GETAvatar: Generative Textured Meshes for Animatable Human Avatars

Xuanmeng Zhang[1,2*] Jianfeng Zhang[2,3*] Rohan Chacko[2]
Hongyi Xu[2] Guoxian Song[2] Yi Yang[4] Jiashi Feng[2]
[1]ReLER, AAII, University of Technology Sydney [2]ByteDance
[3] National University of Singapore [4]ReLER, CCAI, Zhejiang University

(a) Varying camera parameters (b) Varying body poses

Figure 1: GETAvatar generates controllable human avatars with diverse textures and detailed geometries under full control over camera poses and body poses. Please refer to the Appendix for more multi-view and animation results.

Abstract

*We study the problem of 3D-aware full-body human generation, aiming at creating animatable human avatars with high-quality textures and geometries. Generally, two challenges remain in this field: i) existing methods struggle to generate geometries with rich realistic details such as the wrinkles of garments; ii) they typically utilize volumetric radiance fields and neural renderers in the synthesis process, making high-resolution rendering non-trivial. To overcome these problems, we propose **GETAvatar**, a **G**enerative model that directly generates **E**xplicit **T**extured 3D meshes for animatable human **Avatar**, with photo-realistic appearance and fine geometric details. Specifically, we first design an articulated 3D human representation with explicit surface modeling, and enrich the generated humans with realistic surface details by learning from the 2D normal maps of 3D scan data. Second, with the explicit mesh representation, we can use a rasterization-based renderer to perform surface rendering, allowing us to achieve high-resolution image generation efficiently. Extensive experiments demonstrate that GETAvatar achieves state-of-the-art performance on 3D-aware human generation both in appearance and geometry quality. Notably, GETAvatar can generate images at 512^2 resolution with 17FPS and 1024^2 resolution with 14FPS, improving upon previous methods by $2\times$. Our code and models will be at* `https://getavatar.github.io/`.

*Equal contribution.

1. Introduction

Generating high-quality 3D human avatars with explicit control over camera poses, body poses and shapes has been a long-standing challenge in computer vision and graphics. It has wide applications in video games, AR/VR, and movie production. Recently, 3D-aware generative models have demonstrated impressive results in producing multi-view-consistent images of 3D shapes [29, 4, 22, 3, 24, 6]. However, despite their success in modeling relatively simple and rigid objects, it remains challenging for modeling dynamic human bodies with large articulated motions. The main reason is that these 3D GANs are not designed to handle body deformations, such as variations in human shapes and poses. Thus, they struggle to manipulate or animate the generated avatars given the control signals.

Some recent works [23, 2, 9, 35] have incorporated human priors [17] into 3D-aware generative models [4, 3] to

generate animatable 3D human avatars. However, these methods face two challenges. First, the generated human avatars lack fine geometric details, such as cloth wrinkles and hair, which are highly desirable for the photo-realistic 3D human generation. Second, existing methods [23, 2, 9, 35] adopt volumetric neural renderers in the synthesis process, which suffers from high computational costs, making high-resolution rendering non-trivial.

In this work, we propose GETAvatar, a generative model that produces explicit textured 3D meshes with rich surface details (see Fig. 1) for animatable human avatars. Previous methods [23, 2, 9, 35] model the human body with implicit geometry representations, *i.e*., density fields and signed distance fields, which produce either noisy or over-smoothed geometries due to the lack of explicit surface modeling and insufficient geometric supervision. To improve the geometry quality of the generated humans, different from previous methods, we propose to model fine geometric details with a normal field [7], which associates a normal vector with each point on the surface. The direction and magnitude of the normal vector provide crucial geometric information to represent the detailed human body surface.

Specifically, we design a body-controllable articulated 3D human representation with body deformation modeling and explicit surface modeling. The former allows us to deform the generated humans to target pose and shape, and the latter enables us to extract the underlying human body surface as an explicit mesh in a differentiable manner. Based on the extracted meshes, we further construct a normal field to depict the detailed surface of the generated humans. The normal field enables the model to capture realistic geometric details from 2D normal maps that are rendered from available 3D human scans, improving geometry quality and resulting in higher fidelity appearance generation significantly.

Besides, existing methods [23, 2, 9] struggle to render high-resolution images as the volume rendering process require intensive memory and computational costs. The main issue is that volumetric radiance field and neural renders perform volume sampling on both occupied and free regions [19] that do not contribute to the rendered images, resulting in computational inefficiency. In contrast, GETAvatar benefits from the proposed explicit representation and thus can generate textured meshes in a differentiable manner. With the extracted mesh surface, we can render high-resolution images up to 1024^2 with a highly efficient rasterization-based surface renderer [12].

To validate the effectiveness of GETAvatar, we conduct extensive experiments on two 3D human datasets [1, 34]. The quantitative and qualitative results demonstrate that GETAvatar consistently outperforms previous methods in terms of both visual and geometry quality (see Fig. 3). Overall, our work makes the following contributions:

1. We propose a generative model, GETAvatar, that enables high-quality 3D-aware human generation, with full control over camera poses, body shapes, and human poses.

2. We propose to model the complex body surface using a 3D normal field, which significantly improves the geometric details of the generated clothed humans.

3. We design an articulated 3D human representation with differentiable surface modeling. The explicit mesh representation supports 360° free-view, high-resolution image synthesis (1024^2) for the generated avatars, and supports normal map rendering.

4. Our GETAvatar can be applied to a wide range of tasks, such as re-texturing, single-view 3D reconstruction, and re-animation.

2. Related Work

3D-aware Generative Models. In recent years, 3D-aware image generation [29] has gained a surge of interest. To generate objects and scenes in 3D space, 3D-aware generative models [4, 22, 3, 6, 24, 37, 32] incorporate 3D representations into generative adversarial networks, such as point clouds, 3D primitives [14], voxels [21], meshes [6], and neural radiances fields [29]. Among them, NeRF-based generative models [4, 22, 3, 24, 37] have become the dominating direction of 3D generation due to the high-fidelity image synthesis and 3D consistency. EG3D [3] introduces an efficient explicit-implicit framework with a triplane hybrid representation. StyleSDF [24] combines an SDF-based 3D volume renderer and a style-based 2D generator to obtain 3D surfaces. These methods typically perform volume rendering at a low resolution and then adopt a super-resolution module as the 2D decoder to get high-resolution results. Recently, Gao *et al*. [6] propose GET3D for synthesizing textured meshes for static rigid objects. In this work, we take inspiration from GET3D [6] and propose a generative model for animatable human avatars.

3D Human Generation. Recently, some works [23, 2, 9, 35] tackle the 3D human generation by combing 3D GANs with human representations. Noguchi *et al*. introduce ENARF [23] to learn articulated geometry-aware representations from 2D images. Bergman *et al*. propose Generative Neural Articulated Radiance Fields (GNARF) [2] to implement the generation and animation of human bodies. Built on EG3D [3], AvatarGen [35] adopts signed distance fields (SDFs) as geometry proxy to synthesize clothed 3D human avatars. By dividing the human body into local parts, Hong *et al*. propose a compositional NeRF representation for 3D human generation [9]. However, these methods fail to synthesize high-resolution images, and also cannot generate intricate geometric details of garments. In contrast,

GETAvatar exploits an explicit mesh representation and a 3D normal field for human geometry modeling, thus is capable of generating human avatars with realistic details and achieves high-resolution photo-realistic image rendering.

3. Preliminaries

Our method involves triplane representation [3] and SMPL human model [17]. Here we provide a brief introduction to them. More details can be found in the original papers [3, 17].

Triplane 3D Representation. Recently, EG3D [3] proposes an expressive and efficient 3D representation named triplanes for 3D generation. Triplane contains three orthogonal axis-aligned feature planes with a shape of $N \times N \times C$, where N and C denote the spatial resolution and the number of channels respectively. Given any 3D points $\mathbf{x} \in \mathbb{R}^3$, its feature can be extracted by projection and bi-linear lookups on triplanes. Then, we can decode the aggregated triplane features into neural fields, *i.e*., color and signed distance.

SMPL. Skinned Multi-Person Linear model (SMPL) [17] is a parametric human model that represents a wide range of human body poses and shapes. It defines a parameterized deformable mesh $\mathcal{M}(\beta, \theta)$, where a template mesh is deformed by linear blend skinning [13] with θ and β representing articulated pose and shape parameters. It provides an articulated geometric proxy to the underlying dynamic human body.

4. Method

Our goal is to generate animatable human avatars with full control over their camera views, body poses and shapes. To achieve this, we propose GETAvatar for explicit textured 3D human meshes generation with high-quality appearance and rich geometric details (*e.g*., clothing wrinkles and hairs). Different from previous 3D human generation methods [23, 2, 35, 9], GETAvatar adopts an explicit ariticulated 3D representation and thus supports 360° free-view, high-resolution (1024^2) and normal map rendering.

4.1. Overview

Framework. Given two latent codes z_{geo} and z_{tex} randomly sampled from Gaussian distribution, a camera parameter c consists of camera intrinsics and extrinsics, a SMPL [17] parameter $p = (\theta, \beta)$ that includes the human pose θ and shape β parameters, GETAvatar first generates a human avatar mesh with the specified body attributes p, and then synthesizes the corresponding RGB image, normal map, and foreground mask from the view defined by camera c. Here we define the 3D space corresponding to the target human representation with SMPL parameter p as the *deformed space*, and a *canonical space* with a body pose- and shape-independent template human representation.

We formulate the animatable human generation as a "*body deformation and explicit surface modeling*" process. The core idea is to first deform the implicit canonical human representation (triplane-based signed distance field) to the target pose and shape via body deformation, and then we model the body surface with an explicit mesh representation and a normal field in the deformed space. The overview of the proposed framework is shown in Fig. 2. Specifically, we first generate a shape- and pose-independent implicit human representation via two triplane branches [3] in the canonical space. To generate human avatar with the desired body shape and pose, we deform the implicit human representation from the canonical space to the deformed space with the guidance of the SMPL model (Sec. 4.2). To model the body surface with fine details, we extract the explicit 3D human body mesh from the signed distance fields in a differentiable manner [30], and improve the details of human body surface with a normal field (Sec. 4.3). After that, we render the generated human mesh into a 2D mask, normal map, and RGB image via an efficient differentiable surface renderer [12] (Sec. 4.4), and train the whole framework via adversarial training [11] (Sec.4.5).

4.2. Controllable 3D Human Modeling

Canonical Human Generation. GETAvatar is built upon 3D GANs [3, 6]. We model the geometry and texture of the canonical human with two separate triplane branches [3], allowing for the disentanglement of geometry and appearance (see Fig. 2). Given two latent codes z_{geo} and z_{tex} sampled from Gaussian distribution, the geometry and texture mapping networks produce two intermediate latents w_{geo} and w_{tex} to control the generation of the geometry and texture triplanes. For the geometry branch, we model the canonical human representation as a signed distance field (SDF). Specifically, given any 3D points in the canonical space, we query its feature from the geometry triplane, and adopt an MLP conditioned by w_{geo} to decode the queried feature as the signed distance value. However, directly modeling the signed distance field of clothed humans is challenging due to complex pose and shape variations. Therefore, instead of directly predicting the signed distance value, we predict a signed distance offset from the surface of the SMPL template [33]. For the texture branch, similar to the geometry branch, we use an MLP to map the queried texture triplane features to color value. In this process, we condition the MLP on both w_{geo} and w_{tex}, as the texture generation can also be influenced by changes in geometry.

SMPL-guided Deformation. To warp the generated canonical human to a desired pose θ and shape β, we establish a correspondence mapping between the canonical space and the deformed space. For any point $\mathbf{x_d}$ in the deformed space (θ, β), we aim to find its corresponding point $\mathbf{x_c}$ in the canonical space via the body deformation process. It

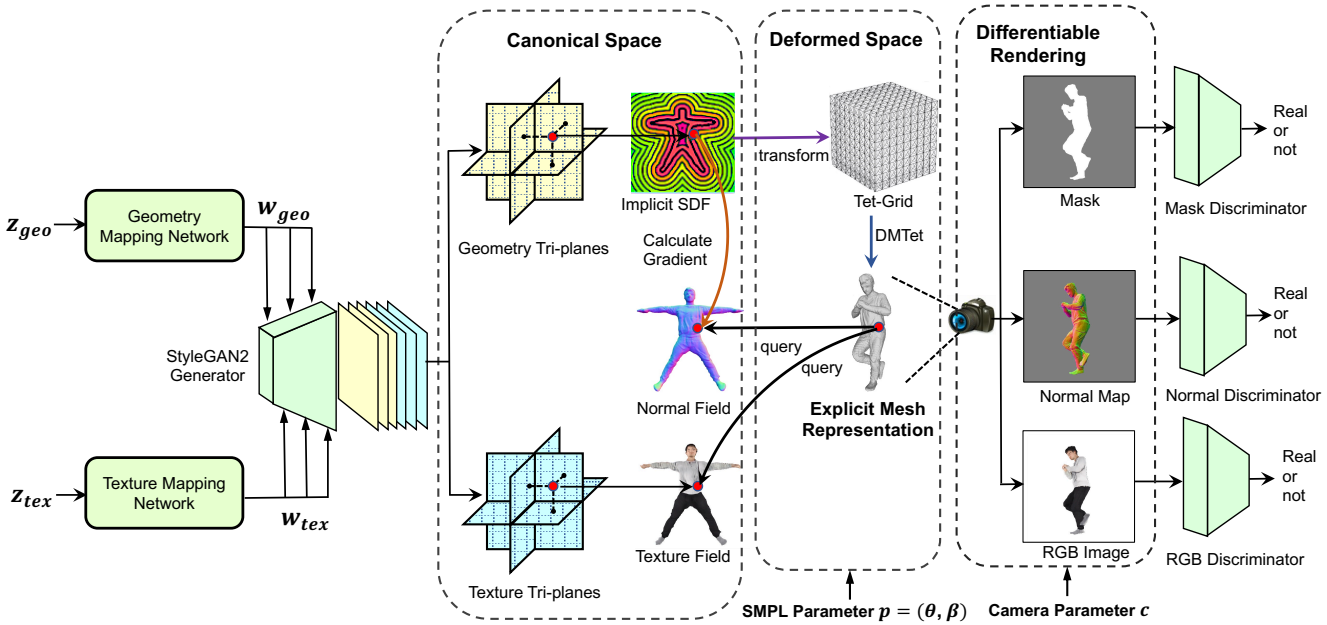

Figure 2: **The pipeline of GETAvatar. I. Generator.** Given latent codes z_{geo} and z_{tex} sampled from Gaussian distribution, GETAvatar generates the geometry triplanes T_{geo} and the texture triplanes T_{tex} via the StyleGAN2 generator backbone. **II. Canonical Space.** We model the canonical human representation with the signed distance, normal, and texture fields. **III. Deformed Space.** To generate the target human avatar defined by SMPL [17] parameter p, we use a SMPL-guided deformation to transform the canonical signed distance field into a deformable tetrahedral grid in the deformed space, and then employ the DMTet [30] to extract the underlying 3D mesh. For every point at the surface of the generated mesh, the color and the normal can be obtained by querying the normal field and texture field from the corresponding location in the canonical space. **IV. Differentiable Rendering.** To achieve the high-resolution image rendering, we adopt a differentiable rasterizer to render the 3D mesh into an RGB image, a normal map, and a mask from camera pose c. **V. Adversarial Training.** We use three discriminators [11] to classify whether the input RGB image, normal map, and mask are real or not.

is intuitive to exploit the 3D human model SMPL [17] as deformation guidance. Specifically, we generalize the linear blend skinning process [13] of the SMPL model from the coarse naked body to our generated clothed human. The core idea is to associate each point with its closest vertex on the deformed SMPL mesh $\mathcal{M}(\theta, \beta)$, assuming they undergo the same kinematic changes between the deformed and canonical spaces. Specifically, for a point $\mathbf{x_d}$ in the deformed space, we first find its nearest vertex v^* in the SMPL mesh. Then we use the skinning weights of v^* to un-warp $\mathbf{x_d}$ to $\mathbf{x_c}$ in the canonical space:

$$\mathbf{x_c} = \left(\sum_{i=1}^{N_j} s_i^* \cdot B_i(\theta, \beta) \right)^{-1} \cdot \mathbf{x_d}, \tag{1}$$

where $N_j = 24$ is the number of joints, s_i^* is the skinning weight of vertex v^* w.r.t. the i-th joint, $B_i(\theta, \beta)$ is the bone transformation matrix of join i. With the SMPL-guided body deformation process, we can deform the canonical human to any desired pose and shape, enabling controllable human generation.

4.3. Explicit Surface Modeling

Although the above pipeline can achieve controllable human generation, the resulting geometry produced by the implicit SDF is often noisy or over-smoothed (See 2nd row of Fig. 4), due to the lack of powerful geometry representation and insufficient geometric supervision. To resolve this issue, we propose utilizing an **explicit mesh representation** for the human geometry modeling, and a **normal filed** build upon this explicit representation for generating the fine geometric details, *e.g.*, hair, face, and cloth wrinkles.

To achieve explicit surface modeling in the deformed space, we extract the mesh of the generated human avatars under the desired poses and shapes through a differentiable surface modeling technique, *i.e.*, Deep Marching Tetrahedra (DMTet) [30]. Specifically, DMTet represents the surface of humans with a discrete signed distance field defined on a deformable tetrahedral grid, where a mesh face will be extracted if two vertices of an edge in a tetrahedron have different signs of SDF values. Here we transform the implicit SDF of canonical space to the deformable tetrahedral grid of deformed space via the SMPL-guided deformation process. Given any vertexes $\mathbf{x_d}$ in the tetrahedral grid under the deformed space, we first find its corresponding point $\mathbf{x_c}$ in canonical space using Eq. 1 and query its signed distance value $d(\mathbf{x_d})$ from the canonical SDF as $d(\mathbf{x_d}) = d(\mathbf{x_c})$. Then, we extract a triangular mesh of the generated human from the tetrahedral grid via the differentiable marching tetrahedra algorithm [30].

To achieve detailed geometric modeling of the generated humans, we further build a normal field on the extracted human mesh. The normal field depicts the fine-grained geometry of the clothed humans by associating each surface point with a normal vector, whose direction and magnitude represent the orientation and curvature of the surface at each point. Following IGR [7], we first construct a canonical normal field by calculating the spatial gradient of the canonical signed distance fields as:

$$n(\mathbf{x_c}) = \nabla_x d(\mathbf{x_c}), \tag{2}$$

where $d(\mathbf{x_c})$ and $n(\mathbf{x_c})$ are the signed distance value and normal vector for canonical point $\mathbf{x_c}$. Similarly, we transform the normal field from the canonical space to the deformed space via the SMPL-guided deformation process. For any points $\mathbf{x_d}$ at the extracted mesh surface in deformed space, we find $\mathbf{x_c}$ in canonical space via Eq. 1 and determine its surface normal vector $n(\mathbf{x_d})$ by:

$$n(\mathbf{x_d}) = \left(\sum_{i=1}^{N_j} s_i^* \cdot R_i(\theta, \beta) \right) \cdot n(\mathbf{x_c}), \tag{3}$$

where s_i^* and $R_i(\theta, \beta)$ are the skinning weights and rotation component of $B_i(\theta, \beta)$ in Eq. 1. With the explict surface normal modeling, we can further render the extracted human mesh into a 2D normal map, enabling the model to learning realistic surface details from the normal maps of 3D scans.

4.4. Efficient Differentiable Rendering

Existing 3D human GANs [23, 2, 35, 9] typically exploit implicit neural representation along with a neural volumetric rendering technique for 3D-aware human generation. However, such neural volumetric rendering is computationally inefficient and GPU memory intensive, making them hardly generate high-resolution images. Differently, our GETAvatar adopts an explicit mesh representation for human modeling, which support highly efficient rasterizer-based rendering [12], and thus can generate images up to 1024^2 resolution.

To render images, we first project the extracted mesh into a 2D mask and a coordinate map using the efficient rasterizer Nvdiffrast [12] given the camera parameter c. Each pixel on the coordinate map stores the corresponding 3D coordinates on the mesh surface. Then, for every pixel on the coordinate map, we un-warp its corresponding 3D coordinate back to the canonical space and query the color value, yielding high-resolution RGB images (see Fig. 2).

Additionally, our method can render normal maps directly from normal fields of the extracted meshes thanks to the differentiable surface modeling and rendering techniques. This enables us to perform adversarial training on normal maps, which helps capture high-frequency geometric details such as wrinkles, hair, and faces from the 2D normal maps of 3D scans, as demonstrated in our experiments.

4.5. Adversarial Training

We train our GETAvatar model from a collection of 2D human images with corresponding SMPL parameters $p = (\theta, \beta)$ and camera parameters c. We adopt a non-saturating GAN objective with $R1$ gradient penalty during the training process. To enable the model to learn 3D shapes, geometric details, and textures sufficiently, we use three StyleGAN2 discriminators [11] to perform adversarial training on 2D masks, normal maps, and RGB images individually [6]. To encourage the surface normal to be an unit 2-norm, we apply Eikonal loss [7, 24] on the surface of generated mesh:

$$\mathcal{L}_{eik} = \sum_{x_i} (||\nabla_x d(x_i)|| - 1)^2, \tag{4}$$

where x_i and $d(x_i)$ denote the surface point and its signed distance value. To eliminate internal geometry and floating faces, we follow [6, 20] to regularize the signed distance values of DMTet [30] using the cross-entropy loss:

$$\mathcal{L}_{ce} = \sum_{i,j \in \mathbb{S}_e} H(\sigma(d_i), sign(d_j)) + H(\sigma(d_j), sign(d_i)), \tag{5}$$

where H, σ, $sign$ denote the binary cross-entropy, sigmoid, sign function, respectively, and $\mathbb{S}_e$ is the set of edges where the signed distance values of two vertices have different signs. Therefore, the overall loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{adv} + \lambda_{eik}\mathcal{L}_{eik} + \lambda_{ce}\mathcal{L}_{ce}, \tag{6}$$

where $\lambda_{eik} = 0.001$, $\lambda_{ce} = 0.01$, and $\mathcal{L}_{adv}$ is the sum of adversarial losses on 2D masks, normal maps, and images.

5. Experiments

Datasets. We conduct experiments on two high-quality 3D human scan datasets: THUman2.0 [34] and RenderPeople [1]. THUman2.0 contains 526 high-quality human scans with a diverse range of body poses captured by a dense DSLR rig, and provides official SMPL fitting results for each scan. For RenderPeople, we utilize 1,190 scans with varying body shapes and textures to prepare the training images, and incorporated SMPL fits from AGORA dataset [25]. For every scan on these datasets, we employ Blender to render 100 RGB images, 2D silhouette masks, and normal maps with randomly-sampled camera poses.

Evaluation Metrics. To evaluate the visual quality and diversity of the generated RGB images, we compute Frechet Inception Distance [8] between 50k generated RGB images and all real RGB images: FID_{RGB}. We evaluate

Figure 3: Quatitative comparison between StyleSDF [24], ENARF [23], GNARF [2] and ours.

the geometry quality of generated human avatars from 3 aspects: the quality of surface details, the correctness of generated poses, and the plausibility of generated depth. First, to evaluate the quality of generated surface details, we measured Frechet Inception Distance [8] the normal maps: FID_{normal}, between 50k generated normal maps and all real normal maps. Second, to measure the correctness of generated poses, we employ the Percentage of Correct Keypoints (PCK) metric, as used in previous animatable 3D human generation methods [23, 2, 9, 35]. To compute PCK, we first use a human pose estimation model [5] to detect the human keypoints on both the generated and real images with the same camera and SMPL parameters. Then, we calculated the percentage of detected keypoints on the generated image within a distance threshold on the real image. Additionally, we evaluate the depth plausibility by comparing the generated depths with the pseudo-ground-truth depth estimated from the generated images by PIFuHD [28]. To assess the rendering speed, we report FPS running on a single NVIDIA V100 GPU.

Baselines. We compare our method against both state-of-the-art 3D-aware image synthesis [24, 6, 3] and 3D human generation [23, 9] methods. It is worth noting that 3D-aware image synthesis models cannot control the human pose in the generated images. Since all compared baselines cannot produce body surface normal maps directly due to their lack of explicit surface modeling, we implement extra post-processing steps to obtain the normal maps from their generated meshes. Specifically, we first reconstruct the 3D shapes from the density fields or SDFs using the marching cube algorithm [18], and then render their normal maps using a PyTorch3D rasterizer [27].

5.1. Results

Qualitative Results. We visualize the generated RGB images and normal maps for qualitative comparison, as shown in Fig. 3. From the results, we can make the following observations. StyleSDF [24] produces distorted body shapes due to its lack of an explicit human body representation. Even though ENARF [23] models the pose prior as a skeletal distribution, it still struggles to generate human avatars with correct target poses due to the inaccurate body deformation modeling. GNARF [2] can render reasonable RGB images but suffers from noisy geometries. Besides, it generates "floating noises" outside the human body with large pose articulations. Additionally, the geometries of generated human bodies are coarse and over-smoothed, lacking geometric details like clothes and hairs. Due to the high rendering cost and representation limitation, the rendered images from ENARF [23] and GNARF [2] have relatively low resolutions, resulting in low-quality rendering results. In contrast, as shown in the right column of Fig. 3, our method generates significantly better human avatars with detailed body geometries, even under large pose articulations like "sitting" and "single-leg standing". Please refer to the Appendix for more visualization results (*e.g.*, animation and multi-view videos).

Quantitative Evaluations. As shown in Tab. 1, GETAvatar consistently outperforms all the baseline methods [24, 6, 3, 23, 2] on both datasets [1, 34] w.r.t. all the met-

Method	Anim.	Res.	FPS ↑	THUman2.0				RenderPeople			
				FID_{RGB} ↓	FID_{normal} ↓	PCK ↑	Depth ↓	FID_{RGB} ↓	FID_{normal} ↓	PCK ↑	Depth ↓
StyleSDF [24]	✗	512^2	-	80.45	266.97	-	1.14	68.65	259.22	-	1.14
		1024^2		94.72	273.43	-	1.41	81.27	227.91	-	1.33
GET3D [6]	✗	512^2	-	73.71	136.53	-	1.16	37.95	124.28	-	1.13
		1024^2	-	65.77	134.53	-	0.92	42.63	106.84	-	0.85
EG3D [3]	✗	512^2	-	63.59	161.85	-	1.37	22.99	109.51	-	1.12
		1024^2	-	75.70	204.70	-	1.15	24.97	156.47	-	1.04
ENARF [23]	✓	128^2	8	124.61	223.72	82.08	1.37	108.59	205.26	75.66	1.26
GNARF [2]	✓	256^2	8	68.31	166.62	94.28	1.44	55.07	132.35	93.28	1.62
Ours	✓	512^2	**17**	**13.54**	**22.31**	**99.61**	0.83	12.65	**34.58**	**99.12**	0.92
		1024^2	14	17.91	55.02	99.39	**0.82**	**11.77**	58.57	98.99	**0.73**

Table 1: Quantitative comparisons with best results in **bold**. "Anim." represents whether the method is animatable or not, and "Res." denotes the image resolution.

rics. We observe that 3D-aware image generation models (StyleSDF [24], GET3D [6], and EG3D [3]) struggle to achieve reasonable FID scores on the THUman2.0 [34] due to its complexity and diverse poses. ENARF [23] and GNARF [2] can only generate relatively low-resolution images,*i.e.*, resolution of 128^2 and 256^2. In contrast, our model produces high-resolution human images (512^2 and 1024^2) with superior visual quality (FID_{RGB}), geometry quality (FID_{normal}), pose controllability (PCK), and depth plausibility (Depth). In terms of inference speed, GETAvatar generates 512^2 images at 17 FPS and 1024^2 images at 14 FPS, while ENARF [23] and GNARF [2] operate at 8 FPS for 128^2 and 256^2 images, respectively, verifying the efficiency of our method.

5.2. Ablation Studies

Geometry Modeling Scheme. We further conduct experiments on THUman2.0 [34] dataset to analyze the impact of different model designs. To begin with, we investigate the effects of SDF modeling, and find that directly predicting the sign distance value without incorporating the geometry prior of the SMPL template led to noticeable artifacts, *e.g.*, strange geometries in the waist region, as illustrated in the first row of Fig. 4. We then examine the effect of the normal field. The second row of Fig. 4 demonstrates that the generated human bodies contain noises and holes when the gradient-based normal field modeling is not included. We also explore an alternative method of modeling the surface normal by predicting the normals from an MLP [31, 15]. However, as shown in the third and fourth row of Fig. 4, we observe that the geometry quality of the predicted normal modeling method is inferior to our gradient-based approach. To further assess the effects of our geometry modeling scheme, we conduct quantitative experiments, the results of which are presented in Tab. 2. Both the quantitative and qualitative findings demonstrate that the normal field modeling and normal map supervision contribute to better geometry and appearance quality.

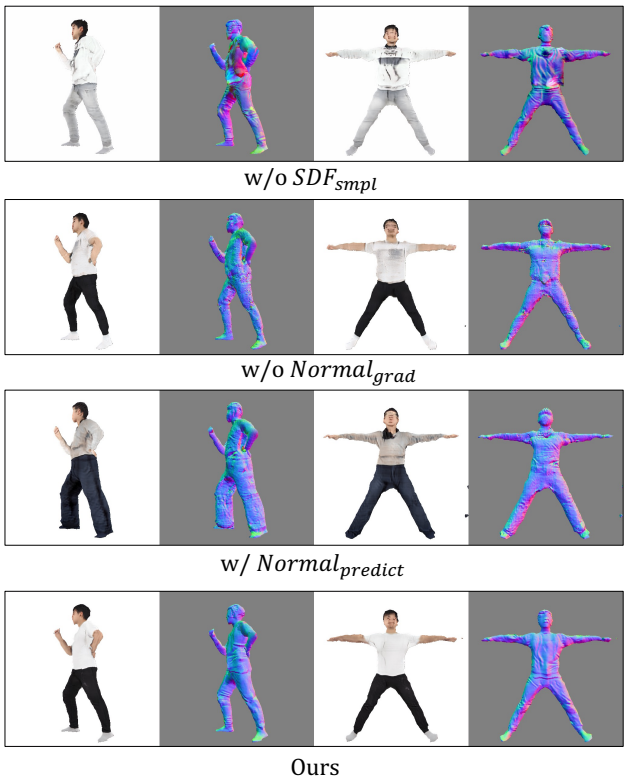

Figure 4: **Abalation on the model designs.** We visualize the human avatars with both deformed and canonical poses.

5.3. Applications

Transfer Learning. Benefiting from the explicit mesh representation, our method is ready for transfer learning. In practice, normal maps may not be available for in-the-wild

Method	$\text{FID}_{RGB} \downarrow$	$\text{FID}_{normal} \downarrow$	PCK $\uparrow$	Depth $\downarrow$
w/o SDF_{smpl}	15.80	29.52	99.47	0.92
w/o Normal_{grad}	20.23	63.49	99.25	0.96
w/ $\text{Normal}_{predict}$	16.93	89.38	99.11	0.88
Ours	**13.54**	**22.31**	**99.61**	**0.83**

Table 2: Ablation studies on geometry modeling scheme.

(a) Directly trained on DeepFashion.

(b) Pretrained on THUman2.0 and finetuned on DeepFashion.

Figure 5: **Transfer learning.** Compared to directly training on DeepFashion dataset [16], we find that pretraining on THUman2.0 [34] leads to much better geometries.

datasets, such as DeepFashion [16]. We observe that training directly on these datasets leads to degenerate results due to insufficient geometric supervision by learning from RGB images alone (see the first row of Fig. 5). To improve the generated geometries, one possible solution is to inherit the human geometry knowledge by learning from the 3D human datasets [16] via transfer training. We first pretrain the model on the THUman2.0 [34] containing 2D normal maps, and then fine-tune on the DeepFashion [16], leveraging both the rich geometry information of the 3D human dataset and the diverse texture information of the 2D fashion dataset. From Fig. 5, we observe that the transfer learning significantly improves the geometries on 2D fashion dataset [16].

Single-view 3D Reconstruction and Manipulation. GETAvatar enables the creation of full-body human avatars using a single-view portrait image. We adopt the optimization approach proposed in [10] to fit the given image. The optimization process involves using a frozen model with camera pose and SMPL parameters estimated from the portrait image, and minimizing the mean squared error (MSE) and perceptual loss [36] between the target and generated images. Once the optimization process is complete, the reconstructed portrait can be manipulated to different camera views and human poses, based on certain controlling signals as shown in Fig. 6.

Re-texturing. GETAvatar utilizes two separate triplane branches to represent the geometry and texture, allowing for the disentanglement of geometry and texture. This disentanglement enables the application of re-texturing by combining a shared geometry latent with various texture codes. As shown in Fig. 7, our approach can alter the textures of the generated humans by modifying different texture codes while preserving the underlying body geometries.

Figure 6: **Inversion.** Given a target image, we reconstruct its 3D human avatar and manipulate to novel camera views and novel poses.

Figure 7: **Retexturing.** GETAvatar supports changing the textures of generated humans while maintaining the underlying geometries of bodies.

6. Conclusion

In this work, we introduce GETAvatar, a 3D-aware generative model that directly generates explicit textured 3D meshes for controllable full-body human avatars. To enrich realistic surface details from 2D normal maps, we perform differentiable surface modeling by extracting the underlying surface as a 3D mesh and further building a normal field to depict the surface details. The explicit mesh representation enables us to achieve high-resolution rendering efficiently when combined with a rasterization-based renderer. Extensive experiments that GETAvatar achieves the state-of-the-art performance on 3D-aware human generation in terms of visual quality, geometry quality, and inference speed.

Limitations. Although our method can generate high-fidelity animatable 3D human avatars, there is still room for improvement. One limitation is that it lacks the ability to control fine motions of the human avatars, such as changes in facial expressions. To address this issue, we could consider using more expressive 3D human models, such as SMPL-X [26], as the guidance of deformation. For ethical considerations, the proposed method could be misused for generating fake imagery of real people, and we do not condone using our model with the intent of spreading disinformation.

References

[1] Renderpeople, 2020. https://renderpeople.com/.

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

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

[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 *CVPR*, 2021.

[5] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020.

[6] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In *NeurIPS*, 2022.

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

[8] 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. In *NeurIPS*, 2017.

[9] 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.

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

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

[12] 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)*, 2020.

[13] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000.

[14] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *CVPR*, 2020.

[15] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022.

[16] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.

[17] 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)*, 2015.

[18] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 1987.

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

[20] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *CVPR*, 2022.

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

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

[23] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *ECCV*, 2022.

[24] 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 *CVPR*, 2022.

[25] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *CVPR*, 2021.

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

[27] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

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

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

[30] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *NeurIPS*, 2021.

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

[32] Eric Zhongcong Xu, Jianfeng Zhang, Junhao Liew, Wenqing Zhang, Song Bai, Jiashi Feng, and Mike Zheng Shou. Pv3d: A 3d generative model for portrait video generation. In *ICLR*, 2023.

[33] Wang Yifan, Lukas Rahmann, and Olga Sorkine-Hornung. Geometry-consistent neural shape representation with implicit displacement fields. *arXiv preprint arXiv:2106.05187*, 2021.

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

[35] 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. *arXiv preprint arXiv:2211.14589*, 2022.

[36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[37] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Pan Pan, and Yi Yang. Multi-view consistent generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2022.