MBARI Monterey Bay Aquarium Research Institute

Opti-Acoustic Fusion for Gaussian Splatting

Conor Gagliardi, Oregon State University

Mentors: Akshay Hinduja, Giancarlo Troni

Summer 2025

Keywords: Gaussian Splatting, Sensor Fusion, Scene Reconstruction

ABSTRACT

Mapping and Scene reconstruction are highly prioritized areas of research interest within the marine domain. 3D Gaussian Splatting (3DGS) is a new technique that is currently dominating the research on object and scene reconstruction [1]. Despite its prevalence, as with many mapping/rendering techniques, it struggles to perform well underwater. In this work, I explore the implementation of a method called Z-Splat [2], which utilizes sonar sensor fusion with RGB imagery to create a high-quality reconstruction that overcomes several marine artifacts, including turbidity, light scattering, and distortion. I compare it with standard 3DGS to identify strengths in the state-of-the-art and potential research gaps to overcome in future work. Notably, while most methods excel in object reconstruction, this method shows great potential for image and sonar fusion for high-quality terrain reconstruction, which is crucial for oceanic mapping.

1 INTRODUCTION

Mapping, 3D Scene rendering, and digital environment reconstruction are all tasks that are currently being pushed for improvement in the marine community. Coral ecosystems, marinas, shipyards, oil rigs, seamounts, mining, defense, and construction are all media that rely on high-quality data for informed decision-making, study, and mission completion.

Enabling these tasks underwater, remote-operated vehicles (ROVs) and autonomous underwater vehicles (AUVs) are commonly employed. For the task, these systems are typically equipped with perception sensor suites, including monocular cameras and/or sonars. These perception systems enable the capture and representation of 3D spatial data through the use of existing techniques, such as photogrammetry, structure from motion (SfM), and neural radiance fields. [3, 4, 5]

3D Gaussian Splatting (3DGS) [1] is a novel view synthesis and differentiable scene reconstruction technique that has gained recent prominence. Most developments have been made in the surface world, and as such, are well-adapted to typical environmental constraints that one would encounter in the air. Underwater environments present additional challenges, including turbidity, light diffusion, and dynamic environments, among others.

Cameras and Sonars each have their own strengths and weaknesses; cameras are great at capturing structure and visual understanding, but lack depth (in reference to the sensor) and are susceptible to turbid and low-light conditions, which can degrade data capture. Sonar, on the contrary, tends to overcome turbidity and low light, being acoustic sensors. It is also good at capturing depth data, while visual structure is lacking compared to RGB monocular camera imagery. The use of each complements the other to mitigate the previously mentioned weaknesses.

Another obstacle is capturing viewing angles for the environment. Many techniques, such as photogrammetry, work best when used with an exhaustive range of view angles. Additional view angles allow the camera data to overcome the depth information shortcoming. [3, 6] This is typically done for the digitization of objects. We can capture images in a full 360-degree view of the object to gain a comprehensive understanding of its structure. The problem faced when attempting to use the same technology for terrain reconstruction is that the viewing angles are typically much less robust for any given scene. A terrain traversal normally has a fixed range and angle relative to the environment.

Naturally, we can reach the conclusion that, while using cameras that lack depth information and having a sensor that captures depth, their fusion should greatly benefit the overall objective of a terrain reconstruction mission.

2 MATERIALS AND METHODS

2.1 DATA COLLECTION

Data were collected for testing using two methods. The first, using Mitsuba [7], simulated data from 3D models capturing both RGB and depth imagery to emulate sonar. Field data were collected from the MBARI test tank using the CoMPAS Lab's MOLA ROV/AUV, augmented with a high-resolution camera and an Oculus m750d forward-looking sonar. Various trajectories and objects were captured, which will be discussed in the following two sections.

2.1.1 SIMULATED DATA

To capture the simulated data, 3D models were gathered from the Stanford 3D model library [8]. In order to extract RGB and sonar views of the model, Mitsuba was used [7]. Given a trajectory for a target-facing orbit, a set of key frames was extracted for RGB and Depth images, from which slices were taken to emulate the range histograms of an FLS. The simulated data is helpful for iterative testing and controlling the exact parameters for viewing angles. We primarily used the LEGO bulldozer model and the Stanford rabbit. The Lego helped because it has color data, and the rabbit has a more curvy structure than the rigid edges of the Lego construct.

2.1.2 FIELD DATA

Field data were captured from the test tank. Facilitating data capture was the MOLA, a ROV/AUV (remotely controlled in this case), outfitted with a Sony 8 K (16-47mm equivalent, f/1.8 lens, 1″ sensor) camera and an Oculus m750d sonar. The sensors were adjacent and parallel, facing the same direction from nearly the same position (functionally the same position for the testing purposes). The first two tests were "lawnmower pattern" rasters of the test tank walls and the southwest corner. The corner was captured to test how well the spatial relativity of the 3D scene reconstruction functioned. The following two data collection tests involved "tidally locked" target-oriented orbits of the MOLA about two objects (a Docking station and a metal panel calibration board) suspended roughly 1.5m underwater. The MOLA maintained a static depth and distance from the objects during the orbits.

2.1.3 Sonar and Camera Sensor Data

These sensor data are each formatted differently, which requires a few extra steps for their fusion. Camera data is traditionally represented as a 2D matrix, with the origin located at the top-left. The coordinate frame has X values increasing along the horizontal axis and Y values increasing along the vertical axis. Z data represents depth, orthogonal to the matrix, away from the view frame. Sonar data is nuanced based on the sensor type. In this case, we use a forward-looking imaging sonar. The FLS data is represented in polar coordinates, with the origin being a single point, while angles are projected onto a cone to capture the information. The information is represented by an azimuth angle (theta) and a range/distance (r). Angle captures X values, distance aligns with Z, and the elevation data, as Y (or phi), is ambiguous, capturing an angle and compressing it onto the plane, as the Z data is in a typical monocular camera image. In simulation, this datatype is emulated by creating a range histogram from a horizontal slice of a depth image.

2.1.4 Key Frame Pairs

Key frame pairs are necessary for ensuring the fusion data represent the same view at the same time. In simulation, data is generated simultaneously, making matching trivial. In the case of field data, the sensors operate at different frequencies. The sonar captured data at a higher frequency, and thus was synced to chosen images that best captured the scene with good quality and low blur (at a roughly equivalent time step between each frame).

2.1.5 Preparing Data for Z-Splat

Due to the camera housing being a dome for waterproofing, image distortion is expected from the data capture. The distortion can cause 3D reconstruction to have lower quality, making straight objects appear curved. To prevent this, the images must be rectified. In this case, we use COLMAP's undistortion algorithm to achieve this [9].

Sonar data also must be preprocessed for use in Z-Splat. The data is represented in a 2D matrix, similar to an RGB camera, with the origin located at the top of the image. Sonars also tend to have low-intensity noise and dense noise right near the sensor. This is overcome by overlaying zero-intensity values near the origin and filtering the remaining noise below a specified threshold.

2.1.6 Structure from Motion

Finally, before optimizing, a set of points is used to seed the initial scene. There are two methods used for this: random initialization and estimated scene initialization. Random initialization uses scattered points within the scene dimensions. For estimated scene initialization, an algorithm such as COLMAP [9] is used. This algorithm takes a set of images and uses feature matching to estimate the camera's position in space. Based on this, it initializes matched features as points in a point cloud, which then seeds the optimization with a good guess of the scene.

2.2 Z-Splat

Z-Splat [2] is based on 3D Gaussian Splatting [1], a state-of-the-art volume rendering technique that also encompasses other methods, such as Neural Radiance Fields (NeRF) [5]. The goal of these algorithms is to capture spatial data in a way such that each view uniquely captures the properties of light and structure.

2.2.1 Gaussian Splatting

Gaussian splatting, as discussed in [1], involves projecting 3D gaussians onto a 2D image. It can do this due to the anisotropic nature of the Gaussians and their included spherical harmonics. The projection is rasterized and rendered as a 2D camera image by projecting rays through the Gaussians from the view frame and using alpha-blending for each pixel. These images can then be compared against ground truth to differentiate a loss, the feature that makes the algorithm differentiable and optimizable (commonly used in machine learning, such as backpropagation and gradient descent).

2.2.2 Optimization Steps

The optimization steps involve three main functions. These are typically referred to as the Clone, Split, and Prune steps. These steps modify the Gaussians within the scene to better align with the ground truth images using the Structural Similarity Index Measure (SSIM) [10]. When a region requires additional information, the Clone step splits a Gaussian, enabling further information to be represented within the specified area. If there is a region being overgeneralized by too large a Gaussian, the split step divides it in half. Finally, if there are Gaussians that provide no useful information, they are culled in the Prune step. Further parameters are also adjusted during each optimization step, including opacity, colors, and size.

3 RESULTS

3.1 Simulated Data Results

3.1.1 RGB Only

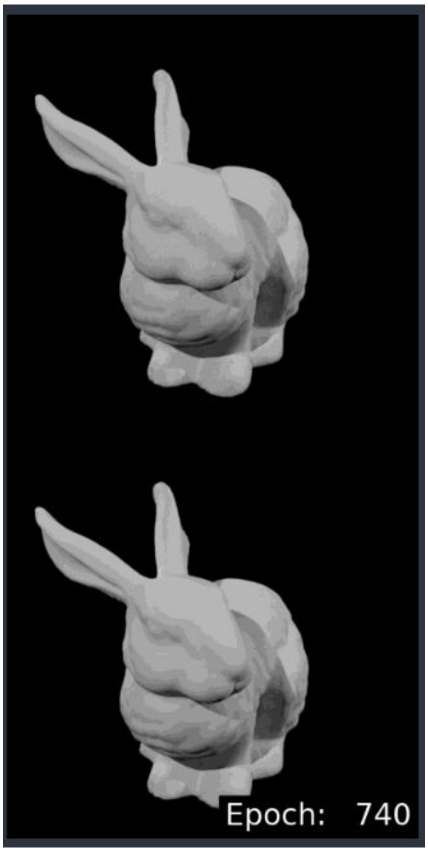

Figure 1: *Rendered scene view from RGB-only reconstruction (splatting is bottom).*

Geometric scores:

Chamfer Distance: 0.5365268203411178

Hausdorff Distance: 1.4812917328902362

Median Distance: 0.05141216006735875

F-score: 0.04518639763503701

Precision: 0.0235

Recall: 0.5855089102980083

Photometric scores:

SSIM: 0.9878209829330444

PSNR: 46.5294189453125

LPIPS: 0.047710977494716644

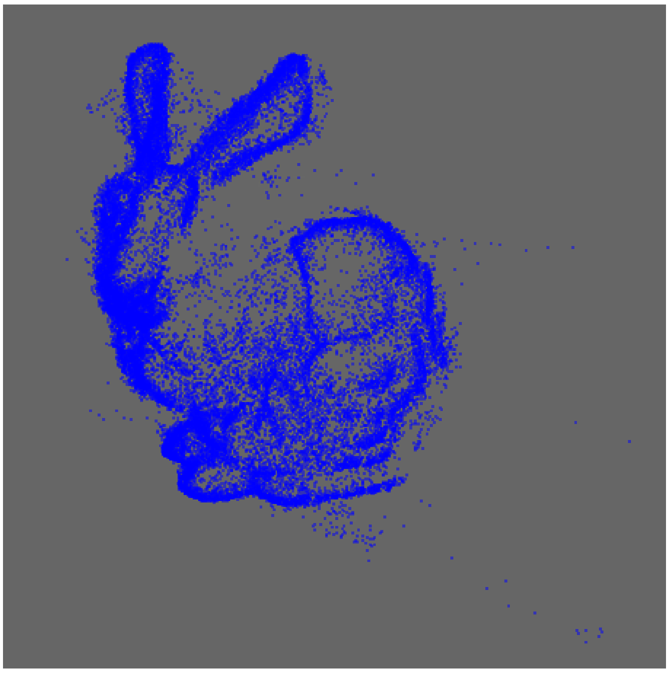

Figure 2: *Point cloud reconstruction using RGB-only input (simulated).*

3.1.2 Fusion

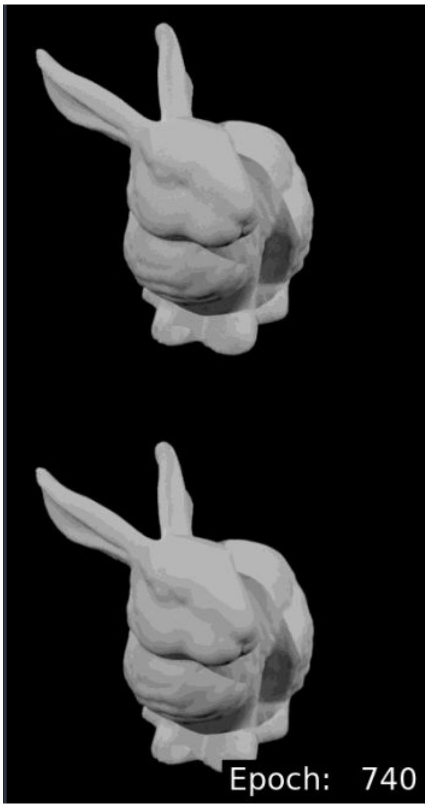

Figure 3: *Rendered scene view from RGB + Sonar fusion (splatting is bottom).*

Geometric scores:

Chamfer Distance: 0.4638771872954217

Hausdorff Distance: 1.4476355190714936

Median Distance: 0.057023677623188475

F-score: 0.06203365005083043

Precision: 0.0329

Recall: 0.5418793033616849

Photometric scores:

SSIM: 0.9832841157913208

PSNR: 42.77613067626953

LPIPS: 0.0637257918715477

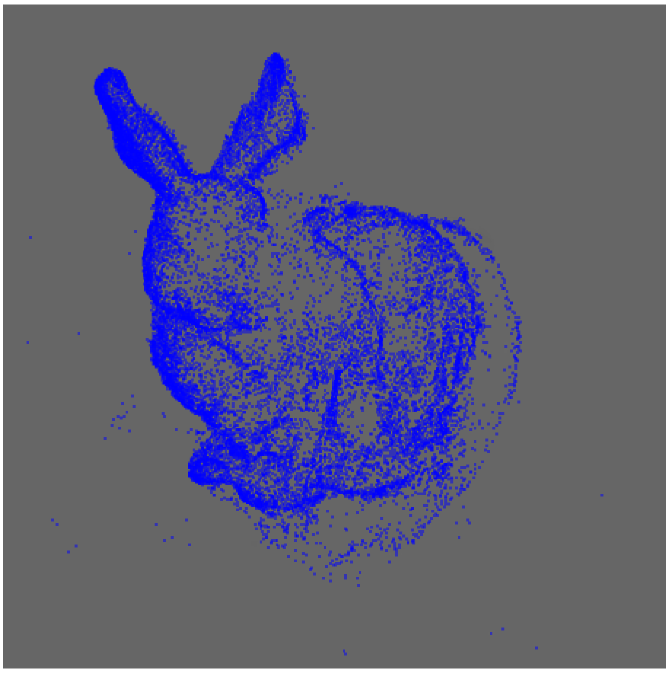

Figure 4: *Point cloud reconstruction using RGB + Sonar fusion (simulated).*

3.2 Field Data Results

3.2.1 RGB Only

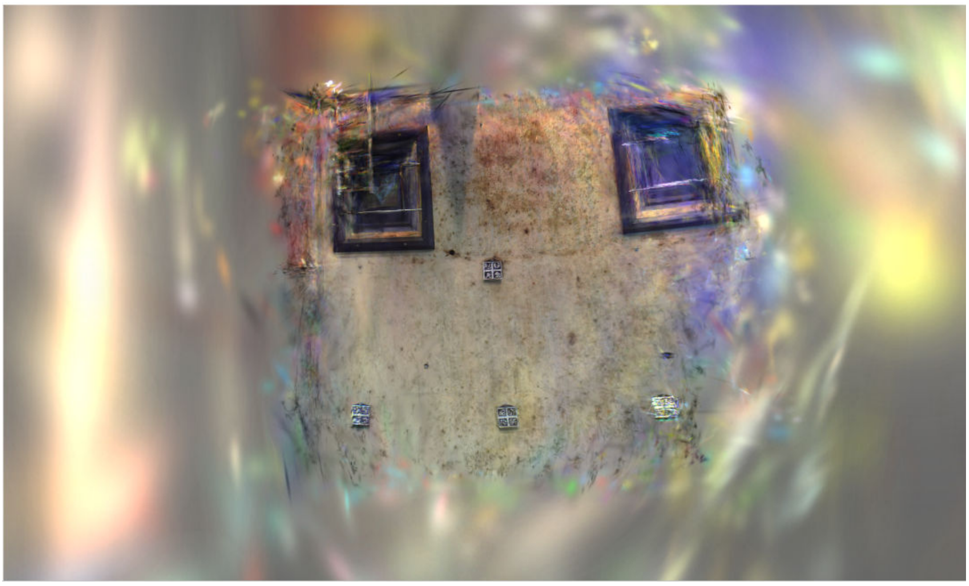

Figure 5: *Rendered scene view from RGB-only reconstruction.*

Photometric scores:

SSIM: 0.9380944967269897

PSNR: 36.92620849609375

LPIPS: 0.20102904736995697

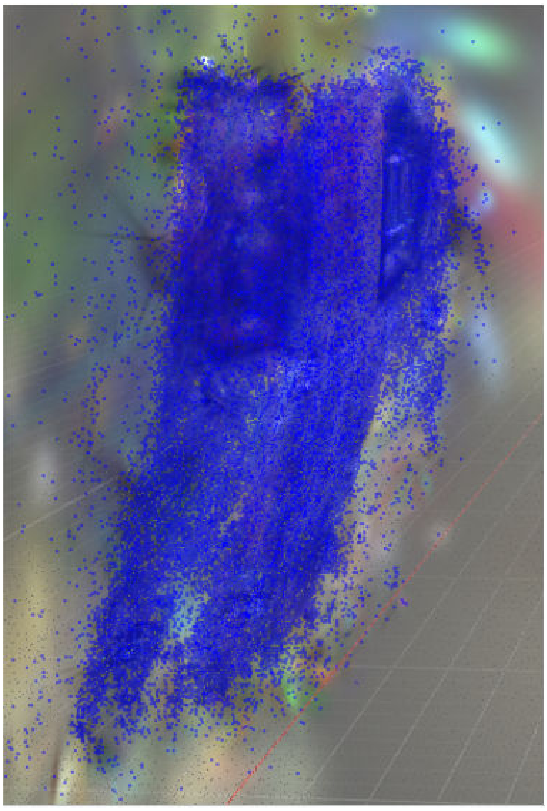

Figure 6: *Point cloud reconstruction using RGB-only input.*

3.2.2 Fusion

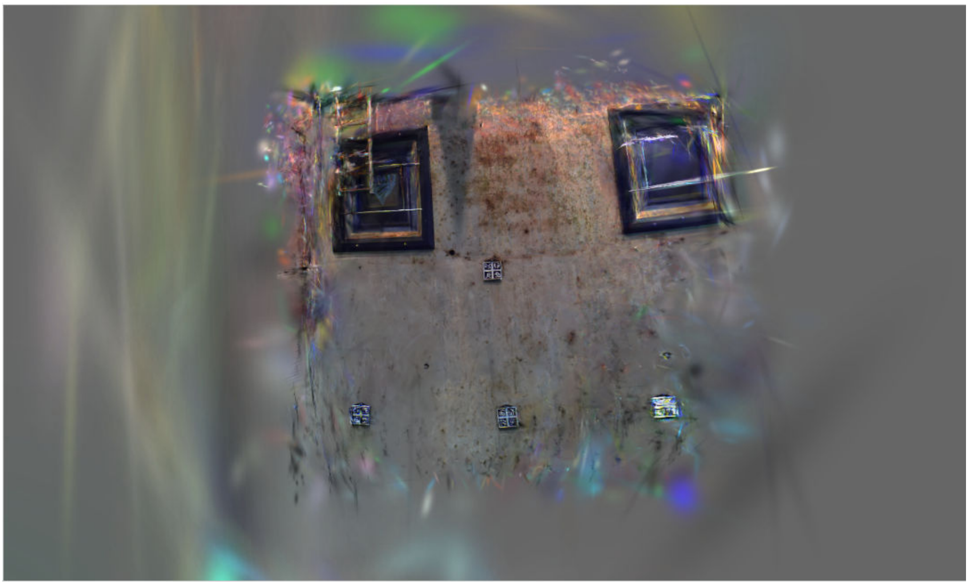

Figure 7: *Rendered scene view from RGB + Sonar fusion.*

Photometric scores:

SSIM: 0.9758889675140381

PSNR: 43.005008697509766

LPIPS: 0.09243087470531464

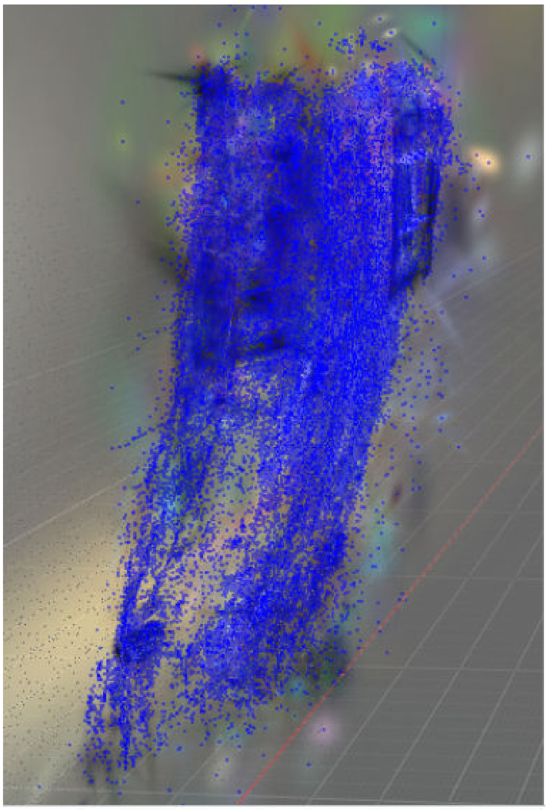

Figure 8: *Point cloud reconstruction using RGB + Sonar fusion.*

Dock Test

RGB

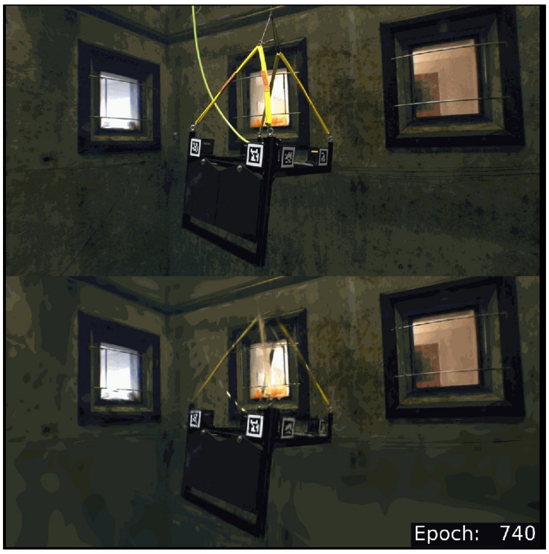

Figure 9: *Rendered scene view of Dock test using RGB-only input.*

Fusion

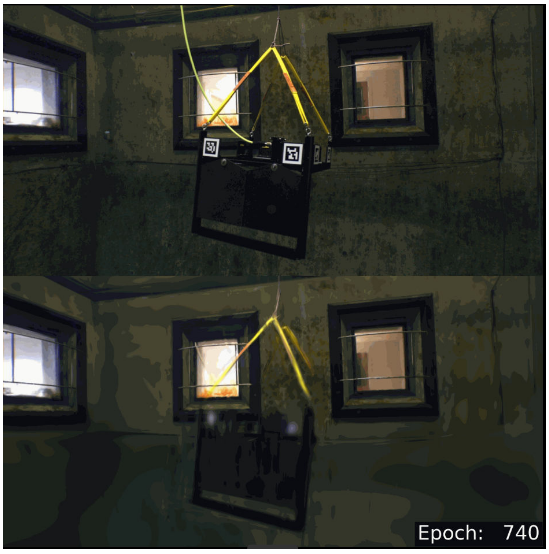

Figure 10: *Rendered scene view of Dock test using RGB + Sonar fusion.*

DISCUSSION

Throughout experiments, it was apparent that including the sonar enabled the pruning of points in space commonly referred to as "floaters." While some of the photometric results were relatively close, empirical results from observation show that there was significant culling of the floating gaussians in 3D space when the reconstructed scene was viewed in a real-time view renderer. This is especially evident in the scene renders from the field test. The renderer used for observations was on the website "superspl.at."

Additionally, final experiments involving object orbits showed that extensive view angles with z-splatting harmed the final result, as it culled portions of the objects in space due to stronger reflections off the test tank walls. In Figures 9 and 10, this can be seen, where the fusion render for the whole object orbit appears to cull some of the object of interest, whereas the RGB-only orbit reconstructs it more accurately.

CONCLUSIONS & RECOMMENDATIONS

In conclusion, Z-splat has proven to be an effective tool in view-constrained data collection scenarios. Its strengths lie in its ability to better estimate depth data than other techniques due to the fusion with sonar. While it excels in cases with limited view angles, it struggles with more comprehensive datasets that encompass broader view angles.

Future work may overcome these object removal issues by exploring background subtraction. There is also existing work that has explored real-time optimization for 3DGS, utilizing depth cameras and relying on the spatial carving strengths of that sensor. This can likely be extended to the marine domain through the use of sonars.

ACKNOWLEDGEMENTS

The MBARI Summer Internship Program is generously supported through a gift from the Dean and Helen Witter Family Fund and the Rentschler Family Fund in memory of former MBARI board member Frank Roberts (1920-2019) and by the David and Lucile Packard Foundation. Additional funding is provided by the Maxwell/Hanrahan Foundation.

REFERENCES

[1] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3d gaussian splatting for real-time radiance field rendering,” *ACM Transactions on Graphics*, vol. 42, no. 4, Jul. 2023, sIGGRAPH 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3592433

[2] Z. Qu, O. Vengurlekar, M. Qadri, K. Zhang, M. Kaess, C. A. Metzler, S. Jayasuriya, and A. Pediredla, “Z-splat: Z-axis gaussian splatting for camera-sonar fusion,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 9, pp. 7255–7267, 2025, special Section on Computational Photography (ICCP 2024). [Online]. Available: https://www.computer.org/csdl/journal/tp/5555/01/10685550/20tj9GlRcje

[3] M. J. Westoby, J. Brasington, N. F. Glasser, M. J. Hambrey, and J. M. Reynolds, “’structure-from-motion’ photogrammetry: A low-cost, effective tool for geoscience applications,” *Geomorphology*, vol. 179, pp. 300–314, 2012. [Online]. Available: https://doi.org/10.1016/j.geomorph.2012.08.021

[4] N. Snavely, S. M. Seitz, and R. Szeliski, “Photo tourism: Exploring photo collections in 3D,” in *ACM SIGGRAPH 2006 Papers*, 2006, pp. 835–846. [Online]. Available: https://dl.acm.org/doi/10.1145/1179352.1141964

[5] 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 *European Conference on Computer Vision (ECCV)*, 2020, arXiv:2003.08934. [Online]. Available: https://arxiv.org/abs/2003.08934

[6] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, “Pixelwise view selection for unstructured multi-view stereo,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. [Online]. Available: https://demuc.de/papers/schoenberger2016mvs.pdf

[7] M. Nimier-David, D. Vicini, T. Zeltner, and W. Jakob, “Mitsuba 2: A retargetable forward and inverse renderer,” *ACM Transactions on Graphics*, vol. 38, no. 6, Dec. 2019, proceedings of SIGGRAPH Asia 2019. [Online]. Available: https://dl.acm.org/doi/10.1145/3355089.3356498

[8] S. C. G. Laboratory, “The stanford 3d scanning repository,” 1996, accessed 2025-09-29. [Online]. Available: https://graphics.stanford.edu/data/3Dscanrep/

[9] J. L. Schönberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 4104–4113.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. [Online]. Available: https://www.cns.nyu.edu/pub/lcv/wang03-preprint.pdf