Rodin: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion

Tengfei Wang[1†*] Bo Zhang[2*] Ting Zhang[2] Shuyang Gu[2] Jianmin Bao[2]
Tadas Baltrusaitis[2] Jingjing Shen[2] Dong Chen[2] Fang Wen[2] Qifeng Chen[1] Baining Guo[2]
[1]HKUST [2]Microsoft Research

Figure 1. Our diffusion model, Rodin, can produce high-fidelity 3D avatars, as shown in the first row. Our model also supports 3D avatar generation from a single portrait or text prompt, while permitting text-based semantic manipulation (second row). See the webpage for video demos.

Abstract

This paper presents a 3D generative model that uses diffusion models to automatically generate 3D digital avatars represented as neural radiance fields. A significant challenge in generating such avatars is that the memory and processing costs in 3D are prohibitive for producing the rich details required for high-quality avatars. To tackle this problem we propose the roll-out diffusion network (Rodin), which represents a neural radiance field as multiple 2D feature maps and rolls out these maps into a single 2D feature plane within which we perform 3D-aware diffusion. The Rodin model brings the much-needed computational efficiency while preserving the integrity of diffusion in 3D by using 3D-aware convolution that attends to projected features in the 2D feature plane according to their original relationship in 3D. We also use latent conditioning to orchestrate the feature generation for global coherence, leading to high-fidelity avatars and enabling their semantic editing based on text prompts. Finally, we use hierarchical synthesis to further enhance details. The 3D avatars generated by our model compare favorably with those produced by existing generative techniques. We can generate highly detailed avatars with realistic hairstyles and facial hair like beards. We also demonstrate 3D avatar generation from image or text as well as text-guided editability.

[*]Equal contribution. [†]Intern at Microsoft Research.

1. Introduction

Generative models [2, 34] are one of the most promising ways to analyze and synthesize visual data including 2D images and 3D models. At the forefront of generative modeling is the diffusion model [14, 24, 61], which has shown phenomenal generative power for images [19,49,52,54] and videos [23, 59]. Indeed, we are witnessing a 2D content-creation revolution driven by the rapid advances of diffusion and generative modeling.

In this paper, we aim to expand the applicability of diffusion such that it can serve as a generative model for 3D digital avatars. We use "digital avatars" to refer to the traditional avatars manually created by 3D artists, as opposed to the recently emerging photorealistic avatars [8, 44]. The reason for focusing on digital avatars is twofold. On the one hand, digital avatars are widely used in movies, games, the metaverse, and the 3D industry in general. On the other hand, the available digital avatar data is very scarce as each avatar has to be painstakingly created by a specialized 3D

artist using a sophisticated creation pipeline [20, 35], especially for modeling hair and facial hair. All this leads to a compelling scenario for generative modeling.

We present a diffusion model for automatically producing digital avatars represented as neural radiance fields [39], with each point describing the color radiance and density of the 3D volume. The core challenge in generating neural volume avatars is the prohibitive memory and computational cost for the rich details required by high-quality avatars. Without rich details, our results will always be somewhat "toy-like". To tackle this challenge, we develop Rodin, the roll-out diffusion network. We take a neural volume represented as multiple 2D feature maps and roll out these maps into a single 2D feature plane and perform 3D-aware diffusion within this plane. Specifically, we use the tri-plane representation [9], which represents a volume by three axis-aligned orthogonal feature planes. By simply rolling out feature maps, the Rodin model can perform 3D-aware diffusion using an efficient 2D architecture and drawing power from the model's three key ingredients below.

The first is the 3D-aware convolution. The 2D CNN processing used in conventional 2D diffusion cannot well handle the feature maps originating from orthogonal planes. Rather than treating the features as plain 2D input, the 3D-aware convolution explicitly accounts for the fact that a 2D feature in one plane (of the tri-plane) is a projection from a piece of 3D data and is hence intrinsically associated with the same data's projected features in the other two planes. To encourage cross-plane communication we involve all these associated features in the convolution and thus bridge the associated features together and synchronize their detail synthesis according to their 3D relationship.

The second key ingredient is latent conditioning. We use a latent vector to orchestrate the feature generation so that it is globally coherent across the 3D volume, leading to better-quality avatars and enabling their semantic editing. We do this by using the avatars in the training dataset to train an additional image encoder which extracts a semantic latent vector serving as the conditional input to the diffusion model. This latent conditioning essentially acts as an autoencoder in orchestrating the feature generation. For semantic editability, we adopt a frozen CLIP image encoder [9] that shares the latent space with text prompts.

The final key ingredient is hierarchical synthesis. We start by generating a low-resolution tri-plane (64×64), followed by a diffusion-based upsampling that yields a higher resolution (256×256). When training the diffusion upsampler, it is instrumental in penalizing the image-level loss that we compute in a patch-wise manner.

Taken together, the above ingredients work in concert to enable the Rodin model to coherently perform diffusion in 3D with an efficient 2D architecture. The Rodin is trained with a multi-view image dataset of 100K avatars of diverse identities, hairstyles, and clothing created by 3D artists [69].

Several application scenarios are thus supported. We can use the model to generate an unlimited number of avatars from scratch, each avatar being different from others as well as the ones in the training data. As shown in Figure 1, we can generate highly-detailed avatars with realistic hairstyles and facial hairs styled as beards, mustaches, goatees, and sideburns. Hairstyle and facial hair are essential for representing people's unique personal identities. Yet, these styles have been notoriously difficult to generate well with existing approaches. The Rodin model also allows avatar customization with the resulting avatar capturing the visual characteristics of the person portrayed in the image or the textual description. Finally, our framework supports text-guided semantic editing. The strong generative power of diffusion shows great promise in 3D modeling.

2. Related Work

The state of generative modeling [5, 14, 15, 28, 50, 65, 75] has seen rapid progress in past years. Diffusion models [14, 24, 61, 73] have recently shown unprecedented generative ability and compositional power. The most remarkable success happens in text-to-image synthesis [19, 40, 49, 52, 54], which serves as a foundation model and enables various appealing applications [21, 53, 66] previously unattainable. While diffusion models have been successfully applied to different modalities [11, 23, 26, 32], its generative capability is much less explored in 3D generation, with only a few attempts on modeling 3D primitives [37, 74, 76].

Early 3D generation works [58] rely on either GAN [17] or VAE [29] to model the distribution of 3D shape representation like voxel grids [6, 70], point clouds [1, 7, 31, 72], mesh [33, 63] and implicit neural representation [43, 60]. However, existing works have not demonstrated the ability to produce complex 3D assets yet. Concurrent to this work, Bautista *et al*. [4] train a diffusion model to generate the latent vector that encodes the radiance field [39] of synthetic scenes, yet this work only produces coarse 3D geometry. In comparison, we propose a hierarchical 3D generation framework with effective 3D-aware operators, offering unprecedented 3D detail synthesis.

Another line of work learns 3D-aware generation by utilizing richly available 2D data. 3D-aware GANs [9, 10, 12, 16, 18, 42, 56, 57, 62, 71, 77] recently attract significant research interest, which are trained to produce radiance fields with image level distribution matching. However, these methods suffer from instabilities and mode collapse of GAN training, and it is still challenging to attain authentic avatars that can be viewed from large angles. Concurrently, there are a few attempts to use diffusion models for the problem. Daniel *et al*. [68] proposes to synthesize novel views with a pose-conditioned 2D diffusion model, yet the results are not intrinsically 3D. Ben *et al*. [46] optimizes a

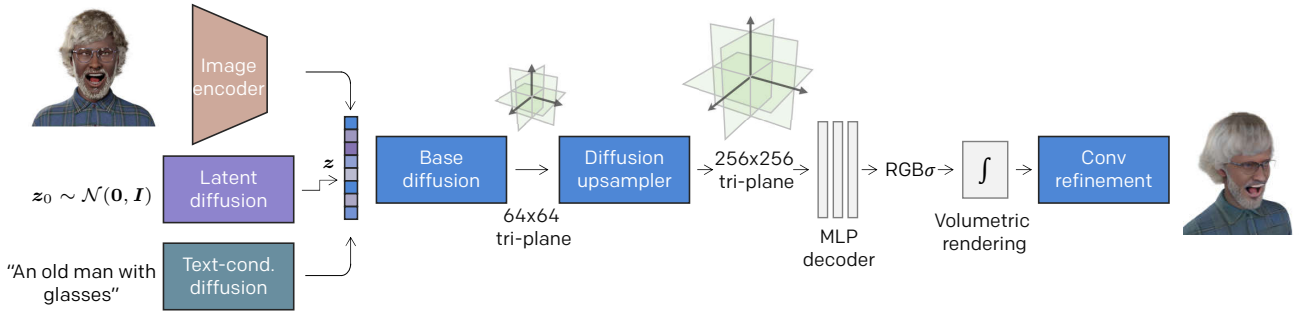

Figure 2. An overview of our Rodin model. We derive the latent $\boldsymbol{z}$ via the mapping from image, text, or random noise, which is used to control the base diffusion model to generate 64×64 tri-planes. We train another diffusion model to upsample this coarse result to 256×256 tri-planes that are used to render final multi-view images with volumetric rendering and convolutional refinement. The operators used in diffusion models are designed to be 3D-aware.

radiance field using the supervision from a pretrained text-to-image diffusion model and produces impressive 3D objects of diverse genres. Nonetheless, pretrained 2D generative networks only offer limited 3D knowledge and inevitably lead to blurry 3D results. A high-quality generation framework in 3D space is still highly desired.

3. Approach

Unlike prior methods that learn 3D-aware generation from a 2D image collection, we aim to learn the 3D avatar generation using the multi-view renderings from the Blender synthetic pipeline [69]. Rather than treating the multi-view images of the same subject as individual training samples, we fit the volumetric neural representation for each avatar, which is used to explain all the observations from different viewpoints. Thereafter we use diffusion models to characterize the distribution of these 3D instances. Our diffusion-based 3D generation is a hierarchical process — we first utilize a diffusion model to generate the coarse geometry, followed by a diffusion upsampler for detail synthesis. As illustrated in Figure 2, the whole 3D portrait generation comprises multiple training stages, which we detail in the following subsections.

3.1. Robust 3D Representation Fitting

To train a generative network with explicit 3D supervision, we need an expressive 3D representation that accounts for multi-view images, which should meet the following requirements. First, we need an explicit representation that is amenable to generative network processing. Second, we require a compact representation that is memory efficient; otherwise, it would be too costly to store a myriad of such 3D instances for training. Furthermore, we expect fast representation fitting since hours of optimization as vanilla NeRF [39] would make it unaffordable to generate abundant 3D training data as required for generative modeling.

Taking these into consideration, we adopt *tri-plane representation* proposed by [9] to model the neural radiance field of 3D avatars. Specifically, the 3D volume is factorized into three axis-aligned orthogonal feature planes, denoted by $\boldsymbol{y}_{uv}, \boldsymbol{y}_{wu}, \boldsymbol{y}_{vw} \in \mathbb{R}^{H \times W \times C}$, each of which has spatial resolution of $H \times W$ and number of channel as C. Compared to voxel grids, the tri-plane representation offers a considerably smaller memory footprint without sacrificing the expressivity. Hence, rich 3D information is explicitly memorized in the tri-plane features, and one can query the feature of the 3D point $\boldsymbol{p} \in \mathbb{R}^3$ by projecting it onto each plane and aggregating the retrieved features, *i.e.*, $\boldsymbol{y}_{\boldsymbol{p}} = \boldsymbol{y}_{uv}(\boldsymbol{p}_{uv}) + \boldsymbol{y}_{wu}(\boldsymbol{p}_{wu}) + \boldsymbol{y}_{vw}(\boldsymbol{p}_{vw})$. With such positional feature, one can derive the density $\sigma \in \mathbb{R}^+$ and view-dependent color $\boldsymbol{c} \in \mathbb{R}^3$ of each 3D location given the viewing direction $\boldsymbol{d} \in \mathbb{S}^2$ with a lightweight MLP decoder $\mathcal{G}_\theta^{\text{MLP}}$, which can be formulated as

$$\boldsymbol{c}(\boldsymbol{p}, \boldsymbol{d}), \sigma(\boldsymbol{p}) = \mathcal{G}_\theta^{\text{MLP}}\big(\boldsymbol{y}_{\boldsymbol{p}}, \xi(\boldsymbol{y}_{\boldsymbol{p}}), \boldsymbol{d}\big). \tag{1}$$

Here, we apply the Fourier embedding operator $\xi(\cdot)$ [64] on the queried feature rather than the spatial coordinate. The tri-plane features and the MLP decoder are optimized such that the rendering of the neural radiance field matches the multi-view images $\{\boldsymbol{x}\}_{N_v}$ for the given subject, where $\boldsymbol{x} \in \mathbb{R}^{H_0 \times W_0 \times 3}$. We enforce the rendered image given by volumetric rendering [38], *i.e.*, $\hat{\boldsymbol{x}} = \mathcal{R}(\boldsymbol{c}, \sigma)$, to match the corresponding ground truth with mean squared error loss. Besides, we introduce sparse, smooth, and compact regularizers to reduce the "floating" artifacts [3] in free space. For more tri-plane fitting details, please refer to the Appendix.

While prior per-scene reconstruction mainly concerns the fitting quality, our 3D fitting procedure should also consider several key aspects for generation purposes. First, the tri-plane features of different subjects should rigorously reside in the same domain. To achieve this, we adopt a shared MLP decoder when fitting distinct portraits, thus implicitly pushing the tri-plane features to the shared latent space recognizable by the decoder. Second, the MLP decoder has to possess some level of *robustness*. That is, the decoder should be tolerant to slight perturbation of tri-plane fea-

(a) 256×256 tri-plane (b) 64×64 tri-plane without random scaling (c) 64×64 tri-plane with random scaling

Figure 3. While 256×256 tri-planes give good renderings (a), the 64×64 variant gives much worse result (b). Hence, we introduce random scaling during fitting so as to obtain a robust representation that can be effectively rendered in continuous scales (c).

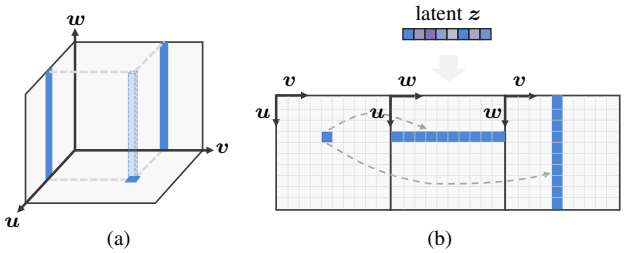

Figure 4. We propose two mechanisms to ensure coherent tri-plane generation. Our 3D-aware convolution considers the 3D relationship in (a) and correlates the associated elements from separate feature planes as shown in (b). In (b), we also visualize the usage of a shared latent code to orchestrate the feature generation.

tures, and thus one can still obtain plausible results even if the tri-plane features are imperfectly generated. More importantly, the decoder should be robust to varied tri-plane sizes because hierarchical 3D generation is trained on multi-resolution tri-plane features. As shown in Figure 3, when solely fitting 256×256 tri-planes, its 64×64 resolution variant cannot be effectively rendered. To address this, we randomly scale the tri-plane during fitting, which is instrumental in deriving multi-resolution tri-plane features simultaneously with a shared decoder.

3.2. Latent Conditioned 3D Diffusion Model

Now the 3D avatar generation is reduced to learning the distribution of tri-plane features, *i.e.*, $p(\boldsymbol{y})$, where $y = (\boldsymbol{y}_{uv}, \boldsymbol{y}_{wu}, \boldsymbol{y}_{vw})$. Such generative modeling is non-trivial since $\boldsymbol{y}$ is highly dimensional. We leverage diffusion models for the task, which have shown compelling quality in complex image modeling.

On a high level, the diffusion model generates $\boldsymbol{y}$ by gradually reversing a Markov forward process. Starting from $\boldsymbol{y}_0 \sim p(\boldsymbol{y})$, the forward process q yields a sequence of increasing noisy latent codes $\{\boldsymbol{y}_t \mid t \in [0, T]\}$ according to $\boldsymbol{y}_t := \alpha_t \boldsymbol{y}_0 + \sigma_t \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is the added Gaussian noise; α_t and σ_t define a noise schedule whose log signal-to-noise ratio $\lambda_t = \log[\alpha_t^2 / \sigma_t^2]$ linearly decreases with the timestep t. With sufficient noising steps, we reach a pure Gaussian noise, *i.e.*, $\boldsymbol{y}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The generative process corresponds to reversing the above noising process, where the diffusion model is trained to denoise $\boldsymbol{y}_t$ into $\boldsymbol{y}_0$ for all t using a mean squared error loss. Following [24], better generation quality can be achieved by parameterizing the diffusion model $\hat{\boldsymbol{\epsilon}}_\theta$ to predict the added noise:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t, \boldsymbol{x}_0, \boldsymbol{\epsilon}} \left[\left\| \hat{\boldsymbol{\epsilon}}_\theta(\alpha_t \boldsymbol{y}_0 + \sigma_t \boldsymbol{\epsilon}, t) - \boldsymbol{\epsilon} \right\|_2^2 \right]. \quad (2)$$

In practice, our diffusion model training also jointly optimizes the variational lower bound loss $\mathcal{L}_{\text{VLB}}$ as suggested in [41], which allows high-quality generation with fewer timesteps. During inference, stochastic ancestral sampler [24] is used to generate the final samples, which starts from the Gaussian noise $\boldsymbol{y}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and sequentially produces less noisy samples $\{\boldsymbol{y}_T, \boldsymbol{y}_{T-1}, \ldots\}$ until reaching $\boldsymbol{y}_0$.

We first train a base diffusion model to generate the coarse-level tri-planes, *e.g.*, at 64×64 resolution. A straightforward approach is to adopt the 2D network structure used in the state-of-the-art image-based diffusion models for our tri-plane generation. Specifically, we can concatenate the tri-plane features in the channel dimension as in [9], which forms $\boldsymbol{y} = (\boldsymbol{y}_{uv} \oplus \boldsymbol{y}_{wu} \oplus \boldsymbol{y}_{vw}) \in \mathbb{R}^{H \times W \times 3C}$, and employ a well-designed 2D U-Net to model the data distribution through the denoising diffusion process. However, such a baseline model produces 3D avatars with severe artifacts. We conjecture the generation artifact comes from the *incompatibility* between the tri-plane representation and the 2D U-Net. As shown in Figure 4(a), intuitively, one can regard the tri-plane features as the projection of neural volume towards the frontal, bottom, and side views, respectively. Hence, the channel-wise concatenation of these orthogonal planes for CNN processing is problematic because these planes are not spatially aligned. To better handle the tri-plane representation, we make the following efforts.

3D-aware convolution. Using CNN to process channel-wise concatenated tri-planes will cause the mixing of theoretically uncorrected features in terms of 3D. One simple yet effective way to address this is to spatially *roll out* the tri-plane features. As shown in Figure 4(b), we concatenate the tri-plane features horizontally, yielding $\tilde{\boldsymbol{y}} = \texttt{hstack}(\boldsymbol{y}_{uv}, \boldsymbol{y}_{wu}, \boldsymbol{y}_{vw}) \in \mathbb{R}^{H \times 3W \times C}$. Such feature roll-out allows independent processing of feature planes. For simplicity, we subsequently use $\boldsymbol{y}$ to denote such input form by default. However, the tri-plane roll-out hampers cross-plane communication, while the 3D generation requires the synergy of the tri-plane generation.

To better process the tri-plane features, we need an efficient 3D operator that performs on the tri-plane rather than treating it as a plain 2D input. To achieve this, we propose *3D-aware convolution* to effectively process the tri-plane features while respecting their 3D relationship. A point on a certain feature plane actually corresponds to an

axis-aligned 3D line in the volume, which also has two corresponding line projections in other planes, as shown in Figure 4(a). The features of these corresponding locations essentially describe the same 3D primitive and should be learned synchronously. However, such a 3D relationship is neglected when employing plain 2D convolution to tri-plane processing. As such, our 3D-aware convolution explicitly introduces such 3D inductive bias by attending the features of each plane to the corresponding row/column of the rest planes. In this way, we enable 3D processing capability with 2D CNNs. This 3D-aware convolution applied on the tri-plane representation, in fact, is a generic way to simplify 3D convolutions previously too costly to compute when modeling high-resolution 3D volumes.

The 3D-aware convolution is depicted in Figure 4(b). Ideally, the compute for $\boldsymbol{y}_{uv}$ would attend to full elements from the corresponding row/column, *i.e.*, $\boldsymbol{y}_{wu}$ and $\boldsymbol{y}_{vw}$, from other planes. For parallel computing, we simplify this and aggregate the row/column elements. Specifically, we apply the axis-wise pooling for $\boldsymbol{y}_{wu}$ and $\boldsymbol{y}_{vw}$, yielding a row vector $\boldsymbol{y}_{wu\to u} \in \mathbb{R}^{1\times W\times C}$ and a column vector $\boldsymbol{y}_{vw\to v} \in \mathbb{R}^{H\times 1\times C}$ respectively. For each point of $\boldsymbol{y}_{uv}$, we can easily access the corresponding element in the aggregated vectors. We expand the aggregated vectors to the original 2D dimension (*i.e.*, replicating the column vectors along row dimension, and vice versa) and thus derive $\boldsymbol{y}_{(\cdot)u}, \boldsymbol{y}_{v(\cdot)} \in \mathbb{R}^{H\times W\times C}$. By far, we can perform 2D convolution on the channel-wise concatenation of the feature maps, *i.e.*, $\texttt{Conv2D}(\boldsymbol{y}_{uv} \oplus \boldsymbol{y}_{(\cdot)u} \oplus \boldsymbol{y}_{v(\cdot)})$. because $\boldsymbol{y}_{uv}$ is now spatially aligned with the aggregation of the corresponding elements from other planes. The compute for $\boldsymbol{y}_{vw}$ and $\boldsymbol{y}_{wu}$ is conducted likewise. The 3D-aware convolution greatly enhances the cross-plane communication, and we empirically observe reduced artifacts and improved generation of thin structures like hair strands.

Latent conditioning. We further propose to learn a latent vector to *orchestrate* the tri-plane generation. As shown in Figure 2, we additionally train an image encoder $\mathcal{E}$ to extract a semantic latent vector serving as the conditional input of the base diffusion model, so essentially the whole framework is an *autoencoder*. To be specific, we extract the latent vector from the frontal view of each training subject, *i.e.*, $\boldsymbol{z} = \mathcal{E}_\theta(\boldsymbol{x}_{\text{front}}) \in \mathbb{R}^{512}$, and the diffusion model conditioned on $\boldsymbol{z}$ is trained to reconstruct the tri-plane of the same subject. We use adaptive group normalization (AdaGN) to modulate the activations of the diffusion model, where $\boldsymbol{z}$ is injected into every residual block, and in this way, the features of the orthogonal planes are synchronously generated according to a shared latent.

The latent conditioning not only leads to higher generation quality but also permits a disentangled latent space, thus allowing semantic editing of generated results. To achieve better editability, we adopt a frozen CLIP image encoder [48] that has shared latent space with text prompts. We will show how the learned model produces controllable text-guided generation results.

Another notable benefit of latent conditioning is that it allows *classifier-free guidance* [25], a technique typically used to boost the sampling quality in the conditional generation. When training the diffusion model, we randomly zero the latent embedding with 20% probability, thus adapting the diffusion decoder to unconditional generation. During inference, we can steer the model toward better generation sampling according to

$$\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{y}, \boldsymbol{z}) = \lambda \boldsymbol{\epsilon}_\theta(\boldsymbol{y}, \boldsymbol{z}) + (1-\lambda)\boldsymbol{\epsilon}_\theta(\boldsymbol{y}), \tag{3}$$

where $\boldsymbol{\epsilon}_\theta(\boldsymbol{y}, \boldsymbol{z})$ and $\boldsymbol{\epsilon}_\theta(\boldsymbol{y})$ are the conditional and unconditional $\boldsymbol{\epsilon}$-predictions respectively, and $\lambda > 0$ specifies the guidance strength.

Our latent conditioned base model thus supports both unconditional generation as well as the conditional generation that is used for portrait inversion. To account for full diversity during unconditional sampling, we additionally train a diffusion model to model the distribution of the latent $\boldsymbol{z}$, whereas the latent $\boldsymbol{y}_T$ describes the residual variation. We include this latent diffusion model in Figure 2.

3.3. Diffusion Tri-plane Upsampler

To generate high-fidelity 3D structures, we further train a diffusion super-resolution (SR) model to increase the tri-plane resolution from 64×64 to 256×256. At this stage, the diffusion upsampler is conditioned on the low-resolution (LR) tri-plane $\boldsymbol{y}^{\text{LR}}$. Different from the base model training we parameterize the diffusion upsampler $\boldsymbol{y}_\theta^{\text{HR}}(\boldsymbol{y}_t^{\text{HR}}, \boldsymbol{y}^{\text{LR}}, t)$ to predict the high-resolution (HR) ground truth $\boldsymbol{y}_0^{\text{HR}}$ instead of the added noise $\boldsymbol{\epsilon}$. The 3D-aware convolution is utilized in each residual block to enhance detail synthesis.

Following prior cascaded image generation works, we apply *condition augmentation* to reduce the domain gap between the output from the base model and the LR conditional input for SR training. We conduct careful tuning for the tri-plane augmentation with a combination of random downsampling, Gaussian blurring and Gaussian noises, making the rendered augmented LR tri-plane resemble the base rendering output as much as possible.

Nonetheless, we find that a tri-plane restoration with a lower ℓ_2 distance to the ground truth may not necessarily correspond to a satisfactory image rendering. Hence, we need to directly constrain the rendered image. Specifically, we obtain the rendered image $\hat{\boldsymbol{x}}^{\text{HR}} \in \mathbb{R}^{256\times256\times3}$ from the predicted tri-plane $\hat{\boldsymbol{y}}_0^{\text{HR}}$ with $\hat{\boldsymbol{x}} = \mathcal{R}(\mathcal{G}_\theta^{\text{MLP}}(\hat{\boldsymbol{y}}_0^{\text{HR}}))$, and we further penalize the perceptual loss [27] between this rendered result and the ground truth, according to

$$\mathcal{L}_{\text{perc}} = \mathbb{E}_{t,\hat{\boldsymbol{x}}} \sum_l \|\Psi^l(\hat{\boldsymbol{x}}) - \Psi^l(\boldsymbol{x})\|_2^2, \tag{4}$$

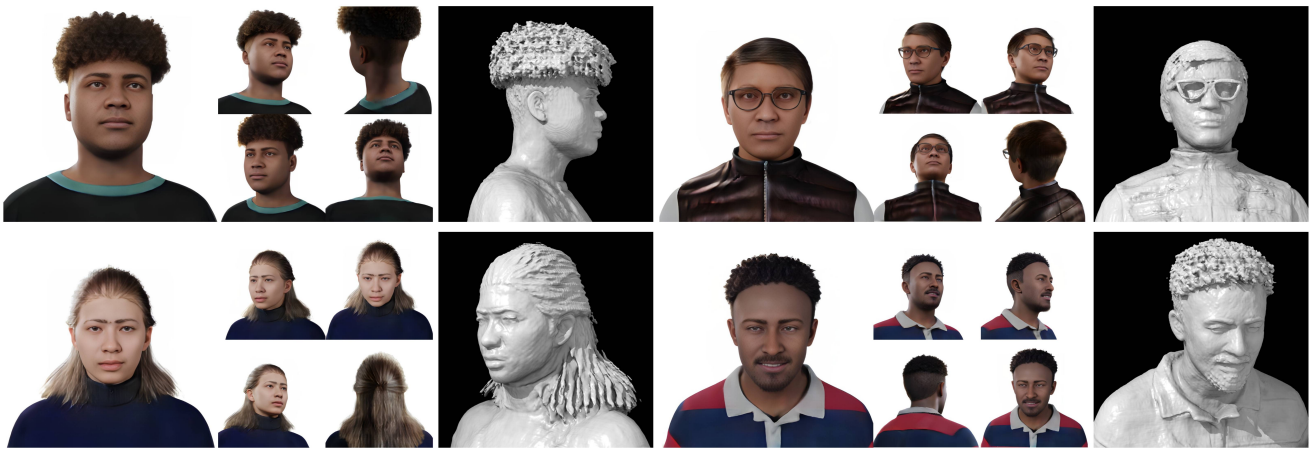

Figure 5. Unconditional generation samples by our Rodin model. We visualize the mesh extracted from the generated density field.

Figure 6. Latent interpolation results for generated avatars.

	Pi-GAN	GIRAFFE	EG3D	Autoencoder	Ours
FID ↓	78.3	64.6	40.5	67.4	**26.1**

Table 1. Quantitative comparison with baseline methods.

Model configuration	FID↓
A. Baseline	39.2
B. + Latent conditioning	37.4
C. + Tri-plane roll-out	28.4
D. + 3D-aware conv	**26.1**

Table 2. Ablation study of the proposed components.

where Ψ^l denotes the multi-level feature extraction using a pretrained VGG. Usually, the volume rendering requires stratified sampling along each ray, which is computationally prohibitive for high-resolution rendering. Hence, we compute $\mathcal{L}_{\text{perc}}$ on random 112×112 image patches with high sampling importance on face region. Compared with prior 3D-aware GANs that require rendering full images, our 3D-aware SR can be easily scalable to high resolutions due to the permit of patchwise training with direct supervision.

Modeling high-frequency detail and thin structures are particularly challenging in volumetric rendering. Thus, at this stage, we jointly train a convolution refiner [67] on our data which complements the missing details of the NeRF rendering, ultimately producing compelling 1024×1024 image outputs.

4. Experiments

4.1. Implementation Details

To train our 3D diffusion, we obtain 100K 3D avatars with a random combination of identities, expressions, hairstyles, and accessories using synthetic engine [69]. For each avatar, we render 300 multi-view images with known camera pose, which are sufficient for a high-quality radiance field reconstruction. The tri-planes for our generation have the dimension of $256 \times 256 \times 32$ in each feature plane. We optimize a shared MLP decoder when fitting the first 1,000 subjects. This decoder consists of 4 fully connected layers and is fixed when fitting the following subjects. Thus different subjects are fitted separately in distributed servers.

Both the base and upsampling diffusion networks adopt U-Net architecture to process the roll-out tri-plane features. We apply full-attention for 8^2, 16^2 and 32^2 scales within the network and adopt 3D-aware convolution at higher scales to enhance the details. While we generate 256^2 tri-planes with the diffusion upsampler, we also render image and compute image loss at 512^2 resolution, with a convolutional refinement further enhancing the details to 1024^2. For more details about the network architecture and training strategies, please refer to our Appendix.

4.2. Unconditional Generation Results

Figure 5 shows several samples generated by the Rodin model, showing the capability to synthesize high-quality 3D renderings with impressive details, *e.g.*, glasses and hairstyle. To reflect the geometry, we extract the mesh from the generated density field using marching cubes, which demonstrates high-fidelity geometry. More uncurated sam-

Pi-GAN [10] GIRAFFE [42] EG3D [9] Our Rodin model

Figure 7. Qualitative comparison with state-of-the-art approaches.

ples are shown in the Appendix. We also explore the interpolation of the latent condition z between two generated avatars, as shown in Figure 6, where we observe consistent high-quality interpolation results with smooth appearance transition.

4.3. Comparison

We compare our method with state-of-the-art 3D-aware GANs, *e.g.*, Pi-GAN [10] and GIRAFFE [42] and EG3D [9], which learn to produce neural radiance field from 2D image supervision. Moreover, we implement an auto-encoder baseline, which leverages the multi-view supervision and reconstructs the radiance field from the latent. We differ in this baseline by using the power diffusion-based decoder with 3D-aware designs. We adapt the official implementation of prior works to 360-degree generation and retrain them using the same dataset.

We use FID score [22] to measure the quality of image renderings. As per [30], we use the features extracted from the CLIP model to compute FID, which we find better correlates the perceptual quality. Specifically, we compute the FID score using 5K generated samples. The quantitative comparison is shown in Table 1, where we see that the Rodin model induces significantly lower FID than others.

The visual comparison in Figure 7 shows a clear quality superiority of our Rodin model over prior arts. Our method gives visually pleasing multi-view renderings with high-quality geometry, *e.g.*, for glasses and hair, whereas 3D-aware GANs produce more artifacts due to the geometry ambiguity caused by the simple use of image supervision.

Base diffusion Diffusion upsampler Convolutional refiner

Figure 8. Hierarchical generation progressively improves results.

4.4. Analysis of the Rodin model

Both 3D-aware convolution and latent conditioning are crucial for 3D synthesis. To prove this, we conduct the ablation study as shown in Table 2. We start from a baseline that uses a plain 2D CNN to process channel-wise concatenated tri-plane features following [9]. With latent conditioning, we achieve a lower FID. Feeding the network with roll-out tri-plane features significantly reduces the FID score because tri-planes are no longer improperly mingled. The proposed 3D-aware convolution further improves the synthesis quality, especially for thin structures like hair and cloth texture. More visual results regarding these ablations can be found in the Appendix.

Hierarchical generation is critical for high-fidelity results. One significant benefit of this approach is that we can train different diffusion models dedicated to different scales in a supervised manner, as opposed to end-to-end synthesis with image loss. This also enables patch-wise training without the need to render full images. Thus hierarchical training allows high-resolution avatar generation without suffering the prohibitive memory issue. Figure 8 shows the progressive quality improvement after the base

(a) (b)

Figure 9. Results of (a) portrait inversion and (b) text-to-avatar generation.

Tri-plane level loss	Image-level loss	Cond. augment	FID ↓
✓			48.5
✓	✓		38.6
✓	✓	✓	**26.1**

Table 3. Ablation study of the tri-plane upsampling strategy.

diffusion, diffusion upsampler, and convolution refinement, respectively. It can be seen that the diffusion upsampler is critical, largely enhancing the synthesis quality, while convolution refinement further adds delicate details.

Diffusion upsampling training strategies. When training the tri-plane upsampler, we parameterize the model to predict the clean tri-plane ground truth at each diffusion reversion step. Meanwhile, conditioning augmentation is of great significance to let the model generalize to the coarse-level tri-plane generated from the base model. Besides, we observe enforcing image-level loss is beneficial to final perceptual quality. The effectiveness of these strategies are quantitatively justified in Table 3.

4.5. Applications

3D portrait from a single image. We can hallucinate a 3D avatar from a single portrait by conditioning the base generator with the CLIP image embedding for that input image. Note that our goal is different from face/head reconstruction [13, 51], but to conveniently produce a personalized digital avatar for users. As shown in Figure 9(a), the generated avatars keep the main characteristics of the portrait, *e.g.*, expression, hairstyle, glass wearing, etc., while being 360-degree renderable.

Text-to-avatar generation. Another natural way to customize avatars is to use language guidance. To do this, we train a text-conditioned diffusion model to generate the CLIP image embedding used to semantically control the avatar generation. We use a subset of the LAION-400M dataset [55] containing portrait-text pairs to train this model.

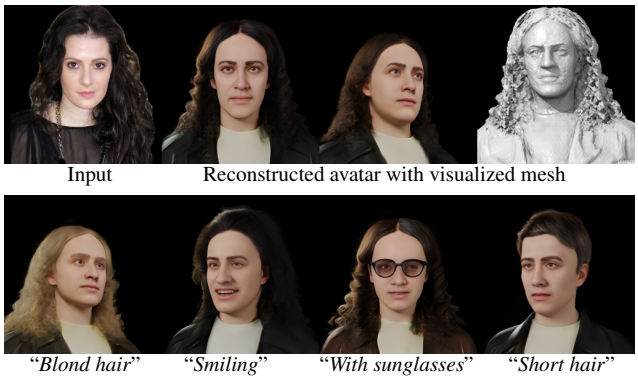

Figure 10. Results of text-guided avatar manipulation.

As shown in Figure 9(b), one can finely customize the avatars using detailed text descriptions.

Text-based avatar customization. We can also semantically edit generated avatars using text prompts. For a generated avatar with the CLIP image embedding $\boldsymbol{z}_i$, we can obtain a direction $\boldsymbol{\delta}$ in the CLIP's text embedding based on prompt engineering [45]. We assume colinearity between the CLIP's image and text embedding, thus we obtain the manipulated embedding as $\boldsymbol{z}_i + \boldsymbol{\delta}$, which is used to condition the generative process. As shown in Figure 10, we can achieve a wide variety of disentangled and meaningful control faithful to the text prompt.

5. Conclusion

From experiments, we observe that the Rodin model is a powerful generative model for 3D avatars. This model also allows users to customize avatars from a portrait or text, thus significantly lowering the barrier of personalized avatar creation. While this paper only focuses on avatars, the main ideas behind the Rodin model are applicable to the diffusion model for general 3D scenes. Indeed, the prohibitive computational cost has been a challenge for 3D content creation. An efficient 2D architecture for performing coherent and 3D-aware diffusion in 3D is an important step toward

tackling the challenge. For future work, it would be fruitful to improve the sampling speed of the 3D diffusion model and study jointly leveraging the ample 2D data to mitigate the 3D data bottleneck.

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. 2

[2] Alankrita Aggarwal, Mamta Mittal, and Gopi Battineni. Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, 1(1):100004, 2021. 1

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 3, 13

[4] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. *arXiv preprint arXiv:2207.13751*, 2022. 2

[5] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *arXiv preprint arXiv:2103.04922*, 2021. 2

[6] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Generative and discriminative voxel modeling with convolutional neural networks. *arXiv preprint arXiv:1608.04236*, 2016. 2

[7] Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snavely, and Bharath Hariharan. Learning gradient fields for shape generation. In *European Conference on Computer Vision*, pages 364–381. Springer, 2020. 2

[8] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 1

[9] 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, 7, 13

[10] 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. 2, 7

[11] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022. 2

[12] 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. 2

[13] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 8

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 2, 12

[15] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[16] 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. *arXiv preprint arXiv:2209.11163*, 2022. 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[18] 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. 2

[19] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 1, 2

[20] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 2

[21] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[22] 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. 7

[23] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben

Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1, 2

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 4

[25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[26] Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. Diff-tts: A denoising diffusion model for text-to-speech. *arXiv preprint arXiv:2104.01409*, 2021. 2

[27] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[28] 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. 2

[29] Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. 2

[30] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022. 7

[31] Ruihui Li, Xianzhi Li, Ka-Hei Hui, and Chi-Wing Fu. Sp-gan: Sphere-guided 3d shape generation and manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021. 2

[32] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*, 2022. 2

[33] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5871–5880, 2020. 2

[34] Ming-Yu Liu, Xun Huang, Jiahui Yu, Ting-Chun Wang, and Arun Mallya. Generative adversarial networks for image and video synthesis: Algorithms and applications. *Proceedings of the IEEE*, 109(5):839–862, 2021. 1

[35] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. 2

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations, ICLR*, 2019. 12

[37] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 2

[38] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 3

[39] 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. 2, 3

[40] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[41] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 4

[42] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2, 7

[43] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[44] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 1

[45] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 8

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

[47] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 13

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

[49] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2, 13

[50] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[51] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5620–5629, 2021. 8

[52] 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. 1, 2

[53] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 2

[54] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 2

[55] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 8

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

[57] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *arXiv preprint arXiv:2206.07695*, 2022. 2

[58] Zifan Shi, Sida Peng, Yinghao Xu, Yiyi Liao, and Yujun Shen. Deep generative models on 3d representations: A survey. *arXiv preprint arXiv:2210.15663*, 2022. 2

[59] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1

[60] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 2

[61] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 2

[62] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Transactions on Graphics*, 2022. 2

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

[64] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 3

[65] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020. 2

[66] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 2

[67] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021. 6

[68] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 2

[69] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. 2, 3, 6, 12

[70] 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. 2

[71] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. *arXiv preprint arXiv:2206.07255*, 2022. 2

[72] Chulin Xie, Chuxin Wang, Bo Zhang, Hao Yang, Dong Chen, and Fang Wen. Style-based point generator with adversarial rendering for point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4619–4628, 2021. 2

[73] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022. 2

[74] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022. 2

[75] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11304–11314, 2022. 2

[76] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021. 2

[77] 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. 2

Appendix

A. Background of Diffusion Models

Diffusion models produce data by reversing a gradual noising process. The forward noising process is a Markov chain that corrupts the data by gradually adding random noises for steps $t = 1, \cdots, T$. Each step in the forward process is a Gaussian transition $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}(\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I})$, where $\{\beta_t\}_{t=0}^T$ are usually predefined variance schedule. Furthermore, the noisy latent variable $\boldsymbol{x}_t$ can be derived from $\boldsymbol{x}_0$ directly as:

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{z}, \quad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{5}$$

where $\alpha_t := \prod_{s=1}^{t}(1-\beta_s)$. When T is large enough, α_T gets closer to 0 and the last latent variable $\boldsymbol{x}_T$ is nearly an isotropic Gaussian distribution.

To sample data from the given distribution, we can reverse the noising process by learning a denoising model $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$. The denoising model $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ starts from the Gaussian noise $\boldsymbol{x}_T$ and iteratively reduces the noise for $t = T-1, \cdots, 0$. Specifically, it takes the noisy latent variable $\boldsymbol{x}_t$ at each timestep t and predicts the corresponding noise $\boldsymbol{\epsilon}$ with a minimal mean square error:

$$\min_\theta \mathbb{E}_{\boldsymbol{x}_0 \sim p(\boldsymbol{x}_0), \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), t} ||\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{z}||_2^2. \tag{6}$$

With the learned denoising model, the data can be sampled with the following reverse diffusion process:

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right) + \sigma_t \boldsymbol{z}, \tag{7}$$

where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is a randomly sampled noise, and σ_t is the variance of the added noise.

B. Implementation Details

B.1. Architectural Design and Training Details

Our base diffusion model adopts the U-Net architecture from [14] with a channel number of 192, while we make several modifications including tri-plane roll-out and 3D-aware convolution, as discussed in Section 3.2. To orchestrate the tri-plane generation and enable semantic editing, we also introduce a condition encoder, a fixed CLIP ViT-B/32 image encoder, to map a reference image to a semantic latent vector. The upsample diffusion model is a U-Net-like model but we apply only one upsample layer that directly upscales the feature maps from 64 to 256 for efficiency, as shown in Figure 11. The tri-plane roll-out and 3D-aware convolution are utilized in each residual block. When training the upsample model, we apply condition augmentation on the tri-planes to reduce the domain gap as described in Section 3.3. Specifically, we degrade the ground-truth 256×256 tri-planes with a random combination of downscale, Gaussian blur, and Gaussian noise.

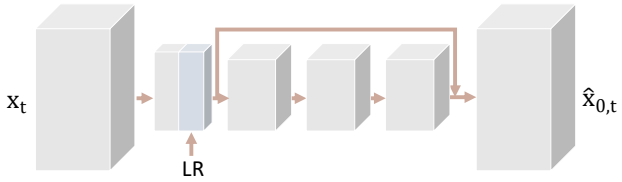

Figure 11. Architecture of the upsample diffusion model.

We utilize AdamW optimizer [36] with a batch size of 48 and a learning rate of $5e$-5 for the base diffusion model, and with a batch size of 16 and a learning rate of $5e$-5 for the upsample diffusion model. We also apply the exponential moving average (EMA) with a rate of 0.9999 during training. We set the diffusion steps as 1,000 for the base model, and 100 for the upsample model, with a linear noise schedule. During inference we sample 100 diffusion steps for both the base model and the upsample model. All the experiments are performed on NVIDIA Tesla 32G-V100 GPUs.

B.2. Tri-plane Fitting

Our framework learns the 3D avatar generation from explicit 3D representations obtained from fitting multi-view images. However, a multi-view consistent, diverse, high-quality and large-scale dataset of face images is difficult and expensive to collect. Images collected from the Web have no guarantee of multi-view consistency and suffer privacy and copyright risks. Regarding this, we turn to synthetic techniques that can randomly render novel 3D portraits by randomly combining assets manually created by artists. We leverage the Blender synthetic pipeline [69] that generates human faces along with random sampling from a large collection of hair, clothing, expression and accessory. Hence, we create 100K synthetic individuals independently and for each render 300 multi-view images with a resolution of 256×256.

For tri-plane fitting, we learn $256 \times 256 \times 32 \times 3$ spatial features for each person along with a lightweight MLP decoder consisting of 4 fully connected layers as shown in Figure 12. We randomly initialize the tri-plane feature and MLP weights. During fitting, we apply random rescaling (downsample to a resolution in $[64, 256]$ followed by an upsampling to 256) to ensure that the MLP decoder is robust to

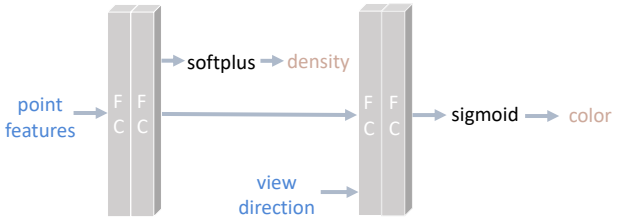

Figure 12. Architecture of the MLP decoder.

multi-resolution tri-plane features. To enable scalable and efficient fitting, we first optimize the shared 4-layer MLP decoder when fitting the first 1,000 subjects, and this decoder is fixed when fitting the following subjects. Thus different subjects are fitted separately in distributed servers.

For multi-view images $\{\boldsymbol{x}\}_{N_v}$ for the given subject, where $\boldsymbol{x} \in \mathbb{R}^{H_0 \times W_0 \times 3}$, we minimize the mean squared error $\mathcal{L}_{\text{MSE}}$ between the rendered image via volumetric rendering, *i.e.*, $\hat{\boldsymbol{x}} = \mathcal{R}(\boldsymbol{c}, \sigma)$ and the corresponding ground truth image. Moreover, we introduce additional regularizers to improve the fitting quality. To be specific, we manage to reduce the "floating" artifact by enforcing the sparsity loss $\mathcal{L}_{\text{sparse}}$ which penalizes the ℓ_1 magnitude of the predicted density, the smoothness loss $\mathcal{L}_{\text{smooth}}$ [9] that encourages a smooth density field, as well as the distortion loss $\mathcal{L}_{\text{dist}}$ [3] that encourages compact rays with localized weight distribution.

B.3. Text-based Avatar Customization

As shown in Section 4.5, the Rodin model can edit generated avatars with text prompts. For a generated avatar with a conditioned latent $\boldsymbol{z}_i$, we can obtain an editing direction $\boldsymbol{\delta} = E_T^{clip}(T_{tgt}) - E_T^{cilp}(T_{src})$ in the text embedding space of CLIP based on prompt engineering. For instance, we can choose the source text T_{src} from some general descriptions such as "a photo of a person" and "a portrait of a person", and use the target text T_{tgt} such as "a photo of a person with blond hair" and "a photo of a smiling person ". As we assume colinearity between the CLIP's image and text embedding, we can obtain the manipulated embedding as $\boldsymbol{z}_i + \boldsymbol{\delta}$, which is used to generate edited avatars.

B.4. Latent Diffusion for Unconditional Sampling

As discussed in Section 3.2, our base diffusion model supports both unconditional generation and conditional generation. To account for full diversity during unconditional sampling, we additionally train a diffusion model to model the distribution of the latent $\boldsymbol{z}$. The latent diffusion adopts a 20-layer MLPs network [47] with the hidden channel of 2048 that iteratively predicts the latent code $\boldsymbol{z} \in \mathbb{R}^{512}$ from random Gaussian noise. We set the diffusion steps as 1,000 with a linear noise schedule. We utilize AdamW optimizer with a batch size of 96 and a learning rate of $4e-5$, and also apply exponential moving average (EMA) with a rate of 0.9999 during training.

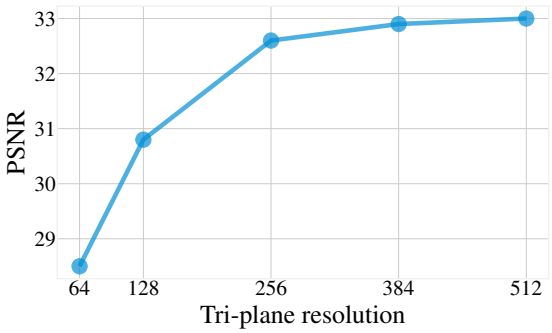

Figure 13. Effect of tri-plane resolution for tri-plane fitting.

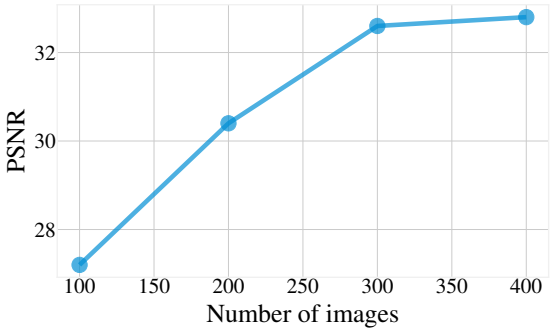

Figure 14. Effect of image numbers for tri-plane fitting.

B.5. Text-to-avatar Generation

As shown in Section 4.5, we perform text-to-avatar generation by training a text-conditioned diffusion model that generates an image embedding from a text embedding in the CLIP space. We adopt the network architecture from [49] and train it on a subset of the LAION-400M dataset, containing 100K portrait-text pairs. We set the diffusion steps as 1,000 with a linear noise schedule. We utilize AdamW optimizer with a batch size of 96 and a learning rate of $4e-5$, and also apply exponential moving average (EMA) with a rate of 0.9999 during training.

C. Additional Ablation Study and Analysis

C.1. Tri-plane Settings

Choices of Tri-plane resolution. To analyze the impact of tri-plane resolution, we experiment with different tri-planes from a set of $\{64, 128, 256, 384, 512\}$ to fit 1024×1024 images and show the results in Figure 13. Overall, the fitting quality increases with the tri-plane resolution. Empirically, we find that the 256×256 tri-plane is strong enough to represent a subject. Considering the memory cost, we thus choose to utilize 256×256 tri-planes in our experiments.

Figure 15. Visualization of intermediate generation results of different time steps.

Figure 16. Both tri-plane roll-out and 3D-aware convolution are crucial for high-fidelity results.

Number of images for fitting. We also explore how many images are needed to achieve a high-quality fitting. As shown in Figure 14, the fitting quality get almost saturated when using 300 different views for the neural tri-plane reconstruction.

C.2. Visualization of Different Diffusion Steps

Diffusion models generate samples by gradually removing noises for $t \in [T, 0]$, and analyzing these intermediate results would reach an in-depth understanding of the generation process. We thus demonstrate the generated results over the reverse process in Figure 15, where we render the predicted tri-plane of our base diffusion, $\hat{x_0}$, at each time step t. Notwithstanding that our diffusion is performed in tri-plane feature space, the reverse process is similar to that in image space, where the coarse structure appears first and fine details appear in the last iterative steps.

C.3. Effect of 3D-aware Convolution

By rolling out tri-plane feature maps and applying 3D-aware convolution, the Rodin model performs 3D-aware diffusion using an efficient 2D architecture. As analyzed in Section 3.2, tri-plane roll-out and 3D-aware convolution are crucial for high-fidelity results, especially for thin structures such as hair strands and clothing details, by enhancing cross-plane communication. To validate the impact of these designs in high-quality tri-plane, we modify the upsample diffusion model with different configurations and remove the convolution refinement with the base diffusion fixed. Figure 16 demonstrates with rollout and 3D-aware convolution, the full model shows a clear improvement compared to the base model.

Scale	w/o CFG	1.2	1.5	3.0	6.0
PSNR	24.06	**24.21**	24.07	24.05	24.15
SSIM	**0.795**	0.794	0.792	0.782	0.775
LPIPS	0.128	**0.121**	0.133	0.141	0.146

Table 4. Quantitative results of conditional avatar reconstruction.

C.4. Nearest Neighbors Analysis

The Rodin model enables a hassle-free creation experience of an unlimited number of avatars from scratch, each avatar being distinct. Figure 17 shows the nearest neighbors of some generated samples in the main paper, which indicates that the model does not simply memorize the training data.

C.5. Conditional Avatar Generation

Quantitative metrics. On top of unconditional generation, we can also hallucinate a 3D avatar from a single reference image by conditioning the base generator with the CLIP image embedding for that input image. We evaluate the conditional generation on 1K test data where each subject contains 300 images from different views. Table 4 reports the metrics between reconstructed images and ground-truth synthetic images.

Classifier-free guidance. Our model supports classifier-free guidance (CFG) sampling when inference, which is a technique typically used to boost the sampling quality in conditional generation. Table 4 evaluates generation quality with different scales of classifier-free guidance in terms of PSNR, SSIM and LPIPS.

D. Additional Visual Results

Figure 18 and Figure 19 show more random samples generated by the Rodin model, showing the capability to synthesize high-quality 3D renderings with impressive details. To reflect the geometry, we also extract the mesh from the generated density field using marching cubes, which demonstrates high-fidelity geometry. Figure 20 gives uncurated generated samples, which possess visually-pleasing quality and diversity. We also explore the interpolation of the latent condition $\boldsymbol{z}$ between two generated avatars, as

Figure 17. Nearest neighbors in the training data according to CLIP feature similarity.

shown in Figure 21, where we observe consistent interpolation results with smooth appearance transition. Figure 21 shows additional results of creating 3D portraits from a single reference image.

E. Societal Impact

The Rodin model aims to enable a low-cost, fast and customizable creation experience of 3D digital avatars that refer to the traditional avatars manually created by 3D artists, as opposed to photorealistic avatars. The reason for focusing on digital avatars is twofold. On the one hand, digital avatars are widely used in movies, games, the metaverse, and the 3D industry in general. On the other hand, the available digital avatar data is very scarce as each avatar has to be painstakingly created by a specialized 3D artist using a sophisticated creation pipeline, especially for modeling hair and facial hair.

Rather than collecting real photos, all our training images are rendered by Blender. Such synthetic data can mitigate the privacy and copyright concerns that existed in real face collection. Another advantage of using synthetic data is that we could have control over the variation and diversity of rendered images, eliminating the data bias in existing face datasets. Also, digital avatars are easier to be distinguished from real people compared with photo-realistic avatars, hindering misuse for impersonating real persons. Nonetheless, the 3D portrait reconstruction and text-based avatar customization could still be misused for spreading disinformation maliciously, like all other AI-based content generation models. We caution that the high-quality renderings produced by our model may potentially be misused and viable solutions so avoid this include adding tags or watermarks when distributing the generated photos.

This work successfully generalizes the power of diffusion models from 2D to 3D and is promising to offer the new design tool for 3D artists which could significantly save the costs of the traditional 3D modeling and rendering pipeline. In the next we intend to explore the possibility of modeling general 3D scenes using the same technique and investigate novel applications such as Lego and architect designs.

Figure 18. Unconditional generation samples by our Rodin model. We visualize the mesh extracted from the generated density field.

Figure 19. Unconditional generation samples by our Rodin model.

Figure 20. Uncurated generation results by our Rodin model.

Figure 21. Latent interpolation results for generated avatars.

Figure 22. Additional results of portrait inversion.