
Efficient 3D Articulated Human Generation with Layered Surface Volumes

Yinghao Xu^{1,2*}

yhxu@stanford.edu
xy119@ie.cuhk.edu.hk

Wang Yifan¹

yifan.wang@stanford.edu

Alexander W. Bergman¹

awb@stanford.edu

Menglei Chai³

mengleichai@google.com

Bolei Zhou⁴

bolei@cs.ucla.edu

Gordon Wetzstein¹

gordonwz@stanford.edu

¹**Stanford**

²**CUHK**

³**Google**

⁴**UCLA**

computationalimaging.org/publications/lsv/

Abstract

Access to high-quality and diverse 3D articulated digital human assets is crucial in various applications, ranging from virtual reality to social platforms. Generative approaches, such as 3D generative adversarial networks (GANs), are rapidly replacing laborious manual content creation tools. However, existing 3D GAN frameworks typically rely on scene representations that leverage either template meshes, which are fast but offer limited quality, or volumes, which offer high capacity but are slow to render, thereby limiting the 3D fidelity in GAN settings. In this work, we introduce layered surface volumes (LSVs) as a new 3D object representation for articulated digital humans. LSVs represent a human body using multiple textured mesh layers around a conventional template. These layers are rendered using alpha compositing with fast differentiable rasterization, and they can be interpreted as a volumetric representation that allocates its capacity to a manifold of finite thickness around the template. Unlike conventional single-layer templates that struggle with representing fine off-surface details like hair or accessories, our surface volumes naturally capture such details. LSVs can be articulated, and they exhibit exceptional efficiency in GAN settings, where a 2D generator learns to synthesize the RGBA textures for the individual layers. Trained on unstructured, single-view 2D image datasets, our LSV-GAN generates high-quality and view-consistent 3D articulated digital humans without the need for view-inconsistent 2D upsampling networks.

1 Introduction

High-quality 3D articulated digital human assets are becoming increasingly important for several industries, such as gaming, VR/AR, and social platforms. Manual authoring of these assets, however, is a laborious task that requires domain expertise and artistic skills. By automating the asset generation, generative 3D networks show great potential in facilitating this content creation process.

Immense progress has recently been seen in 3D-aware generative adversarial networks (GANs) [36, 7, 10, 8, 18, 38, 62, 40]. However, articulated humans synthesized by 3D GANs still suffer

*work done when Yinghao was a visiting student at Stanford University



Figure 1: Trained using unstructured, single-view image collections, such as StyleGAN-Human [13], our GAN framework leverages a new layered surface volume representation to generate high-quality 3D human bodies in a canonical pose (left), which can be rendered from different camera perspectives (center), and animated using articulated motion (right).

from limited diversity and quality [17, 5, 63, 52, 39, 23, 22]. These limitations can be primarily attributed to either the limited representational capacity or computational inefficiency of existing 3D network architectures. For instance, some recent approaches [17, 63, 52] generate 2D features using application-specific template meshes, such as SMPL [35] for human bodies. These templates, unfortunately, cannot adequately model fine details like hair, clothes, or accessories. On the other hand, 3D GANs utilizing volumetric representations [5, 39, 23, 22] have the potential to capture off-surface details, but the required volume rendering is often slow. This computational inefficiency can fundamentally limit the quality and diversity of the network since training a 3D GAN necessitates rendering tens of millions of images, which quickly becomes computationally infeasible. Consequently, upsampling networks have been widely adopted, but they often lead to significant degradation of generated shape quality and adversely impact multi-view consistency of the synthesized assets [18, 38, 62, 8].

In this work, we aim to combine the advantages of efficient template meshes with the high representational capacity of volumetric scene representations. To this end, we introduce the concept of surface volumes, i.e., volumetric manifolds with non-zero thickness centered around the surface of a template mesh. A surface volume encapsulates off-surface details and volumetric structures like hair or accessories that exist close to a template but are not adequately modeled by an infinitesimally thin surface. Contrary to conventional volumetric representations, surface volumes do not waste capacity and resources in empty space. More importantly, we can approximate these volumetric manifolds using a set of layered isosurfaces, each represented as an appropriately deformed version of the original template mesh. These layers are textured with color and transparency, allowing for fast rasterization instead of slow volumetric ray casting to render an image. The RGBA textures of our layered surface volumes (LSVs) can be synthesized using conventional 2D generators. Remarkably, rendering an LSV is so efficient that the GAN training is now bottlenecked by texture generation rather than neural rendering, eliminating the need for view-inconsistent upsampling networks. LSVs draw inspiration from multiplane images [70, 56] and manifolds [10, 59], but they are aligned with application-specific template meshes tailored to digital humans. By leveraging this novel representation in a 3D GAN setting, we demonstrate state-of-the-art quality, diversity, and multi-view consistency in generating articulated 3D humans on the DEEPFASHION [34] and StyleGAN-Human (SHHQ) [13] datasets.

2 Related Work

In this section, we briefly review the most relevant 3D generative approaches. For a recent survey of the larger field of neural scene representation and rendering, we refer to [55].

3D-aware Generative Models. Recent works on 3D GANs extend 2D image-based GANs [44, 25–27, 12, 13] by learning to generate 3D-aware multi-view-consistent objects or scenes from collections of unstructured, single-view 2D images in an unsupervised manner. The choice of neural scene representations has played a crucial role in the success of these 3D GANs. For example, some methods use meshes [53, 32], dense [58, 15, 71, 20, 36, 37, 61] or sparse [19, 46] voxel grids, 2D feature planes [8, 11, 49, 4, 50, 1] or manifolds [10, 59, 68], fully implicit networks [45, 7, 40, 69, 41, 51, 54, 47, 67, 9, 64], or a combination of low-resolution voxel grids combined with 2D CNN-based image upsampling layers [18, 38, 62, 60]. Very recently, diffusion models have also been explored as a platform for generating 3D objects or scenes [43, 57, 33, 48].

Among these approaches, 3D GANs building on 2D feature planes or manifolds produce state-of-the-art multi-view-consistent image quality, approaching photorealism. Ours is most closely related to these methods, but rather than uniformly distributing the limited capacity of the 2D feature manifolds in 3D space, the proposed LSV representation focuses its capacity on a thin volume that is aligned with the surface of a template mesh for digital humans.

Generating Articulated 3D Digital Humans. 3D-aware GANs have been proposed to generate 3D digital humans whose body pose can be explicitly controlled by a user after generating the identity. Many recent approaches in this category generate a feature volume using a global [5, 39, 23, 65] or local [22] triplane-based representation, which contains the human in a canonical pose and can be deformed using a target body pose. Another class of methods generates 2D textures or features on the surface of a human template mesh [17, 63, 52, 3]. Instead of generating only the appearance with a fixed template mesh, both shape and appearance can also be generated simultaneously [16]. Shape generation followed by text-guided texture optimization, for example using CLIP [21] or diffusion models [6, 66], is also an emerging topic for digital human generation. Note that several of these works are concurrently developed to ours, without public code available (e.g., [23, 65, 23, 6, 66, 3]).

Similar to many of these approaches, ours uses a template mesh for digital humans to enable post-generation articulation. Rather than using a single shell of such a template mesh, which limits its ability to represent hair, accessories, and other details, we introduce LSVs as a layered template mesh that combines the computational efficiency of mesh-based rendering with the flexibility of local volumes, allocated where needed, to represent fine detail.

3 Method

Existing 3D GANs for articulated humans [5, 39, 23, 22, 65] often employ inefficient voxel representations and volume rendering, hindering performance during GAN training. In this section, we first introduce our layered surface volumes (LSVs) and the associated fast rasterization-based rendering pipeline as a way to alleviate these shortcomings. We then discuss how to use LSVs as a backbone in a 3D GAN setting for generating digital humans.

Code and pre-trained checkpoints will be made public.

3.1 Layered Surface Volumes

Representation. We utilize a parametric mesh template to capture the generic shape of a human body, and leverage the pre-computed skinning weights and UV mapping to efficiently articulate and generate appearance of the human from texture maps.

In particular, we adopt SMPL [35], which models the vertex locations of a human template mesh $M = (V, F)$ for body pose $\theta \in \mathbb{R}^{3J}$ and shape $\beta \in \mathbb{R}^B$, where $V \in \mathbb{R}^{3V}$ and $F \in \mathbb{R}^{3F}$ denote the vertex coordinates and face indices of the mesh, and J and B denote the number of joints and shape bases, respectively. Formally, the SMPL model can be written as a mapping $V = M(\theta, \beta)$ with $M : \mathbb{R}^{|\theta|} \times \mathbb{R}^{|\beta|} \mapsto \mathbb{R}^{3V}$. M includes a shape-dependent articulation step based on linear blend skinning (LBS) [30] that deforms the mesh vertices from the T-pose to an arbitrary target pose:

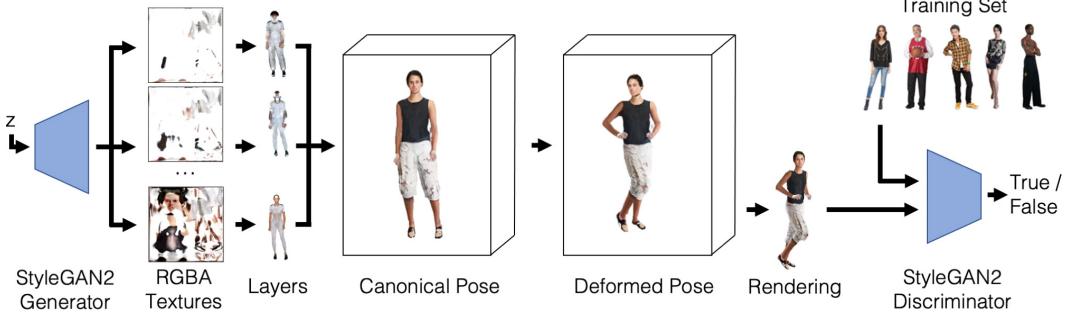


Figure 2: LSV-GAN pipeline. A latent code z is fed into a 2D StyleGAN2 generator network, which outputs N RGBA textures. These are applied to the individual mesh layers. All textured layers together are deformed into the target pose distribution and rendered using fast, differentiable rasterization before being fed into a camera- and body-pose-conditioned StyleGAN2 discriminator. An additional face discriminator is used but not shown.

$V = \text{LBS}(\mathbf{M}^T(\boldsymbol{\beta}), J(\boldsymbol{\beta}), \theta)$, where $\mathbf{M}^T(\boldsymbol{\beta})$ and $J(\boldsymbol{\beta})$ denote the T-pose mesh and the regressed joint locations, respectively. Given the fixed topology and its parametrization, the appearance of a SMPL shape can be modeled through texture maps in the 2D UV space. Each vertex \mathbf{v} is assigned to a unique position in the 2D texture map \mathbf{T} , i.e., $\{\mathbf{c}, o\} = \mathbf{T}(\text{UV}(\mathbf{v}))$, where \mathbf{c} and o are the retrieved color and opacity, UV defines the mapping from vertex to coordinates on the texture map, which is pre-computed and fixed for the SMPL mesh.

The key idea of LSVs is to augment the base SMPL mesh \mathbf{M}^T with a small numbers of SMPL meshes $\{\mathbf{M}_n^T\}_{n=1}^N$, namely *layers*, wrapping around the base mesh. Each layer is equipped with its own texture map $\mathbf{T}_n \in \mathbb{R}^{H \times W \times 4}$, encoding color $\mathbf{c} \in \mathbb{R}^3$ and opacity $o \in \mathbb{R}^{\geq 0}$, which are composited together in the rendering stage (detailed in the next section) to capture geometry variations that can not be covered by a single SMPL base mesh.

Since the mesh topology stays the same, we can obtain the colors \mathbf{c}_n and opacity o_n of each layer from the texture map \mathbf{T}_n using the same UV mapping. Let \mathcal{T} denote all layers in LSVs, and $[\mathbf{C}, \mathbf{O}]$ be color and opacity samples for vertex \mathbf{v} on all layers,

$$[\mathbf{C}, \mathbf{O}] = \mathcal{T}(\text{UV}(\mathbf{v})). \quad (1)$$

Similarly, the same skinning weights and joint regressor can be applied to simultaneously deform all mesh layers:

$$\mathbf{V}_n = \text{LBS}(\mathbf{M}_n^T(\boldsymbol{\beta}), J(\boldsymbol{\beta}), \theta). \quad (2)$$

To obtain layers from a base mesh, we keep the connected topology in SMPL and inflate (and shrink) the mesh along the vertex normals \mathbf{n} for a fixed thickness $t_n \in \mathbb{R}$:

$$\mathbf{v}_n = \mathbf{v} + t_n \mathbf{n}. \quad (3)$$

We obtain N SMPL layers $\{\mathbf{M}_n\}_{n=0}^{N-1}$, from the smallest to the largest, by setting $t_n = t_{\min} + n(t_{\max} - t_{\min})/(N - 1)$, with $t_{\max} = 0.01$ and $t_{\min} = -0.01$.

Rendering. One key advantage of our representation lies in its rendering efficiency. Instead of sampling the entire 3D volume hundreds of times along each ray, as done in volumetric rendering, we can apply differentiable rasterization [28], which is much more efficient and has been highly optimized in hardware-accelerated graphics pipelines. Specifically, to render an LSV model, we rasterize each layer independently. For each pixel \mathbf{p} on the final rendered image, the rasterizer finds the projected polygon (if any) with each layer, and evaluates the corresponding depth $z_n(\mathbf{p})$, $\mathbf{c}_n(\mathbf{p})$ and opacity $o_n(\mathbf{p})$ by interpolating values from the polygon vertices. In what follows, we omit the pixel index \mathbf{p} for the sake of readability.

We propose the following composition function to obtain the final color \mathbf{c} at each pixel:

$$\mathbf{c} = \sum_{n=1}^N w_n o_n \mathbf{c}_n, \quad (4)$$

where the compositing weights w_n depend on the relative depth of the layers and their opacity, similar to [29]:

$$w_n = \frac{o_n \exp(-o_n \bar{z}_n / \gamma)}{\sum_{n=1}^N o_n \exp(-o_n \bar{z}_n / \gamma)}, \text{ with } \bar{z}_n = \frac{z_n - \min_n(z_n)}{\max_n(z_n) - \min_n(z_n)}. \quad (5)$$

3.2 3D GAN Framework

An overview of our generation framework can be found in Fig. 2. In this section, we elaborate on how to efficiently adopt LSVs into a 3D GAN framework for articulated humans.

Generator. The textures of LSVs in our 3D GAN are generated by a StyleGAN2-based architecture, which is tasked with generating color and opacity values for each layer, leading to a total of $4N$ output channels, instead of 3. The generator does not use camera pose or human pose conditioning, to prevent the rendered images from being overly dependent on the input view and body pose, which helps to ensure that the generated textures are robust across different camera views and human poses when performing animation or camera movement.

Discriminator. Our framework also leverages the discriminator $D(\cdot)$ of StyleGAN2 for adversarial training. Like EG3D [8] and GNARF [5], the discriminator is conditioned on camera and body poses, which enforces the synthesized images to be well-aligned with the given camera and pose condition instead of just being in the correct distribution.

Face Discriminator. Given that the face occupies a small portion of the rendered image, the discriminator provides weak learning signal to the face region, which often times leads to inferior face quality. We use the joints of SMPL to estimate a coarse bounding box of the face and crop the face patch from the whole frame. A face discriminator $D_{\text{face}}(\cdot)$ is then introduced on the cropped face patch to help the texture generator learn better facial details.

Progressive Training. High-fidelity texture maps are critical to the final rendering quality. However, synthesizing high-resolution textures poses a significant challenge to the generator. To address this issue, we adopt a progressive strategy for our GANs training. Unlike previous approaches that progressively add new blocks to both the generator and discriminator [24], our method fixes the architecture of the texture generator while gradually increasing the resolution of the rendered image. This approach allows us to reuse the generator’s parameters without requiring re-optimization when the resolution changes, resulting in more stable training. With our progressive training, the generator initially learns to synthesize coarse textures from low-resolution images, which then serve as good initialization for optimizing the textures into high-quality ones with fine-grained details.

Hand Regularization. Rendering realistic hands is another challenging task, particularly because the SMPL model cannot accurately simulate the distribution of real hands in the datasets (see supplement). Even if we deform the hand mesh, it cannot perfectly fit the actual hand pose and shape, resulting in artifacts such as translucent fingers. To overcome this issue, we reduce the deformation scale of the hand and process the hand textures T_{hand} independently using the UV atlas. To prevent the texture of the hand from learning to be transparent, we regularize the alpha map of the hand to be as opaque as possible with l_1 loss $\mathcal{L}_{\text{hand}} = |\mathbf{O}_{T_{\text{hand}}} - 1|$. With this regularizer, we encourage hand textures to produce realistic and coherent results.

Training Details. We first sample SMPL pose parameters $\mathbf{p} = (\beta, \gamma)$ from the dataset and generate the mesh layers according to our LSV representation. Next, we use a differentiable rasterizer [28] to render each textured isosurface into 2D images \mathbf{I}_f and compose them as mentioned before. The entire image generation process is formulated as $\mathbf{I}_f = G(\mathbf{p}, \mathbf{z}, \xi)$, where the generator $G(\cdot)$ takes as input SMPL parameters \mathbf{p} , a latent code \mathbf{z} sampled from $\mathcal{N}(0, 1)$, and a camera pose ξ to synthesize the image \mathbf{I}_f . During training, we randomly sample \mathbf{p} , \mathbf{z} , and ξ , while the real image \mathbf{I}_r is sampled from the dataset. Operating on the output of the generator, we use a discriminator $D(\cdot)$ to guarantee the global coherence of the rendered human and a face discriminator $D_{\text{face}}(\cdot)$ on cropped faces to improve face fidelity.

The generator and discriminators are jointly trained using \mathcal{L}_{GAN} , the non-saturating GAN loss with R1 regularization on both discriminators. Additionally, we use the aforementioned hand regularizer with weight λ to improve the realism of hands. The overall loss function is defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GAN}} + \lambda \mathcal{L}_{\text{hand}}$.

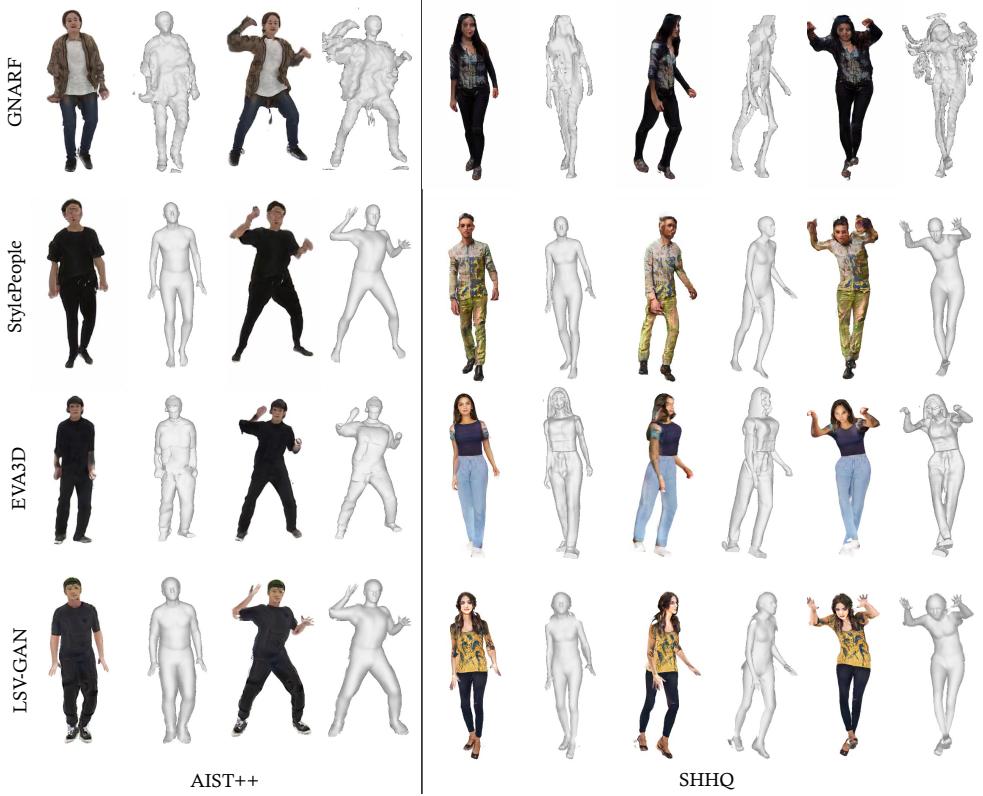


Figure 3: Qualitative results and comparisons. We compare the results of several baselines, including GNARF, a representative implementation of StylePeople, and EVA3D, with our LSV-GAN using the AIST++ and SHHQ datasets. In each case, we show an image and the shape rendered from the generated, canonical pose on the left in addition to one or two additional deformed body poses on the right. Our approach generates high-quality 3D humans with more detailed faces and more accurate shapes than the baselines.

4 Experiments

4.1 Settings

We first outline the settings of our experiments before evaluating the proposed LSV-based 3D GAN framework for articulated human generation. More implementation details can be found in the supplement.

Datasets. We evaluate LSV-GAN on three human datasets: AIST++ [31], DEEPFASHION [34], and SHHQ [14]. AIST++ is a large dataset consisting of 10.1M images covering 30 different performers in 9 camera views. Each frame is annotated with a camera pose and fitted SMPL body poses. We filter out the noisy samples with inaccurate SMPL annotations and collect a subset of 360k images. We also use the annotated bounding box to perform center cropping and then resize all images to a resolution of 512×512 . DEEPFASHION and SHHQ are single-view image datasets consisting of 8k and 40k identities, respectively. We adopt SMPLify-X [42] to estimate SMPL parameters and camera parameters. All images from these two datasets are resized to 512×256 for GAN training.

Baselines. We compare LSV-GAN with several baselines: EG3D [8] and StyleSDF [40] are state-of-the-art methods for 3D-aware object synthesis; ENARF [39] and GNARF [8] are methods which perform deformation on triplane representation to achieve articulated human generation; and EVA3D [22] is a method which uses a compositional signed distance function for articulated human generation. We also include a comparison with our unofficial implementation of StylePeople [17], combining a single mesh layer and a neural feature decoder for human generation. We implement

Table 1: Quantitative evaluation. We compare several baselines (left) using three different datasets. The quality and diversity, as measured by the FID score, are best for our LSV-GAN for the larger DEEPFASHION and SHHQ datasets. Multi-view consistency is evaluated using the PCK metric; our approach consistently outperforms baselines in this metric. The training time (TR., measured in days on a single A6000 GPU) is the lowest for our method among all the high-resolution GANs operating at a resolution of 512^2 . The rendering time at inference (INF., measured in ms/image) is by far the lowest for our approach. * numbers adopted from [22]; \dagger representative implementation of 2D texture generation on SMPL template mesh with feature-based upsampling, such as StylePeople.

Model	AIST++		DEEPFASHION		SHHQ		Comp. Cost	
	FID \downarrow	PCK \uparrow	FID \downarrow	PCK \uparrow	FID \downarrow	PCK \uparrow	TR. \downarrow	INF. \downarrow
EG3D* (512^2)	34.76	—	26.38	—	32.96	—	56	38
StyleSDF* (512^2)	199.5	—	92.40	—	14.12	—	65	32
ENRAF* (128^2)	73.07	42.85	77.03	43.74	80.54	40.17	5	104
GNARF (512^2)	11.13	96.11	33.85	97.83	14.84	98.96	24	72
EVA3D* (512^2)	19.40	83.15	<u>15.91</u>	87.50	<u>11.99</u>	88.95	40	200
StylePeople \dagger (1 layer, 512^2)	18.97	96.96	17.72	98.31	14.67	98.58	<u>20</u>	28
LSV-GAN (12 layers, 512^2)	<u>17.05</u>	98.95	12.02	99.47	11.10	99.44	<u>20</u>	28

StylePeople with our 1-layer surface volume rasterized at a resolution of 128×128 and then adopt a $4 \times$ upsampler [8] to get the 512×512 output².

Metrics. We employ the Fréchet Inception Distance (FID) score to assess the quality and diversity of our generated images. Specifically, we compute the FID score between 50,000 generated samples and all real images. In addition, we use the Percentage of Correct Keypoints (PCK@0.5) [2] to evaluate the quality of animations and view-consistency of generated results. We use 5,000 samples to evaluate this, following the protocol in EVA3D.

4.2 Results

Qualitative Evaluation. Fig. 3 compares our method with all baselines at a resolution of 512×256 . We synthesize the images for all methods with the same SMPL pose for a fair comparison. When using GNARF, which is trained at a low resolution and uses an image-based upsampler to achieve the target resolution, we observe inconsistency when synthesizing the same person with a variety of poses. StylePeople also performs rasterization to map the neural textures onto a single-layer surface and then uses a neural decoder to generate human images. This method achieves a limited quality and its decoder network, which models complex hair and accessories in 2D image space, introduces view-inconsistent artifacts. While EVA3D is capable of generating high-quality images, the rendering of hands is not very detailed and we also witness blending artifacts when the arms are close to the body. Additionally, rendering humans from a side view and performing large animations on EVA3D can result in artifacts on the face, hair, and hands, which is likely due to inaccurate geometry. In comparison, our method generates 3D human images with high-fidelity appearance and holds better 3D consistency across different body poses. Moreover, our model is able to handle challenging cases with large camera or body motion. More detailed comparisons can be found in the supplementary material.

Quantitative Evaluation. As shown in Tab. 1, LSV-GAN consistently outperforms baselines in terms of all quantitative metrics. EG3D and StyleSDF are not designed to handle the large diversity of body poses in the training data and humans generated with these approaches cannot be articulated. ENARF is trained at a low resolution, heavily relying on a view-inconsistent upsampling network, which results in low image quality. In comparison to methods that adopt neural radiance fields (GNARF), our LSVs can directly generate high-resolution images without using any upsamplers or neural decoders, leading to superior image quality and multi-view consistency. The large computational overhead of volumetric ray casting in EVA3D makes their training and inference cost much larger than other methods, despite good performance on image quality. Moreover,

²The official repository of StylePeople only contains inference code. We implement and train it on new datasets for a fair comparison.



Figure 4: Latent code interpolation of our approach trained on SHHQ.

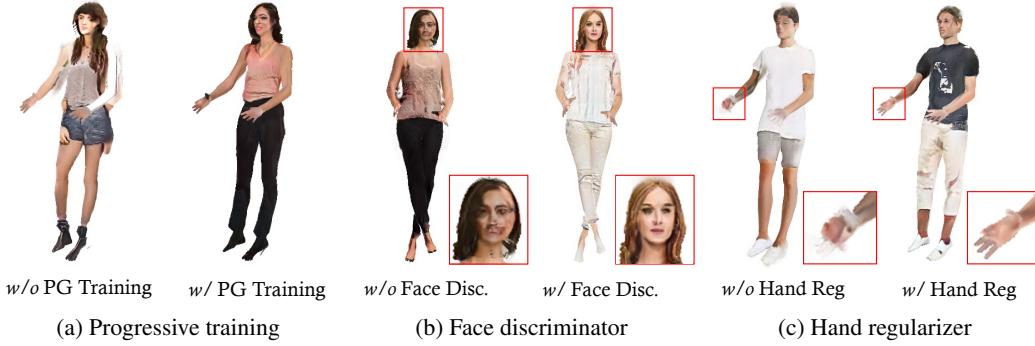


Figure 5: Qualitative comparison for ablations on progressive training (a), face discriminator (b), and hand regularizer (c).

the compositional representation used by EVA3D can lead to inconsistencies between the conditioned SMPL pose and the synthesized humans, as reflected by the PCK metric. In contrast, the PCK scores of our method are nearly 100% across all datasets, demonstrating excellent multi-view consistency. StylePeople also achieves high PCK scores, which are slightly impacted by the employed 2D feature decoder. Additionally, our model can render full images without the need for an upsampling network, while maintaining better or comparable training and inference efficiency compared to other baselines operating at the same resolution. This is made possible by our layered surface volumes representation, which uses fast rasterization instead of slow volumetric rendering. Note that the metric values of all models denoted with * in Tab. 1 are adopted from [22]. Using the same evaluation procedure, we trained and evaluated the GNARF and StylePeople baselines as well as our method from scratch.

Latent Code Interpolation. In Fig. 4, we show example renderings of the interpolation between the latent codes of three different identities in the rest pose. This experiment validates the high quality of the latent space learned by LSV-GAN.

4.3 Ablation Study

We ablate the main components of LSV-GAN to better understand their individual contributions. Besides the FID score, we also include another metric, FID_{face} , to evaluate the quality of generated faces. All ablations are performed on SHHQ with the same training schedule.

Number of Surface Layers. The number of layers of our surface volumes is a crucial factor in synthesizing realistic humans using our framework, as shown in Tab. 2a. The single-layer case denotes a very basic setting where only the textured base SMPL layer is rasterized to render a human image. However, the SMPL mesh does not account for deformed clothes, hair, and accessories, resulting in rendered images that deviate significantly from the real dataset distribution and produce poor FID and FID_{face} scores. As we increase the number of layers, the quality and diversity of both images and faces improve significantly, as shown by the decreasing FID and FID_{face} scores. As we

Table 2: Ablation study					
(a) Layers of Surface Volume			(b) Ablations on LSV-GAN components		
#LSV	FID	FID _{face}	Model	FID	FID _{face}
1	101.3	124.3	N-layer base	12.5	32.3
3	20.5	31.9	+ progressive training	11.8	29.6
6	15.8	27.7	+ face discriminator	11.0	24.7
12	11.1	24.6	+ hand regularizer	11.1	24.6

show in the supplement, layered surface volumes can capture details and volumetric structures, such as hair and clothes, while a single-layer volume tends to generate very thin people and struggles to synthesize realistic human images.

Progressive Training. In Fig. 5a, we present a visual comparison between our model and the one trained without progressive training. We observed that training the model at full resolution often leads to “white texture” artifacts, where the textures around the border of the arm or leg tend to learn the background color using opaque surface volumes to deceive the discriminator. Progressive training, which starts at a low resolution, enables the optimization of a coarse texture initially. This is beneficial for initializing a high-fidelity texture map later and also helps alleviate “white texture” artifacts. Tab. 2b shows that progressive training also achieves better image quality than the base model, primarily due to the improvement of the texture quality.

Face Discriminator. In Tab. 2b, we present the effects of face discrimination. We observe a significant improvement in face quality when face discrimination is applied. The overall FID score also improves thanks to the good face fidelity. The visual samples in Fig. 5b demonstrate that the model with face discriminator models hair as well as facial features with greater detail.

Hand Regularizer. We also ablate the hand regularizer to study its effects. As shown in Fig. 5c, certain parts of the fingers are learned to be translucent to simulate complex hand poses without the hand regularization. However, when utilizing a texture atlas to model hand textures independently and regularizing the alpha channel to be opaque, the rendered hands appear more natural. Nevertheless, the quantitative results show a very slight drop in performance as the rendered hands do not fit the distribution of the real data as accurately, as shown in Tab. 2b. We nevertheless prefer using the hand regularizer, as it leads to more natural-looking results.

5 Discussion

Limitations and Future Work. Our work is limited in several ways. Although the quality, diversity, and view consistency of results generated with our approach are quantitatively better than the baselines, the level of detail for all 3D human GAN approaches is still relatively low. This limitation is primarily due to the limited resolution of 512^2 , which simply provides too few pixels for important body parts, such as faces. The resolution of 3D human GANs should be significantly improved, which could potentially be achieved using LSVs in combination with texture atlases or varying levels of detail in the generated textures. Accurate hand pose estimation from in-the-wild training images is challenging and inaccurate, which also degrades the quality and diversity of generated humans. Better pose estimation algorithms would help alleviate this issue. Although the textured layers generated by our method could in principle be directly imported into conventional graphics pipelines, we did not explore this direction. Finally, while the linear blend skinning approach used to animate our humans is fairly standard, it does not enable the realistic motion of hair, clothes, or other accessories. Combining LSVs with differentiable physical simulation engines could be an interesting avenue of future research.

Ethical Considerations. GANs, such as ours, could be misused for generating edited imagery of real people. Such misuse of image synthesis techniques poses a societal threat, and we do not condone using our work with the intent of spreading misinformation or tarnishing reputation. We also

recognize a potential lack of diversity in our results, stemming from implicit biases of the datasets we process.

Conclusion. We propose layered surface volumes (LSVs), a novel 3D representation for articulated digital humans, which combines the advantages of efficient template meshes with the high representational capacity of volumetric scene representations. Integrated with a 2D generator network architecture, our LSV-GAN overcomes the computational burden of neural volume rendering by leveraging fast rasterization, and is able to generate high-quality and view-consistent 3D articulated digital humans without the need for view-inconsistent 2D upsampling networks. These and other benefits of our framework enable us to take an important step towards generating photorealistic 3D digital human assets that can be articulated, which is a capability vital to the visual effects industry, virtual or augmented reality systems, and teleconferencing among other applications.

Acknowledgements. We thank Thabo Beeler, Sida Peng, Jianfeng Zhang, Fangzhou Hong, Ceyuan Yang for fruitful discussions and comments about this work.

References

- [1] S. An, H. Xu, Y. Shi, G. Song, U. Ogras, and L. Luo. Panohead: Geometry-aware 3d full-head synthesis in 360°. *arXiv preprint arXiv:2303.13071*, 2023.
- [2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.
- [3] S. Aneja, J. Thies, A. Dai, and M. Nießner. ClipFace: Text-guided Editing of Textured 3D Morphable Models. In *ArXiv preprint arXiv:2212.01406*, 2022.
- [4] M. A. Bautista, P. Guo, S. Abnar, W. Talbott, A. Toshev, Z. Chen, L. Dinh, S. Zhai, H. Goh, D. Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. *Advances in Neural Information Processing Systems*, 35:25102–25116, 2022.
- [5] A. W. Bergman, P. Kellnhofer, W. Yifan, E. R. Chan, D. B. Lindell, and G. Wetzstein. Generative neural articulated radiance fields. In *NeurIPS*, 2022.
- [6] Y. Cao, Y.-P. Cao, K. Han, Y. Shan, and K.-Y. K. Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models, 2023.
- [7] E. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein. pi-GAN: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2021.
- [8] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. Guibas, J. Tremblay, S. Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022.
- [9] A. Chen, R. Liu, L. Xie, Z. Chen, H. Su, and J. Yu. Sofgan: A portrait image generator with dynamic styling. *ACM Transactions on Graphics (TOG)*, 41(1):1–26, 2022.
- [10] Y. Deng, J. Yang, J. Xiang, and X. Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *CVPR*, 2022.
- [11] T. DeVries, M. A. Bautista, N. Srivastava, G. W. Taylor, and J. M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *CVPR*, pages 14304–14313, 2021.
- [12] A. Frühstück, K. K. Singh, E. Shechtman, N. J. Mitra, P. Wonka, and J. Lu. Insetgan for full-body image generation. In *CVPR*, pages 7723–7732, 2022.
- [13] J. Fu, S. Li, Y. Jiang, K.-Y. Lin, C. Qian, C. C. Loy, W. Wu, and Z. Liu. Stylegan-human: A data-centric odyssey of human generation. In *ECCV*, pages 1–19, 2022.
- [14] J. Fu, S. Li, Y. Jiang, K.-Y. Lin, C. Qian, C. C. Loy, W. Wu, and Z. Liu. Stylegan-human: A data-centric odyssey of human generation. In *ECCV*, 2022.
- [15] M. Gadelha, S. Maji, and R. Wang. 3D shape induction from 2D views of multiple objects. In *3DV*, 2017.
- [16] J. Gao, T. Shen, Z. Wang, W. Chen, K. Yin, D. Li, O. Litany, Z. Gojcic, and S. Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In *NeurIPS*, 2022.

- [17] A. Grigorev, K. Iskakov, A. Ianina, R. Bashirov, I. Zakharkin, A. Vakhitov, and V. Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *CVPR*, pages 5151–5160, 2021.
- [18] J. Gu, L. Liu, P. Wang, and C. Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. In *Int. Conf. Learn. Represent.*, 2022.
- [19] Z. Hao, A. Mallya, S. Belongie, and M.-Y. Liu. GANcraft: Unsupervised 3D neural rendering of minecraft worlds. In *ICCV*, 2021.
- [20] P. Henzler, N. J. Mitra, and T. Ritschel. Escaping Plato’s cave: 3D shape from adversarial rendering. In *ICCV*, 2019.
- [21] F. Hong, M. Zhang, L. Pan, Z. Cai, L. Yang, and Z. Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022.
- [22] F. Hong, Z. Chen, Y. LAN, L. Pan, and Z. Liu. EVA3d: Compositional 3d human generation from 2d image collections. In *ICLR*, 2023.
- [23] S. Jiang, H. Jiang, Z. Wang, H. Luo, W. Chen, and L. Xu. Humangen: Generating human radiance fields with explicit priors. In *CVPR*, 2023.
- [24] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- [25] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.
- [26] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020.
- [27] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021.
- [28] S. Laine, J. Hellsten, T. Karras, Y. Seol, J. Lehtinen, and T. Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020.
- [29] C. Lassner and M. Zollhofer. Pulsar: Efficient sphere-based neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1440–1449, 2021.
- [30] J. P. Lewis, M. Cordner, and N. 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*, pages 165–172, 2000.
- [31] R. Li, S. Yang, D. A. Ross, and A. Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021.
- [32] Y. Liao, K. Schwarz, L. Mescheder, and A. Geiger. Towards unsupervised learning of generative models for 3D controllable image synthesis. In *CVPR*, 2020.
- [33] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022.
- [34] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.
- [35] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.
- [36] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y.-L. Yang. HoloGAN: Unsupervised learning of 3d representations from natural images. In *ICCV*, 2019.
- [37] T. Nguyen-Phuoc, C. Richardt, L. Mai, Y.-L. Yang, and N. Mitra. BlockGAN: Learning 3D object-aware scene representations from unlabelled images. In *NeurIPS*, 2020.
- [38] M. Niemeyer and A. Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021.
- [39] A. Noguchi, X. Sun, S. Lin, and T. Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *ECCV*, pages 597–614, 2022.

- [40] R. Or-El, X. Luo, M. Shan, E. Shechtman, J. J. Park, and I. Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *CVPR*, 2022.
- [41] X. Pan, X. Xu, C. C. Loy, C. Theobalt, and B. Dai. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. In *NeurIPS*, 2021.
- [42] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.
- [43] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- [44] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [45] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger. GRAF: Generative radiance fields for 3d-aware image synthesis. In *NeurIPS*, 2020.
- [46] K. Schwarz, A. Sauer, M. Niemeyer, Y. Liao, and A. Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *arXiv preprint arXiv:2206.07695*, 2022.
- [47] Z. Shi, Y. Shen, J. Zhu, D.-Y. Yeung, and Q. Chen. 3d-aware indoor scene synthesis with depth priors. In *ECCV*, pages 406–422. Springer, 2022.
- [48] J. R. Shue, E. R. Chan, R. Po, Z. Ankner, J. Wu, and G. Wetzstein. 3d neural field generation using triplane diffusion. In *CVPR*, 2023.
- [49] I. Skorokhodov, S. Tulyakov, Y. Wang, and P. Wonka. Epigraf: Rethinking training of 3d gans. *arXiv preprint arXiv:2206.10535*, 2022.
- [50] M. Son, J. J. Park, L. Guibas, and G. Wetzstein. Singraf: Learning a 3d generative radiance field for a single scene. *arXiv preprint arXiv:2211.17260*, 2022.
- [51] J. Sun, X. Wang, Y. Zhang, X. Li, Q. Zhang, Y. Liu, and J. Wang. Fenerf: Face editing in neural radiance fields. In *CVPR*, 2022.
- [52] J. Sun, X. Wang, L. Wang, X. Li, Y. Zhang, H. Zhang, and Y. Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *CVPR*, 2023.
- [53] A. Szabó, G. Meishvili, and P. Favaro. Unsupervised generative 3D shape learning from natural images. *arXiv preprint arXiv:1910.00287*, 2019.
- [54] A. Tewari, X. Pan, O. Fried, M. Agrawala, C. Theobalt, et al. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *CVPR*, 2022.
- [55] A. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, W. Yifan, C. Lassner, V. Sitzmann, R. Martin-Brualla, S. Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, volume 41, pages 703–735. Wiley Online Library, 2022.
- [56] R. Tucker and N. Snavely. Single-view view synthesis with multiplane images. In *CVPR*, 2020.
- [57] H. Wang, X. Du, J. Li, R. A. Yeh, and G. Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022.
- [58] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016.
- [59] J. Xiang, J. Yang, Y. Deng, and X. Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. *arXiv preprint arXiv:2206.07255*, 2022.
- [60] Y. Xu, M. Chai, Z. Shi, S. Peng, I. Skorokhodov, A. Siarohin, C. Yang, Y. Shen, H.-Y. Lee, B. Zhou, et al. Discoscene: Spatially disentangled generative radiance fields for controllable 3d-aware scene synthesis. *arXiv preprint arXiv:2212.11984*, 2022.
- [61] Y. Xu, S. Peng, C. Yang, Y. Shen, and B. Zhou. 3d-aware image synthesis via learning structural and textural representations. In *CVPR*, 2022.

- [62] Y. Xue, Y. Li, K. K. Singh, and Y. J. 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.
- [63] Z. Yang, S. Li, W. Wu, and B. Dai. 3dhumangan: Towards photo-realistic 3d-aware human image generation. *arXiv preprint arXiv:2212.07378*, 2022.
- [64] J. Zhang, E. Sangineto, H. Tang, A. Siarohin, Z. Zhong, N. Sebe, and W. Wang. 3d-aware semantic-guided generative model for human synthesis. In *ECCV*, pages 339–356. Springer, 2022.
- [65] J. Zhang, Z. Jiang, D. Yang, H. Xu, Y. Shi, G. Song, Z. Xu, X. Wang, and J. Feng. Avatargen: A 3d generative model for animatable human avatars. *ArXiv*, 2023.
- [66] L. Zhang, Q. Qiu, H. Lin, Q. Zhang, C. Shi, W. Yang, Y. Shi, S. Yang, L. Xu, and J. Yu. Dreamface: Progressive generation of animatable 3d faces under text guidance. *arXiv preprint arXiv:2304.03117*, 2023.
- [67] X. Zhang, Z. Zheng, D. Gao, B. Zhang, P. Pan, and Y. Yang. Multi-view consistent generative adversarial networks for 3d-aware image synthesis. In *CVPR*, pages 18450–18459, 2022.
- [68] X. Zhao, F. Ma, D. Güera, Z. Ren, A. G. Schwing, and A. Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *ECCV*, pages 18–35. Springer, 2022.
- [69] P. Zhou, L. Xie, B. Ni, and Q. Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021.
- [70] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM. Trans. Graph. (SIGGRAPH)*, 2018.
- [71] J.-Y. Zhu, Z. Zhang, C. Zhang, J. Wu, A. Torralba, J. B. Tenenbaum, and W. T. Freeman. Visual object networks: image generation with disentangled 3D representations. In *NeurIPS*, 2018.

A Single Scene Overfitting

In addition to the layer ablation experiment described in the main paper for the GAN setting, we also perform a single-scene overfitting experiment. We use a textured model downloaded from SketchFab and rendered 400 360-degree views of the model. We use 300 views as training data and the remaining images as testing data. We apply the same training settings as described in the main paper and present our method’s results with varying numbers of layers. The last two columns in the presented table show the results of InstantNGP and the ground truth. To create the mesh layers, we first fitted a SMPL model. However, as the ground-truth mesh had a rather cartoonish body proportion, we manually refined the fitted SMPL model in Blender to approximately match the ground-truth mesh.



Figure A1: Single-scene overfitting result. We show the results of our method with different number of layers. The last two columns are the results of InstantNGP and ground truth.

B Implementation Details

B.1 Generator

Our approach employs the generator architecture of StyleGAN2 [26], which consists of two components: a mapping network and a convolutional backbone. The generator takes a 512-dimensional Gaussian noise input and conditions it using an eight-layer mapping network of 512



Figure A2: Visualization of real images and corresponding estimated SMPL mesh.

hidden units. We do not condition the generator on camera pose or body pose. The mapping network produces a 512-dimensional latent code, which modulates the layers of the StyleGAN2 convolutional backbone. The resulting output is a high-resolution image with 48 channels at 1024×1024 resolution. To facilitate further processing, we reshape this output into 12 texture planes consisting of RGB and alpha channels, each of shape $1024 \times 1024 \times 4$. Our architecture is trained from scratch, without using any pretrained networks.

B.2 Discriminator

In contrast to EG3D [8] and GNARF [5], our framework does not use a dual discriminator because we do not use upsampling to produce the final output images. Instead, we condition the discriminator on the expected body pose by including the body pose parameters in addition to the camera parameters as input to the mapping network. This allows the discriminator to ensure that the applied deformation matches the specified pose. To ensure stable training, we add 0.5 standard deviation of Gaussian noise to the body pose parameters before passing them to the discriminator. This prevents the discriminator from overfitting to specific poses and cameras in the ground truth data.

B.3 Training Details

We defaultly use 12 layers of the surface volume for generation and the deformation scale is ranges from 0 to 0.05. We use the Adam optimizer for both the generator and discriminator during optimization, with a learning rate of 2.5×10^{-3} for the generator and 2×10^{-3} for the discriminator.

During training, we set the loss weight for R1 regularization to 5 to penalize the gradients of the discriminator and the loss weight of hand regularizer is empirically set to 1. For the face discriminator, we pad the cropped face into a square shape and resize it to 80×80 resolution.

Our models are trained for 4 days on 4 NVIDIA A6000 GPUs, with a batch size of 32. At test time, our model runs at 36 FPS on one NVIDIA A6000 GPU.

C Additional Results

Hand Distribution. As shown in Fig. A2, we observed that most fingers have a natural curve in dataset, but the SMPL model itself is unable to represent this pose. Most SMPL hand poses are



Figure A3: Qualitative comparison for ablations on number of layers.

naturally open, and as a result, even with deformation, the hand mesh cannot accurately fit the hand pose distribution of dataset. This can lead to some fingers appearing translucent. To address this issue, we reduce the original deformation scale and aim to learn an opaque texture to ensure the photorealistic appearance of the hand. In future work, we plan to use a more precise SMPL-H model to further improve the representation of the hand pose.

Number of Layers. In Fig. A3, we present visualizations of the generated humans at different numbers of layers. As the number of layers increases, we observe significant improvements in the quality and diversity of both the images and faces generated by our model. In the case with only one layer, we rasterize the texture onto the original SMPL surface, resulting in very thin people and limited ability to synthesize realistic human images. This approach struggles to handle complex hair and clothing structures. However, as the number of layers increases, the layered surface volumes can capture more details and volumetric structures, such as hair and clothing, resulting in more realistic and diverse human images.