$ by regressing them from interpolated vertices. We then project $\mathcal{P}_{3d}$ onto 2D poses $\mathcal{P}_i, i = 1, \ldots, k$ from $\mathcal{K}$ horizontally uniformly sampled viewpoints φ. In this way, paired 2D pose images and their corresponding camera parameters $\{\mathcal{P}_i, \varphi_i\}$ are formulated.

Controlled 2D image synthesis. With the pose image $\mathcal{P}_i$, we feed it into off-the-shelf ControlNet [54] as the pose condition to guide diffusion models [44] to synthesize human images in desired poses (i.e., views). The text prompt T is also used for diverse contents. Given a prompt T, instead of generating a human image $\mathcal{I}_s : \mathcal{I}_s = \mathcal{C}(\mathcal{P}_i, T)$ independently for each view φ_i, we horizontally concatenate $\mathcal{K}$ pose images $\mathcal{P}_i \in R^{H\times W\times 3}$, resulting in $\mathcal{P}'_i \in R^{H\times KW\times 3}$ and feed $\mathcal{P}'_i$ to $\mathcal{C}$, along with a prompt hint of 'multi-view' in T. In this way, multi-view human images $\mathcal{I}'_s$ are synthesized with roughly coherent appearance. We split $\mathcal{I}'_s$ to single view images $\mathcal{I}_\varphi$ under specific views φ. This concatenation strategy facilitates the convergence of distributions in synthetic multi-views, thus easing the learning of common 3D representation meeting multi-view characteristics.

Generalizable 3D representation learning. With synthetic data of paired $\{\mathcal{I}_\varphi, \varphi\}$, we learn the 3D generative module $\mathcal{G}_{3d}$ from them to produce diverse 3D-aware human images with realistic appearance. Inspired by EG3D [5], we employ a triplane-based generator to produce a generalizable representation $\mathcal{T}$ and introduce a patch-composed neural renderer to learn intricate human representation efficiently. Specifically, instead of uniformly sampling 2D pixels on the image $\mathcal{I}$, we decompose patches in the ROI region including human bodies, and only emit rays towards pixels in these patches. The rays are rendered into RGB color with opacity values via volume rendering. Based on the decomposed rule, we decode rendered colors to multiple patches and re-combine these patches for full feature images. In this way, the representation is composed of effective human body parts, which directs the attention of the networks towards the human subject itself. This design facilitates fine-grained local human learning while maintaining computational efficiency.

For the training process, we employ two discriminators, one for RGB images and another for silhouettes, which yields better disentanglement of foreground objects with global geometry. The training loss for this module L_{3d} consists of the two adversarial terms:

$$\mathcal{L}_{3d} = \mathcal{L}_{adv}(\mathcal{D}_{rgb}, \mathcal{G}_{3d}) + \lambda_s \mathcal{L}_{adv}(\mathcal{D}_{mask}, \mathcal{G}_{3d}), \quad (1)$$

where λ_s denotes the weight of silhouette item. $\mathcal{L}_{adv}$ is computed by the non-saturating GAN loss with R1 regularization [32].

With the trained $\mathcal{G}_{3d}$, we can synthesize 3D-aware human images $\mathcal{I}_g^\varphi$ with view control, and extract coarse 3D shapes $\mathcal{M}_c$ from the density field of neural renderer using the Marching Cubes algorithm [30].

3.2. Geometric sculpting

Our 3D generative module can produce high-quality and 3D-consistent human images in view controls. However, its training solely relies on discriminations made using 2D renderings, which can result in inaccuracies in capturing the inherent geometry, especially for complex human bodies. Therefore, we integrate the geometric sculpting, an optimization module leveraging geometric information from high-quality multi-views to carve surface details. Combined with a hybrid 3D representation and a differentiable rasterizer, it can rapidly enhance the shape quality within seconds.

DMTET adaption. Owing to the expressive ability of arbitrary topologies and computational efficiency with direct shape optimization, we employ DMTET as our 3D representation in this module and adapt it to the coarse mesh $\mathcal{M}_c$ via an initial fitting procedure. Specifically, we parameterize DMTET as an MLP network Ψ_g that learns to predict the SDF value $s(v_i)$ and the position offset δv_i for each vertex $v_i \in VT$ of the tetrahedral grid (VT, T). A point set $P = \{p_i \in R^3\}$ is randomly sampled near $\mathcal{M}_c$ and their SDF values $SDF(p_i)$ can be pre-computed. We adapt the parameters ψ of Ψ_g by fitting it to the SDF of $\mathcal{M}_c$:

$$\mathcal{L}_{ada} = \sum_{p_i \in P} ||s(p_i; \psi) - SDF(p_i)||_2. \quad (2)$$

Geometry refinement. Using the adapted DMTET, we leverage the highly-detailed normal maps $\mathcal{N}$ derived from realistic multi-view images as a guidance to refine local surfaces. To obtain the pseudo-GT normals $\mathcal{N}_\varphi$, we extract

them from $\mathcal{I}_g^\varphi$ using a pre-trained normal estimator [50]. For the rendered normals $\hat{\mathcal{N}}_\varphi$, we extract the triangular mesh $\mathcal{M}_{tri}$ from (VT, T) using the Marching Tetrahedra (MT) layer in our current DMTET. By rendering the generated mesh $\mathcal{M}_{tri}$ with differentiable rasterization, we obtain the resulting normal map $\hat{\mathcal{N}}_\varphi$. To ensure holistic surface polishing that takes into account multi-view normals, we randomly sample camera poses φ that are uniformly distributed in space. We optimize the parameters of Ψ_g using the normal loss, which is defined as:

$$\mathcal{L}_{norm} = ||\hat{\mathcal{N}}_\varphi - \mathcal{N}_\varphi||_2. \tag{3}$$

After rapid optimization, the final triangular mesh $\mathcal{M}_{tri}$ can be easily extracted from the MT layer. If the hands exhibit noise, they can be optionally replaced with cleaner geometry hands from SMPL-X, benefiting from the alignment of the generated body in canonical space with the underlying template body.

3.3. Explicit texturing

With the final mesh, the explicit texturing module aims to disentangle a UV texture map from multi-view renderings $\mathcal{I}_g^\varphi$. This intuitive module not only facilitates the incorporation of high-fidelity textures but also enables various editing applications, as verified in Section 4.4.

Given the polished triangular mesh $\mathcal{M}_{tri}$ and multi-views $\mathcal{I}_g^\varphi$, we model the explicit texture map T_{uv} of $\mathcal{M}_{tri}$ with a semantic UV partition and optimize T_{uv} using a differentiable rasterizer $\mathcal{R}$ [25]. Specifically, leveraging the canonical properties of synthesized bodies, we semantically split $\mathcal{M}_{tri}$ into γ components and rotate each component vertically, thus enabling effective UV projection for each component with cylinder unwarping. We then combine the texture partitions together for the full texture T_{uv}. We optimize T_{uv} from a randomly initialized scratch using the texture loss, which consists of a multi-view reconstruction term and a total-variation (tv) term:

$$\mathcal{L}_{tex} = \mathcal{L}_{rec} + \lambda_{tv}\mathcal{L}_{tv}, \tag{4}$$

where λ_{tv} denotes the weight of the tv loss.

Multi-view guidance. To ensure comprehensive texturing in the 3D space, we render the color images $\mathcal{R}(\mathcal{M}_{tri}, \varphi)$ and silhouettes $\mathcal{S}$ using $\mathcal{R}$ and optimize T_{uv} utilizing multi-view weighted guidance. Their pixel-alignment distances to the original multi-view renderings $\mathcal{I}_g^\varphi$ are defined as the reconstruction loss:

$$\mathcal{L}_{rec} = \sum_{\varphi \in \Omega} w_\varphi ||\mathcal{R}(\mathcal{M}_{tri}, \varphi) \cdot \mathcal{S} - I_g^\varphi \cdot \mathcal{S}||_2, \tag{5}$$

where Ω is the set of viewpoints $\{\varphi_i, i = 1, ..., k\}$ and w_φ denotes weights of different views. w_φ equals to 1.0 for $\varphi \in \{front, back\}$ and 0.2 otherwise.

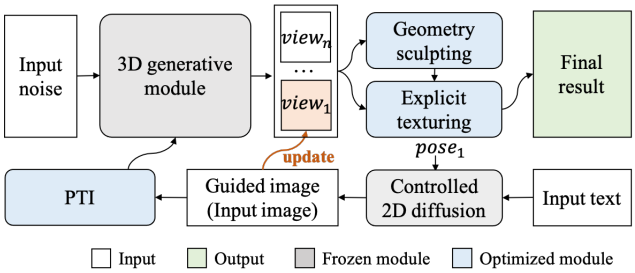


Figure 3. The visualized flowchart of our method that synthesize textured 3D human avatars from input noises, texts or images.

Smooth constraint. To avoid abrupt variations and smooth the generated texture T_{uv}, we utilize the total-variation loss $\mathcal{L}_{tv}$ which is computed by:

$$\mathcal{L}_{tv} = \frac{1}{h \times w \times c} ||\nabla_x(T_{uv}) + \nabla_y(T_{uv})||, \tag{6}$$

where x and y denote horizontal and vertical directions.

3.4. Inference

Built upon the above modules, we can generate high-quality 3D human avatars from either random noises or guided inputs such as texts or images. The flowchart for this process is shown in Figure 3. For input noises, we can easily obtain the final results by sequentially using the 3DGM, GS and ET modules. For text-guided synthesis, we first convert the text into a structured image using our controlled diffusion $\mathcal{C}$, and then inverse it to the latent space using PTI [43]. Specially, the GS and ET modules provide an interface that accurately reflects viewed modifications in the final 3D objects. As a result, we utilize the guided image to replace the corresponding view image, which results in improved fidelity in terms of geometry and texture. The same process is applied for input images as guided images.

4. Experimental Results

Implementation details. Our process begins by training the 3D generative module (3DGM) on synthetic data. During inference, we integrate the geometric sculpting (GS) and explicit texturing (ET) as optimization modules. For 3DGM, we normalize the template body to the $(0, 1)$ space and place its center at the origin of the world coordinate system. We sample $7(\mathcal{K} = 7)$ viewpoints uniformly from the horizontal plane, ranging from 0° to 180° (front to back), with a camera radius of 2.7. For each viewpoint, we generate $100K$ images using the corresponding pose image. To ensure diverse synthesis, we use detailed descriptions of age, gender, ethnicity, hairstyle, facial features, and clothing, leveraging a vast word bank. To cover 360° views, we horizontally flip the synthesized images and obtain 1.4 million human images at a resolution of 512^2 in total. We train

Figure 4. Results of synthesized 3D human avatars at 512^2.

the 3DGM for about 2.5M iterations with a batch size of 32, using two discriminators with a learning rate of 0.002 and a generator learning rate of 0.0025. The training takes 8 days using 8 NVIDIA Tesla-V100 GPUs. For GS, we optimize ψ for 400 iterations for DMTET adaption and 100 iterations for surface carving (taking about 15s in total on 1 NVIDIA RTX 3090 GPU). For ET, we set $\lambda_{uv} = 1$ and optimize T_{uv} for 500 iterations (around 10 seconds). We split $\mathcal{M}_{tri}$ into $5(\gamma = 5)$ body parts (i.e., trunk, left/right arm/leg) with cylinder UV unwarping. We use the Adam optimizer with learning rates of 0.01 and 0.001 for Ψ_g and T_{uv}, respectively. Detailed network architectures can be found in the supplemental materials (Suppl).

4.1. 3D human generation

Figure 4 showcases several 3D human avatars synthesized by our pipeline, highlighting the image quality, geometry accuracy, and diverse outputs achieved through our method. Additionally, we explore the interpolation of the latent conditions to yield smooth transitions in appearance, leveraging the smooth latent space learned by our generative model. For more synthesized examples and interpolation results, please refer to the Suppl.

4.2. Comparisons

Qualitative comparison. In Figure 5, we compare our method with three baselines: EVA3D [15] and AG3D [9], which are state-of-the-art methods for generating 3D humans from 2D images, and EG3D [5], which serves as the foundational backbone of our method. The results of first two methods are produced by directly using source codes and trained models released by authors. We train EG3D using our synthetic images with estimated cameras from scratch. As we can see, EVA3D fails to produce 360° humans with reasonable back inferring. AG3D and EG3D are able to generate 360° renderings but both struggle with photorealism and capturing detailed shapes. Our method synthesizes not only higher-quality, view-consistent 360° images but also higher-fidelity 3D geometry with intricate details, such as irregular dresses and haircuts.

Quantitative comparison. Table 1 provides quantitative results comparing our method against the baselines. We measure image quality with Frechet Inception Distance (FID) [14] and Inception Score [45] for 360° views (IS-

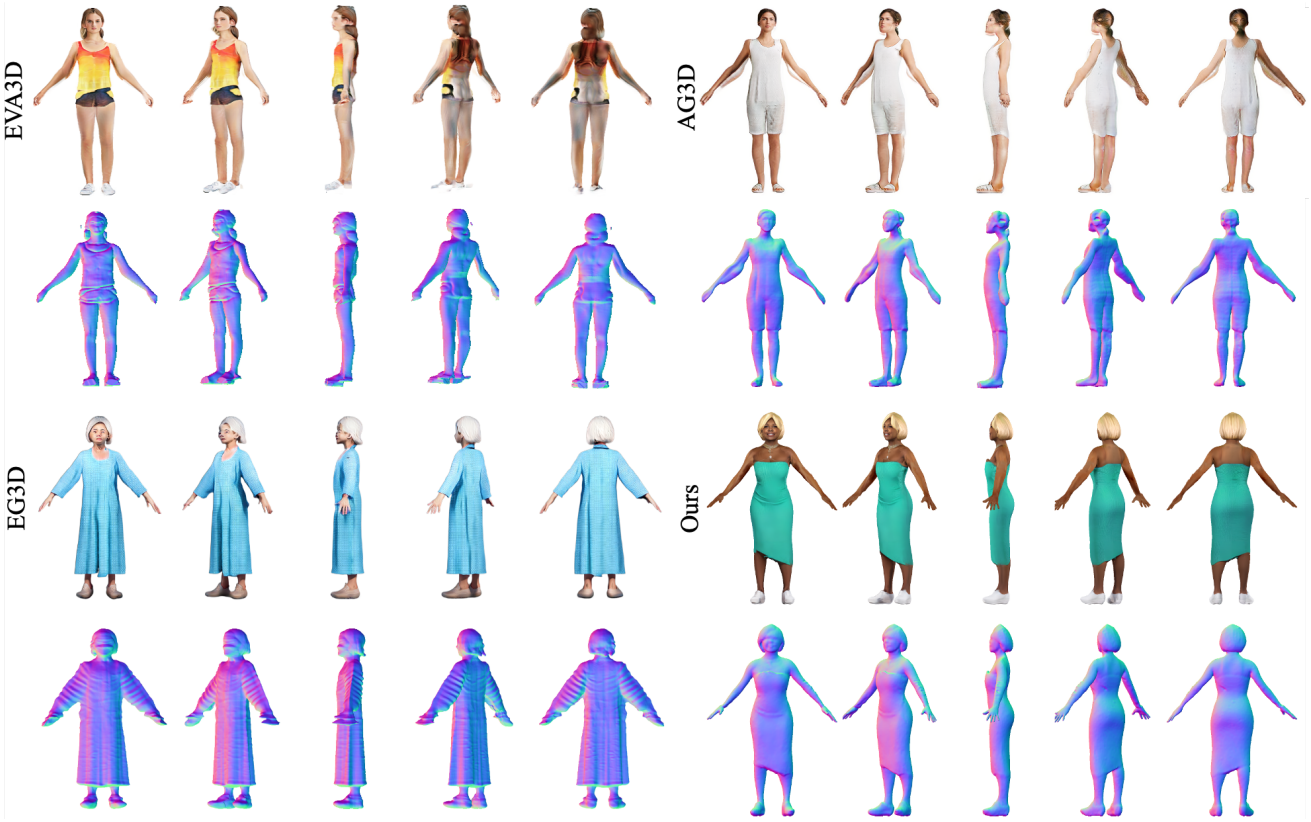


Figure 5. Qualitative comparison with three state-of-the-art methods: EVA3D [15], AG3D [9] and EG3D [5].

Table 1. Quantitative evaluation using FID, IS-360, normal accuracy (Normal) and identity consistency (ID).

Method	FID ↓	IS-360 ↑	Normal ↓	ID↑
EVA3D [15]	15.91	3.19	30.81	0.72
AG3D [9]	10.93	3.28	20.83	0.69
EVA3D-syn	6.27	3.25	9.57	0.72
AG3D-syn	8.93	3.22	8.64	0.71
EG3D [5]	7.48	3.26	12.74	0.71
Ours	**2.73**	**3.43**	**5.62**	**0.74**

360). FID measures the visual similarity and distribution discrepancy between 50k generated images and all real images. IS-360 focuses on the self-realism of generated images in 360° views. For shape evaluation, we compute FID between rendered normals and pseudo-GT normal maps (Normal), following AG3D. The FID and Normal scores of EVA3D and AG3D are directly fetched from their reports. Additionally, we access the multi-view facial identity consistency using the ID metric introduced by EG3D. For a clearer comparison, we also provide metric results of EVA3D and AG3D trained on our synthetic data (-syn). Overall, our method demonstrates significant improvements in FID and Normal, bringing the generative human model to a new level of realistic 360° renderings with delicate geometry while maintaining state-of-the-art view consistency.

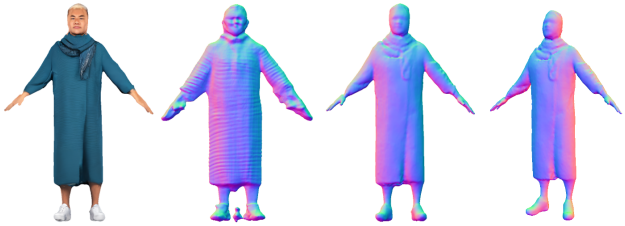
Table 2. Results of models trained by replacing physical parameters with estimated ones (w/o SYN-P) or removing patch-composed rendering (w/o PCR).

	Ours	Ours-w/o SYN-P	Ours-w/o PCR
FID ↓	**2.73**	4.28	3.26
IS-360 ↑	**3.43**	3.31	3.35

4.3. Ablation study

Synthesis flow and patch-composed rendering. We assess the impact of our carefully designed synthesis flow by training a model with synthetic images but with camera and pose parameters estimated by SMPLify-X [40] (w/o SYN-P). As Table 2 shows, the model w/o SYN-P results in worse FID and IS-360 scores, indicating that the synthesis flow contributes to more accurate physical parameters for realistic appearance modeling. By utilizing patch-composed rendering (PCR), the networks focus more on the human region, leading to more realistic results.

Geometry sculpting module (GS). We demonstrate the importance of this module by visualizing the meshes before and after its implementation. Figure 6 (b) shows that the preceding module yields a coarse mesh due to the complex



(a) Rendering (b) w/o GS (c) Full method

Figure 6. Effects of the GS module to carve fine-grained surfaces.

(a) Guided image (b) w/o ET (c) Full method

Figure 7. Effects of the ET module for guided synthesis.

human anatomy and the challenges posed by decomposing ambiguous 3D shapes from 2D images. The GS module utilizes high-quality multi-view outputs and employs a more flexible hybrid representation to create expressive humans with arbitrary topologies. It learns from pixel-level surface supervision, leading to a significant improvement in shape quality, characterized by smooth surfaces and intricate outfits (Figure 6 (c)).

Explicit texturing module (ET). This intuitive module not only extracts the explicit UV texture for complete 3D assets but also enables high-fidelity results for image guided synthesis. Following the flowchart in Figure 3, we compare the results produced with and without this module. Our method without ET directly generates implicit renderings through PTI inversion, as shown in Figure 7 (b). While it successfully preserves global identity, it struggles to synthesize highly faithful local textures (e.g., floral patterns). The ET module offers a convenient and efficient way to directly interact with the 3D representation, enabling the production of high-fidelity 3D humans with more consistent content including exquisite local patterns (Figure 7 (a, c)).

4.4. Applications

Avatar animation. All avatars produced by our method are in a canonical body pose and aligned to an underlying 3D skeleton extracted from SMPL-X. This alignment allows for easy animation and the generation of motion videos, as demonstrated in Figure 1 and Suppl.

Texture doodle and local editing. Our approach benefits from the explicit disentanglement of geometry and texture, enabling flexible editing capabilities. Following the flowchart of text or image guided synthesis (Section 3.4),

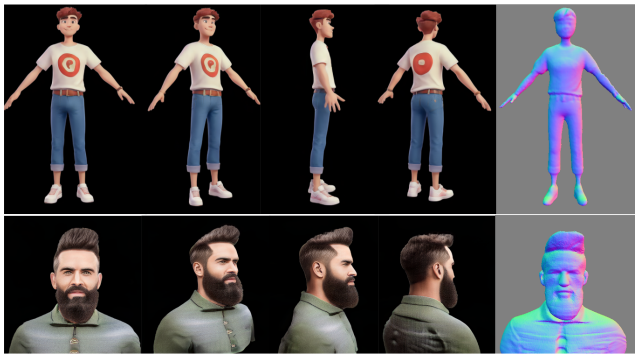


Figure 8. Results synthesized by adapting our method to various styles (e.g., Disney cartoon characters) or contents (e.g., portrait heads).

users can paint any pattern or add text to a guided image. These modifications can be transferred to 3D human models by inputting modified views into the texture module (e.g., painting the text 'hey' on a jacket as shown in Figure 1 (d)). Our approach also allows for clothing editing by simultaneously injecting edited guide images with desired clothing into the GS and ET modules (e.g., changing a jacket and jeans to bodysuits in Figure 1 (e)). More results can be found in Suppl.

Content-style free adaption. Our proposed scheme is versatile and can be extended to generate various types of contents (e.g., portrait heads) and styles (e.g., Disney cartoon characters). To achieve this, we fine-tune our model using synthetic images from these domains, allowing for flexible adaptation. We showcase the results in Figure 8. More results and other discussions (e.g., limitations, negative impact, etc.) can be found in the Suppl.

5. Conclusion

We introduced En3D, a novel generative scheme for sculpting 3D humans from 2D synthetic data. This method overcomes limitations in existing 3D or 2D collections and significantly enhances the image quality, geometry accuracy, and content diversity of generated 3D humans. En3D comprises a 3D generative module that learns generalizable 3D humans from synthetic 2D data with accurate physical modeling, and two optimization modules to carve intricate shape details and disentangle explicit UV textures with high fidelity, respectively. Experimental results validated the superiority and effectiveness of our method. We also demonstated the flexibility of our generated avatars for animation and editing, as well as the scalability of our approach for synthesizing portraits and Disney characters. We believe that our solution could provide invaluable human assets for the 3D vision community. Furthermore, it holds potential for use in common 3D object synthesis tasks.

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. PMLR, 2018. 3

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 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] 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*, pages 5799–5809, 2021. 3

[5] 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, 4, 6, 7

[6] 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. 2

[7] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10673–10683, 2022. 3

[8] 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. 1, 2

[9] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. Ag3d: Learning to generate 3d avatars from 2d image collections. *arXiv preprint arXiv:2305.02312*, 2023. 2, 3, 6, 7

[10] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 1, 2

[11] 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 *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 3

[12] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 3

[13] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9984–9993, 2019. 3

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[15] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 2, 3, 6, 7

[16] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. *arXiv preprint arXiv:2310.01406*, 2023. 3

[17] Suyi Jiang, Haoran Jiang, Ziyu Wang, Haimin Luo, Wenzheng Chen, and Lan Xu. Humangen: Generating human radiance fields with explicit priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12543–12554, 2023. 3

[18] Kyungmin Jo, Wonjoon Jin, Jaegul Choo, Hyunjoon Lee, and Sunghyun Cho. 3d-aware generative model for improved side-view image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22862–22872, 2023. 2

[19] 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

[20] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2

[21] 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. 3

[22] 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 8110–8119, 2020. 3

[23] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 3

[24] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[25] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 5

[26] 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. 3

[27] 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. 3

[28] 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 transactions on graphics (TOG)*, 40(6):1–16, 2021. 2

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6), 2015. 2

[30] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 4

[31] 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

[32] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018. 4

[33] 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. 3

[34] 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. 3

[35] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *Advances in neural information processing systems*, 33:6767–6778, 2020. 3

[36] 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. 3

[37] 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, 3

[38] 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*, pages 598–613. Springer, 2020. 2

[39] 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*, pages 12695–12705, 2021. 2

[40] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 4, 7

[41] 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. 1, 2

[42] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[43] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1):1–13, 2022. 5

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4

[45] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 6

[46] 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. 3

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

[48] 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. 1, 2, 3

[49] 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, 2016. 3

[50] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from nor-

mals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 5

[51] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe hd: A high-resolution 3d-aware generative model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18440–18449, 2022. 3

[52] Zhuoqian Yang, Shikai Li, Wayne Wu, and Bo Dai. 3dhumangan: 3d-aware human image generation with 3d pose mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23008–23019, 2023. 3

[53] 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 *European Conference on Computer Vision*, pages 668–685. Springer, 2022. 3

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 4

[55] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021. 3