RenderDiffusion: Image Diffusion for 3D Reconstruction, Inpainting and Generation
Supplementary Material

S1. Overview

In this supplementary material, we provide additional architecture details (Section S2), additional results for unconditional generation (Section S3), additional results for 3D-aware inpainting (Section S4), an additional experiment where generate multiple ShapeNet categories with a single model (Section S5), and additional results for reconstruction (Section S6).

S2. Architecture Details

Here we summarise the architecture of the denoiser network. Code, training configurations and datasets are publicly available at `https://github.com/Anciukevicius/RenderDiffusion`.

Triplane encoder. The triplane encoder transforms the input image of size $M \times M \times 3$ into a triplane representation of size $N \times N \times 3n_f$. We choose $M = 64$ and $N = 256$ for our experiments on ShapeNet, as we found that the increased triplane resolution improves the quality of our results, and $M = 32$, $N = 32$ for CLEVR1. Similar to other 2D diffusion models [5], we use a UNet [10] architecture for the triplane encoder. The UNet consists of 8 down and up blocks [10]. Each block consists of 2 ResNet blocks [3] that additionaly take a timestep embedding, and linear attention. If the triplane has larger resolution than the input image, we append additional up blocks to the UNet that upsample the image to the triplane resolution. These have the same architecture as the other UNet blocks, except that they do not use skip connections, as there is no down block in the UNet with the corresponding resolution.

Triplane renderer. To render triplanes, we mostly follow EG3D [1]; however, we use explicit volumetric rendering that samples points along the ray and queries a 2-layer fully-connected neural network to output color and a density [1]. The network takes as input 32-dimensional sum-pooled interpolations of triplane features. Unlike EG3D, we also use a positional embedding of the 3D sample position [9] as input to the network, which allows the network to represent parts of the ground plane that extend beyond the triplanes with a single constant feature.

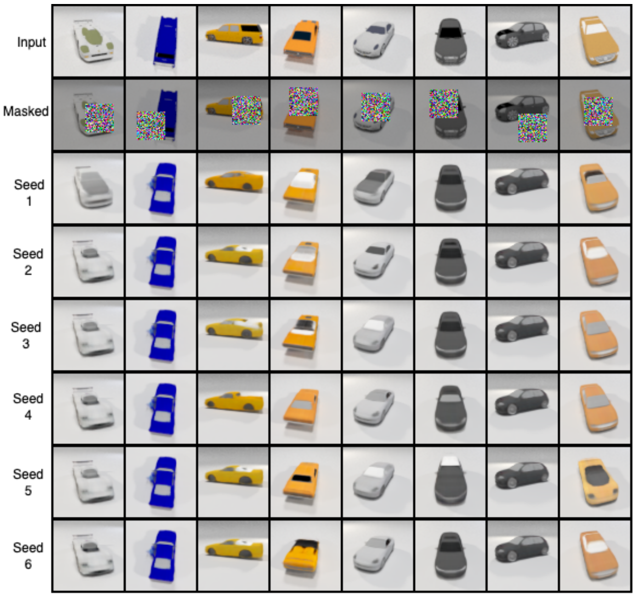

Figure S1. 3D reconstructions over 6 random seeds given masked input image from `car` test set. Notice the diversity in predictions. Results were selected randomly (non-cherry-picked). Since there are many possible inpaintings, Tab. S3 reports quantitative results by taking best-performing reconstruction.

S3. Additional Unconditional Generation Results

Quantitative evaluation We evaluate the distributions of generated scenes quantitatively using four metrics. FID$_r$ is the Fréchet Inception Distance (FID) [4] computed between training views of the generated scenes and training views of all training set scenes. FID$_t$ is the FID computed between test views of the generated scenes and test views of all scenes from the test set. The coverage metric (cov.) [7] measures how well the generated distribution covers the data distribution. It is defined as the fraction of training set images with neighborhoods that contain at least

Table S1. 3D generation performance for our model RenderDiffusion, and baselines GIRAFFE [8], pi-GAN [2] and EG3D [1], on ShapeNet and CLEVR1 datasets. We report FID for train viewpoints for all methods, and also for test viewpoints with ours and EG3D, as well as coverage (*cov.*) and density (*dens.*) [7]

	ShapeNet																CLEVR1			
	car				plane				chair				average							
	FID$_r$	FID$_t$	cov.	dens.	FID$_r$	FID$_t$	cov.	dens.	FID$_r$	FID$_t$	cov.	dens.	FID$_r$	FID$_t$	cov.	dens.	FID$_r$	FID$_t$	cov.	dens.
GIRAFFE [8]	30.5	–	0.11	0.03	56.1	–	0.45	0.41	35.2	–	0.39	0.30	40.6	–	0.32	0.25	–	–	–	–
pi-GAN [2]	25.6	–	0.07	0.02	33.5	–	0.16	0.08	41.4	–	0.14	0.05	33.5	–	0.12	0.15	36.0	–	0.04	0.002
EG3D [1]	**14.4**	**17.9**	**0.70**	**1.32**	**15.0**	**20.9**	0.61	1.02	**10.5**	**14.2**	0.71	1.18	**13.3**	**17.7**	0.67	**1.17**	**15.6**	**19.6**	0.97	**0.90**
Ours	42.1	46.5	0.46	0.26	38.5	43.5	**0.84**	**1.31**	48.0	53.3	**0.85**	**1.46**	42.8	47.8	**0.72**	1.01	15.7	**19.6**	**0.99**	0.65

Table S2. 3D generation performance for our model RenderDiffusion, and baselines GIRAFFE [8], pi-GAN [2] and EG3D [1], on FFHQ (faces) and AFHQ (cats). We report FID for train viewpoints, as well as coverage (*cov.*) and density (*dens.*) [7]. For GIRAFFE and pi-GAN FID, we use the results from [1]; for EG3D we use resolution 64 × 64, i.e. the same as ours; we omit pi-GAN coverage and density due to lack of a publicly-available checkpoint on which to calculate these.

	FFHQ			AFHQ		
	FID$_r$	cov.	dens.	FID$_r$	cov.	dens.
GIRAFFE [8]	31.5	0.66	1.17	16.1	0.07	0.20
pi-GAN [2]	29.9	–	–	16.0	–	–
EG3D [1]	19.8	0.68	1.20	23.7	0.37	0.94
Ours	59.3	0.31	1.01	18.0	0.26	0.37

one generated sample, with a neighborhood defined based on the 3-nearest neighbors. Similarly, the density metric (dens.) [7] measures how close generated samples are to the data distribution, by calculating the average number of real samples whose neighborhoods contain each generated sample. Neighborhoods are defined in the feature space of a VGG-16 network (last hidden layer) that was applied to all training set views of a generated scene. The latter two metrics are similar to the *improved recall and precision* metrics of [6], but avoid certain pathological behaviors [7].

Results on the synthetic datasets are presented in Tab. S1. We see that EG3D performs well on the FID metric, with ours second for CLEVR and pi-GAN second for ShapeNet. Our approach tends to perform better on the coverage and density metrics, while pi-GAN is particularly poor on these. To interpret our quantitative performance relative to EG3D, we refer to the qualitative results shown in Figure 5 of the main paper, where we see that our shapes are slightly more blurry than EG3D, but exhibit more variety and similar shape quality, apart from the blurriness. We hypothesize that the blurriness introduces a bias that the FID metric is highly sensitive to (as the blurriness may affect the feature average that the FID is based on). Coverage and density are less sensitive to the blurriness, as they don't rely on an average over all samples and instead provide a more detailed comparison of the sample distributions by working with sample neighborhoods. This interpretation of the quantitative results suggests that, leaving aside the blurriness, our method generates samples that better cover the data distribution, at a comparable sample quality. This is in line with current understanding of the differences between diffusion models like RenderDiffusion, and GANs like EG3D. Note that our models were not fully converged at the time of measuring these results, and we observed that the blurriness gradually decreases over the course of the training, making it likely that the blurriness can be reduced with additional training. Tab. S2 shows quantitative results on the real datasets FFHQ (photos of human faces) and AFHQ (photos of cat faces); see also the qualitative results in the main paper.

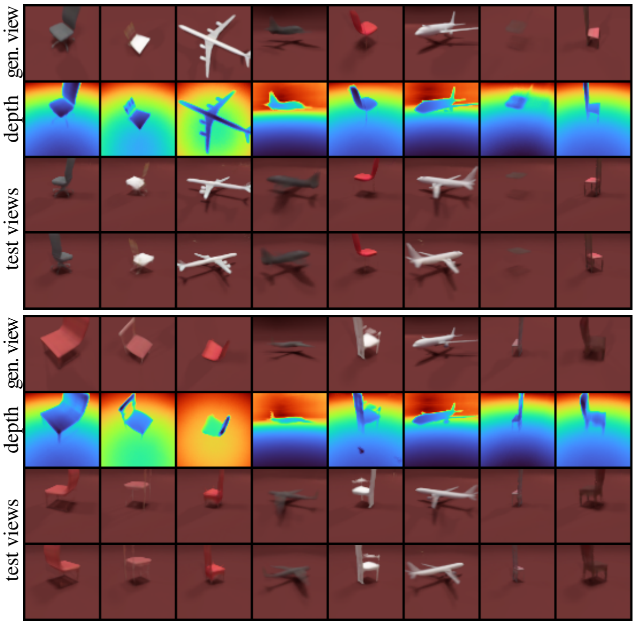

Figure S2. Multi-category generation results. We show generated scenes from a single RenderDiffusion model trained on both `chair` and `airplane` categories.

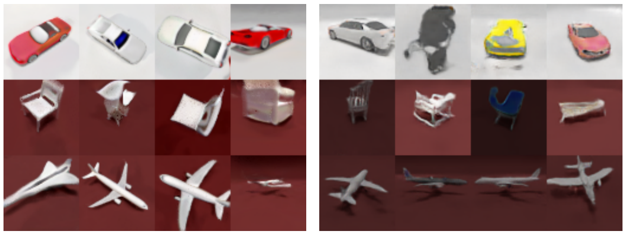

Figure S3. Uncurated samples from pi-GAN (left four columns) and GIRAFFE (right four columns), on the three ShapeNet classes. Compare with results from ours and EG3D in the main paper.

Additional qualitative results The supplementary video shows additional uncurated (not cherry-picked) qualitative results for unconditional generation, shown from a camera that rotates around the object. Fig. S3 shows uncurated generated samples from GIRAFFE and pi-GAN, on the three ShapeNet classes.

S4. Additional Inpainting Results

To quantitatively measure how well our generative model inpaints 3D scenes, we treat inpainting as a 3D reconstruction task with occlusions, where the mask is the occluder. Similar to the unoccluded case, we compare renders of the reconstructed scene from test set viewpoints to ground truth renders using PSNR and SSIM as metrics. Since there are often many plausible inpaintings (i.e. the task is ambiguous), we sample K different inpaintings with our model and select the best matching one. For CLEVR we choose $K = 25$ as the mask often hides majority of the object (increasing the degree of ambiguity), while for ShapeNet we choose $K = 10$. This gives as an indication if the distribution of generated scenes for a given masked input image includes the ground truth scene. To choose the masked-out region of each image, we use a square with width and height equal to 40% of the image resolution (e.g. for an image of size 64×64 the mask will be of size 26×26). The mask is placed uniformly at random within a square region of side length $\frac{5}{16}$ of the image size, itself centered in the image. This ensures the mask always covers part of the foreground object, not just the background. Illustration of masked inputs and diversity in RenderDiffusion predictions is shown in Fig. S1.

Quantitative results on this task are given in Tab. S3. We compare the reconstruction performance with and without masked input. We can see that in most cases, the performance for the two cases is comparable, indicating that a scene resembling the ground truth is contained in the output distribution. We show additional qualitative results with multiple seeds in the supplementary video.

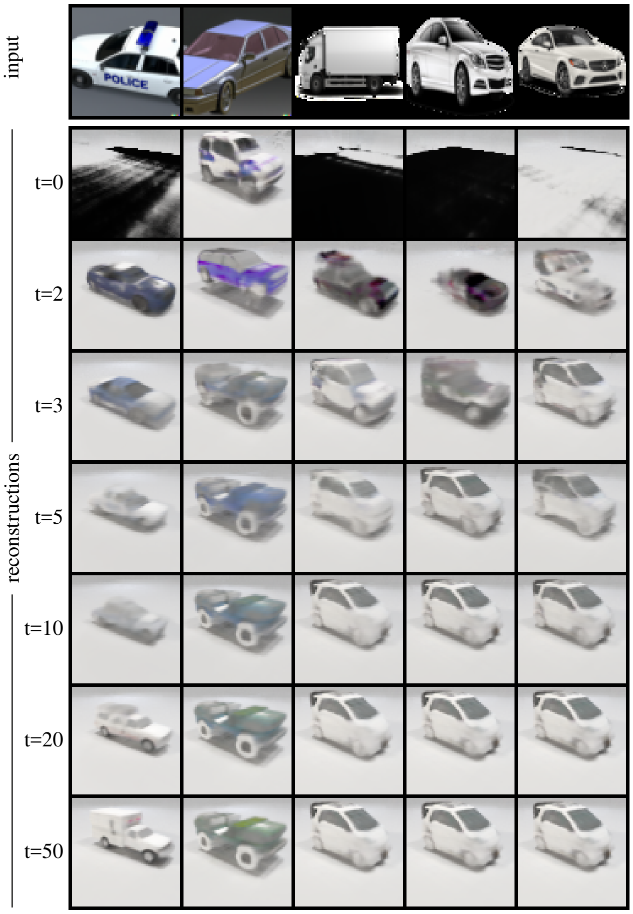

Figure S4. Additional reconstructions from out-of-distribution images. We show reconstructions with different amounts of added noise for out-of-distribution input. We use the same random seed for all reconstructions. Note how the amount of noise trades of between reconstruction quality and fidelity to the input image.

S5. Multi-Category Generation Results

To further demonstrate that our model can represent complex, multi-modal distributions, we perform an additional experiment where a single model is trained jointly on multiple ShapeNet categories. Specifically, we train RenderDiffusion on the union of the `chair` and `plane` categories, otherwise using the same architecture and training protocol as described in the main paper.

In Fig. S2, we show qualitative results from this model. We see that RenderDiffusion has successfully captured both modes of the dataset, sampling plausible chairs and airplanes. As in the single-category experiments in the main paper, the samples are 3D-consistent, exhibit plausible depth-maps, and look realistic from novel test viewpoints.

S6. Additional Reconstruction Results

In Figure S5, we show addditional reconstruction results for CLEVR1 and ShapeNet `chair` datasets. In Fig-

Table S3. 3D reconstruction performance when part of the input image is masked. For easier comparison, we copy the unmasked results from the table in the main paper.

	ShapeNet								CLEVR1	
	car		plane		chair		average			
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
unmasked input	**25.4**	**0.805**	26.3	0.834	**26.6**	**0.830**	26.1	0.823	**39.8**	**0.976**
masked input	24.7	0.790	**27.6**	**0.870**	26.2	0.820	**26.2**	**0.827**	38.9	0.970

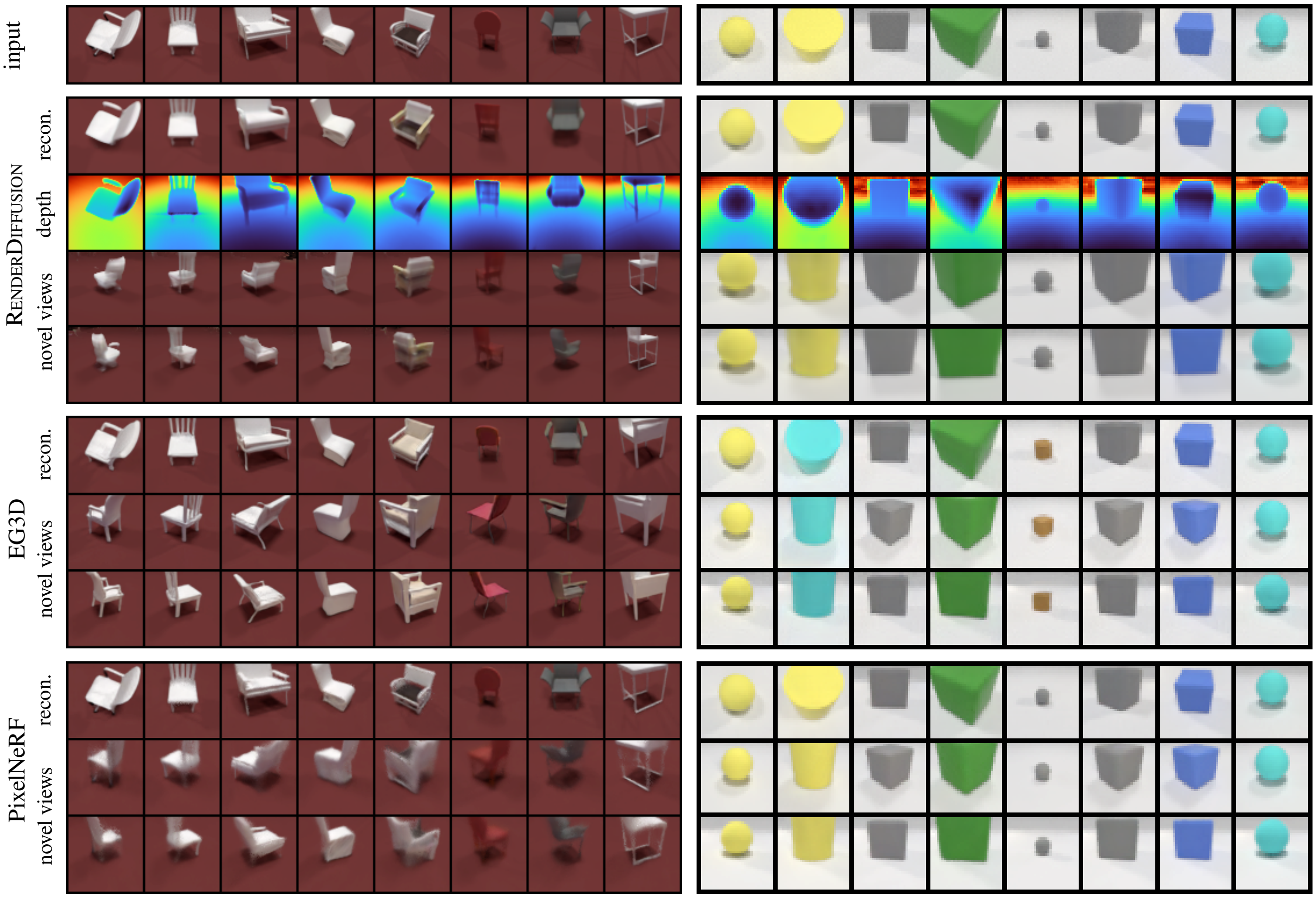

Figure S5. Reconstruction results for ShapeNet `chair` and CLEVR1. Similar to the results on `car` and `plane` datasets in the main paper, our reconstructions better preserve shape identity than EG3D, and are sharper and more detailed than PixelNeRF.

ure S4, we show reconstruction from out-of-distribution images with different amounts of added noise, ranging from no noise at $t = 0$ to 50 noise steps at $t = 50$. Adding larger amounts of noise results in reconstructions that are more generic and increasingly diverge from the input image, as the generative model fills in details covered by the noise, but also show increasingly higher-quality shapes.

References

[1] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 1, 2

[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*, 2021. 2

[3] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017. 1

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

[6] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[7] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020. 1, 2

[8] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 2

[9] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020. 1

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 1