BUILD-A-VOLCAP: AUTOMATED SYNTHETIC VALIDATION OF REAL-WORLD VOLUMETRIC CAPTURE QUALITY

Antonella Rech[1], *Giulia Martinelli*[1,2], *Nicola Garau*[1,2], *Nicola Conci*[1,2], *Francesco De Natale*[1,2]

[1]University of Trento, Via Sommarive 14, 38123 Trento TN, Italy
[2]CNIT – Consorzio Nazionale Interuniversitario per le Telecomunicazioni, Italy

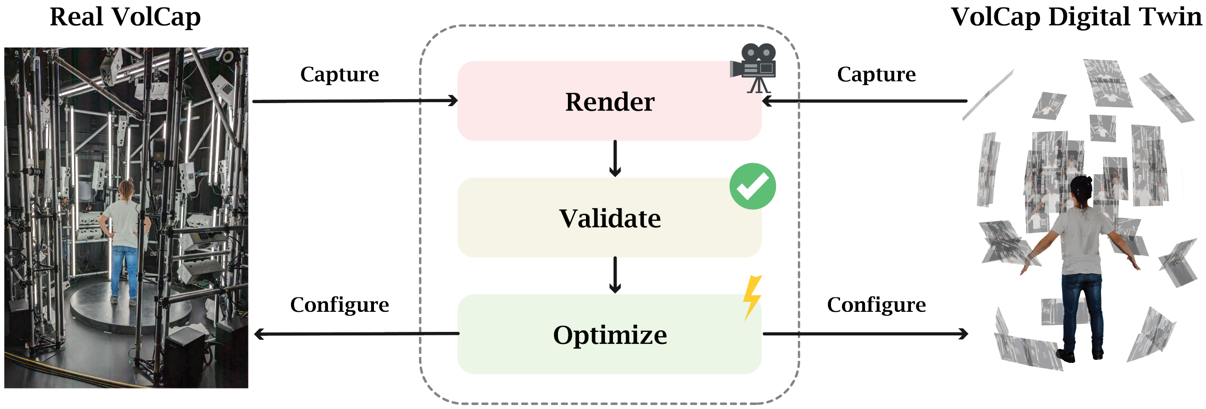

Fig. 1: We present **Build-A-Volcap**, an automated pipeline for validating real-world Volumetric Capture (VC) setups using a synthetic Digital Twin. Left to right: scenes from a real world volumetric capture are captured and rendered to images. The original intrinsic and extrinsic parameters are extracted and transformed to build a Digital Twin of the volumetric capture. Images from the synthetic VC are then re-rendered and compared against the real ones. Exploiting the obtained measurements, we optimize the cameras positions and validate them once again, until a desired quality in the reconstruction is reached. Finally, after the synthetic VC converges to a desired configuration, the real VC can be adjusted accordingly.

ABSTRACT

Volumetric capture techniques have significantly advanced in recent years, enabling detailed 3D reconstructions of dynamic real-world scenes. However, predicting their performance prior to real-world deployment remains challenging, often leading to costly experimental setups and uncertain outcomes. In this paper, we present Build-A-Volcap, an automated synthetic validation environment designed to predict the quality of real-world volumetric capture systems. Our solution constructs comprehensive synthetic datasets and environments that faithfully simulate practical capture conditions. This enables rigorous testing and benchmarking of volumetric methods such as photogrammetry, NeRF, Gaussian Splatting and Radiant Foam without the need for physical setups. Through extensive experiments, we demonstrate how Build-A-Volcap effectively identifies methodological strengths and limitations, significantly reducing the synthetic-to-real gap. Our pipeline enables researchers and companies to optimize and validate their volumetric capture setups virtually, ensuring robust performance upon real-world deployment. More information and code release on our project page: `https://mmlab-cv.github.io/Build-A-Volcap/`.

***Index Terms*—** Volumetric Capture, Photogrammetry, Volume Rendering, Computer Vision, Computer Graphics

1. INTRODUCTION

In recent years, volumetric capture (VC) technologies have seen a remarkable progress, enabling high-resolution reconstruction of static and dynamic scenes across various domains, such as virtual and mixed reality (VR, MR), virtual production, cultural heritage and many more. Traditionally, photogrammetry has been used to reconstruct point clouds and meshes from a set of object-centric pictures. Novel methods, such as Neural Radiance Fields (NeRF), Gaussian Splatting, and Radiant Foam have emerged as an alternative for capturing and rendering complex 3D environments, with some additions such as the inclusion of view-dependent effects, such as transparency and reflections. Despite these advancements, accurately predicting and evaluating the performance of volumetric capture setups before real-world implementation remains a significant challenge. Experimental setups for volumetric capture are often resource-intensive, costly, and susceptible to uncertainties arising from environmental conditions, hardware limitations, and methodological constraints. Even in real-world conditions, changing the camera configuration for a volumetric capture system is time consuming due to the need for re-calibration, capturing and testing.

To address these challenges, we introduce Build-A-Volcap, a novel automated synthetic validation framework that provides a robust simulation environment designed specifically to emulate realistic volumetric capture scenarios. Our proposed solution can be used to validate different volumetric capture configurations prior to building the physical setup, using multiple different reconstruction methods. By bridging the synthetic-to-real gap, Build-A-Volcap enables detailed analysis and benchmarking of method performance, helping identify strengths and limitations early in the development process.

Through extensive experimental setups and ablation studies, we show that Build-A-Volcap can be a highly valuable resource for research institutions and companies willing to build a volumetric capture studio, especially in space-constrained scenarios or with a limited budget. Furthermore, we show some results on a digital twin of a real VC setup, proving that Build-A-Volcap can be used to precisely replicate real, high fidelity setups.

2. RELATED WORK

2.1. Volumetric Capture Setups

Recent advances in VC technology have enabled detailed 3D reconstructions through advanced multi-camera setups, including mono, stereo, depth and ToF cameras. Most of the VC setups use large arrays of synchronized RGB cameras and distributed GPU processing to achieve a high-fidelity capture [1]. Other systems, such as the Panoptic Studio [2] are based on a combination of low and high resolution RGB images as well as depth data. Most of the VC setups require entire rooms or large areas in general; only some commercially available solutions offer a professional high-resolution setup that can fit in a limited space. In recent years, monocular volumetric performance capture methods [3] using deep learning, structure from motion (SfM), or a combination of both have also emerged; however, they suffer from limitations and are not ideal for capturing volumetric videos, but rather static scenes. [4] provides a comprehensive review of VC setups, highlighting different system architectures, as well as the challenges in camera calibration, frame synchronization and data management.

2.2. Volumetric Capture Methods

Traditional photogrammetry and multi-view stereo (MVS) techniques remain fundamental baselines in volumetric reconstruction, particularly represented by established frameworks like COLMAP [5, 6] and newer commercially available solutions, such as Epic Games' Reality Capture[1]. Neural Radiance Fields (NeRF) [7] have substantially impacted the field by representing scenes with neural networks trained to model continuous volumetric density and color fields. Early NeRF implementations faced computational constraints, which have been partially solved from subsequent methods such as Instant-NGP [8] and Mip-NeRF 360 [9], which significantly accelerated rendering and improved scalability for larger, real-world environments. Further improvements in real-time rendering were achieved through Gaussian Splatting [10], which employs optimized Gaussian primitives for efficient rasterization-based rendering. Radiant Foam [11] exploits a decades-old volumetric mesh ray tracing algorithm to introduce hybrid explicit-implicit representations through differentiable volumetric meshes, effectively combining the advantages of traditional geometry and neural rendering.

3. METHOD

For all the real-world experiments, we collect data using a custom version of a Mantis Vision VC setup[2], with 32 units, each containing a stereo cameras of resolution 2456x2054, an infrared camera and projector, as well as an Intel NUC mini PC. For more information on the VC setup, please refer to [12]. After building the digital twin of this setup, as well as two comparable configurations (Fig. 3), we choose Reality Capture as a photogrammetry baseline, together with NeRF, Gaussian Splatting and Radiant Foam, due to their popularity in the research community.

3.1. Real-to-synth

We start by processing the intrinsic and extrinsic parameters of the Mantis volumetric capture system, which transforms are relative to the first infrared (IR) camera unit. Specifically, these parameters are represented as camera matrices, defined as:

$$\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}], \tag{1}$$

where $\mathbf{K}$ is the intrinsic matrix comprising focal lengths and principal point coordinates, and $[\mathbf{R}|\mathbf{t}]$ represents the extrinsic (rotation and translation) parameters of the camera. These parameters are first transformed from the local coordinate system to a unified global coordinate frame, enabling consistent comparisons and conversions among different reference systems (Blender, Reality Capture, Nerfstudio). To accurately

[1] https://www.capturingreality.com/

[2] https://mantis-vision.com/3d-studio-3iosk/

align captured images within a unified global coordinate system, we apply the transformations described in Algorithm 1. Each stereo pair includes a primary and a secondary RGB camera. The primary camera pose is directly transformed to the global frame. For the secondary camera, which is defined relative to the infrared (IR) reference, we first compute the global IR-to-world transformation $\mathbf{T}^{\text{IR}}_{\text{world}}$. Its inverse, $\mathbf{T}^{\text{world}}_{\text{IR}}$, is used to map poses into the global frame. A 180° rotation around the z-axis ($\mathbf{R}_z$) corrects orientation, and translation values are scaled from millimeters to meters. The final transformation $\mathbf{T}^{\text{world}}_{\text{RGB}}$ aligns each camera for consistent real-to-synthetic comparison.

Algorithm 1 Camera Transformation

```
**Input camera parameters:** $\mathbf{K}$, $\mathbf{T}_{\text{RGB}}$
**Output:** $\mathbf{T}^{\text{world}}_{\text{RGB}}$
**if** secondary camera **then**
    $\mathbf{T}^{\text{world}}_{\text{RGB}} \leftarrow R_z \cdot \text{inv}(\mathbf{T}_{\text{RGB}}) \cdot \mathbf{T}^{\text{world}}_{\text{IR}} \cdot \text{inv}(\mathbf{T}^{\text{world}}_{\text{IR}})$
**else**
    $\mathbf{T}^{\text{world}}_{\text{RGB}} \leftarrow \mathbf{R}_z \cdot \text{inv}(\mathbf{T}_{\text{RGB}}) \cdot \text{inv}(\mathbf{T}^{\text{world}}_{\text{IR}})$
**end if**
$\mathbf{T}^{\text{world}}_{\text{RGB}}[1:3,4] \leftarrow \mathbf{T}^{\text{world}}_{\text{RGB}}[1:3,4]/1000$
**return** $T^{world}_{RGB}$
```

The resulting transformation $\mathbf{T}^{\text{world}}_{\text{RGB}}$ provides a consistent coordinate system suitable for accurate comparisons and evaluations within synthetic reconstruction environments.

Now that the transformation is computed, the scene can be rendered within the chosen toolkit. The resulting reconstruction can then be compared against the original captured images to quantitatively and qualitatively assess the fidelity and accuracy of the synthetic pipeline.

3.2. Synth-to-real

After creating the digital twin of the physical VC setup, our goal is to optimize the camera configuration without manually adjusting the cameras in the real world. As illustrated in Figure 1, this is achieved by virtually reconfiguring the camera positions within the synthetic Build-A-Volcap environment, re-rendering the same target scene, and comparing the outputs against the original images. This iterative procedure enables validation and refinement of the configuration, eventually leading to improved capture quality, as discussed in Section 4.

To ensure robustness, we validate our pipeline using both synthetic and real subjects. These subjects exhibit a wide range of characteristics, including varying textures, levels of detail, and sizes. This diversity allows us to assess the performance of Build-A-Volcap under different conditions, ensuring that the final camera setup achieves the desired quality and consistency.

Upon completing the optimization phase, the resulting configuration can be directly transferred to the physical setup. Cameras in the real-world studio can then be re-arranged according to the optimized layout without requiring a trial-and-error process. In this sense, Build-A-Volcap acts as a powerful simulation and planning environment that facilitates the design and adjustment of volumetric capture systems before physically implementing them.

	PSNR↑	**SSIM↑**	**LPIPS↓**
Chair	28.89	0.920	0.129
Fruits	27.7	0.910	0.177
Mixamo	29.74	**0.970**	**0.043**
Subject1	**31.51**	0.968	0.061
Subject2	29.49	0.966	0.057

Table 1: NeRF quantitative results for the volumetric reconstruction of 5 subjects in the Mantis VC setup.

	PSNR↑	**SSIM↑**	**LPIPS↓**
Chair	**30.31**	0.911	0.098
Fruits	25.54	0.891	0.167
Mixamo	29.25	**0.973**	**0.041**
Subject1	30.25	0.958	0.067
Subject2	29.78	0.965	0.057

Table 2: Gaussian Splatting quantitative results for the volumetric reconstruction of 5 subjects in the Mantis VC setup.

4. RESULTS AND ABLATION STUDIES

The main goal of this paper is to provide a solution to replicate, validate and improve existing volumetric capture systems. To do so, we first run state-of-the-art volumetric reconstruction methods on multiple subjects and extract some relevant metrics. Subsequently, we provide ablations on both the physical setup of the VC and the reconstruction pipelines.

For the baseline results, we run several volume reconstruction pipeline on five different objects, both real and synthetic: a chair, a basket full of fruit, a synthetic character from the Mixamo dataset [3] and two real characters captured using

	PSNR↑	**SSIM↑**	**LPIPS↓**
Chair	36.33	0.949	0.078
Fruits	30.27	0.930	0.116
Mixamo	36.89	**0.987**	**0.024**
Subject1	**39.20**	0.981	0.040
Subject2	38.03	0.980	0.039

Table 3: Radiant Foam quantitative results for the volumetric reconstruction of 5 subjects in the Mantis VC setup.

[3] https://www.mixamo.com/

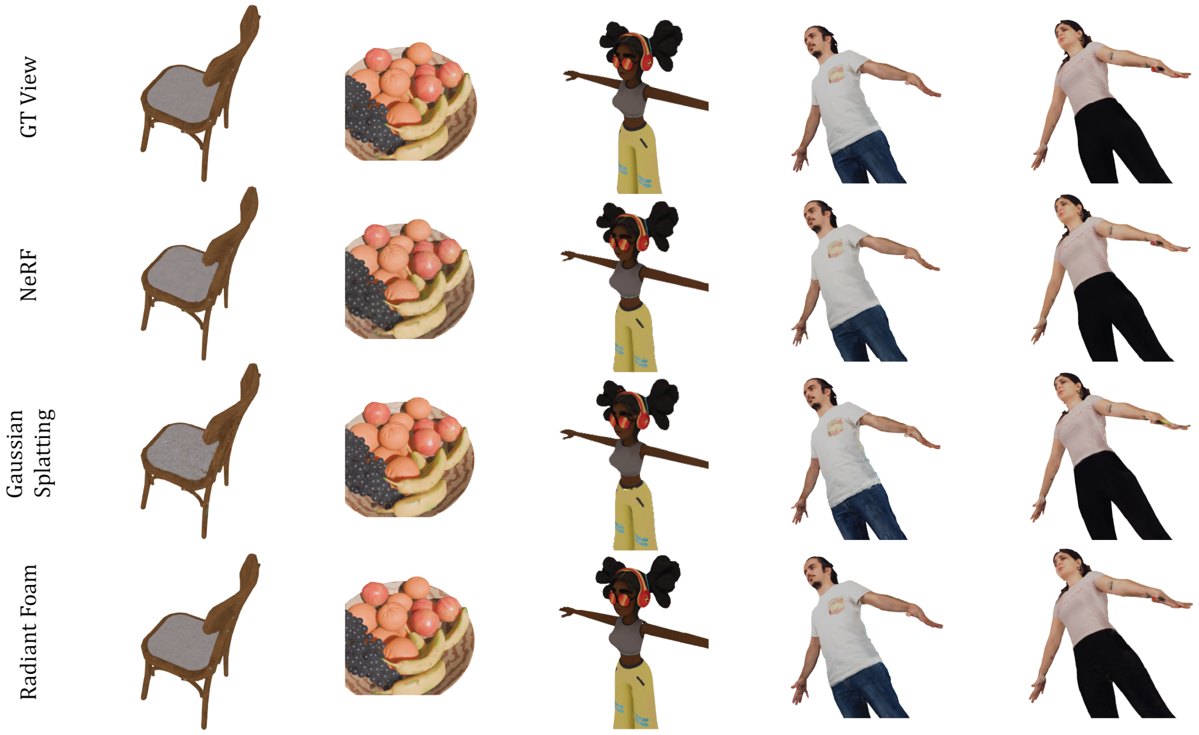

Fig. 2: Qualitative results for the five chosen characters using the selected volumetric reconstruction methods on the *Mantis* configuration. All the methods perform very similarly, although they produce different kinds of barely perceptible artifacts. The last two columns are being compared to real-world scenes captured using a physical VC setup by Mantis Vision.

the Mantis VC setup. As for the choice of the reconstruction pipelines, we choose Reality Capture as the only photogrammetry method, while for the "neural-like" methods, we choose NeRF [7], Gaussian Splatting [10] and Radiant Foam [11]. For the first two, we rely on the Nerfstudio implementation [13] for convenience, while for Radiant Foam we use the original implementation (v1).

We show the results in terms of PSNR, SSIM and LPIPS in Tables 1,2,3. In all our experiments we showcase consistent results over all the subjects, with little variation between the three methods, with Radiant Foam achieving a slightly better overall result compared to NeRF and Gaussian Splatting. In Figure 2 we show some qualitative results on all the reconstructed objects from the same viewpoint. All the chosen methods achieve comparable results in terms of perceived quality.

4.1. Ablation Studies

We conducted extensive ablation studies for our pipeline, that can be seen in Tables 4 and 5. Figure 3 shows the three VC setups that we considered for our experiment. *Mantis* is the exact digital twin of a customized Mantis Vision VC system. Cylinder refers to a simple setup composed of two rings of 16 cameras each, and a similar reasoning is applied to the Cube setup.

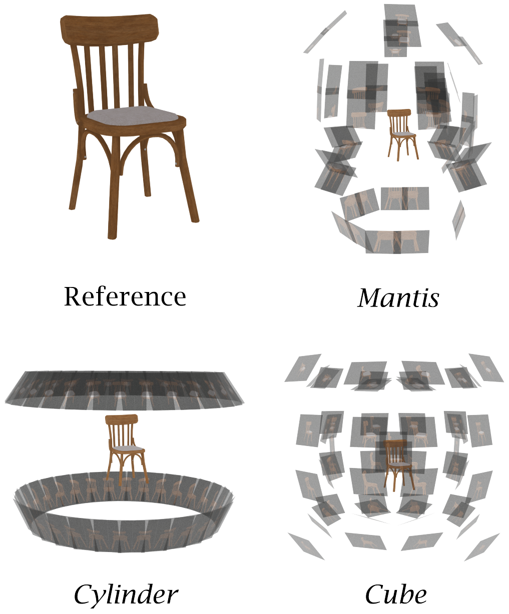

Fig. 3: Different configurations for the VC setup: Mantis (exact digital twin), Cylinder (prioritize density) and Cube (prioritizing uniformity).

	PSNR↑			**SSIM↑**			**LPIPS↓**		
	NeRF	***GS***	***RF***	***NeRF***	***GS***	***RF***	***NeRF***	***GS***	***RF***
Mantis	29.49	29.78	**38.03**	0.970	0.960	**0.980**	0.057	0.057	**0.039**
Cylinder	32.83	32.89	**39.97**	0.970	0.970	**0.980**	0.045	0.042	**0.024**
Cube	28.36	28.37	**41.25**	0.960	0.960	**0.988**	0.065	0.065	**0.022**

Table 4: Ablation studies for the different configurations of the cameras in the synthetic VC. In all the configurations, Radiant Foam (RF) consistently outperforms the other two methods, while the Cube configuration is the best overall capturing setup of the three, most likely because of its uniformily distributed views.

	PSNR $\mu\uparrow$	**PSNR** $\sigma\downarrow$	**SSIM** $\mu\uparrow$	**SSIM** $\sigma\downarrow$	**LPIPS** $\mu\downarrow$	**LPIPS** $\sigma\downarrow$
Reality Capture	29.36	1.253	0.963	0.006	0.043	**0.010**
Nerfacto	31.51	0.924	0.968	**0.005**	0.061	0.019
Nerfacto-big	28.23	**0.833**	0.940	0.048	0.070	0.069
Nerfacto-huge	27.34	1.191	0.933	0.014	0.078	0.019
Splatfacto	30.26	1.961	0.958	0.013	0.067	0.028
Splatfacto-big	**40.48**	2.084	**0.985**	0.007	**0.016**	0.011
Radiant Foam	39.20	2.367	0.981	0.009	0.039	0.022

Table 5: Ablation studies for the chosen reconstruction pipelines. Please refer to Nerfstudio [13] for exact model sizes and parameters. Splatfacto-big appears to be the best model for this particular configuration (Mantis), closely followed by Radiant Foam.

In this ablation study, we keep the camera parameters fixed in all the configurations. Specifically, we consistently use 60 training cameras, and 12 test cameras for each configuration, both with a fixed 35mm focal length. In particular, we conducted a thorough analysis of each algorithm's behavior with respect to camera pose and spatial proximity within the scene. For instance, we use the *Mantis* configuration to represent a sparse view setup, the *Cylinder* configuration as a dense representation with multiple closely spaced views, and the *Cube* setup to prioritize views uniformity. The results associated to this study can be seen in Table 4.

In addition, we perform further ablation studies on the type and size of volumetric reconstruction pipeline, as seen in Table 5. It is worth noticing that the larger models do not always lead to better performances.

5. CONCLUSIONS

In this work, we presented Build-A-Volcap, an automated synthetic validation framework designed to predict and optimize real-world volumetric capture quality prior to physical deployment. By simulating a realistic volumetric capture setup and generating high-resolution multi-view datasets, our method allows detailed analysis of various volumetric reconstruction techniques, including photogrammetry, NeRF, Gaussian Splatting, and Radiant Foam. Extensive quantitative experiments and ablation studies demonstrated that Build-A-Volcap effectively identifies methodological strengths and limitations, significantly bridging the synthetic-to-real gap. Additionally, our synthetic simulations can precisely replicate complex real-world capture setups by creating a proper digital twin of the physical VC. Build-A-Volcap thus provides a valuable tool for researchers and companies to efficiently test and optimize their volumetric capture solutions, reducing both costs and uncertainty associated with the deployment of traditional setup. Furthermore, the proposed framework is highly customizable, allowing the developer to arbitrarily manipulate the cameras parameters and design novel capturing setups.

6. ACKNOWLEDGEMENTS

This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART") and by the Italian Ministry of University and Research in the framework of the Call for Proposals for scrolling of final rankings of the PRIN 2022 call - Protocol no. 2022KCKYA2. This work was carried out with the support of Centre for Immersive Visual Technologies (CIVIT) research infrastructure, Tampere University, Finland[4].

[4] `https://civit.fi/`

7. REFERENCES

[1] Jonathan Heagerty, Sida Li, Eric Lee, Shuvra Bhattacharyya, Sujal Bista, Barbara Brawn, Brandon Feng, Susmija Jabbireddy, Joseph JaJa, Hernisa Kacorri, David Li, Derek Yarnell, Matthias Zwicker, and Amitabh Varshney, "HoloCamera: Advanced volumetric capture for cinematic-quality vr applications," *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[2] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh, "Panoptic studio: A massively multiview system for social motion capture," 2015.

[3] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li, "Monocular real-time volumetric performance capture," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer, 2020, pp. 49–67.

[4] Yili Jin, Kaiyuan Hu, Junhua Liu, Fangxin Wang, and Xue Liu, "From capture to display: A survey on volumetric video," *arXiv preprint arXiv:2309.05658*, 2023.

[5] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[6] Johannes Lutz Schönberger and Jan-Michael Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] 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," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[8] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, July 2022.

[9] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," *CVPR*, 2022.

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, July 2023.

[11] Shrisudhan Govindarajan, Daniel Rebain, Kwang Moo Yi, and Andrea Tagliasacchi, "Radiant foam: Real-time differentiable ray tracing," *arXiv:2502.01157*, 2025.

[12] Guillaume Gautier, Alexandre Mercat, Louis Fréneau, Mikko Pitkänen, and Jarno Vanne, "Uvg-vpc: voxelized point cloud dataset for visual volumetric video-based coding," in *2023 15th international conference on quality of Multimedia experience (QoMEX)*. IEEE, 2023, pp. 244–247.

[13] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa, "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, SIGGRAPH '23.