Seeing a Rose in Five Thousand Ways
–Supplementary Material–

Yunzhi Zhang
Stanford University

Shangzhe Wu
University of Oxford

Noah Snavely
Cornell Tech

Jiajun Wu
Stanford University

1. Implementation Details

Network Architectures. The generator consists of a SDF network and an albedo network, both modulated by FC layers with SIREN [8] activations, and an optimizable scalar parameter s (Equation 6 in the main paper). The architectures of the two networks are identical to those in StyleSDF [7], except that we disable view-dependence for the albedo network. The discriminator architecture is identical to that of GIRAFFE [6], except that the image discriminator additionally outputs 6 channels for pose prediction, which is used to compute the regularization loss term (Equation 7 in the main paper).

Scale and Shift Augmentations. Recall that we apply scale and shift augmentations to prevent the discriminator from overfitting and to improve robustness towards estimation error in the prior pose distribution, as discussed in Section 3.3. The shift augmentation is implemented as integer offsets along height and width, uniformly sampled between $[-0.125, 0.125]$ relative to the resolution of image crops. The scale factor s is sampled from $\log_2 s \sim \mathcal{N}(0, 0.2)$. An ablation study shows the benefit of data augmentation, as shown in Table 1.

Discriminator inputs. As introduced in Section 3.3, we use two discriminators, D_η and $D_{\eta_{\text{mask}}}$, which take RGB values and masks of instances as real inputs, respectively. Examples of these real inputs are shown in Figure 1.

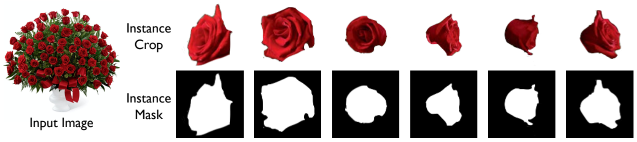

Figure 1. Instances are cropped from the input image as inputs to the discriminators. As shown are examples of RGB values of instance crops (first row) and corresponding instance masks (second row).

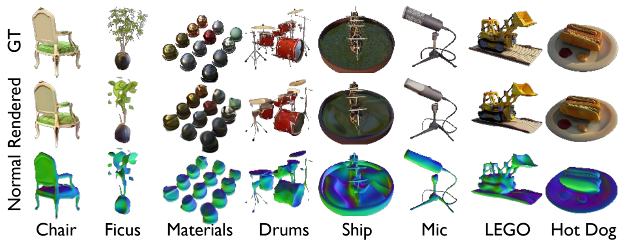

Figure 2. Results on all scenes from the dataset from NeRF [5]. Our method robustly reconstructs the scene geometry training on multi-view images with *unknown* camera poses.

2. Learning from Multiview Images

Dataset. While our method is designed to learn a generative model from observations of similar, non-identical instances, to evaluate its robustness, we additionally evaluate our method to learn from multiple views of one instance with known foreground masks and unknown camera poses.

We use the synthetic dataset from NeRF [5], which contains 8 multi-view scenes and 100 training images for each scene. During training, poses are randomly sampled as oriented towards the upper hemisphere. The light direction is initialized as collocated with the camera. We train our method for 300K iterations for all scenes. While there is no instance variance among multiview observations, we use identical configurations as experiments in the main paper for consistency and keep the 64-dimensional latent space introduced in Section 3.1.

Results. Figure 2 compares rendered novel views with the ground truth. In this visualization, for each ground truth image from the held-out test split, we render 1000 images with randomly sampled poses and latent vectors, and visualize the one with the lowest LPIPS error. Results show that our method robustly recovers the 3D geometry in all scenes.

3. Ablation Studies

Architecture Designs. We conduct ablation studies on several design choices introduced in Section 3.3 in the paper, namely the effects of scale and shift augmentations, the

	Normal	Depth	View Synthesis		
	Angle(°)↓	MSE↓	PSNR↑	SSIM↑	LPIPS↓
No-aug	14.50	13.12	27.04	0.89	0.06
No-alb	22.15	**0.03**	27.24	0.90	0.05
No-D_{mask}	16.26	0.11	28.94	**0.93**	**0.04**
Fix-bg	19.57	0.07	28.89	0.92	**0.04**
Full	**14.35**	0.05	**29.16**	**0.93**	**0.04**

Table 1. Ablation Studies on synthetic data. We compare with the full model ("Full") with the following model variants: without data augmentations ("No-aug"), without intrinsics decomposition ("No-alb"), without mask discriminator ("No-D_{mask}"), and one without background randomization ("Fix-bg"). The full model achieves better reconstruction quality both in geometry and appearance synthesis.

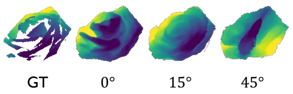

Figure 3. Rendered depth.

	0°	±15°	±45°
MSE	**2.34**	3.07	3.25

Table 2. Depth error ($\times 10^{-4}$).

explicit shading model, the mask discriminator, and background randomization.

We compare performance of a model variant with no augmentation (denoted as "No-aug" in Table 1), one that predicts the final color instead of albedo ("No-alb"), one with no mask discriminator ("No-D_{mask}"), one with fixed background ("Fix-bg"), and the full model ("Full"). All variants are evaluated on two of the four scenes, rendered with the 3D asset from [1]. As shown in Table 1, each of these module components improves shape reconstruction quality measured by normal angle errors and depth errors, and improves view synthesis results.

Robustness to Inaccurate Pose Distribution Estimation. We show that the model is robust against a certain amount of estimation error in the prior pose distribution. Specifically, we perturb the estimated pose distribution by perturbing the camera elevation with ±15°, ±45°, and rerun the model on the scene of roses in Figure 6 of the main paper. Figure 3 and Table 2 show that the model still learns reasonable shapes with a ±15° perturbation.

4. Qualitative Results on Real-Captured Data

More qualitative comparisons with the baseline method on real-captured data are shown in Figure 4. Our method produces results with significantly better view consistency compared to the baseline method.

5. Qualitative Results on Synthetic Data

In Figure 5, together with Figure 7. from the main paper, we show comparisons with baseline methods on all four scenes from the synthetic dataset. In Table 3, we compare with NeRD [2] and Neural-PIL [3] using the same image crops as training inputs as used in our method. These two methods originally assume that camera intrinsics are consistent across training views, but image crops of instances have different principal point offsets, violating this assumption. Therefore in Table 3 from the main paper, we re-render each instance re-centered to the origin as training data for the baselines. In either case, our method produces better intrinsics decomposition and reconstruction results compared to the baseline methods, which have additional access to ground truth camera poses.

6. Qualitative Results on In-the-Wild Data

The proposed method effectively learns object intrinsics across a range of objects from in-the-wild data, and can be applied to applications such as view synthesis, relighting, and novel instance generations, as shown in Figures 6 and 7.

References

[1] 3D model bun. https://www.blendswap.com/blend/18213. 2

[2] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *ICCV*, 2021. 2, 3

[3] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan T. Barron, and Hendrik P.A. Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. In *NeurIPS*, 2021. 2, 3

[4] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *ICCV*, 2021. 3

[5] 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*, 2020. 1

[6] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 1

[7] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. In *CVPR*, 2022. 1

[8] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020. 1

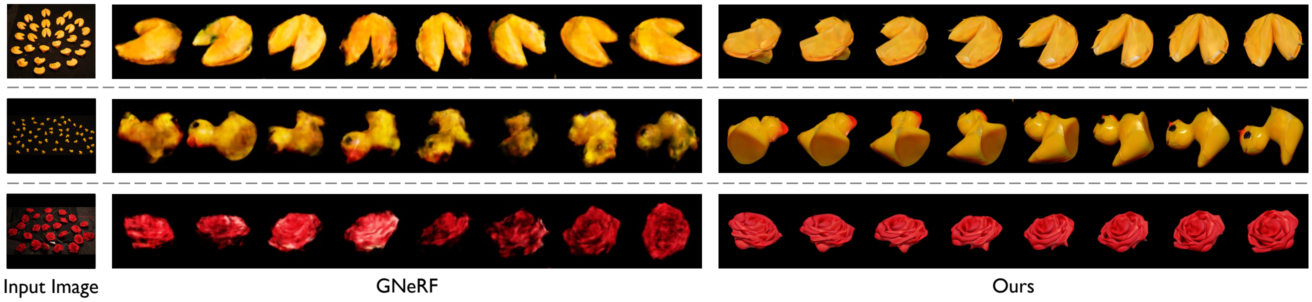

Figure 4. View synthesis results on three real-captured scenes. Each row contains the input image along with images under eight different viewpoints synthesized by GNeRF [4], and by our method. Compared with GNeRF, our method synthesizes images with significantly better view consistency and higher fidelity.

	Normal	Depth	Albedo			View Synthesis			Relighting		
	Angle(°)↓	MSE↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRD* [2]	65.94	95.33	14.19	0.62	0.35	18.61	0.64	0.32	18.28	0.63	0.32
Neural-PIL* [3]	75.84	77.68	13.66	0.58	0.36	19.91	0.65	0.32	19.76	0.64	0.32
Ours	**22.69**	**1.10**	**22.48**	**0.87**	**0.10**	**29.13**	**0.93**	**0.04**	**26.03**	**0.91**	**0.06**

Table 3. Results on synthetic data. Both baselines use ground truth poses (denoted as *) and use the same image crops as in our method as training inputs. Our method produces better reconstruction results compared to the baselines.

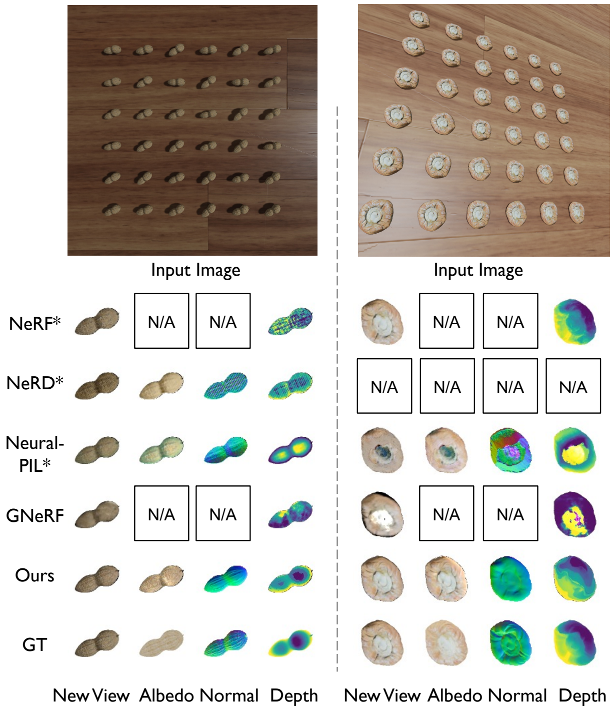

Figure 5. Results of intrinsic decomposition on two of the four scenes from the synthetic dataset. The remaining two scenes are shown in Figure 7 of the main paper.

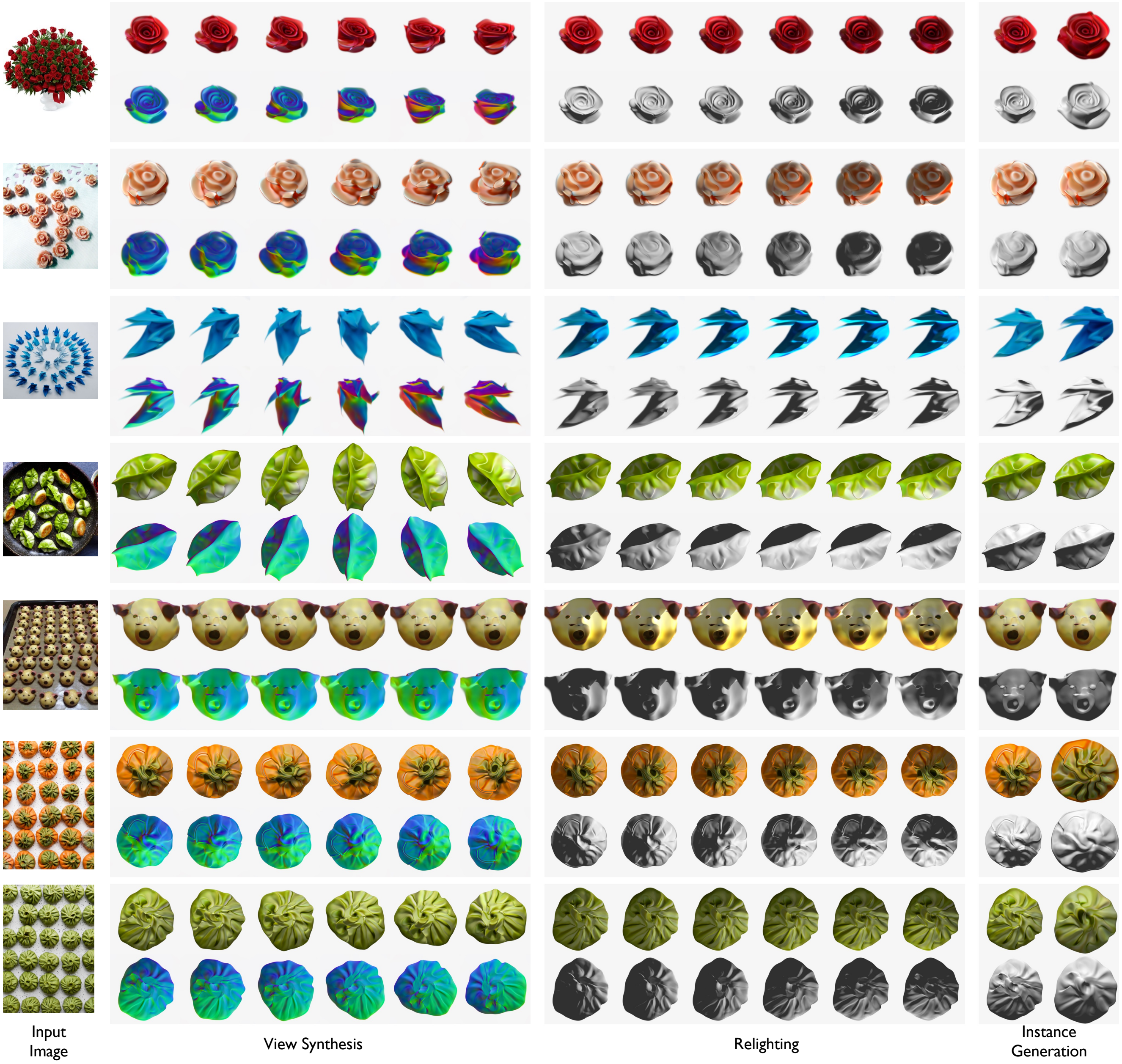

Figure 6. Results for view synthesis, relighting and instance generation on in-the-wild images. Each row corresponds to predicted appearance, normal or shading maps.

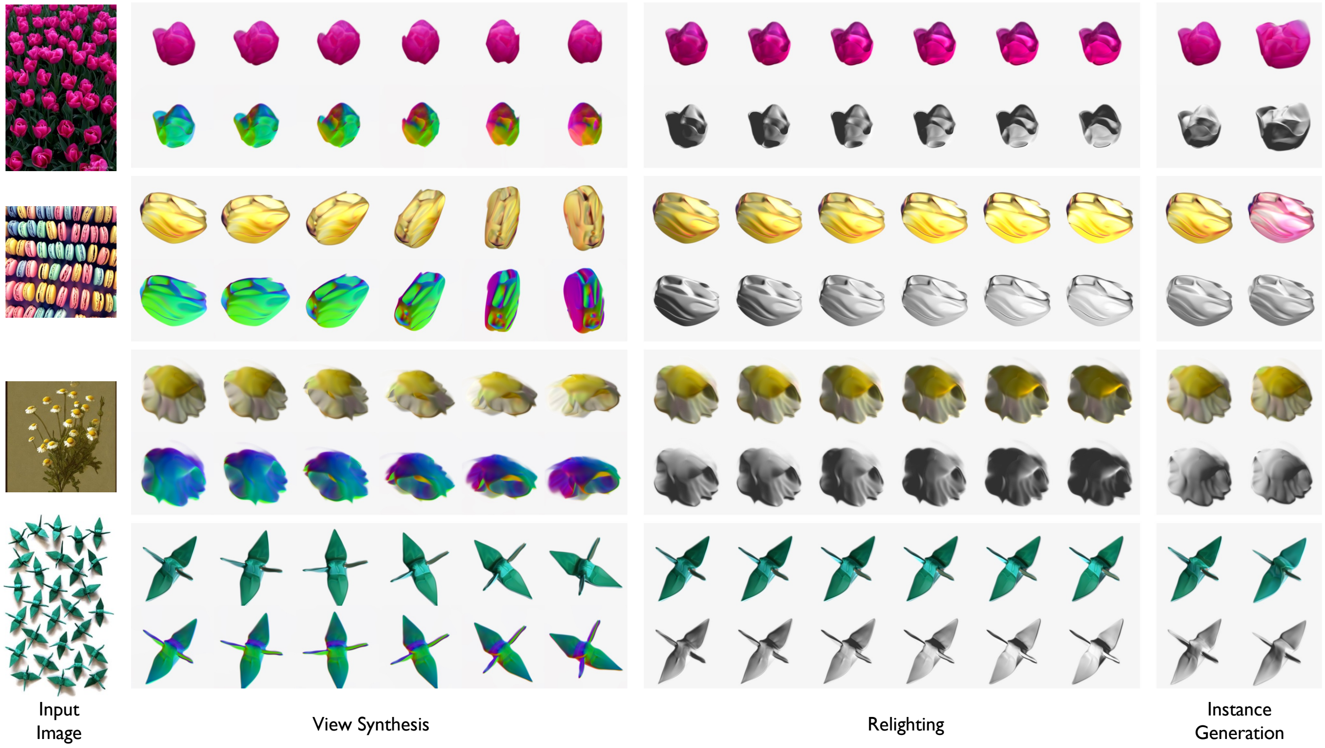

Figure 7. Results for view synthesis, relighting and instance generation on in-the-wild images. Each row corresponds to predicted appearance, normal or shading maps.