RodinHD: High-Fidelity 3D Avatar Generation with Diffusion Models
Supplementary Material

Bowen Zhang[1*], Yiji Cheng[2*], Chunyu Wang[3†], Ting Zhang[3], Jiaolong Yang[3], Yansong Tang[2], Feng Zhao[1], Dong Chen[3], and Baining Guo[3]

[1] University of Science and Technology of China
[2] Tsinghua University
[3] Microsoft Research Asia

A Additional Implementation Details

Triplane fitting. We split the triplane fitting into two stages to reduce computation costs. In the first stage, we jointly train the MLP decoder and the triplanes on a subset of 64 avatars. During each inner loop iteration, 8192 rays are randomly sampled for loss calculation. We optimize one avatar per GPU for each outer loop iteration due to large GPU memory consumption. The detailed hyper-parameters of the first stage are listed in Tab. 1, including ablation studies. In the second stage, we fix the decoder's weights and fine-tune the triplanes of $46K$ avatars independently. The fitting iteration for each avatar is set to 25000. For rendering efficiency, an occupancy grid of 128^3 resolution is maintained [9] to skip ray marching steps in empty space. Since we do not have the occupancy grid of the diffusion generated triplane, we update the occupancy grid 16 times from zero initialization before performing volumetric rendering.

Diffusion training. For triplane $\boldsymbol{x} = (\boldsymbol{x}_{uv}, \boldsymbol{x}_{wu}, \boldsymbol{x}_{vw})$ of shape $\mathbb{R}^{3\times H\times W\times C}$, we perform triplane roll-out $\overline{\boldsymbol{x}} = \texttt{hstack}(\boldsymbol{y}_{uv}, \boldsymbol{y}_{wu}, \boldsymbol{y}_{vw}) \in \mathbb{R}^{H\times 3W\times C}$ in order to employ the well-designed 2D UNet model in diffusion [6, 10] following [12]. We also leverage 3D-aware convolution [12] for cross-plane feature communication. The portrait image is resized to 256×256 and the resulting multi-scale features have the resolution of 128×128, 64×64, and 32×32. The conditional features are injected to the base diffusion model at layers with resolutions of 32×32, 16×16, and 8×8, respectively, through cross attention. The upsample diffusion model only uses the conditional features of 128×128, which are injected to the middle latent features.

For our base diffusion model, we adopt the UNet model architecture from [6]. We train our base model using AdamW optimizer [8] with a learning rate $1e-5$. To condition on the multi-scale image features of input portrait as illustrated in Sec. 3.2, we perform cross attention at resolutions $(32, 16, 8)$. Our optimized noise schedule is based on the cosine schedule mentioned in [3], and we further adjust its hyper-parameters for 3D diffusion training. We provide the detailed configurations of the model and diffusion below.

---

* Interns at Microsoft Research Asia. Equal contribution. †Corresponding author.

Table 1: Hyper-parameters for the first stage of fitting, including ablation studies.

	Rodin (512)	+ Task relay	+ Wight decay	Ours
Inner loop iterations	15000	5000	5000	5000
Outer loop iterations per avatar	1	30	30	30
Loss Weight of TV regularization	1e-2	1e-2	1e-2	1e-2
Loss Weight of L2 regularization	1e-4	1e-4	1e-4	1e-4
Loss Weight of IWC regularization	0	0	0	0.1
Loss Weight of weight decay	0	0	1e-4	0
Triplane learning rate	2e-3	2e-3	2e-3	2e-3
Decoder learning rate	2e-4	2e-4	2e-4	2e-4
Ray batch size	8192	8192	8192	8192
Samples per ray	1024	1024	1024	1024

For our upsample diffusion model, we also adopt the UNet model architecture from [6]. We train our upsample model using Adam optimizer [7] with a learning rate $1e-5$. We remove self-attention due to unaffordable computation cost at high resolutions, and perform cross attention at resolution 128 for conditioning on input portrait features. Our optimized noise schedule for upsample diffusion is based on the sigmoid noise schedule in [3], then we carefully adjust the hyper-parameters for 3D diffusion training. The detailed configurations of the model and diffusion are shown below.

```
# 128x128 Base diffusion
UNet configuration = {
    "channels": 192,
    "channel_mult": (1, 1, 2, 3, 4),
    "embed_dim": 768,
    "num_res_blocks": (3, 3, 3, 3, 3),
    "attn_resolutions": (32, 16, 8),
    "ms_vae_feature_cross_attn_res": (32, 16, 8),
    "3D_aware_conv_res": (128),
    "dropout": 0,
    "feature_pooling_type": "attention",
    "use_scale_shift_norm": True
}

Diffusion configuration = {
    "Training steps": 1000,
    "Noise schedule": Cosine(start=0.2, end=1, tau=3),
    "Inference steps": 10,
    "Inference sampler": "DDPM"
}
```

```
# 128x128 -> 512x512 Upsample diffusion
UNet configuration = {
    "channels": 128,
    "channel_mult": (1, 2, 4),
    "embed_dim": 512,
    "num_res_blocks": (2, 2, 6),
    "ms_vae_feature_cross_attn_res": (128),
    "3D_aware_conv_res": (512, 256, 128),
    "dropout": 0,
    "feature_pooling_type": "attention",
    "use_scale_shift_norm": False
}

Diffusion configuration = {
    "Training steps": 100,
    "Noise schedule": Sigmoid(start=0, end=3, tau=0.1),
    "Inference steps": 10,
    "Inference sampler": "DDPM"
}
```

Table 2: Comparison fitting quality (PSNR) of Triplane Resolution and Channel.

Res. \ *Ch.*	4	8	16	32
128	30.15	30.71	31.21	31.59
256	30.24	31.01	31.44	31.67
512	30.38	31.31	31.60	**31.71**

B Additional Analysis and Visualization

Choices of triplane resolution and channel. We argue that both triplane resolution and channel affect the preservation of high-frequency information in renderings. To validate this argument, we experiment with different triplanes from a resolution set of $\{128, 256, 512\}$ and a channel set of $\{4, 8, 16, 32\}$ to fit 1024×1024 images of one subject and show the results in Tab. 2 and Fig. 1. Overall, the fitting quality increases with the triplane resolution and channel. High-resolution triplane can render high-frequency detail, and low-resolution triplane tends to produce blurring results. On the other hand, the triplane with more channels can keep high-fidelity appearance without introducing noisy pattern, but low-resolution triplane with more channels cannot achieve better high-frequency detail preservation than high-resolution triplane. We thus choose to utilize $512 \times 512 \times 32$ triplanes in our experiments.

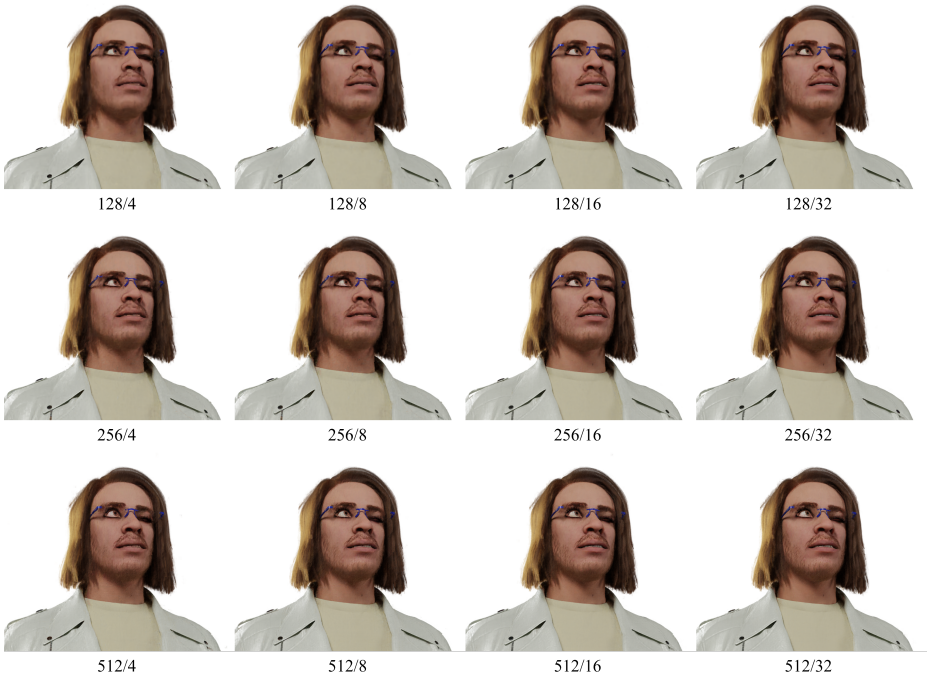

Fig. 1: Comparison on Triplane Resolution and Channel. Zoom in for better visualization.

Visualization of intermediate results in the denoising process. During inference, our model starts from isotropic Gaussian noise and progressively reduces the noise to obtain the final high-quality triplanes. We visualize the renderings of generated results $\boldsymbol{x}_t$ of intermediate timesteps $t \in (0, 1)$ in the denoising process to provide a comprehensive understanding of the triplane diffusion procedure. From Fig. 2, we observe that our model establishes the global structure of the avatar, and subsequently adds more detail, which is similar to [11, 12].

C Additional Comparison

More quantitative comparison with Rodin. We additionally compare our method with Rodin [12] on conditional avatar generation using more evaluation metrics. We evaluate the cosine similarity of identity embedding derived from ArcFace [4] between generated avatars and ground-truths (CSIM), as well as between paired renderings of generated avatars from different camera viewpoints (CSIM-CrossView). We also include Average Expression Distance (AED), Average Pose Distance (APD) and Average Shape Distance (ASD) between the

Fig. 2: Visualization of intermediate results in the denoising process.

Table 3: Additional comparison of conditional avatar generation. **The subscript * indicates that 2D refinement is applied to the rendered images**.

Models	FID↓	PSNR↑	CSIM↑	CSIM-CrossView↑	AED↓	APD%↓	ASD↓
Rodin	33.20	18.28	0.64	**0.85**	0.21	2.21	**0.44**
Rodin*	**20.51**	17.31	0.63	0.83	0.20	2.21	**0.44**
Ours	26.49	**20.33**	**0.68**	**0.85**	**0.18**	**1.78**	**0.44**

reconstructed 3D faces [5] of generated avatars and ground-truth avatars. The results presented in Tab. 3 demonstrate that our model excels in preserving identity and accurately generating expression, pose, and geometry. While Rodin's 2D refinement achieves lower FID scores, it struggles to maintain the identity and expression details of the conditioned portraits.

Comparison of 3D consistency with EG3D. We additionally compare our 3D consistency with SOTA 3D-aware GANs, EG3D. We evaluate the 3D consistency of unconditional generated results in Fig. 3, which is similar to Fig. 9 of main paper. Since EG3D also utilizes a 2D super-resolution module, the results in Fig. 3 yield obvious texture flickering, whereas our method leads to a natural and smooth texture pattern. We also provide numerical comparison in Tab. 5 by fitting a NeuS model from generated multi-views following Tab. 3 of main paper. Our generated results achieve significantly better metrics due to multi-view consistency.

D Additional Results

Conditional avatar generation. We provide more renderings of generated avatars conditioned on the single portraits from our test set in Fig. 5. Our model is capable of creating high-fidelity avatars with compelling details and vivid expressions, demonstrating the strong capability of the proposed model.

Unconditional avatar generation. Fig. 6 show more unconditional avatars created by our model. Our model is able to produce diverse high-quality avatars with rich details, including complex clothing and hairstyles.

Avatar creation from in-the-wild portrait. In Fig. 7, we present additional generated avatars conditioned on real-world images. Our methodology demonstrates a higher fidelity in preserving the identity of the subjects when compared with [12]. Furthermore, our results exhibit a remarkable ability to retain intricate details such as hairstyle and clothing attributes.

Table 4: Average ranking of user study in conditional generation.

Models	ID Similarity↓	3D Consistency↓	Visual Fidelity↓
Rodin	2.54	2.12	2.61
Rodin*	2.06	2.77	**1.26**
Ours	**1.39**	**1.11**	2.13

EG3D **Our RodinHD**

Fig. 3: Visual comparison of 3D consistency akin to the Epipolar Line Images [2]. Our model yields smooth and natural texture, whereas EG3D produces obvious texture flickering, indicating the 3D inconsistency with 2D refinement.

Text-to-avatar creation. We provide more samples of high-quality text-to-avatar creation in Fig. 8. We first convert the text prompt to reference portrait by our finetuned 2D text-to-image diffusion models, thereafter generate a high-fidelity avatar conditioned on the reference portrait. It is worth noticing that the trigger word we used "Blender Synthetic Avata" is not necessarily needed to be added in the prompts since we can omit it and perform cropping and alignment to the generated images, similar to how we handle realistic image inputs.

User study. We further conduct user study to measure the identity similarity (ID), 3D consistency and visual fidelity. We ask 15 subjects to rank different methods with 20 sets of comparisons in each study. The average ranking in Tab. 4 shows that our method earns user preferences the best in identity preservation and 3D consistency, only slightly worse than Rodin adding 2D refinement (Rodin*) in fidelity. We think more follow-up research can be conducted to further improve the visual quality while ensuring the 3d consistency.

Additional video results. We also provide an additional video in the website including the above generation results and ablation of 3D consistency. The video

Table 5: Quantitative comparison of 3D consistency with EG3D.

	PSNR↑	SSIM↑	LPIPS↓
EG3D	29.51	0.962	0.052
Ours	**33.39**	**0.967**	**0.043**

(a) (b)

Fig. 4: Failure cases.

also demonstrates that our model is able to create high-fidelity avatars with strong 3D consistency.

E Responsible AI Considerations

Our model is trained on the synthetic dataset [13] of 3D digital avatars akin to those crafted by artists, as opposed to photo-realistic humans. This approach to training data selection alleviates privacy and copyright concerns associated with the use of real human face collections. Despite these precautions, it is important to acknowledge that 3D avatar created by our model from real-world images could potentially be exploited for the dissemination of disinformation, similar to other generative models. We must therefore emphasize the importance of responsible use of our technology. As a safeguard against misuse, we recommend the implementation of measures such as embedding visible tags or watermarks into the distributed renderings produced by our model.

F Limitation

As illustrated in Fig. 4, our model still has some limitations. Floating points occasionally appear in the generated avatars as shown in Fig. 4 (a), which are typical NeRF artifacts [1]. Handling glasses remains challenging due to limited training data in Fig. 4 (b).

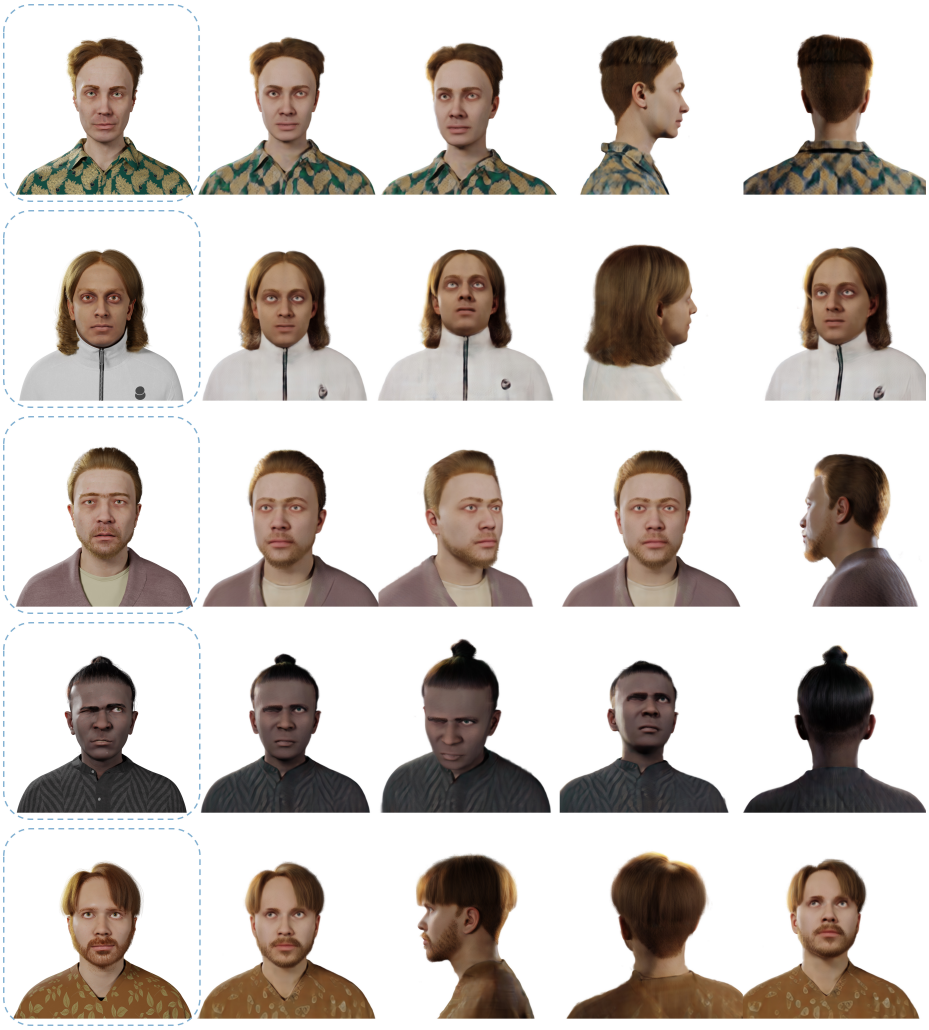

Fig. 5: Conditional generation samples by our model. Reference portraits are shown in dashed boxes.

Fig. 6: Unconditional generation samples by our model.

Fig. 7: Samples of generated avatars conditioned on single in-the-wild portraits. Compared with Rodin, our method preserves more details of identity and clothing.

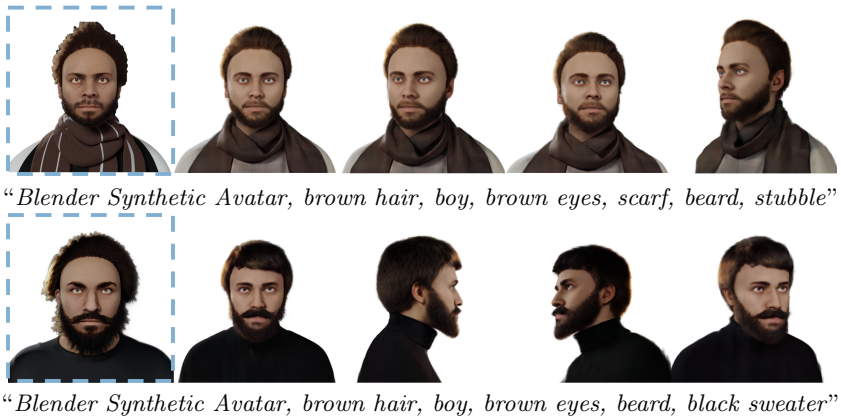

Fig. 8: Samples of text-to-avatar creation using our model. The leftmost reference portraits are first created by based finetuned 2D diffusion model given the text prompts. Then our 3D diffusion model is able to create high-fidelity avatars conditioned on the generated reference portraits.

References

1. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5470–5479 (2022)
2. Bolles, R.C., Baker, H.H., Marimont, D.H.: Epipolar-plane image analysis: An approach to determining structure from motion. International Journal of Computer Vision **1**(1), 7–55 (1987)
3. Chen, T.: On the importance of noise scheduling for diffusion models. arXiv preprint arXiv:2301.10972 (2023)

4. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
5. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: 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. pp. 0–0 (2019)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations, ICLR (2019)
9. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
10. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
11. Shue, J.R., Chan, E.R., Po, R., Ankner, Z., Wu, J., Wetzstein, G.: 3d neural field generation using triplane diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20875–20886 (2023)
12. Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., et al.: Rodin: A generative model for sculpting 3d digital avatars using diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4563–4573 (2023)
13. Wood, E., Baltrušaitis, T., Hewitt, C., Dziadzio, S., Cashman, T.J., Shotton, J.: 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. pp. 3681–3691 (2021)