Journal of Advanced Marine Engineering and Technology, Vol. 49, No. 3, pp. 187~193, 2025
J. Advanced Marine Engineering and Technology (JAMET)
https://doi.org/10.5916/jamet.2025.49.3.187

ISSN 2234-7925 (Print)
ISSN 2765-4796 (Online)
Original Paper

3D reconstruction of underwater objects using NeRF

YeEun Lim[1] · Nakwan Kim[2] · Joohyun Woo†

(Received June 14, 2025 ; Revised June 17, 2025 ; Accepted June 18, 2025)

Abstract: This paper presents a method for reconstructing the 3D shape of underwater objects using Neural Radiance Fields. Underwater reconstruction is often hindered by image degradation caused by environmental noise. To address this, we utilize the model's ability to generate continuous volumetric representations from sparse and incomplete data, selectively training it on low-noise underwater images to reduce distortion effects. While standard view synthesis models offer high-quality and realistic reconstructions, they require long training times, limiting their usability in real-time applications. To overcome this, we employ Instant Neural Graphics Primitives, which significantly reduces training time while maintaining visual fidelity. For validation, we captured image data of the same anchor object in both terrestrial and underwater environments and trained the models separately. Reconstruction quality was assessed using Peak Signal-to-Noise Ratio, Structural Similarity Index Measure, and Learned Perceptual Image Patch Similarity metrics. The terrestrial data achieved an average Peak Signal-to-Noise Ratio of 28.51, Structural Similarity Index Measure of 0.97, and Learned Perceptual Image Patch Similarity of 0.07, while underwater data yielded 23.23, 0.71, and 0.15, respectively. Despite the inherent challenges in underwater imaging, the results demonstrate that the proposed method can achieve natural and reliable 3D reconstructions in both settings.

Keywords: Neural radiance fields, Instant neural graphics primitives, View synthesis, Underwater 3D reconstruction

1. Introduction

In recent years, the application of advanced technologies in underwater environments has gained significant momentum, driving innovation across various fields such as marine science, underwater robotics, and ocean resource exploration. To overcome the complex conditions inherent in underwater environments, new technical approaches have been emphasized to enable more accurate data acquisition, thereby enhancing the efficiency of underwater exploration and monitoring. Among these advancements, 3D shape reconstruction of underwater objects has emerged as a key area of interest, as it enables more precise analysis of marine environments. However, accurately reconstructing 3D shapes underwater presents numerous challenges due to environmental factors such as light absorption and scattering, motion blur, color distortion, and differences in perceived distance compared to in-air environments. These issues degrade visibility and introduce data distortion, making it difficult to obtain high-quality training data, which ultimately reduces reconstruction reliability **[1]**.

Conventional methods have relied heavily on sonar-based sensing and traditional voxel or mesh-based representations **[2]-[4]**. Camera-based approaches also require complex preprocessing due to sensor configurations and underwater noise, often resulting in increased cost and time **[5]-[7]**.

To address these limitations, recent studies have explored deep learning techniques for underwater image modeling and rendering **[8]-[10]**. While GAN-based models have shown promise in removing noise from low-quality images, they face limitations in generating high-resolution images and often suffer from unstable training.

This study proposes a novel approach using Neural Radiance Fields (NeRF), a view synthesis model capable of generating novel views from images captured at discrete viewpoints **[11]**. We trained the model using selectively chosen underwater images with minimal noise, which helps compensate for the limited availability of high-quality data. By extracting spatial

† Corresponding Author (ORCID: http://orcid.org/0000-0002-9495-3245): Professor, Division of Naval Architecture and Ocean Systems Engineering, Korea Maritime & Ocean University, 727, Taejong-ro, Yeongdo-gu, Busan 49112, Korea, E-mail: jhwoo@kmou.ac.kr, Tel: +82-410-4306

1 Researcher, Ocean and Maritime Digital Technology Research Division, Korea Research Institute of Ships & Ocean Engineering (KRISO), E-mail: mimok@kriso.re.kr, Tel: +82-42-866-3613

2 Research Professor, Future Innovation Institute, Seoul National University, E-mail: nwkim@snu.ac.kr, Tel: +82-31-5176-2363

relationships between images, NeRF generalizes well across unseen views, making it effective for underwater scenarios.

Despite NeRF's ability to produce sharp and realistic 3D reconstructions, its long training time poses a barrier for real-time applications. To overcome this, we employ Instant NeRF, which significantly reduces training time to within a few minutes while maintaining reconstruction quality **[12]**.

For evaluation, we captured monocular image data of the same anchor object in both underwater and terrestrial environments and performed 3D reconstruction. The results were quantitatively assessed using PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index Measure), and LPIPS (Learned Perceptual Image Patch Similarity) by comparing reconstructed views with original images from the same viewpoint.

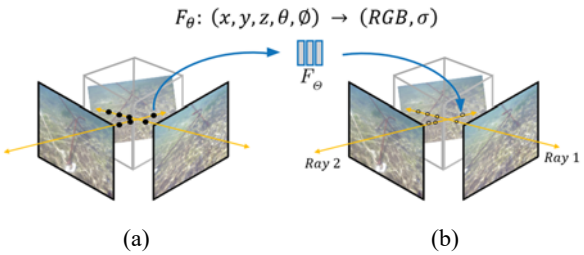
This study presents an efficient and reliable real-time 3D reconstruction method for underwater applications. The proposed approach holds strong potential for enhancing the perceptual capabilities of remotely operated underwater vehicles (ROVs), and for supporting various underwater inspection and maintenance tasks.

2. Network Structure

2.1 NeRF (Neural Radiance Fields)

In this study, the shape reconstruction technique employed is Neural Radiance Fields (NeRF), a neural network-based method that learns color information and volume density representing the interaction between objects and light in 3D space. A key feature of NeRF is its ability to take a small set of multi-view images as input and generate novel views of the object from unseen angles, effectively reconstructing a 3D shape. As illustrated in **Figure 1**, NeRF estimates the color and volume density at any point in 3D space based on the input 3D position (x, y, z) and the viewing direction (θ, ϕ) of the camera. Volume density is defined as the inverse of transparency, where higher values indicate more opacity and lower values indicate greater transparency. To generate the training data, NeRF uses rays that pass-through pixels in the input images, which are cast into the 3D scene.

As shown in **Equation (1)**, a ray is defined as a set of points extending from the camera's focal origin (o) in a specific direction (d), scaled by a distance (t). One ray is generated per pixel, and each ray is uniformly divided into n segments. Sample points along each ray are selected as input, and the 3D position and viewing direction at each point are used for training.



(a) (b)

Figure 1: An overview of neural radiance field scene representation: (a) Sampling 5D coordinates (spatial location and viewing direction) from input images; (b) Feeding the sampled coordinates into an MLP to estimate color and volume density.

$$r(t) = o + td \tag{1}$$

Once the input data is generated through rays, the shape reconstruction is learned using an MLP (Multi-Layer Perceptron). As illustrated in **Figure 2**, the MLP consists of a total of nine fully connected (FC) layers with ReLU (Rectified Linear Unit) activation functions. ReLU is a nonlinear function that outputs zero for negative inputs and returns the input value itself for positive inputs. It introduces nonlinearity into the network, enabling the learning of more complex patterns. Due to its simplicity, ReLU offers high computational efficiency and helps mitigate the vanishing gradient problem, thereby accelerating the training process.

During training, the spatial position information is first passed through eight FC layers to predict the volume density. Then, using the viewing direction, the color value is estimated. This separation accounts for non-Lambertian effects, where an object's appearance varies depending on the viewing angle. As shown in **Equation (2)**, each point along a ray contributes to the final pixel color based on its volume density: points with higher density are assigned higher weights, while those with lower density receive lower weights. These weights are used to modulate the color values, which are then aggregated to determine the final pixel value.

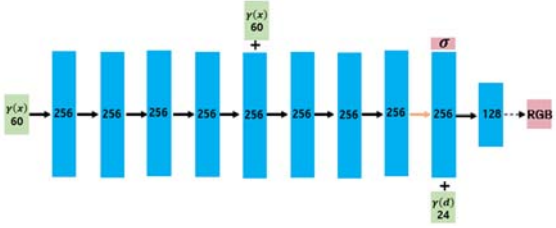


Figure 2: Network architecture of NeRF: Volume density is predicted using spatial location information, followed by color estimation based on viewing direction.

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma\big(r(t)\big)c(r(t), d)dt$$

$$T(t) = \exp\left(-\int_{t_n}^{t_f} \sigma\big(r(s)\big)ds\right) \quad (2)$$

Each ray is uniformly divided into n segments, and specific points along the ray are selected as input data. These sample points may lie within the object but can also fall into empty space where no object exists. Training the model with more focus on regions where objects are present, rather than on empty space, can lead to more accurate results. To address this, NeRF employs a two-stage training strategy as shown in **Figure 3**: a coarse network first processes the full ray uniformly, and then a fine network refines the learning by focusing on regions with high volume density inferred from the coarse stage.

The loss function is defined in **Equation (3)**. For both the coarse and fine networks, the rendered colors are compared with the ground truth pixel colors, and the respective losses are computed. The final loss is then obtained by summing the losses from both stages.

However, training the network directly with the 5D input (x, y, z, θ, ϕ) — comprising spatial location and viewing direction—can be challenging due to the limited amount of input data. To address this issue, positional encoding is employed. As shown in **Equation (4)**, this technique transforms the spatial location into a 60-dimensional vector and the viewing direction into a 24-dimensional vector for training.

$$L = \sum_{r\in R}\left[\left\|\hat{C}_c(r) - C(r)\right\|_2^2 + \left\|\hat{C}_f(r) - C(r)\right\|_2^2\right] \quad (3)$$

$$\gamma(p) = (\sin(2^0\pi p), \cos(2^0\pi p), \ldots, \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p)) \quad (4)$$

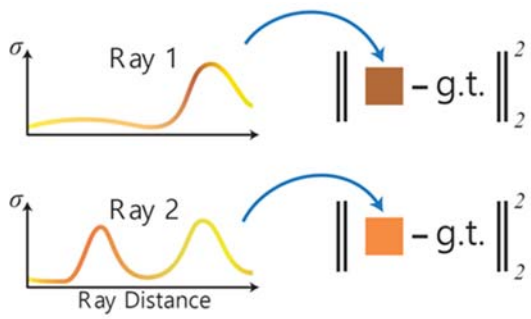


Figure 3: Two-stage training: Initial sampling along the full ray, followed by fine sampling in high-density regions; losses from both stages are computed and combined.

sampled point along the ray, the model learns to estimate its color and volume density. In other words, a single ray determines the color of a single pixel. By aggregating the results of all rays, a complete 3D shape is reconstructed.

2.2 Instant NGP(Neural Graphics Primitives)

Unlike the original NeRF, which uses a positional encoding scheme, Instant-NGP adopts a Multi-Resolution Hash Encoding approach to significantly accelerate training. While NeRF is suitable for generating high-quality 3D models—such as converting object image data into detailed 3D models for inspection and analysis—Instant-NGP is better suited for real-time 3D model generation. For instance, it is ideal for scenarios where the 3D structure of an object must be visualized instantly to support rapid decision-making or planning. Instant-NGP achieves over 1,000 times faster training speed compared to the original NeRF. Furthermore, it supports transfer learning by utilizing pre-trained data, allowing effective training even with a smaller number of input images. To leverage these advantages, this study adopts Instant-NGP—one of the NeRF-based approaches—for 3D shape reconstruction of underwater data.

Instant-NGP reduces training time by using Multi-Resolution Hash Encoding, which divides the 3D scene into grids of varying resolution depending on its spatial complexity. As illustrated in **Figure 4**, each vertex of the grid cells that contain a given point has a learnable feature vector of dimension F. These feature vectors are stored in a hash table, and each vertex coordinate is hashed to retrieve its corresponding feature vector. The vectors from the surrounding grid vertices are then linearly interpolated to generate an encoded input, which is passed into the neural network.

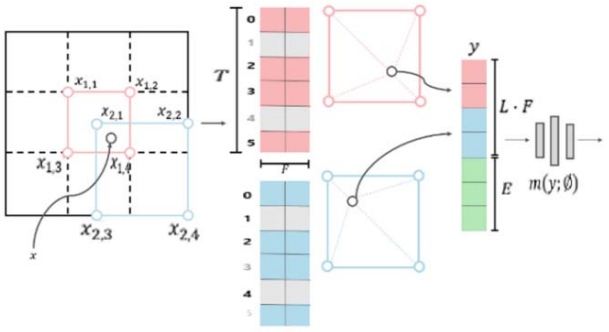


Figure 4: Input generation process: Feature vectors are retrieved by hashing grid vertex coordinates around a location, linearly interpolated, and combined with view direction to form the network input.

Additionally, the viewing direction of the camera is provided as a supplementary input. Thus, the final input to the network consists of the spatial position and viewing direction, while the output includes the volume density and radiance (color) of each pixel.

3. 3D Reconstruction

3.1 Data Acquisition

In this study, data were collected using a monocular camera after placing the same anchor, shown in **Figure 5**, in both underwater and terrestrial environments. The anchor measures 51 cm in width and 68 cm in height. Underwater data were acquired using a remotely operated underwater vehicle (ROV) in coastal waters near Korea Maritime & Ocean University. **Figure 6** (Top) shows the area where the underwater data were collected, and **Figure 6** (Bottom) presents a photograph of the actual sea site. Terrestrial data were collected indoors under controlled conditions.

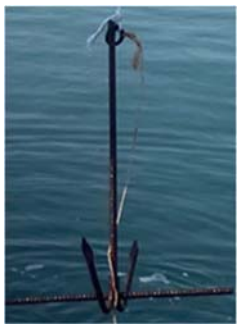


Figure 5: Anchor used as the target object for 3D reconstruction in both underwater and terrestrial environments.

Figure 6: Underwater anchor data acquisition at Korea Maritime and Ocean University: (Top) Data acquisition area; (Bottom) Photograph of the actual marine environment.

Figure 7: Remotely operated underwater robot used for data acquisition

Table 1: Specifications of the Blue Robotics BlueROV2

Parameter	Measurement
Length	0.45 m
Width	0.33 m
Height	0.25 m
Weight in air	12 kg

The remotely operated underwater vehicle (ROV) used for data acquisition was BlueROV2 from Blue Robotics, as shown in **Figure 7**, and its specifications are listed in **Table 1**. The ROV was equipped with six T200 thrusters, also from Blue Robotics. Additional onboard sensors included a monocular camera, a barometer, a 3-degree-of-freedom gyroscope, and pressure, depth, and temperature sensors. However, this study utilized only the monocular camera data for the 3D reconstruction process.

3.2 Training and Performance Evaluation

Underwater training data of the anchor were acquired using a monocular camera mounted on the remotely operated underwater vehicle. Video footage was recorded at a depth of approximately 1.2 meters, from which 190 frames were extracted and used as training images. Using the underwater dataset, Instant-NGP was trained. The top image in **Figure 8** shows the original input at a specific time step, while the bottom image shows the reconstruction result at the same viewpoint.

To quantitatively evaluate the underwater reconstruction results, the same anchor was placed in an on-land environment. The terrestrial dataset was then used to train Instant-NGP, following the same procedure as the underwater case. The top image in **Figure 9** shows the original input at a specific time step, while the bottom image shows the corresponding reconstruction result at the same viewpoint.

A purely visual comparison between the underwater and terrestrial NeRF reconstruction results has inherent limitations. Therefore, this study conducted a quantitative evaluation using

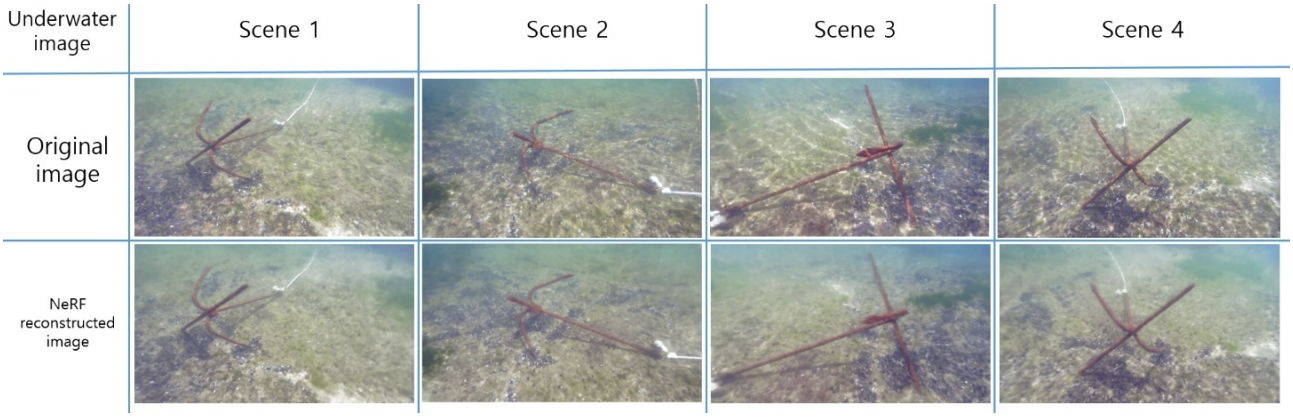


Figure 8: Comparison between original underwater images and NeRF reconstruction results for four scenes using Instant-NGP

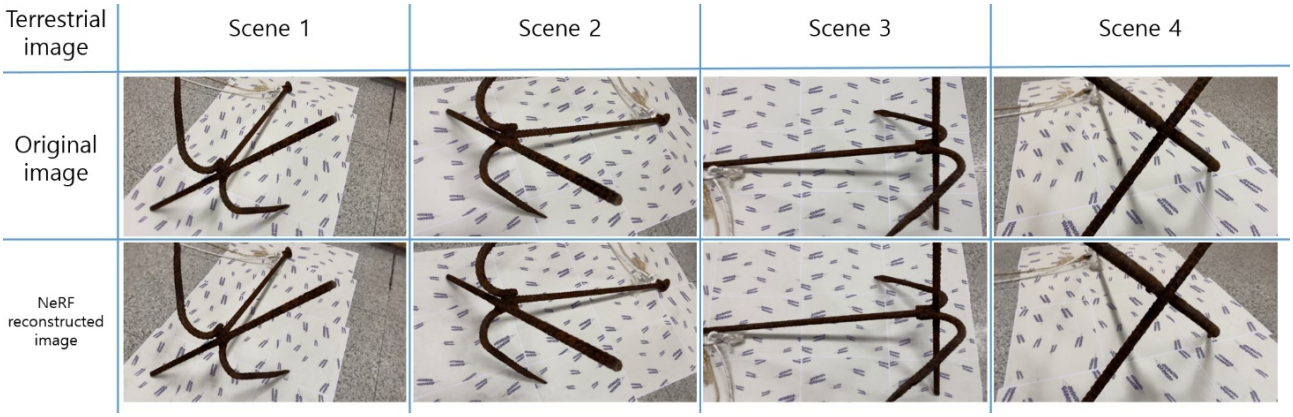


Figure 9: Comparison between original terrestrial images and NeRF reconstruction results for four scenes using Instant-NGP

Table 2: Quantitative comparison of reconstruction performance on terrestrial and underwater images

Terrestrial Image	PSNR	SSIM	LPIPS	Underwater Image	PSNR	SSIM	LPIPS
Scene 1	26.16	0.84	0.09	Scene 1	22.35	0.69	0.16
Scene 2	27.46	0.86	0.1	Scene 2	22.58	0.73	0.14
Scene 3	28.66	0.9	0.07	Scene 3	23.18	0.7	0.16
Scene 4	29.37	0.92	0.06	Scene 4	23.82	0.71	0.15
Scene 5	30.66	0.94	0.03	Scene 5	23.01	0.72	0.16
Scene 6	28.29	0.91	0.06	Scene 6	23.19	0.71	0.16
Scene 7	28.77	0.92	0.06	Scene 7	23.56	0.71	0.14
Scene 8	29.11	0.91	0.07	Scene 8	23.3	0.69	0.17
Scene 9	29.57	0.92	0.05	Scene 9	23.31	0.71	0.13
Scene 10	27.19	0.83	0.1	Scene 10	24.47	0.73	0.11
Scene 11	29.98	0.91	0.07	Scene 11	23.13	0.71	0.14
Scene 12	27.94	0.92	0.06	Scene 12	22.86	0.7	0.16
Average	**28.51**	**0.97**	**0.07**	**Average**	**23.23**	**0.71**	**0.15**

objective performance metrics. Inspired by the original NeRF paper, three commonly used evaluation metrics were adopted to assess the similarity between images: PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index Measure), and LPIPS (Learned Perceptual Image Patch Similarity). These metrics were used to compare the original and reconstructed images at the same viewpoints, as illustrated in **Figure 8** and **Figure 9.**

PSNR is a traditional metric used to assess the quality degradation in reconstructed or compressed images, expressed in decibels (dB). It typically ranges from 0 to above 30, where higher values indicate better reconstruction fidelity. Values above 30 dB suggest differences that are hardly perceptible to the human eye.

SSIM improves upon PSNR by considering luminance, contrast, and structural similarity between images, ranging from 0 to 1. A higher SSIM value indicates greater structural similarity. LPIPS, on the other hand, is based on human visual perception. It computes feature similarity between image patches using a pre-trained VGG network. LPIPS ranges from 1 to 0, where lower values represent better perceptual similarity.

Twelve scenes were evaluated in both underwater and terrestrial environments, with the results summarized in **Table 2**. On average, the terrestrial data achieved better reconstruction performance across all metrics: PSNR (28.51), SSIM (0.97), and LPIPS (0.07), compared to the underwater data with PSNR (23.23), SSIM (0.71), and LPIPS (0.15). These results suggest that NeRF achieves higher reconstruction quality in terrestrial environments. This discrepancy is likely due to the physical characteristics of underwater imaging. In underwater conditions, scattering and absorption of light often degrade image sharpness and color information. Additionally, light scattering varies over time due to water movement, which violates NeRF's assumption that the scene remains static during capture, potentially introducing temporal inconsistency errors.

The degradation in reconstruction accuracy can also be observed through artifacts present in the underwater results. As shown in **Figure 10**, the fog-like region above the anchor is an artifact generated during NeRF reconstruction. Artifacts refer to visually distorted shapes or colors that do not exist in the original input images but are erroneously produced by the model. These tend to appear more frequently when image resolution is low or when viewpoint sampling is sparse. In the underwater case, the limited visibility makes object boundaries less clear and distorts light paths, disrupting the ray continuity assumed during NeRF training. This, in turn, reduces the model's ability to estimate accurate volume densities and colors. In contrast, terrestrial conditions offer sharper object boundaries and more consistent lighting, which likely contributed to more stable and reliable reconstruction performance.

Figure 10: Artifacts observed in underwater images reconstructed using NeRF

4. Conclusion

This study proposed a method for reconstructing the 3D shape of underwater objects using Neural Radiance Fields (NeRF). The approach enables not only reconstruction from the original training viewpoints but also from novel viewpoints not included in the training data. To validate the reconstruction performance, the same object was also trained and evaluated in a terrestrial environment for comparison. While visual inspection showed clear and realistic reconstructions in both cases, quantitative evaluation was conducted using three commonly used metrics: PSNR, SSIM, and LPIPS. The terrestrial dataset yielded average values of PSNR 28.51, SSIM 0.97, and LPIPS 0.07, whereas the underwater dataset achieved PSNR 23.23, SSIM 0.71, and LPIPS 0.15, confirming that the reconstruction quality in terrestrial conditions was superior.

The performance gap is attributed to the unique challenges of underwater environments. Factors such as suspended particles and surface ripples cause light scattering and absorption, resulting in varying appearances of the same object over time. This violates NeRF's core assumption of a static scene, thereby degrading the model's learning accuracy.

Through this validation study, it was confirmed that applying standard NeRF methods directly to underwater environments has limitations. These findings highlight the need for customized 3D reconstruction approaches that account for the physical characteristics of underwater settings. Future research should explore advanced preprocessing techniques, domain-specific adaptations, and multi-sensor fusion to develop more robust underwater 3D reconstruction systems.

Acknowledgement

This research was supported by Korea Institute of Marine Science & Technology Promotion(KIMST) funded by the Ministry of Ocean and Fisheries, Korea(RS-2024-00432366), Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (RS-2021-KI002493,The Competency Development Program for Industry Specialist), "Regional

Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE) (2023RIS-007).

Author Contributions

Conceptualization, Y. E. Lim and J. H. Woo; Methodology, Y. E. Lim and J. H. Woo; Software, Y. E. Lim; Formal Analysis, Y. E. Lim and J. H. Woo; Investigation, Y. E. Lim; Resources, N. W. Kim and J. H. Woo; Data Curation Y. E. Lim; Writing-Original Draft Preparation, Y. E. Lim; Writing-Review & Editing, Y. E. Lim and J. H. Woo; Visualization, Y. E. Lim; Supervision, J. H. Woo; Project Administration, J. H. Woo; Funding Acquisition, N. W. Kim and J. H. Woo.

References

[1] A. V. Sethuraman, M. S. Ramanagopal, and K. A. Skinner, "WaterNeRF: Neural radiance fields for underwater scenes," OCEANS 2023 – MTS/IEEE US Gulf Coast, pp. 1–7, 2023.

[2] T. Guerneve, K. Subr, and Y. Petillot, "Three-dimensional reconstruction of underwater objects using wide-aperture imaging SONAR," Journal of Field Robotics, vol. 35, no. 6, pp. 890–905, 2018.

[3] Z. Li, R. Lan, Z. Chen, X. Luo, J. Li, Z. Huang, and L. Qian, "Underwater high-precision panoramic 3D image generation," 2020 8th International Conference on Digital Home (ICDH), pp. 39–44, 2020.

[4] I. P. Maurell, M. M. dos Santos, P. J. D. de Oliveira Evald, B. H. Justo, J. Arigony-Neto, A. W. Vieira, and P. L. Drews, "Volume change estimation of underwater structures using 2- D sonar data," IEEE Sensors Journal, vol. 22, no. 23, pp. 23380–23392, 2022.

[5] J. -W. Bae, D. -H. Seo, and J. -H. Seong, "Two-dimensional camera and TOF sensor-based volume measurement system for automated object volume measuring," Journal of Advanced Marine Engineering and Technology (JAMET), vol. 47, no. 6, pp. 419-426, 2023.

[5] A. El Saer, C. Stentoumis, I. Kalisperakis, L. Grammatikopoulos, P. Nomikou, and O. Vlasopoulos, "3D reconstruction and mesh optimization of underwater spaces for virtual reality," the International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences, vol. 43, pp. 949–956, 2020.

[6] H. Song, L. Chang, Z. Chen, and P. Ren, "Enhancement-registration-homogenization (ERH): A comprehensive underwater visual reconstruction paradigm," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 6953–6967, 2022.

[7] S. Bi, Z. Xu, P. Srinivasan, B. Mildenhall, K. Sunkavalli, M. Hašan, and R. Ramamoorthi, "Neural reflectance fields for appearance acquisition," arXiv preprint arXiv:2008.03824, 2020. [Online]. Available: https://arxiv.org/abs/2008.03824.

[8] P. P. Srinivasan, B. Deng, X. Zhang, M. Tancik, B. Mildenhall, and J. T. Barron, "NeRV: Neural reflectance and visibility fields for relighting and view synthesis," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7495–7500, 2021.

[9] D. Derksen and D. Izzo, "Shadow neural radiance fields for multi-view satellite photogrammetry," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1152–1161, 2021.