Benchmarking Modern Novel View Synthesis Methods: Comparing Efficiency and Quality Across Indoor and Outdoor Scenes on Consumer-Grade Hardware

Safuan Iusupov
Mid Sweden University
Sundsvall, Sweden
saiu2500@student.miun.se

***Abstract*—Recent progress in novel view synthesis has produced a wide variety of implicit, explicit, and hybrid 3D representations, yet it remains unclear how these methods perform under realistic hardware constraints or across different types of real-world scenes. Existing benchmarks typically rely on high-end GPUs and focus on small or synthetic environments, leaving open questions about practical deployment on consumer-grade hardware. In this work, we systematically evaluate four representative NVS models: 3D Gaussian Splatting, Nerfacto, Instant-NGP, and TensoRF across a diverse set of real-world scenes categorized as indoor and outdoor. Using a unified training and evaluation pipeline, we analyze reconstruction quality and computational efficiency on two consumer-grade GPUs (GTX 1080Ti and RTX 2070S). A statistical analysis shows that scene type (indoor vs. outdoor) has no effect on reconstruction quality for any of the evaluated models. However, scene type does influence computational efficiency for Nerfacto, TensoRF, and 3D Gaussian Splatting. When comparing methods directly, 3D Gaussian Splatting delivers the highest reconstruction quality.**

***Index Terms*—Novel view synthesis, consumer-grade GPUs, reconstruction quality, computational efficiency, indoor, outdoor scenes, NeRF, 3D Gaussian Splatting, TensoRF, benchmarking.**

I. Introduction

Synthesizing novel views of a scene from a set of 2D images is a central problem in 3D computer vision. Recent advances in novel-view synthesis (NVS) have introduced a wide range of approaches, including implicit representations such as Neural Radiance Fields (NeRF) [1] and TensoRF [2], fast multi-resolution methods like Instant-NGP [3], and explicit representations such as 3D Gaussian Splatting [4]. These methods have enabled substantial progress in applications including virtual reality [5], robotics [6], urban modeling [7].

However, as the landscape of NVS techniques continues to expand, a practical question remains unresolved: how should one choose an approach given limited computational resources? Existing benchmarks typically evaluate models on indoor scenes and rely on powerful hardware such as A100 or multi-GPU setups. As a result, it is unclear how these methods compare when deployed on consumer-grade GPUs, which remain the most common hardware for researchers, hobbyists, and practitioners outside big institutions.

Another important but often overlooked factor is the type of scene being reconstructed. In particular, indoor and outdoor environments pose fundamentally different challenges for novel view synthesis. Outdoor scenes are typically unbounded and may include views of the sky or horizon, exhibit larger depth ranges, and suffer from significant image-to-image variations in illumination and appearance. As discussed in [8], these characteristics often require additional modeling components such as appearance embeddings, background handling, or semantic cues. Indoor scenes, in contrast, are usually bounded, more compact, and exhibit more controlled lighting conditions.

Despite the widespread use of this categorization in prior literature, most existing benchmarks do not explicitly analyze how indoor versus outdoor scenes affect reconstruction quality and computational efficiency, especially under consumer-grade hardware constraints.

Given these considerations, this study investigates how modern novel view synthesis methods perform under limited computational resources. Specifically, we ask:

1) **RQ1: How do indoor and outdoor scenes affect reconstruction quality**, measured by PSNR, SSIM, and LPIPS?
2) **RQ2: How do indoor and outdoor scenes affect computational efficiency**, including training time, rendering FPS, and rays-per-second throughput on consumer-grade hardware?

To answer these questions, we benchmark representative NVS models across datasets covering a wide range of indoor and outdoor scenes. Our goal is to provide practical insights into the efficiency–quality trade-offs of modern NVS models when used in realistic, resource-constrained settings.

II. Related Works

Since the introduction of NeRF in 2020 [1], the field of novel view synthesis (NVS) has rapidly expanded, giving rise to implicit, hybrid, and explicit 3D representations. This progress has motivated numerous surveys, benchmark frameworks, and domain-specific evaluations. Below we summarize the works most relevant to this study.

A. Comparative Studies and Surveys

Comprehensive surveys, such as Gao et al. [8], provide broad overviews of NeRF-based and post-NeRF NVS models, covering implicit radiance fields, tensor factorization methods, and explicit point-based approaches such as 3D Gaussian Splatting. This work focuses primarily on taxonomy, theoretical background, and high-level trends rather than controlled empirical evaluations under constrained hardware.

Other comparative studies investigate specific sub-domains, including robotics [6], UAV imagery [9], and robustness to image perturbations [10]. While valuable, these works remain domain-specific and do not address cross-scene general behavior or computational trade-offs.

B. Frameworks for Unified Evaluation

Nerfstudio [11] introduced a modular and user-friendly framework that streamlined the development, training, and visualization of NeRF-based methods by unifying data processing, scene reconstruction pipelines, and real-time tools. Building on this direction, NerfBaselines [12] further advanced standardization by providing a unified benchmarking interface based on official implementations of NeRF and 3D Gaussian Splatting methods, ensuring consistent evaluation protocols and reproducible results across models. Both frameworks conducted their experiments on high-performance GPUs and do not analyze how indoor and outood scenes or consumer-grade hardware affects reconstruction quality or computational performance.

C. Limitations of Existing Benchmarks

Across major NVS papers, including TensoRF [2], Instant-NGP [3], and 3D Gaussian Splatting [4], experimental evaluations rely heavily on high-performance GPUs such as NVIDIA A100, V100, or RTX 3090/4090. These platforms provide 16-80 GB of Video Random Access Memory (VRAM) and high compute throughput, making it unclear how these methods perform on hardware more commonly available to researchers.

Furthermore, although dataset comparisons are prevalent [13][14], scene types is generally characterized only by the *number of images*, rather than the indoor and outdoor scene. As a result, existing studies do not examine whether reconstruction quality or computational efficiency systematically differs between indoor object-centric scenes and outdoor real-world environments.

In summary, while prior surveys and benchmarking frameworks provide valuable foundations, no existing work systematically examines how modern NVS methods behave across indoor and outdoor scenes under consumer-grade hardware constraints. This study addresses this gap by comparing several representative NVS methods across indoor and outdoor scenes, with an emphasis on practical efficiency-quality trade-offs.

TABLE I
SCENES USED IN THE BENCHMARK, THEIR ASSIGNED TYPE, AND SOURCE DATASET.

Scene	Scene Type	Source
Kazan cathedral	outdoor	SfM
Sri Veeramakaliamman	outdoor	SfM
Round church	outdoor	SfM
Doge Palace	outdoor	SfM
Fine arts palace	outdoor	SfM
Cyprus	outdoor	SfM
GBG	outdoor	SfM
Pantheon	outdoor	SfM
Nikolai1	outdoor	SfM
Plaza de armas	outdoor	SfM
Pumpkin	indoor	SfM
Flower	indoor	LLFF
Fortress	indoor	LLFF
Orchids	indoor	LLFF
Bottle	indoor	NeRFBK
Glass	indoor	NeRFBK
Industrial A	indoor	NeRFBK
Metopa	indoor	NeRFBK
Vase	indoor	NeRFBK
Pinecone	indoor	NeRF-360
Vasedeck	indoor	NeRF-360

III. EXPERIMENTAL SETUP

A. Datasets

Existing NVS benchmarks such as LLFF [13], Mip-NeRF 360 [14], or synthetic Blender scenes from [1] focus primarily on indoor, object-centric environments or on clean, synthetic data. To study the effect of scene type on reconstruction performance, we combine scenes from several publicly available datasets covering both indoor and outdoor environments:

- SfM dataset [15]: outdoor structures, facades, statues, and urban landmarks.
- NeRFBK [16]: diverse indoor scenes, reflective objects, glass, and challenging material conditions.
- LLFF dataset [13]: small-to-medium indoor and outdoor scenes.
- NeRF-360 [1]: real-world 360° captures with high-resolution images.

All images were further downscaled so that image resolutions remained near 1920×1080. This step was necessary due to hardware constraints and for ensuring fair comparison across models with different memory footprints. For each scene, images were randomly split into 75% training and 25% validation. Additionally, each scene was manually assigned to either an indoor or outdoor category based on the capture environment, following common practice in prior NVS literature. This categorization reflects known differences in scene boundedness, and depth range rather than relying on explicit physical scale measurements, which are not consistently available across datasets. Table I lists all used scenes, types and source datasets.

B. Models

Since our goal is to provide practical recommendations for users with limited hardware resources, all experiments are conducted using the *Nerfstudio* framework [11], which provides unified data loading, camera pose estimation, training procedures, and evaluation tools.

We evaluate representative NVS models spanning different design paradigms:

- 3D Gaussian Splatting (3DGS) [4] — an explicit point-based representation enabling real-time rendering. We use Splatfacto, the official Nerfstudio implementation.
- Nerfacto [11] — a modern NeRF variant with several practical improvements, used as a strong implicit baseline.
- Instant-NGP [3] — a hash-grid accelerated NeRF optimized for fast training.
- TensoRF [2] — a tensor-factorized radiance field model.

Together, these models cover the main families of contemporary NVS approaches.

C. Evaluation Metrics

Reconstruction quality is evaluated using PSNR, SSIM, and LPIPS, while computational efficiency is measured by training time, rendering FPS, and rays-per-second throughput. These metrics capture both visual fidelity and practical performance on consumer-grade hardware.

D. Hardware and Training Protocol

All experiments were conducted on two consumer-grade setups:

1) NVIDIA GTX 1080Ti (12 GB) with 32 GB RAM.
2) NVIDIA RTX 2070 Super (8 GB) with 64 GB RAM.

Data preprocessing follows the Nerfstudio pipeline, using COLMAP for camera pose estimation. Each model was trained for a fixed 30,000 iterations, following the default training schedule of the Nerfstudio framework. Evaluation was performed on held-out validation views.

Importantly, 3D Gaussian Splatting (Splatfacto) requires a CUDA-ready GPU with Compute Capability ≥ 7.0, which is not supported by the GTX 1080Ti. Therefore, Splatfacto was trained and evaluated only on the 2070S setup, while the 1080Ti experiments include only Nerfacto, Instant-NGP, and TensoRF.

E. Statistical Analysis

To determine whether scene type (*indoor* vs. *outdoor*) has a significant effect on reconstruction quality and efficiency metrics, we perform several complementary statistical tests:

- MANOVA (Pillai's trace) [17] — used as a global test to assess whether the factors *type*, *model*, and their interaction jointly affect all metrics. Pillai's trace is chosen for its robustness to violations of normality and covariance homogeneity.
- Brunner–Munzel rank test [18] — applied as a follow-up test when MANOVA indicates significant effects. It identifies, for each model and metric, whether big vs. small scenes differ. The test is nonparametric and robust to unequal variances and non-normal distributions.
- Benjamini–Hochberg FDR correction [19] — used to control the false discovery rate across multiple pairwise comparisons, ensuring reliability when testing many metrics.

IV. Results

Figures 1 and 2 summarize the distribution of all metrics across small and big scenes for each model on setups 1 and 2. A MANOVA using Pillai's trace showed no significant main effect of scene type on the multivariate set of metrics, but revealed a significant *scene type* $\times$ *model* interaction, indicating that the influence of type differs across methods.

To localize this interaction, we compared indoor vs. outdoor scenes within each model using the Brunner–Munzel test with Benjamini–Hochberg correction. For both hardware setups, scene type showed *no significant effect* on the three reconstruction metrics (PSNR, SSIM, LPIPS) for any model.

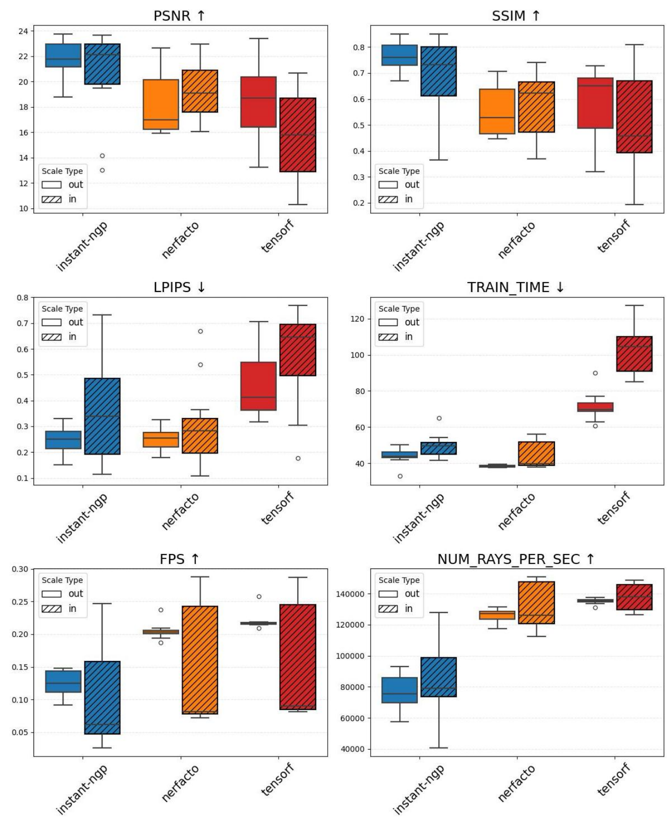

Fig. 1. Distribution of all metrics across outdoor and indoor scenes for each model on the 1080Ti setup.

For computational efficiency, scene type affected performance only in specific cases:

- TensoRF: significantly longer training times on indoor scenes on both the 1 and 2 setups.
- Nerfacto: significantly longer training times on indoor scenes on the 1 setup.
- Splatfacto: significantly higher rays-per-second throughput on outdoor scenes on the 2 setup.

V. Conclusion

This study examined how modern novel view synthesis methods behave under consumer-grade hardware constraints

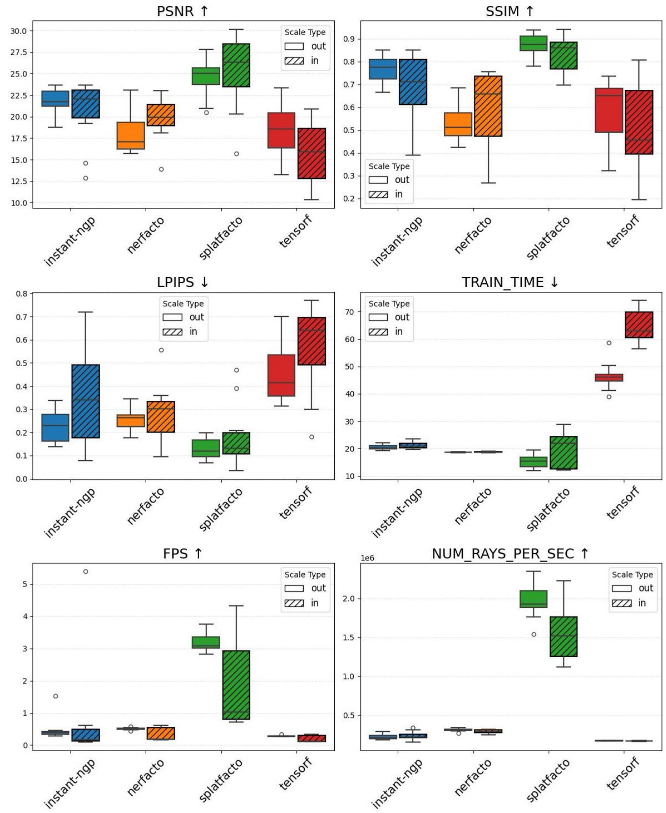

Fig. 2. Distribution of all metrics across outdoor and indoor scenes for each model on the 2070S setup.

and across indoor and outdoor scenes. By benchmarking 3D Gaussian Splatting, Nerfacto, Instant-NGP, and TensoRF on diverse real-world scenes, we provide practical insights into their quality and efficiency trade-offs.

A MANOVA revealed no overall main effect of scene type on reconstruction metrics, but a significant scene type $\times$ model interaction. Follow-up Brunner–Munzel tests showed that scene type does not affect PSNR, SSIM, or LPIPS for any method, but influences computational efficiency in method-specific ways: TensoRF trains more slowly on indoor scenes on both setups, Nerfacto trains more slowly on indoor scenes on the 1080Ti setup, while Splatfacto renders faster on outdoor scenes on the 2070S.

Overall, the results directly answer our research questions. **RQ1** shows that indoor and outdoor scene types do not influence reconstruction quality, whereas **RQ2** indicates that scene type affects computational efficiency in a model and hardware dependent manner. These findings suggest that, on consumer-grade hardware, model selection should prioritize the desired balance between reconstruction quality and efficiency rather than scene indoor and outdoor scene types.

While this study considers indoor and outdoor scene types as a high-level categorization, the observed differences in computational efficiency suggest that scene *complexity* may be a more informative factor than scene type alone. Future work should therefore investigate how more fine-grained measures of scene complexity influence reconstruction quality and efficiency across different NVS representations.

References

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[2] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "Tensorf: Tensorial radiance fields," in *European Conference on Computer Vision (ECCV)*, 2022.

[3] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

[4] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[5] N. Deng et al., "Fov-nerf: Foveated neural radiance fields for virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 11, pp. 3854–3864, 2022.

[6] Y. Ming et al., "Benchmarking neural radiance fields for autonomous robots: An overview," *Engineering Applications of Artificial Intelligence*, vol. 140, p. 109 685, 2025.

[7] X. Fu et al., "Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation," in *2022 International Conference on 3D Vision (3DV)*, IEEE, 2022, pp. 1–11.

[8] K. Gao, Y. Gao, H. He, D. Lu, L. Xu, and J. Li, "Nerf: Neural radiance field in 3d vision, a comprehensive review," *arXiv preprint arXiv:2210.00379*, 2022.

[9] M. E. Atik, "Comparative assessment of neural radiance fields and 3d gaussian splatting for point cloud generation from uav imagery," *Sensors*, vol. 25, no. 10, p. 2995, 2025.

[10] C. Wang, A. Wang, J. Li, A. Yuille, and C. Xie, "Benchmarking robustness in neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2926–2936.

[11] M. Tancik et al., "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, ser. SIGGRAPH '23, 2023.

[12] J. Kulhanek and T. Sattler, "Nerfbaselines: Consistent and reproducible evaluation of novel view synthesis methods," *arXiv preprint arXiv:2406.17345*, 2024.

[13] B. Mildenhall et al., "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Transactions on Graphics (ToG)*, vol. 38, no. 4, pp. 1–14, 2019.

[14] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5470–5479.

[15] O. Enqvist, F. Kahl, and C. Olsson, "Non-sequential structure from motion," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, IEEE, 2011, pp. 264–271.

[16] Z. Yan, G. Mazzacca, S. Rigon, E. M. Farella, P. Trybala, F. Remondino, et al., "Nerfbk: A holistic dataset for benchmarking nerf-based 3d reconstruction," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 48, no. 1, pp. 219–226, 2023.

[17] C. Andrade, A. P. Madhavan, and M. Kishore, "Testing logical memory using a complex passage: Development and standardization of a new test," *Indian Journal of Psychiatry*, vol. 43, no. 3, pp. 252–256, 2001.

[18] U. Munzel and E. Brunner, "An exact paired rank test," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 44, no. 5, pp. 584–593, 2002.

[19] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.