Supplementary Material

Youjia Zhang[1] Zikai Song[1] Junqing Yu[1] Yawei Luo[2] Wei Yang[†1]

[1] Huazhong University of Science and Technology
[2] Zhejiang University
https://youjiazhang.github.io/USD/

1 Unbiased Sampling of Multi-view Diffuser

We rectify the unconditional noise in Formula 4 of the main paper:

$$\epsilon_{\phi}^{\mathbf{CFG}}(\mathbf{z_t}, c_I, c_P) = \omega[\epsilon_{\phi}(\mathbf{z_t}, c_I, c_P) - \epsilon_{\phi}(\mathbf{z_t}, \varnothing, \varnothing)] + \epsilon_{\phi}(\mathbf{z_t}, \varnothing, \varnothing)$$
$$\epsilon_{\phi,\psi}^{\mathbf{CFG}}(\mathbf{z_t}, c_I, c_P) = \omega[\epsilon_{\phi}(\mathbf{z_t}, c_I, c_P) - \epsilon_{\psi}(\mathbf{z_t}, \varnothing)] + \epsilon_{\psi}(\mathbf{z_t}, \varnothing)$$

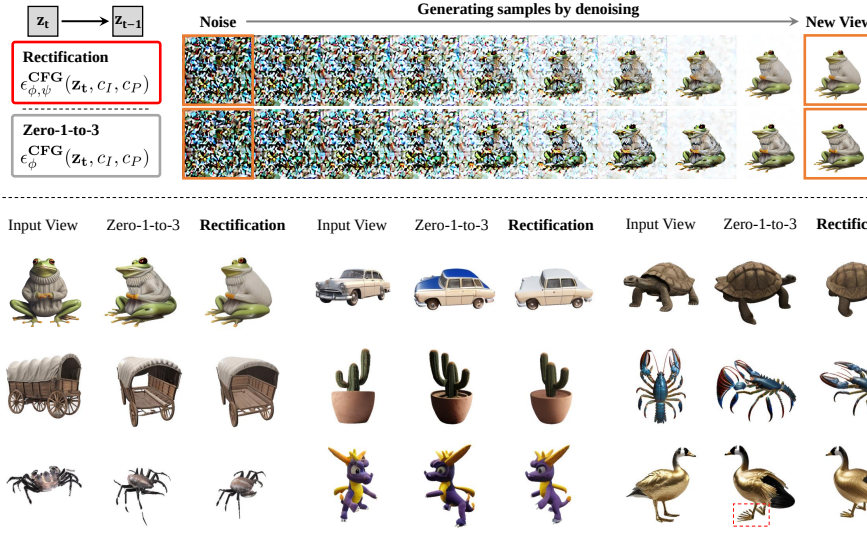

Fig. 1: Novel view synthesis on in-the-wild images. Comparison between Zero-1-to-3 [5] and our rectification. Starting from the input view, the task is to generate an image of the object under a specific camera pose transformation.

We found that rectifying the bias in Zero-1-to-3 [5] achieves significantly better novel views from in-the-wild inputs. Fig. 1 shows examples from One-2-3-45++ [4], Wonder3D [7] and SyncDreamer [6]. Our rectification is able to generate novel views that are more consistent with the input view. Additionally, rectification is able to generate novel views from input view while keeping the original style as well as object geometric details. These examples show the effectiveness of our rectification.

2 Two-stage Specified Diffusion Details

DreamBooth [11] provides a network fine-tuning strategy to adapt a given text-to-image denoising network to generate images of a specific subject.
Low Rank Adaptation (LoRA) [2] provides a memory-efficient and faster technique for DreamBooth. Priors work show that this low-rank residual fine-tuning is an effective technique that preserves several favorable properties of the original DreamBooth while also being memory-efficient as well as fast.

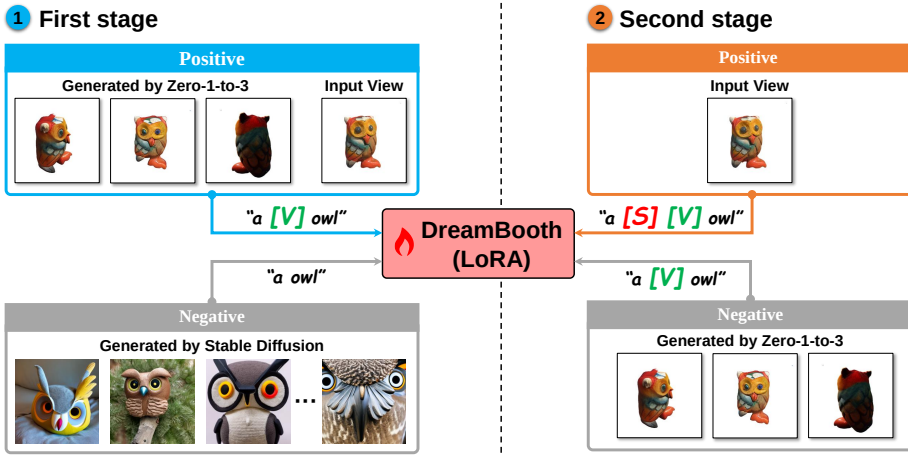

Fig. 2: Fine-tuning. As a demonstration, we use the '*owl*' image as a toy example.

We implement our DreamBooth on the Stable Diffusion [10] V2.1 diffusion model and we predict the LoRA weights for all cross and self-attention layers of the diffusion U-Net [12]. Fig. 2 illustrates the model fine-tuning with the class-generated samples. Fig. 3 shows a collection of generation results to illustrate how our method can generate novel images for a specific subject in different contexts with descriptive prompts.

Fig. 3: Results for re-contextualization of a '*owl*' subject instances.

NeRF [8] uses a volume rendering method to learn a volumetric radiance field for novel view synthesis. However, NeRF architectures are prone to cloudy artifacts (*floaters*), which it is difficult to extract a high-quality surface, as shown

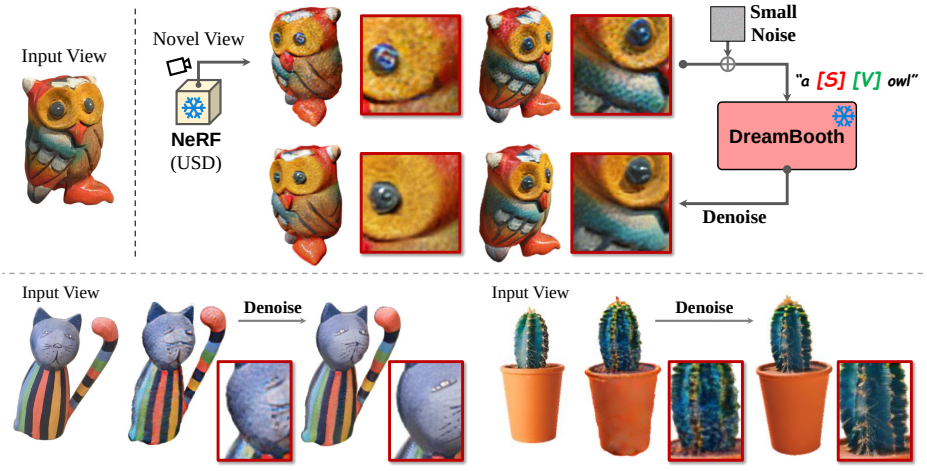

Fig. 4: With the specialization diffusion model, we add a small noise, Stable Diffusion [10] scheduler $t = 200$, to the NeRF [8] render images and conduct the denoising process.

in Fig. 4. To address this issue, we denoise multi-view renderings from the trained NeRF using the fine-tuned DreamBooth model.

3 3D Gaussian Splatting Representation

Recently, DreamGaussian [13] and GaussianDreamer [14] utilizes 3D Gaussians as an efficient 3D representation that supports real-time high-resolution rendering via rasterization. We adapt 3D Gaussian Splatting [3] into the generative setting with Unbiased Score Distillation (USD). Our method is implemented in PyTorch [9], based on threestudio [1].

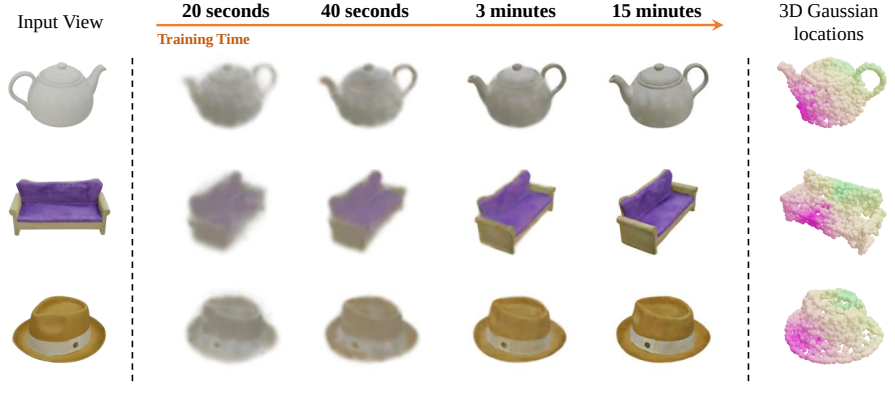

Fig. 5: Optimization Progress. The shape to initializa the 3D Gaussians as sphere.

References

1. Guo, Y.C., Liu, Y.T., Shao, R., Laforte, C., Voleti, V., Luo, G., Chen, C.H., Zou, Z.X., Wang, C., Cao, Y.P., Zhang, S.H.: threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio (2023) 3
2. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021) 2
3. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4), 1–14 (2023) 3
4. Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Wei, X., Chen, H., Zeng, C., Gu, J., Su, H.: One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. arXiv preprint arXiv:2311.07885 (2023) 1
5. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: ICCV (2023) 1
6. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023) 1
7. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. arXiv preprint arXiv:2310.15008 (2023) 1
8. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020) 2, 3
9. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019) 3
10. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 2, 3
11. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023) 2
12. Ryu, S.: Low-rank adaptation for fast text-to-image diffusion fine-tuning (2022), https://github.com/cloneofsimo/lora 2
13. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023) 3
14. Yi, T., Fang, J., Wang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., Wang, X.: Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In: CVPR (2024) 3