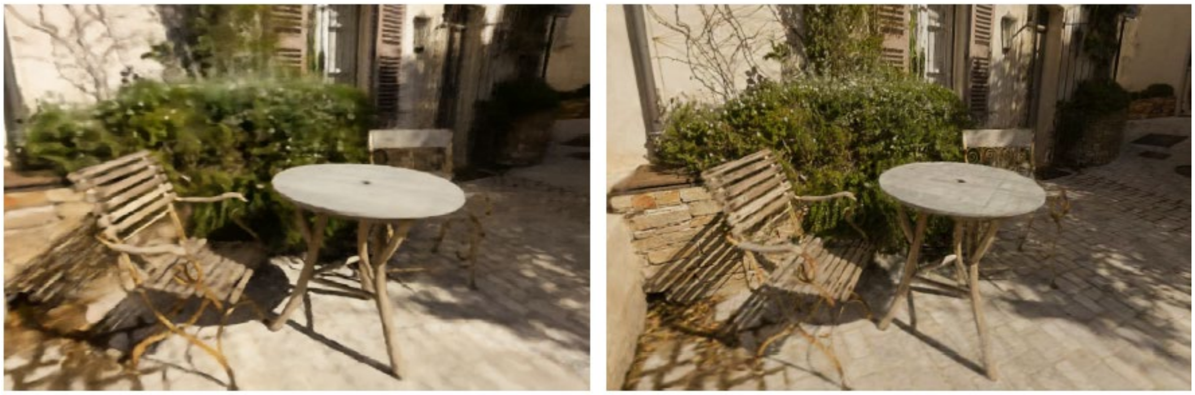
A Methodology and Dataset for Evaluation of Novel View Synthesis and Neural Rendering

(Masters 1 / Engineering internship)

George Drettakis, GRAPHDECO, Inria Sophia Antipolis (France)
http://team.inria.fr/graphdeco

George.Drettakis@inria.fr
http://www-sop.inria.fr/members/George.Drettakis/



(a) NeRF [3] *(b) Point-based Neural Rendering[1]*

Figure 1: (a) Neural Radiance Fields – NeRF [2] achieve impressive results in neural rendering, as does (b) Point-based Neural Rendering [14]. Evaluating the quality of each solution is a very challenging problem.

Context and goal

Neural Rendering has seen an explosion in interest in recent years [1]. Nowadays it is possible to render novel views and video-like camera paths as if you have a perfect 3-d model of the scene, simply using a set of photographs as input. This topic has been approached from many different directions; the recent introduction of Neural Radiance Fields (NeRF [2]) which proposed implicit scene representations has resulted in a vast number of followup with projects, many different variations and applications [3][4][5][6][7][8][9][10]. In addition to NeRF-based solutions, more traditional scene representations such as meshes [11][12] or point clouds [13][14] have shown promising results with recent solutions. Both directions of research have their own pros and cons, and an important scientific question is evaluating the benefits and drawbacks of each approach.

High-quality evaluation of each proposed method is a central part of scientific research. This is especially true for deep-learning, where datasets and evaluation play a central role. The most well-known example is the Image-Net [15] dataset which contributed to the most significant advances of machine learning and computer vision in recent years. It provided a robust and good quality dataset that allowed for concrete and solid conclusions and allowed the research community to quickly assess and compare between methods. In the field of Image-Based rendering or novel view synthesis, understanding different visual artifacts can easily be detected and understood by human observers, but it is still very hard to quantitatively assess the precise (and overall) perceptual performance of different algorithms. The most popular evaluation methods are Leave-One-Out Rendering[14] and train/test-set approaches [11][12][13], typically used in conjunction with commonly used metrics such as PSNR, SSIM[15], L-PIPS[16]. These metrics operate on single images, and thus do not capture temporal artifacts in videos or rendered "paths" with a moving camera.

These approaches and metrics have drawbacks that do not always provide satisfactory scientific results, often producing different rankings of algorithms according to the dataset, evaluation method or metric. Test-set approaches need big datasets that are not always available, and typically involve images "in between" train images, while Leave-One-Out needs either an infeasible amount of compute to produce or is ambiguous since often overfitted neural rendering techniques can "cheat". Meanwhile all the evaluation metrics are in image-space making it impossible to capture temporal artifacts. These problems are also augmented by the fact that different capturing styles and placements of cameras can influence the quality of the results. There have been some attempts in the past ("Tanks and Temples" [17]) but they do not address the shortcoming listed above.

The goal of this project is to create a reference dataset with a solid evaluation method and metric(s) that will allow meaningful and fair evaluation of different novel-view synthesis/neural rendering algorithms. The intern will work in collaboration with G. Kopanas, Ph.D. student in the group and first author of [14].

Approach

We will start by clearly defining the capture process we target. This will typically involve a combination of "inside-out" and "outside-in" captures of the same scene (i.e., the "capture cameras"), together with the capture of "camera paths" (either with video or burst-mode capture) that are completely independent from the "capture cameras" used for reconstruction and/or training. Care will also be taken to minimize exposure and transient object artefacts. This will involve several iterations over test datasets to ensure that the process is the best possible.

In a second step, we will capture a set of diverse scenes (5-10), of indoors and outdoors environments, using the equipment available in the lab, or purchased specifically for this project.

In parallel with the actual capture, a suite of tools will be provided, based on existing code bases that allow simple comparison to these methods, and providing a simple API that will allow new methods to compare in a "plug-and-play" manner. We will also develop a new error metric that also evaluates temporal artifacts for video sequences, to allow this important aspect to also be evaluated.

The dataset and the evaluation tools will be published and provided to the community, in the style of the "Tanks and Temples" dataset [17], but addressing many of it's shortcomings.

Work environment and requirement

The internship will take place at Inria Sophia Antipolis in the GRAPHDECO group (http://team.inria.fr/graphdeco). Inria will provide a monthly stipend of around 1100 euros for EU citizens in their final year of masters, and ~600 euros for other candidates.

Candidates should have strong programming and mathematical skills as well as knowledge in computer graphics, geometry processing and machine learning, with experience in C++, OpenGL and GLSL on the graphics side, and pytorch for deep learning.

References

[1] Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Wang, Y., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S. and Simon, T., 2021. Advances in neural rendering. arXiv preprint arXiv:2111.05849.

[2] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R. and Ng, R., 2020, August. Nerf: Representing scenes as neural radiance fields for view synthesis. In European conference on computer vision (pp. 405-421). Springer, Cham.

[3] BOSS M., BRAUN R., JAMPANI V., BARRON J. T., LIU C., LENSCH H. P. A.: NeRD: Neural reflectance decomposition from image collections. ICCV (2021).

[4] BARRON J. T., MILDENHALL B., TANCIK M., HEDMAN P., MARTIN-BRUALLA R., SRINIVASAN P. P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. ICCV (2021). 12

[5] REISER C., PENG S., LIAO Y., GEIGER A.: KiloNeRF: Speeding up Neural Radiance Fields with Thousands of Tiny MLPs. URL: http://arxiv.org/abs/2103.13744, arXiv:2103.13744.

[6] YU A., LI R., TANCIK M., LI H., NG R., KANAZAWA A.: PlenOctrees for real-time rendering of neural radiance fields. In arXiv (2021).

[7] ZHANG X., SRINIVASAN P. P., DENG B., DEBEVEC P., FREEMAN W. T., BARRON J. T.: NeRFactor: Neural factorization of shape and reflectance under an unknown illumination. SIGGRAPH Asia (2021).

[8] YU A., YE V., TANCIK M., KANAZAWA A.: pixelnerf: Neural radiance fields from one or few images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021).

[9] Park, Keunhong, et al. "Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields." arXiv preprint arXiv:2106.13228 (2021).

[10] Lombardi, Stephen, et al. "Mixture of volumetric primitives for efficient neural rendering." arXiv preprint arXiv:2103.01954 (2021).

[11] RIEGLER G., KOLTUN V.: Free view synthesis. In European Conference on Computer Vision (2020), Springer, pp. 623–640

[12] RIEGLER G., KOLTUN V.: Stable view synthesis. In CVPR (2021).

[13] Rückert, Darius, Linus Franke, and Marc Stamminger. "Adop: Approximate differentiable one-pixel point rendering." arXiv preprint arXiv:2110.06635 (2021).

[14] Georgios Kopanas, Julien Philip, Thomas Leimkühler, George Drettakis, Point-Based Neural Rendering with Per-View Optimization Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering), Volume 40, Number 4 - June 2021 https://repo-sam.inria.fr/fungraph/differentiable-multi-view/

[15] Wang, Zhou, et al. "Image quality assessment: from error visibility to structural similarity." IEEE transactions on image processing 13.4 (2004): 600-612.

[16] Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[17] Knapitsch, Arno, et al. "Tanks and temples: Benchmarking large-scale scene reconstruction." ACM Transactions on Graphics (ToG) 36.4 (2017): 1-13.