StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation Supplementary Material

Roy Or-El[1] Xuan Luo[1] Mengyi Shan[1] Eli Shechtman[2]
Jeong Joon Park[3] Ira Kemelmacher-Shlizerman[1]
[1]University of Washington [2]Adobe Research [3]Stanford University

Appendix

In this appendix we provide additional qualitative and quantitative results on our approach, along with the technical details that supplement the main paper. In Sec. Appendix A, we discuss possible societal impacts of our technology. Then, we present additional experiments on the view-consistency of our RGB renderings via image reprojection (Sec. Appendix B). We further demonstrate the quality of our 3D shapes in Sec. Appendix C. In Sec. Appendix D, we describe the content of the supplementary videos and introduce a geometry-aware noise injection procedure to reduce flickering. Next, we discuss implementation details and the merits of our proposed sampling strategy in (Sec. Appendix E and Appendix F respectively. We conduct ablation studies on our approach in Appendix G). Finally, we continue our discussion on our method's limitations in Sec. Appendix H.

A. Societal Impacts

Image and 3D model-generating technologies (e.g., deepfakes) could be used for spreading misinformation about existing or non-existing people [3, 8]. Our proposed technology allows generating multi-view renderings of a person, and might be used for creating more realistic fake videos. These problems could potentially be addressed by developing algorithms to detect neural network-generated content [12]. We refer readers to [11] for strategies of mitigating negative social impacts of neural rendering. Moreover, image generative models are optimized to follow the training distribution, and thus could inherit the ethnic, gender, or other biases present in the training dataset. A possible solution is creating a more balanced dataset, e.g., as in [4].

B. View Consistency of RGB Renderings

B.1. Volume Rendering Consistency

The consistent volume rendering from our SDF-based technique naturally leads to high view consistency of our RGB renderings. To show the superior 3D-consistency of our SDF-based volume rendering, we measure the reprojection error when a side view pixels are warped to the frontal view. We randomly sample 1,000 identities and render the depth and RGB images at 256×256 and set the side view to be $1.5\times$ the standard deviation of the azimuth distribution in training (which is 0.45 radians for FFHQ and 0.225 radians for AFHQ). We reproject the side-view RGB renderings to the frontal view using the side-view depth, and we do not ignore occluded pixels. We measure color inconsistency with the median of pixel-wise L1 error in RGB (0 - 255), averaged over the 1,000 samples. The use of median effectively removes the large errors coming from occlusions. Note that since PiGAN is trained with center-cropped FFHQ images (resized to 320×320 and center-cropped to 256×256), we apply the same transformation on our results before computing the median.

Dataset:	FFHQ	AFHQ
PiGAN [2]	14.7	16.5
Ours (volume renderer)	**2.9**	**2.6**

Table 1. Quantitative view-consistency comparison of the RGB renderings. We evaluate the color error of the RGB volume renderings between the frontal view and the reprojection from a fixed side view. The error is measured as the median of the per-pixel mean absolute difference (0 - 255). We average the color inconsistency over 1,000 samples for each dataset. Our underlying SDF geometry representation promotes superior 3D consistency. (also see Fig. 1).

As shown in Tab. 1, StyleSDF presents significantly improved color consistency compared to the strongest current baseline, PiGAN [2]. Fig. 1 shows the sample depth and color rendering pairs used for the evaluation, along with the pixelwise error maps. The error maps demonstrate that our volume RGB renderings have high view consistency, as the large reprojection errors are mostly in the occluded regions. On the other hand, PiGAN's reprojections do not align well with the frontal view, showing big errors also near the eyes,

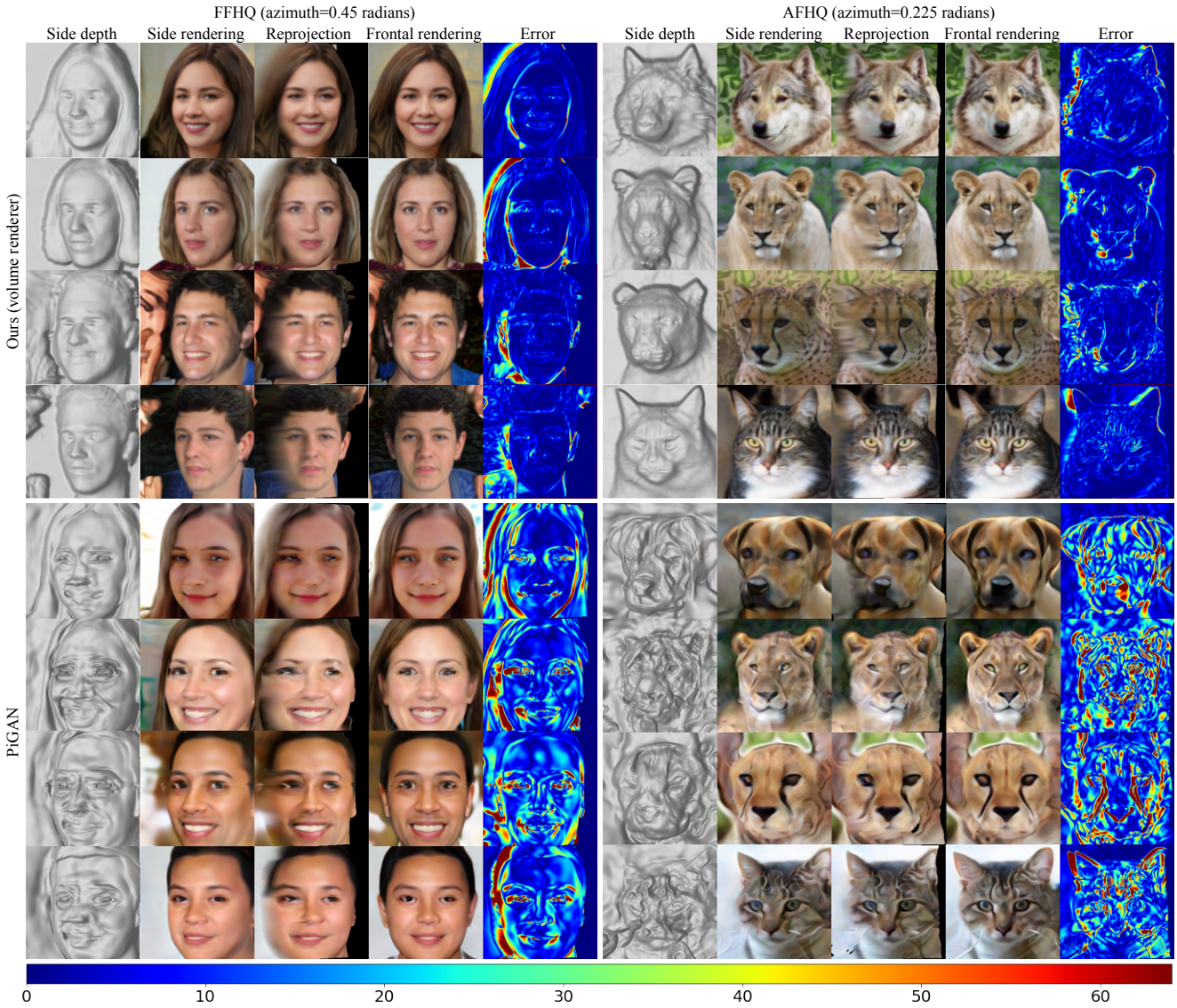


Figure 1. Qualitative view consistency comparison of RGB renderings. We project the rendering from a side view using its corresponding depth map to the frontal view. We compare the reprojection to the frontal-view rendering and compute the error map showing mean absolute difference in RGB channels (0 - 255). Our SDF-based technique generate superior depth quality and significantly improves the view-consistency of the RGB renderings. Most of our errors concentrate on the occlusion boundaries whereas PiGAN's errors spread across the whole subject (e.g., eyes, mouth, specular highlights, fur patterns).

mouth, in presence of specular highlights, etc.

B.2. High-Resolution RGB Consistency

In Fig. 2, we present the reprojection experiment results using our high-resolution RGB outputs. As in the volume rendering consistency experiment, we reproject the RGB pixels from non-frontal views (with varying azimuth and elevation) to the frontal views. The results demonstrate the strong 3D-consistency of our high-resolution images, as the reprojected non-frontal images are similar to the frontal renderings. However, as mentioned in the limitation section of the main paper, the current implementation of StyleGAN2 comes with significant aliasing of the high-frequency components, resulting in noticeable pixel errors on regions with high-frequency details, e.g., hair, ears, eyes, etc. To identify the errors in the high-frequency details, we visualize the mean reprojection images. I.e. we project non-frontal views and average the pixel values across views. As can be seen in Fig. 3, the mean reprojection images closely replicate the identities and important structures of the frontal view, demonstrating strong view-consistencies. The error map confirms that most of the errors are concentrated on

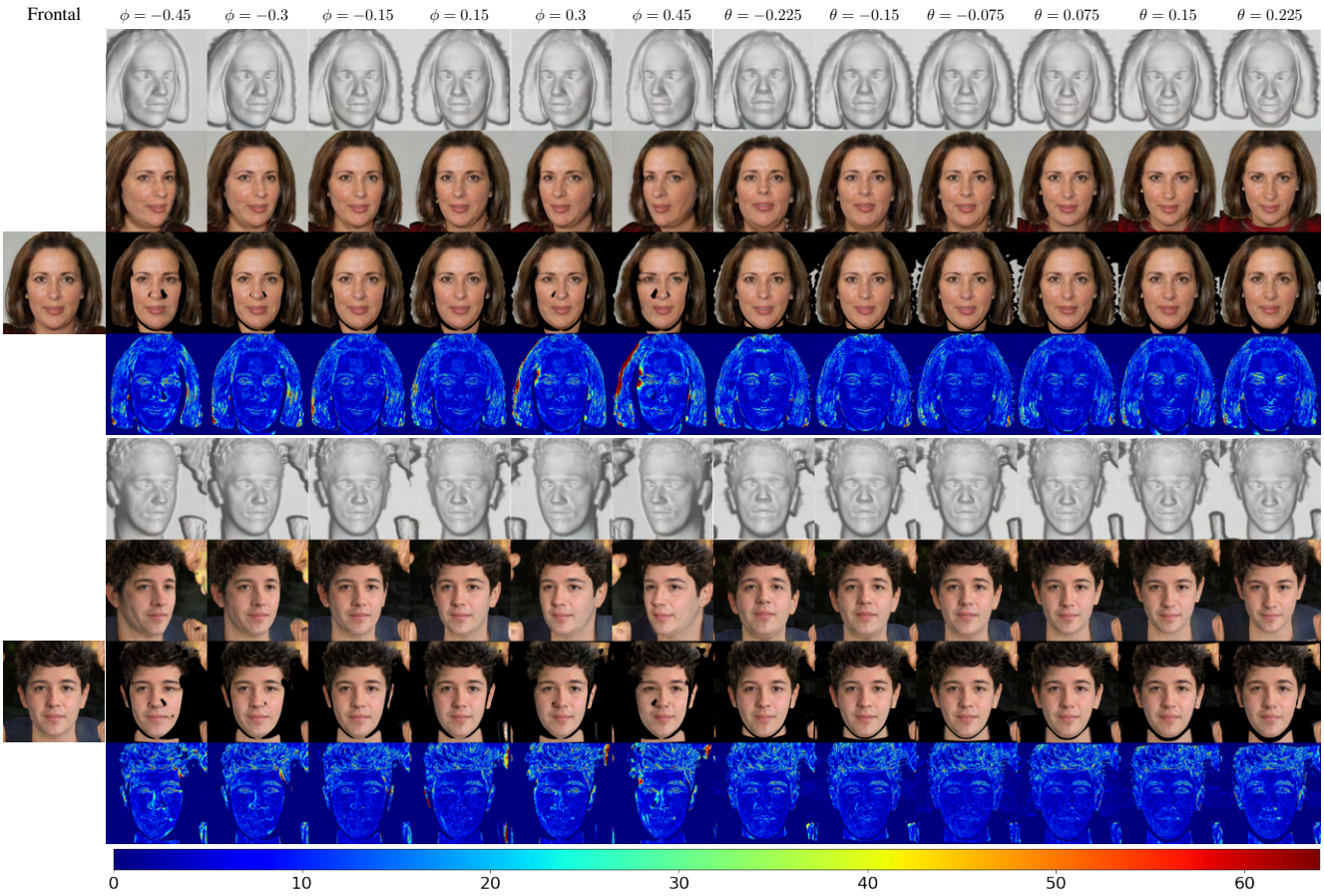


Figure 2. View-consistency visualization of high-resolution renderings. We use the side-view depth maps (first rows) to warp the side-view RGB renderings (second rows) to the frontal view (first column). The reprojected pixels that pass the occlusion testing are shown in the third row. We compare the reprojections with the frontal-view renderings and show the per-pixel error maps (fourth rows). Our reprojections well align with the frontal view with errors mostly in the occlusion boundaries and high-frequency details.

the high-frequency noise of the StyleGAN generator.

C. Qualitative 3D results

We demonstrate the consistency of our 3D representation by overlaying the point clouds from the frontal and side view depth maps (Fig. 4b). The visualization, shown in two different colors, clearly shows high consistency between the depth maps. To show the quality and plausibility of our 3D models, we extract meshes on our SDFs via marching cubes and visualize them in extreme angles (Fig. 4c).

D. Video Results

Since our 3D-consistent high-resolution image generation can be better appreciated with videos, we have attached 24 sequences in the supplementary material, featuring view-generation results on the two datasets using two different camera trajectories. For each identity, we provide two videos, one for RGB and another for depth rendering. The videos are presented in the project's website.

D.1. Geometry-Aware StyleGAN Noise

Even though the images shown in the main paper on multi-view RGB generation look highly realistic, we note that for generating a video sequence, the random noise of StyleGAN2 [6], when naïvely applied to 2D images, could result in severe flickering of high-frequency details between frames. The flickering artifacts are especially prominent for the AFHQ dataset due to high-frequency textures from the fur patterns.

Therefore, we aim at reducing this flickering by adding the Gaussian noise in a 3D-consistent manner, i.e., we want to attach the noise on the 3D surface. We achieve this by extracting a mesh (at 128 resolution grid) for each sequence from our SDF representation and attach a unit Gaussian noise to each vertex, and render the mesh using vertex coloring. Since higher resolution intermediate features require up to 1024×1024 noise map, we subdivide triangle faces of

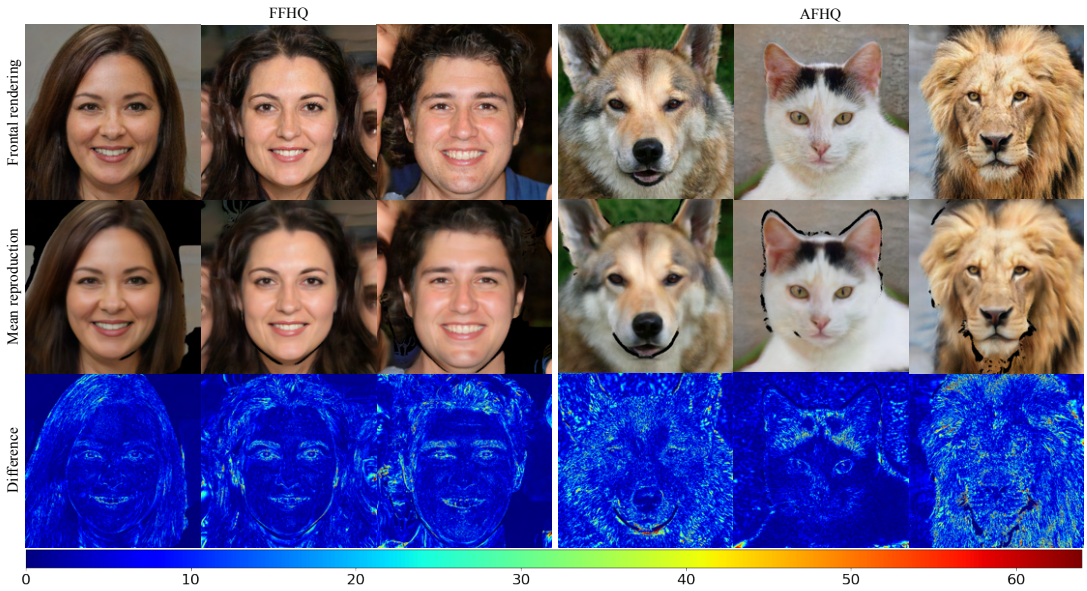
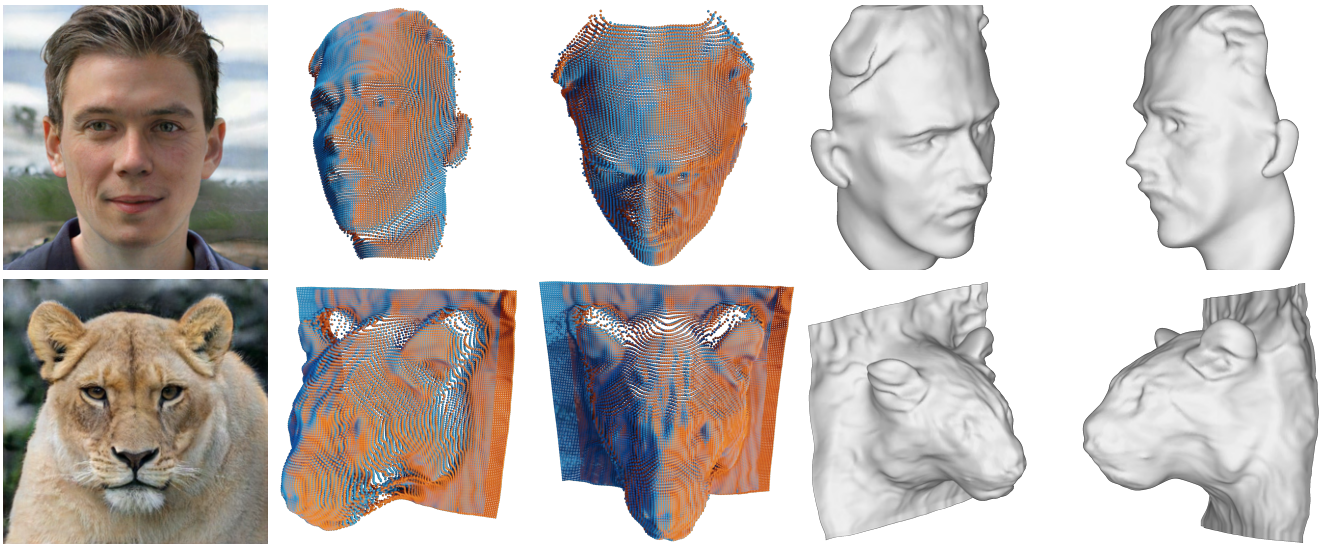


Figure 3. Color consistency visualization with mean faces. We reproject the RGB renderings from the side views to the frontal view (as in Fig. 2). We show the mean reprojections that pass the occlusion testing and their differences to the frontal-view renderings. The mean reprojections are well aligned with the frontal rendering. The majority of the errors are in the high-frequency details, generated from the random noise maps in the StyleGAN component. This demonstrates the strong view consistency of our high-resolution renderings.



(a) Color (b) Overlaid Depth Maps (c) Extracted Marching Cubes

Figure 4. Consistent and plausible 3D shapes. (a) Color images. (b) Overlaid point clouds extracted from frontal and side view depth maps. (c) Marching cubes meshes, rendered from extreme angles.

the extracted mesh once every layer, starting from 128×128 feature layers. The video results show that the geometry-aware noise injection reduce the flickering problem on the AFHQ dataset, but noticeable flickering still exist. Furthermore, we observe that the geometry-aware noise slightly sacrifices individual frame's image quality, presenting less pronounced high-frequency details, likely due to the change of the Gaussian noise distribution during the rendering process. The videos rendered with geometry-aware noise can be viewed at the project's website.

E. Implementation Details

E.1. Dataset Details

FFHQ: We trained FFHQ with R1 regularization loss of 10. The camera field of view was fixed to 12° and its azimuth and elevation angles are sampled from Normal distributions with zero mean and standard deviations of 0.3 and 0.15 re-

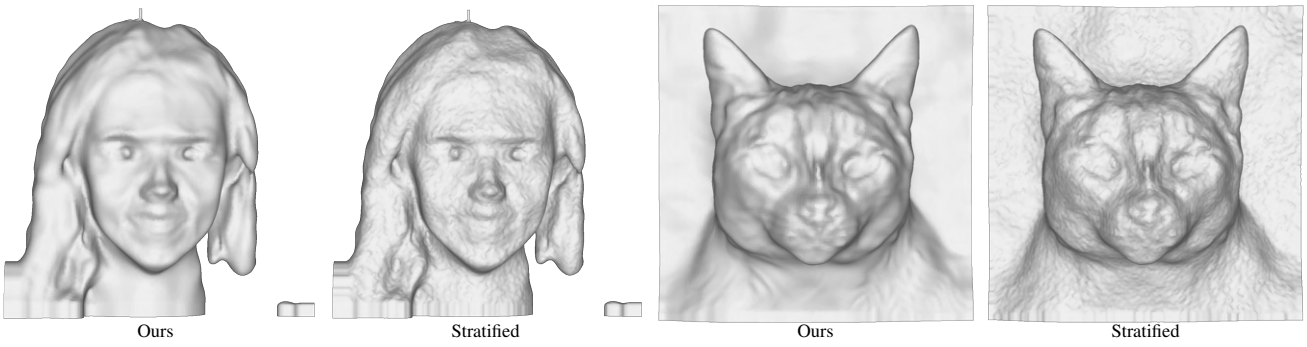


Figure 5. We compare extracted meshes using our sampling strategy vs. stratified sampling. Note the noise induced by stratified sampling. (zoom in for details)

spectively. We set the near and far fields to $[0.88, 1.12]$ and sample 24 points per ray during training .We trained our volume renderer for $200k$ iterations and the 2D-Styled generator for $300k$ iterations.

AFHQ: The AFHQ dataset contains training and validation sets for 3 classes, cats, dogs and wild animals. We merged all the training data into a single training set. We apply R1 regularization loss of 50. Both azimuth and elevation angles are sampled from a Gaussian distribution with zero mean and standard deviation of 0.15 and a camera field of view of $12°$. The near and far fields as well as the number of samples per ray are identical to the FFHQ setup. Our volume renderer as well as the 2D-Styled generator were trained for $200k$ iterations.

E.2. Training Details

Sphere Initialization: During our experiments we have noticed that our SDF volume renderer can get stuck at a local minimum, which generates concave surfaces. To avoid this optimization failure, we first initialize the MLP to generate an SDF of a sphere centered at the origin with a fixed radius. We analytically compute the signed distance of the sampled points from the sphere and fit the MLP to match these distances. We run this procedure for $10k$ iterations before the main training. The importance of sphere initialization is discussed in Appendix G.

Training setup: Our system is trained in a two-stage strategy. First, we train the backbone SDF volume renderer on 64×64 images with a batch size of 24 using the ADAM [9] optimizer with learning rates of $2 \cdot 10^{-5}$ and $2 \cdot 10^{-4}$ for the generator and discriminator respectively and $\beta_1 = 0, \beta 2 = 0.9$. We accumulate gradients in order to fit to the GPU memory constraints. For instance, a setup of 2 NVIDIA A6000 GPUs (a batch of 12 images per GPU) requires the accumulation of two forward passes (6 images per forward pass) and takes roughly 3.5 days to train. We use an exponential moving average model during inference.

In the second phase, we freeze the volume renderer weights and train the 2D styled generator with identical setup to StyleGAN2 [7]. This includes ADAM optimizer with 0.002 learning rate and $\beta_1 = 0, \beta_2 = 0.99$, equalized learning rate, lazy R1 and path regularization, batch size of 32, and exponential moving average. We trained the styled generator on 8 NVIDIA TeslaV100 GPUs for 7 days.

F. Sampling Strategy

NeRF [10], along with existing 3D-aware GANs like PiGAN [2], rely on hierarchical sampling strategy for obtaining more samples near the surface. Our use of SDFs allows sampling the volume with smaller number of samples without sacrificing the surface quality, thereby reducing the memory footprints and simplifying the implementation.

Stratified sampling randomizes the distance between adjacent samples along each ray, adding undesired noise to the volume rendering (Fig. 5). The randomness also amplifies flickering in RGB videos. Our sampling strategy ensures that the integration intervals are of the same length, which eliminates the noise and results in smoother volume rendering outputs.

G. Ablation studies

We perform two ablation studies to show the necessity of the minimal surface loss (see main paper) and the sphere initialization. As can be seen in Figure 6, on top of preventing spurious and non-visible surfaces from being formed, the minimal surface loss also helps to disambiguate between shape and radiance. Penalizing values that are close to zero essentially minimizes the surface area and makes the network prefer smooth SDFs.

In Figure 7, we show the importance of the sphere initialization in breaking the concave/convex ambiguity. Without properly initializing the weights, the network gets stuck at a local minimum that generates concave surfaces. Although concave surfaces are physically incorrect, they can perfectly

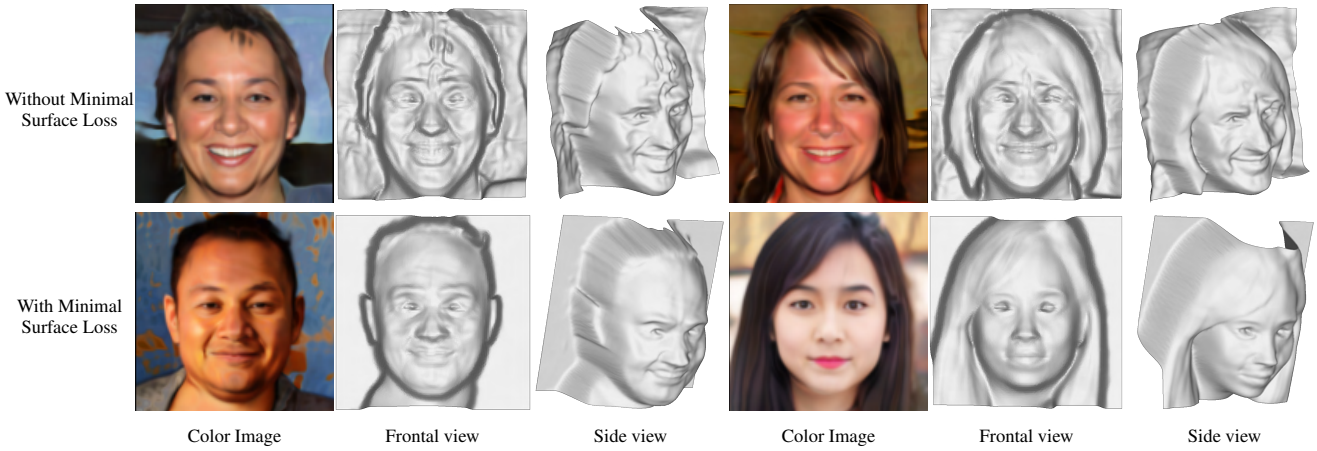


Figure 6. Minimal surface loss ablation study. We visualize the volume rendered RGB and depth images from volume renderers trained with and without the minimal surface loss. The Depth map meshes are visualized from the front and side views. Note how a model trained with the minimal surface loss generates smoother surfaces and is less prone to shape-radiance ambiguities, e.g., specular highlights are baked into the geometry.

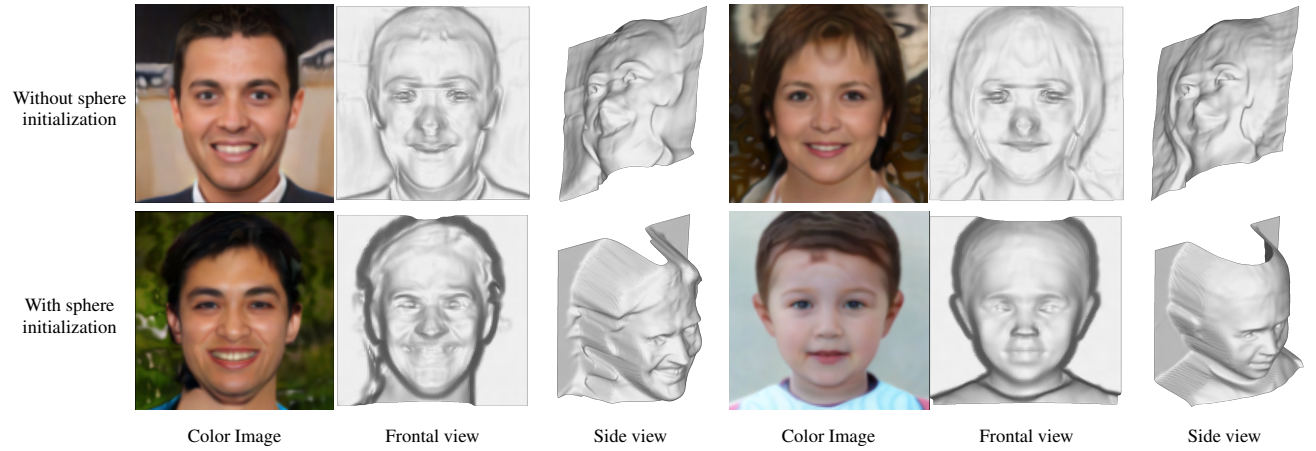


Figure 7. Sphere initialization ablation study. We visualize volume-rendered RGB and depth images from volume renderers trained with and without sphere initialization. The Depth map meshes are visualized from the front and side views. Note how a model trained without model sphere initialization generates concave surfaces.

explain multi-view images, as they are essentially the "mirror" surface. Concave surfaces cause the images to be rendered in the opposite azimuth angle, an augmentation that the discriminator cannot detect as fake. Therefore, the generator cannot recover from the this local minima.

H. Limitations (continued)

As mentioned in the main paper, our high-resolution generation network is based on the implementation of StyleGAN2 [6], and thus might experience the same aliasing and flickering at regions with high-frequency details (e.g., hair), which are recently addressed in Alias-free GAN [5] or Mip-NeRF [1]. Moreover, we observe that the reconstructed geometry for human eyes contain artifacts, characterized by concave, instead of convex, eye balls. We believe that these artifacts often lead to slight gaze changes along with the camera views. As stated in the main paper, our current implementation of volume rendering during inference uses fixed frontal view directions for RGB queries $\mathbf{c}(\mathbf{x}, \mathbf{v})$, and thus cannot express moving specular highlights along with the camera.

I. Additional Results

We show uncurated set of images generated by our networks (Fig. 8).

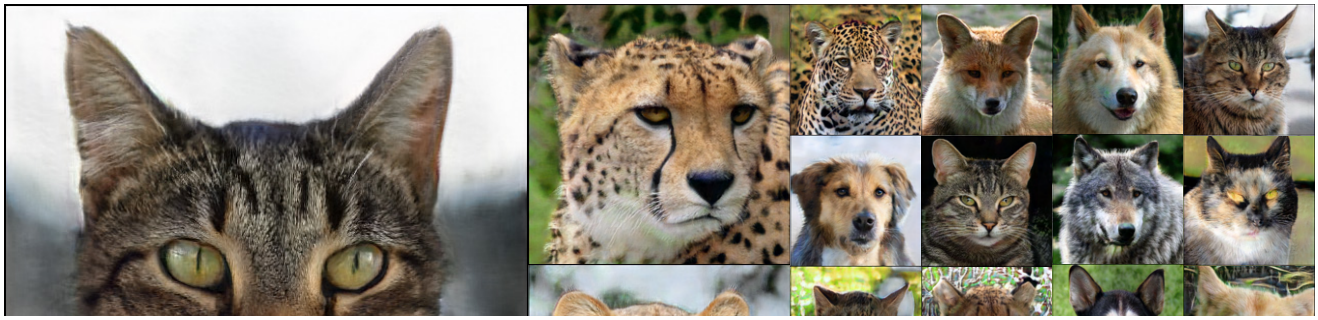


Figure 8. Uncurated high-resolution RGB images that are randomly generated by StyleSDF.

References

[1] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, pages 5855–5864, October 2021. 6

[2] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, pages 5799–5809, 2021. 1, 5

[3] K Hill and J White. Designed to deceive: Do these people look real to you. *New York Times*, 2020. 1

[4] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, pages 1548–1558, 2021. 1

[5] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *arXiv preprint arXiv:2106.12423*, 2021. 6

[6] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 3, 6

[7] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 5

[8] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 1

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for

stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[10] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421. Springer, 2020. 5

[11] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, volume 39, pages 701–727. Wiley Online Library, 2020. 1

[12] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, pages 8695–8704, 2020. 1