A Feasibility Study on Novel View Synthesis of Underwater Structures using Neural Radiance Fields

Yuen Min Too[a], Hari Vishnu[a], Mandar Chitre[a,b], Bharath Kalyan[a], Luyuan Peng[a,b], Rajat Mishra[a]
[a] *Acoustic Research Laboratory, Tropical Marine Science Institute, National University of Singapore*
[b] *Department of Electrical & Computer Engineering, National University of Singapore*

***Abstract*—We explore Neural Radiance Fields (NeRFs) for synthesizing novel views of underwater structures. This learning-based approach relies on a sparse set of camera views to model the 3D geometry of underwater structures and scenes. Real-world underwater scenes exhibit significant temporal variations, introducing challenges in maintaining visual consistency. We investigate three NeRF implementations: 1) nerfacto, which represents the baseline; 2) nerfacto with transient embeddings and 3) nerfacto with a robust loss, which are designed to deal with scene inconsistencies. We evaluate these implementations using datasets collected in 1) a controlled environment and 2) a real underwater setting. The modified implementations consistently outperform the original nerfacto across both datasets. The performance improvements are particularly pronounced in the dataset obtained from the real underwater setting where scene inconsistencies are more prevalent. This underscores the importance of robustifying NeRF implementations to ensure consistent performance in the challenging underwater environments.**

I. Introduction

Significant effort is expended to monitor and maintain underwater structures in freshwater or oceanic environments, both natural and man-made. Natural underwater structures such as coral reefs and underwater caves provide important habitats for marine life and play an important role in the ocean ecosystem. Man-made underwater structures such as offshore oil and gas platforms and submarine cables are essential for the global economy and have a crucial impact on human life. Mapping of these structures using photogrammetry is one of the underpinning technologies enabling accurate and high resolution inspections, such as identifying damage and tracking changes in the structures over time [1], [2], [3]. Traditional photogrammetry constructs a 3D model of a target structure by deriving measurements from a dense set of overlapping images. The 3D model explicitly stores color, reflectivity, texture, and other properties of the structure in the discrete 3D space. Projecting the discrete representations along camera rays onto a 2D plane enables the synthesis of a novel view of the structure. However, for realistic and complex scenes, such a model can consume a significant amount of memory to produce a representation with sufficient fidelity. This is especially problematic on vehicular platforms, such as remotely operated vehicles and autonomous underwater vehicles due to limited memory and computational power.

Neural radiance fields (NeRFs) offer a computationally efficient alternative for synthesizing novel views of scenes. NeRFs store and represent the scene in terms of a continuous volumetric scene function that typically represents the color and density of light at every point in space, and can be trained using a sparse set of images [4]. Novel views can be synthesized by querying the function along camera rays and rendering the corresponding images using classical volume rendering techniques. Despite its potential as a powerful tool for synthesizing photorealistic images, NeRFs may sometimes perform poorly on some types of real-world scenes due to different issues. The main issue that can affect NeRFs is inconsistencies in the different input views of a scene [5]. This creates significant challenges in underwater environments, where dynamic changes in the scenes due to movement of objects such as plankton, sediment, algae, fish, and variable illumination due to lighting changes are particularly common. These factors may lead to inaccuracies in NeRF-based novel view synthesis.

In this study, we investigated the feasibility of novel view synthesis for underwater structures using NeRFs. To collect data for the study, we conducted an experiment in a large state-of-the-art Deepwater Ocean Basin (DOB) at the Technology Centre for Ocean and Marine, Singapore [6] in which an underwater vehicle was made to inspect an underwater structure while capturing images of it. The experiment also had transient objects in the scene and variations in illumination. Subsequently, we trained different models of the scene using improved NeRF variants with the acquired dataset, and assessed the novel view synthesis performance of each model. For further investigation in a realistic environment, we also trained and evaluated the models using a real-world dataset obtained from a marine environment.

Section II of this paper briefly introduces the formulation of NeRFs and their different flavors. Section III describes the experiments and data used by us to assess NeRFs in this paper, and the results. Section IV concludes the paper.

II. Neural Radiance Fields

Let $\mathbf{p} = [x, y, z]$ be a 3D point, $\mathbf{d} = [d_x, d_y, d_z]$ be a unit-norm camera viewing direction, $\mathbf{c} = [r, g, b]$ be color in red green and blue, and σ be a density. NeRFs leverage multilayer perceptrons (MLPs) to map ($\mathbf{p}$, $\mathbf{d}$) to ($\mathbf{c}$, σ). By aggregating colors and densities along a camera ray, denoted by $\mathbf{r}$, through a pixel on the camera plane, the model predicts the color of the pixel, represented by $\hat{\mathbf{C}}(\mathbf{r})$. The model is typically trained by minimizing an L2 reconstruction loss [4]:

$$\mathcal{L} = \|\hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|_2^2 \tag{1}$$

where $\mathbf{C}(\mathbf{r})$ is the observed pixel color of ray $\mathbf{r}$ from an input image. As a baseline model for our comparative study, we employed nerfacto, which is a modular NeRF implementation that adopts recent advancements to improve computational efficiency and handle unbounded scenes [7]. Figure 1 illustrates the nerfacto model structure. Nerfacto integrates per-image appearance embeddings to effectively address the impact of diverse lighting conditions. Each appearance embedding, which is denoted by l^{a}, is a trainable real-valued vector of length n^{a}.

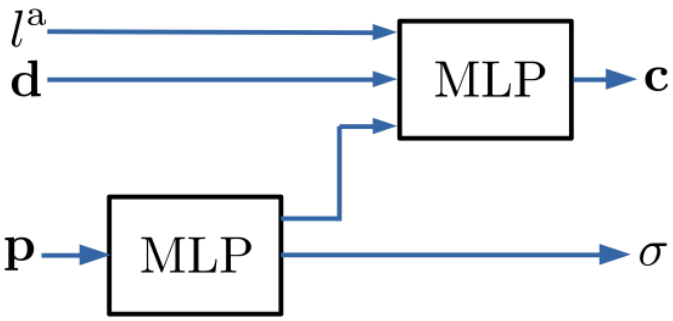

Fig. 1. Model structure of nerfacto.

To make nerfacto more robust to image inconsistencies, we explored two of its variants, namely nerfacto with transient embeddings [8] and nerfacto with a robust loss function [9]. While the model structure of nerfacto with transient embeddings is similar to NeRF in the Wild (NeRF-W) [8], as shown in Fig. 2, the processing pipeline of nerfacto with transient embeddings is rooted in the framework of nerfacto. Each transient embedding, which is denoted by l^{t}, is a trainable real-valued vector of length n^{t}. The transient head emits a field of uncertainty, denoted as β, enabling the model to adaptively adjust its reconstruction loss function by ignoring pixels and 3D points that are likely to involve occluders. The color of the pixel is calculated by aggregating not only the static components ($\mathbf{c}$, σ) but also the transient components ($\mathbf{c}^{\mathrm{t}}$, σ^{t}). The loss function of nerfacto with transient embeddings is written as

$$\mathcal{L}^{\mathrm{te}} = \frac{\|\hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|_2^2}{2\beta(\mathbf{r})^2} + \log(\beta(\mathbf{r})) + \lambda_u g(\mathbf{r}) \qquad (2)$$

where $\beta(\mathbf{r})$ is obtained by accumulating β and σ^{t} along $\mathbf{r}$, $g(\mathbf{r})$ represents the average of σ^{t} along r, weighted by a non-negative scalar denoted as λ_u. In (2), the second term is to balance the reconstruction loss and the third term with a multiplier λ_u enforces sparsity on the transient density.

Instead of incorporating transient embeddings into the nerfacto model structure for representing transient objects in the scene, nerfacto with a robust loss merely replaces (1) with

$$\mathcal{L}^{\mathrm{rl}} = w(\mathbf{r})\|\hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|_2^2 \qquad (3)$$

where $w(\mathbf{r})$ is a binary weight function of $\mathbf{r}$ [9]. The robust loss function, characterized as a squared sum of trimmed entries, automatically distinguishes inconsistent image regions

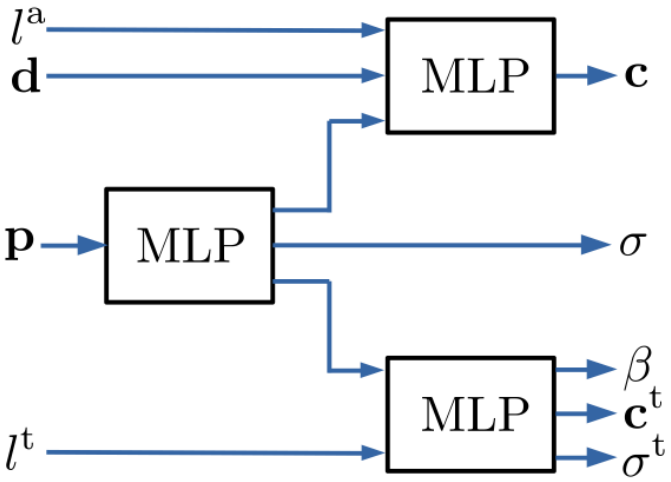

Fig. 2. Model structure of nerfacto with transient embeddings.

and treats them as optimization outliers during the training process which can be neglected. Specifically, the weight function $w(\mathbf{r})$ dynamically adjusts during model fitting, facilitating the rapid learning of fine-grained image details that are not considered outliers. It categorizes pixels along the rays as either inliers or outliers, guided by an inductive bias towards the smoothness of the outlier process. The weight function captures the spatial smoothness of the outlier process, taking into account that inliers generally exhibit large magnitude residuals but limited spatial extent, whereas outliers tend to have weaker residuals but extend over larger spatial regions. It is essential to note that this assumption may not hold true for small transient objects, such as marine snow in underwater scenes.

III. Experiments and Results

We developed a hybrid underwater vehicle named Hydra, equipped with a wide-angle, low-light camera using commercially available, off-the-shelf components. The vehicle was deployed within the DOB, and navigated around a bottom-mounted rectangular structure constructed from steel drums, with the camera directed towards the structure. It captured a series of images during the run emulating a real-world inspection mission. During the mission, there were occasional transient objects picked up at the scene apart from the structure. For example, there were instances where the tether connecting the vehicle to the top-side controller was incidentally captured on the camera (see Fig. 3, row 3, column 1), which is a challenge that may occur in a regular underwater inspection mission. Additionally, to study the robustness of NeRFs to variable illumination conditions as is often encountered in real-world scenarios, the DOB lighting was dimmed during some of the runs, and we sporadically toggled the vehicle's own lighting on and off.

To evaluate the performance of nerfacto and its variants, we conducted training and testing on novel view synthesis tasks, focusing on the underwater structure during the in-

spection mission. The NeRF models were either built upon or adapted from nerfstudio version 1.0.0 [7]. We used the hyperparameters $n^{\mathrm{a}} = 32$, $n^{\mathrm{t}} = 32$, $\lambda_u = 0.1$ to train the models, respectively. The full set of DOB data, consisting of 899 images, underwent a division into training and test sets using a $90 - 10$ split. The estimation of camera poses for the images was accomplished using COLMAP, a structure-from-motion package. [10], [11]. Number of rays per batch used for training iterations is 16384. The training set was used to train the NeRFs while the test set was used for performance evaluation.

Fig. 3 illustrates the novel view synthesis performance of the models on the test dataset collected at the DOB. The first column contains ground-truth images captured by the camera, while the subsequent columns display images synthesized by different NeRF models from the same camera viewpoints. We evaluated the models across four scenarios:

- vehicle light on,
- vehicle light off,
- vehicle light off with the tether present,
- vehicle light on with the tether present,

presented respectively in the four rows in the figure. All the models demonstrate robust performance under varying lighting conditions. In the third and last scenario, nerfacto exhibits artifacts in the image when the vehicle tether is present. These artifacts likely stem from the presence of the tether, as it was incidentally captured in some training images taken by the vehicle's camera in close proximity to the pose corresponding to the capture of the third and last scenario's image. In contrast, the models equipped with transient embeddings or a robust loss function outperform nerfacto by consistently generating photorealistic images without the artifact for all four test scenarios. For quantitative analysis, we compared the camera images with the synthesized images based on PSNR, MS-SSIM [12], and LPIPS [13] as summarized in Table I. Given that camera images may include variable lighting and transient objects that might not be present in the synthesized counterparts, we took several preprocessing steps. Firstly, we standardized each synthesized image based on the mean and standard deviation of the corresponding camera image. Subsequently, to remove any transient objects present in the camera image, we applied a manually labeled binary mask on both images. This approach mitigates the potential differences in lighting and transient elements between the two sets of images. The resulting modification to the nerfacto showcases a slight enhanced image quality compared to the original nerfacto. The improvement is marginal as inconsistencies in the scene represent only a small fraction of the overall training data.

It is worth noting that the DOB represents a controlled environment with clear water. In real underwater settings, conditions tend to be murkier with the presence of suspended particles in the water, often referred to as marine snow. We applied the same preprocessing, training and evaluation procedures to the Torpedo Boat Wreck (TBW) dataset, comprising publicly available images obtained from the Hybrid Remotely Operated Vehicle survey of the wreck of a torpedo boat [14], [15]. The TBW data consists of 442 images, with 90% allocated for training and 10% for testing. The performance gap between nerfacto and the models with improved features becomes even more apparent with this dataset, as shown in Fig. 4. Nerfacto with transient embeddings yields sharper results compared to nerfacto with the robust loss technique. However, the former method tends to produce images with a darker tone which are not true to the scene. The qualitative results based on the TBW test dataset are shown in Table II. The enhanced nerfacto models demonstrate a superior ability to handle the presence of marine snow in the training images, resulting in the generation of synthesized images with significantly higher quality compared to the original nerfacto model.

TABLE I
QUALITATIVE RESULTS ON THE DOB TEST DATASET.

	PSNR ↑	MS-SSIM ↑	LPIPS ↓
nerfacto	29.382	0.915	0.468
nerfacto + transient embeddings	30.074	0.934	0.446
nerfacto + robust loss	30.679	0.929	0.458

TABLE II
QUALITATIVE RESULTS ON THE TBW TEST DATASET.

	PSNR ↑	MS-SSIM ↑	LPIPS ↓
nerfacto	26.269	0.863	0.507
nerfacto + transient embeddings	30.068	0.930	0.434
nerfacto + robust loss	29.926	0.896	0.447

IV. CONCLUSION

We have conducted a feasibility study focused on mapping structures for intricate underwater operations utilizing NeRFs. Being able to incorporate or ignore scene inconsistencies is crucial for enhancing the robustness of NeRF models, enabling the synthesis of high-quality, novel views of underwater structures. This capability presents exciting possibilities in the realms of underwater inspection and intervention. In particular, with a known vehicle pose, we can now generate NeRF-rendered images to serve as priors for compressing the respective camera image before transmission. This could substantially reduce data transmission requirements, paving the way for virtual tethering in underwater vehicles [16].

V. ACKNOWLEDGEMENT

This research project is supported by A*STAR under its RIE2020 Advanced Manufacturing and Engineering (AME) Industry Alignment Fund - Pre-Positioning (IAF-PP) Grant No. A20H8a0241.

REFERENCES

[1] J. Leatherdale and D. Turner, "Underwater photogrammetry in the north sea," *The Photogrammetric Record*, vol. 11, no. 62, pp. 151–167, 1983.

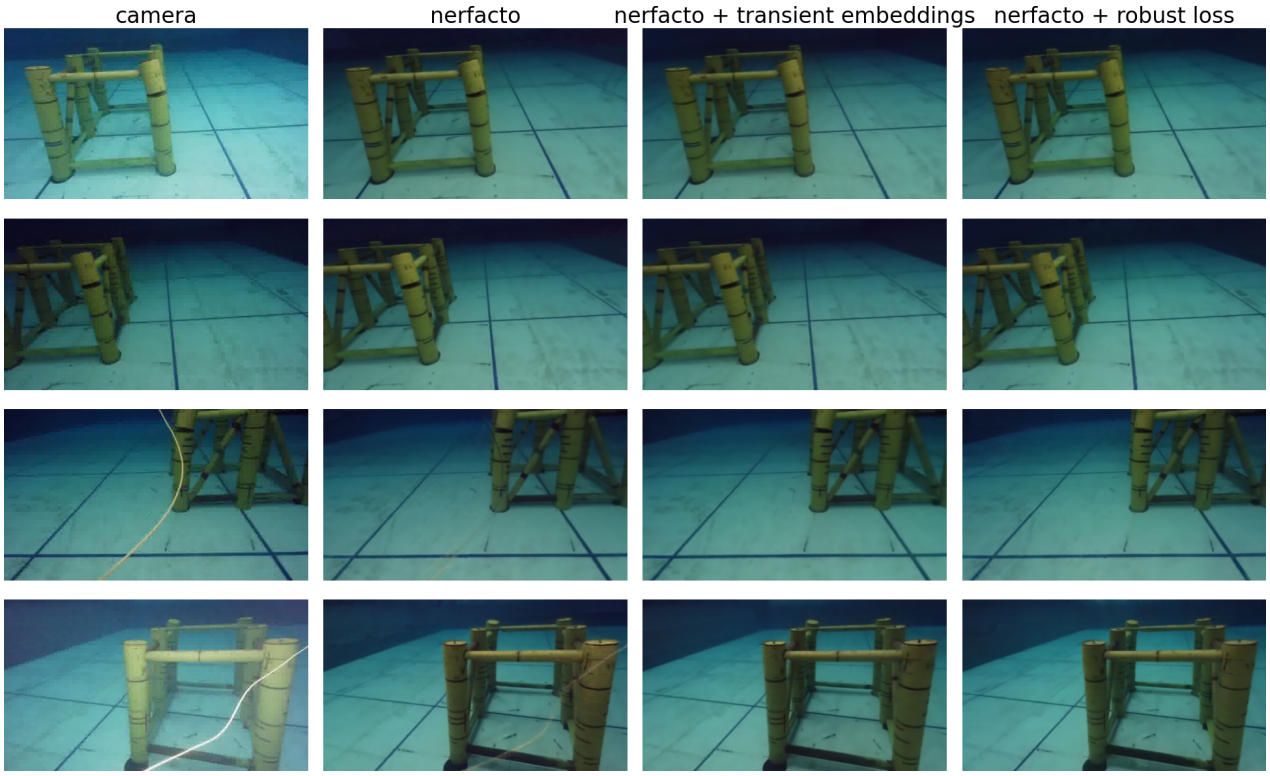

Fig. 3. Evaluation on the test set of the DOB scene in terms of synthesized novel views from different camera viewpoints, compared to camera ground truth (column 1). The rows indicate the four scenarios, and columns 2-4 show synthesized views from the 3 NeRF approaches.

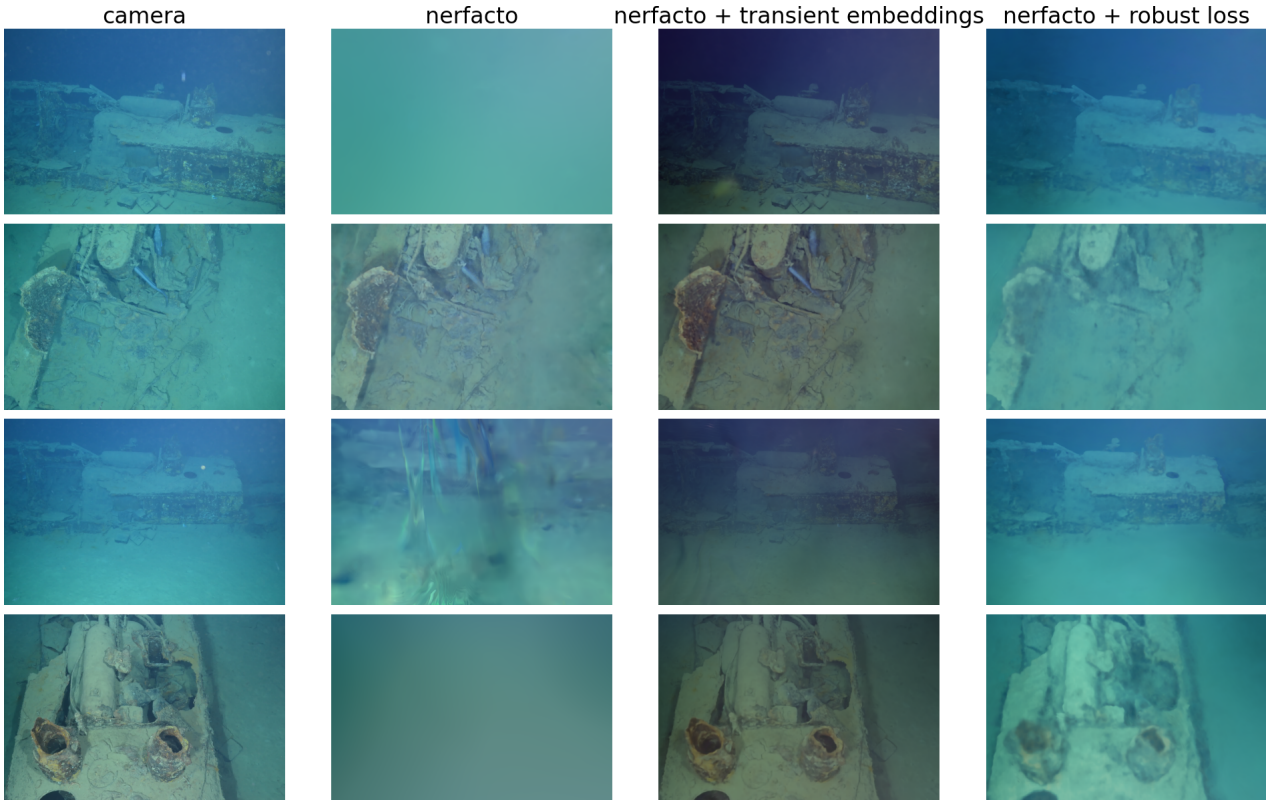

Fig. 4. Evaluation on the test set of the TBW scene. The rows indicate different camera viewpoints.

[2] J. Henderson, O. Pizarro, M. Johnson-Roberson, and I. Mahon, "Mapping submerged archaeological sites using stereo-vision photogrammetry," *International Journal of Nautical Archaeology*, vol. 42, no. 2, pp. 243–256, 2013.

[3] W. Figueira, R. Ferrari, E. Weatherby, A. Porter, S. Hawes, and M. Byrne, "Accuracy and precision of habitat structural complexity metrics derived from underwater photogrammetry," *Remote Sensing*, vol. 7, no. 12, pp. 16 883–16 900, 2015.

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

[5] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "Nerf in the wild: Neural radiance fields for unconstrained photo collections," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7210–7219.

[6] "TCOMS Research Development." [Online]. Available: https://www.tcoms.sg/research-development/

[7] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa, "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, ser. SIGGRAPH '23, 2023.

[8] R. Martin-Brualla, N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections," in *CVPR*, 2021.

[9] S. Sabour, S. Vora, D. Duckworth, I. Krasin, D. J. Fleet, and A. Tagliasacchi, "Robustnerf: Ignoring distractors with robust losses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 626–20 636.

[10] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[12] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[13] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[14] A. Arnaubec and R. Ewen, "Torpedo boat wreck (mediterranean, 43.124n;6.523e): Imagery and 3d model," SEANOE, https://doi.org/10.17882/79028, 2021.

[15] A. Arnaubec, M. Ferrera, J. Escartín, M. Matabos, N. Gracias, and J. Opderbecke, "Underwater 3d reconstruction from video or still imagery: Matisse and 3dmetrics processing and exploitation software," *Journal of Marine Science and Engineering*, vol. 11, no. 5, p. 985, 2023.

[16] R. Mishra, M. Chitre, B. Kalyan, Y. M. Too, H. Vishnu, and L. Peng, "An architecture for virtual tethering of rovs," in *2024 OCEANS-MTS/IEEE Singapore*. IEEE, 2024.