Ellis Summer School

18-22 September 2023

Disorienting NeRF: Assessing Robustness in Noisy Environments

Authors:
Wei Haoyu
Balloni Emanuele
Bianchi Lorenzo
Boldo Michele
Caddeo Gabriele Mario
Simoni Alessandro

Introduction and Method

Introduction

Neural Radiance Fields (NeRF) [1] represent the state of the art on rendering scenes from novel points of view. These methods optimize a volumetric scene on a sparse set of images-poses pairs leading to impressive results, even in fine-grained features such as specularity or transparency. These approaches require as input a set of images and the related 5D ground truth poses of the camera, with three coordinates ([x, y, z]) to define the position and two coordinates ([θ, ϕ]) to define the direction of the incident ray coming out the camera. Figure 1 describes the NeRF pipeline.

However, all that glitters is not gold. At the moment, almost no real-world on-the-fly application exploits NeRF-based methods, due to high training time and dataset feature requirements. Even if recently some works, such as *Instant Neural Graphics with a Multiresolution Hash Encoding* (Instant-NGP) [2], are trying to reduce the training time, NeRF-based methods still overfit on a single scene and cannot generalize to different scenes. Moreover, precise ground truth poses are required along with a non-negligible number of non-noisy images. While obtaining non-noisy images can represent a surmountable limit, having a precise ground truth pose definitely is an issue in every real-world application.

In this project, we focus on the highly demanding dataset requirements and we highlight how the performances are affected by injecting noise in the input images and in their related camera poses.

Method

We decided to simulate the real-world scenario by corrupting the input of the model, trying to simulate the effects and the problems that would be faced. In particular:

- at **image** level, we augment the images by *masking* the image, simulating the possible occlusions, with a *coarse pepper* type of noise;
- at **pose** level, we corrupt the ground truth poses by adding random rotations in a range between [-15; 15] degrees and random translation in a range between [-0.15; 0.15] meters.

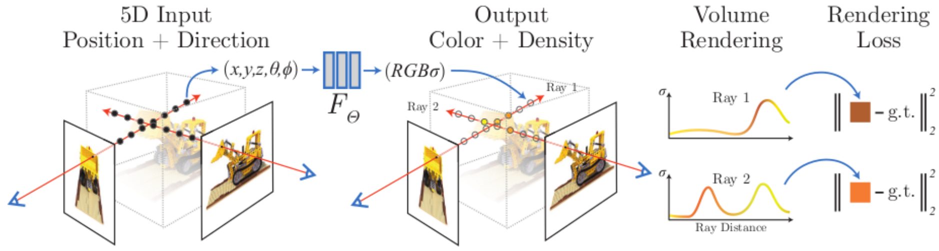
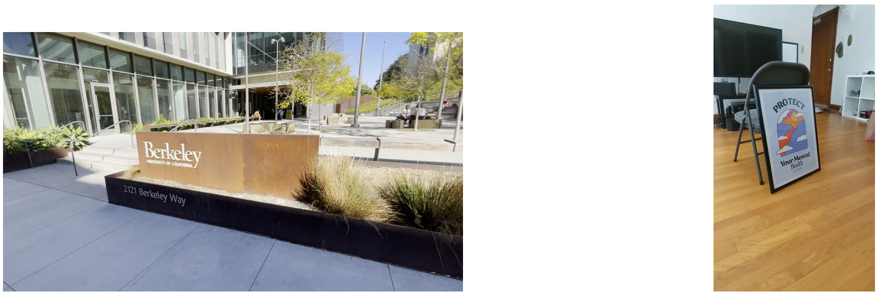


Figure 1: Description of NeRF[1] pipeline. 5D coordinates are sampled along the camera rays and fed to an MLP which outputs a color and a volume density to create images through a volume rendering technique. The loss minimizes the difference between the synthesized image and the ground truth. .

Experiments and Conclusions

Experiments

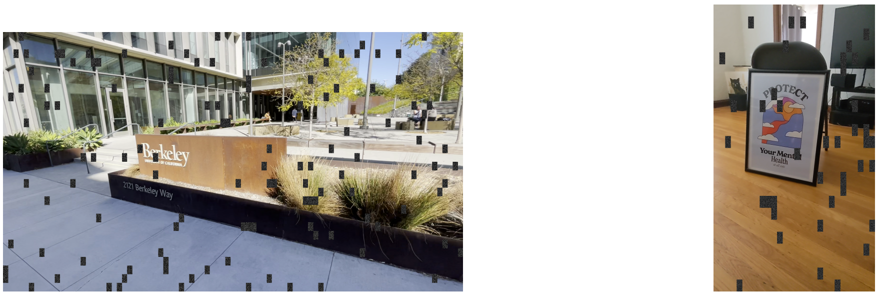
The experiments are carried out on two scenes, representing the two main categories of real-world scenarios, namely one outdoor scene and one indoor scene. The first scene is composed of 386 images of a garden and is recorded outside the Berkley University; the second scene consists of a room with a chair and poster in the middle. Fig. 3 shows some examples of images of the 2 scenes considered. We run the experiments with two backbones, Nerfacto and Instant-NGP provided by *nerfstudio* framework [3]. We evaluate the experiments with three metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [4] and LPIPS [5].



(a) Example of outdoor scene

(b) Example of indoor scene

Figure 2: Examples of images in the dataset



(a) Example of outdoor scene

(b) Example of indoor scene

Figure 3: Examples of masked images in the dataset

Table 1 shows the results of our experiments. The GT method represents our starting point, showing results on the original data sets without any corruption. We apply the *coarse pepper* noise on the 25%, 50%, and 75% of the training set. As expected on the *GT* dataset both NF and I-NGP achieve better results and the error increases proportionally to the increasing percentage of occluded images.

Table 2 highlights the effects of the random noise injected into the camera poses. We evaluate the I-NGP method with or without the camera pose refinement. As we

Table 1: Comparison between Nerfacto (NF) and Instant-NGP (I-NGP) when adding noise to the images.

	Metric	PSNR ↑		SSIM ↑		LPIPS ↓	
Dataset	**Method**	NF	I-NGP	NF	I-NGP	NF	I-NGP
outdoor	Vanilla	22.77 ±1.88	23.70 ±1.70	0.72 ±0.07	0.76 ±0.05	0.35 ±0.03	0.31 ±0.03
	Noisy images 25%	21.68 ±2.54	22.11 ±2.50	0.69 ±0.07	0.74 ±0.06	0.41 ±0.05	0.37 ±0.05
	Noisy images 50%	20.09 ±2.69	20.26 ±2.59	0.67 ±0.07	0.70 ±0.06	0.48 ±0.06	0.43 ±0.07
	Noisy images 75%	18.85 ±2.02	19.04 ±2.09	0.64 ±0.06	0.68 ±0.05	0.53 ±0.05	0.48 ±0.06
indoor	Vanilla	21.35 ±4.39	22.59 ±5.18	0.87 ±0.05	0.90 ±0.05	0.28 ±0.08	0.27 ±0.08
	Noisy images 25%	20.44 ±4.43	20.91 ±4.96	0.84 ±0.07	0.87 ±0.07	0.36 ±0.12	0.36 ±0.11
	Noisy images 50%	19.58 ±3.95	20.09 ±4.41	0.82 ±0.06	0.85 ±0.06	0.41 ±0.09	0.41 ±0.09
	Noisy images 75%	18.66 ±3.53	19.01 ±3.77	0.79 ±0.06	0.82 ±0.07	0.49 ±0.09	0.48 ±0.09

Table 2: Comparison between Instant-NGP with or without pose refinement optimization when corrupting the poses.

	Metric	PSNR ↑		SSIM ↑		LPIPS ↓	
Dataset	**I-NGP w/ opt.**	✓	✗	✓	✗	✓	✗
outdoor	Noisy poses	21.36 ±1.71	17.82 ±1.13	0.68 ±0.07	0.57 ±0.06	0.37 ±0.03	0.65 ±0.03
indoor	Noisy poses	19.13 ±3.56	17.72 ±2.04	0.81 ±0.07	0.78 ±0.06	0.42 ±0.09	0.56 ±0.07

would anticipate, this optimization increases the robustness of the reconstructions, with consistent improvements in every metric.

Figure 4 and Figure 5 depict some qualitative results for the indoor and outdoor datasets, highlighting the effects of adding noise to the input images or to the camera poses respectively.

Conclusions

In this project, we investigated the behavior of state-of-the-art approaches in scene rendering when corrupting the dataset. The results obtained are coherent with the expectations, with a high degradation of the performances following a small perturbation of the inputs. The outcome suggests the NeRF-based methods considered are still not robust when applied to a real scenario. As future work, more baselines need to be considered, along with different sources of noise and a larger dataset.

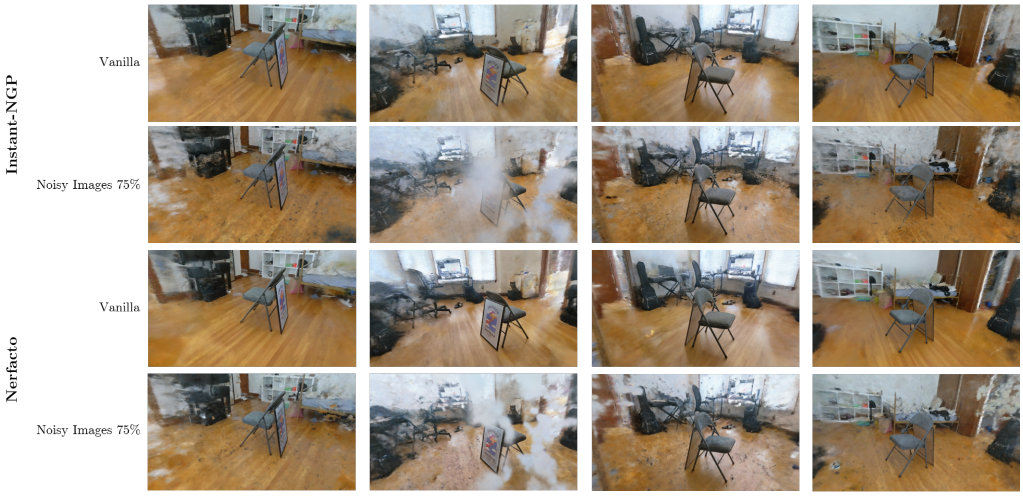


Figure 4: Some qualitative results for the indoor scene highlight the effects of compromising the image quality with coarse pepper noise.

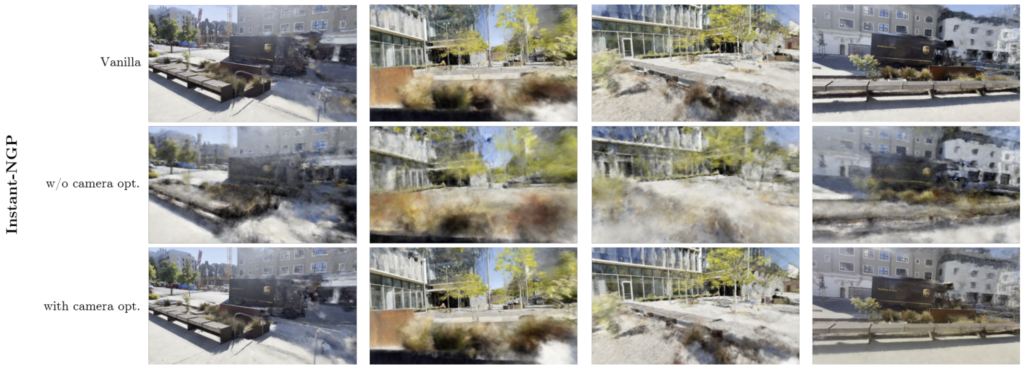


Figure 5: Some qualitative results for the outdoor scene highlight the effects of compromising the camera poses with Gaussian noise.

References

[1] Ben Mildenhall et al. “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis”. In: *ECCV*. 2020.

[2] Thomas Müller et al. “Instant Neural Graphics Primitives with a Multiresolution Hash Encoding”. In: *ACM Trans. Graph.* 41.4 (July 2022), 102:1–102:15. DOI: 10.1145/3528223.3530127. URL: https://doi.org/10.1145/3528223.3530127.

[3] Matthew Tancik et al. “Nerfstudio: A Modular Framework for Neural Radiance Field Development”. In: *ACM SIGGRAPH 2023 Conference Proceedings*. SIGGRAPH ’23. 2023.

[4] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. “Multiscale structural similarity for image quality assessment”. In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Ieee. 2003, pp. 1398–1402.

[5] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.