An Architecture for Virtual Tethering of ROVs

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

***Abstract*—We introduce an architecture designed for wireless Remote Operated Vehicle (ROV) operations during intervention surveys. Using the recent advancements in 3D modeling, visual odometry, video compression, and underwater communication, our approach aims to facilitate fully wireless ROV operations. The proposed architecture comprises several modular blocks, each with versatile applications spanning diverse domains, including multiplayer gaming, remote rover control and video transmission over a low-bandwidth connection. This comprehensive framework signifies progress in ROV operations, replacing the traditional physical tether with a software-defined virtual tether.**

I. Introduction

ROVs play a crucial role in facilitating underwater inspections and interventions in offshore operations. These vehicles are typically operated using a tether, requiring substantial onsite infrastructure such as tether management systems and support vessels with ROV operators aboard. While tethers ensure comprehensive control over ROV operation, they are susceptible to entanglement with subsea obstacles and pose an operational risk.

Autonomous Underwater Vehicles (AUVs) offer the advantage of operating without tethers, suitable for tasks like site surveys or seabed scanning in certain offshore operations. However, intricate subsea tasks still require human-in-the-loop control, such as manipulating valves or maneuvering precisely near subsea infrastructure. The necessity for a solution drove the ongoing development of Hybrid ROVs (HROVs), which aim to combine the advantages of tetherless operation with human-controlled precision.

Early attempts at HROV development involved substituting wired tethers with acoustic links [2]. However, acoustic modems' lower data throughput posed challenges for real-time video streaming and limited Frames Per Second (FPS) using off-the-shelf compression algorithms. While offering wireless connectivity, this approach fell short in delivering real-time video due to inherent acoustic communication limitations.

Integration of optical modems with an acoustic modem aimed to overcome the throughput limitations [3], enabling low-latency wireless connections between base stations and underwater vehicles. This technology combination demonstrated ROV operation without wired tethers [4], using optical modems for real-time video transmission and acoustic modems for manipulator control. However, the operational radius remained constrained as the optical modem's throughput diminished with increasing turbidity.

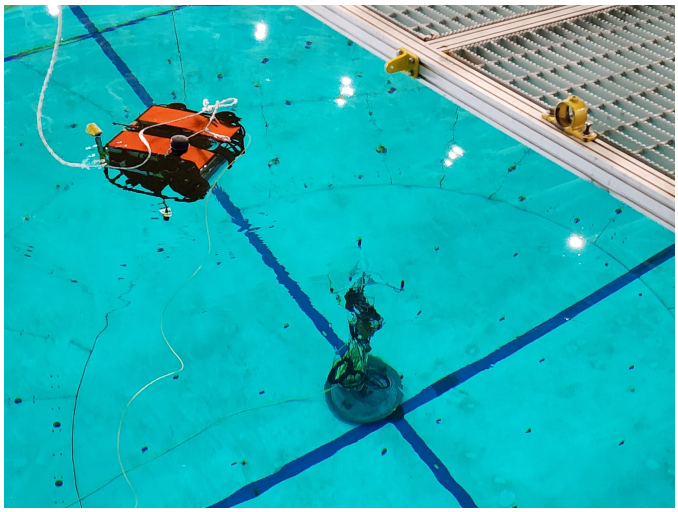


Fig. 1: Our hybrid ROV and the base station, equipped with optical and acoustic modems, used during the experiments at the ocean basin in the Technology Centre for Offshore and Marine, Singapore (TCOMS) [1].

The challenge of replacing wired tethers with alternatives offering extensive operational areas and responsive control remains unresolved.

The offshore inspection and intervention missions utilizing ROVs typically cover the same geographical area in successive runs. During the initial inspection run, the ROV is equipped with additional instrumentation, such as acoustic modems for positioning and tethers for power, control commands, and communication. With these instruments, the ROV can gather extensive information to characterize the baseline state of the survey area and conduct thorough inspections of target infrastructure. The collected data usually includes visual and sonar profiles, along with navigational information. In contrast, intervention missions occur more frequently, often requiring the ROV to execute precise maneuvers. Wireless operation during these intervention missions becomes imperative, and our primary focus is on enabling wireless operations during these interventions.

In recent years, machine learning algorithms have gained significant traction in various fields, including robotics and computer vision. These algorithms have demonstrated remarkable capabilities in tasks such as generating 3D models using Neural Radiance Fields (NeRFs) [5]–[7], estimating positions [8]–[10], and even image compression [11] using solely image

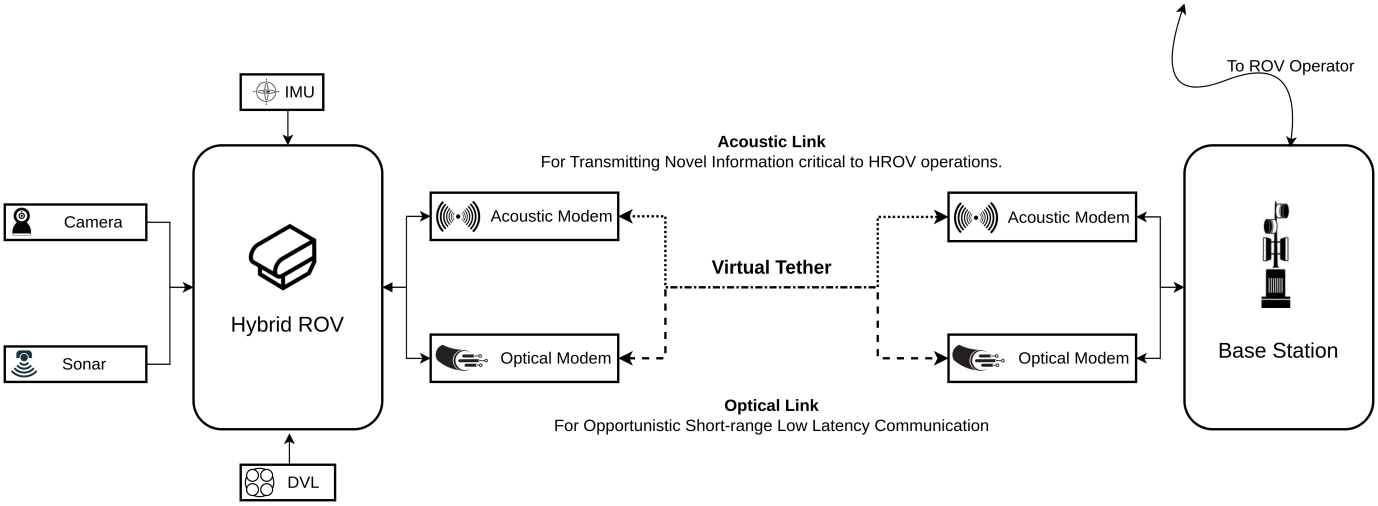


Fig. 2: A graphical depiction of our architecture for virtual tethering of ROVs.

data. Building upon these recent advancements, we introduce an architecture that can function as a virtual tether, providing operational benefits comparable to a wired connection and facilitating remote human-in-the-loop wireless operations.

Our proposed architecture processes streaming data from individual components such as cameras and positioning systems, extracting only novel information from this stream and transmitting solely this condensed data, which is smaller compared to transmitting the entire data stream directly. This processing is aimed at optimizing data packet size for efficient transmission over acoustic links, thus expanding the operational area while maintaining control over the ROV. Additionally, we integrate an auto-switch mechanism to transition to an optical link when the vehicle approaches a base station and requires low-latency communication. Following are the contributions of this work -

1) Use of NeRF model as a Database of Views.
2) Compressing and reconstructing camera view for transmitting it over low-bandwidth communication link.
3) Demonstration of the virtual tether framework through field experiments.
4) An auto-switch mechanism for opportunistic short-range and low-latency communication.

Details of this architecture are provided in Section II, while implementation specifics and experimental results are discussed in Section III. Moreover, the versatility of this architecture and its constituent components extends beyond wireless ROV operations to applications such as remotely controlling robots with limited connectivity and video transmission over constrained bandwidths.

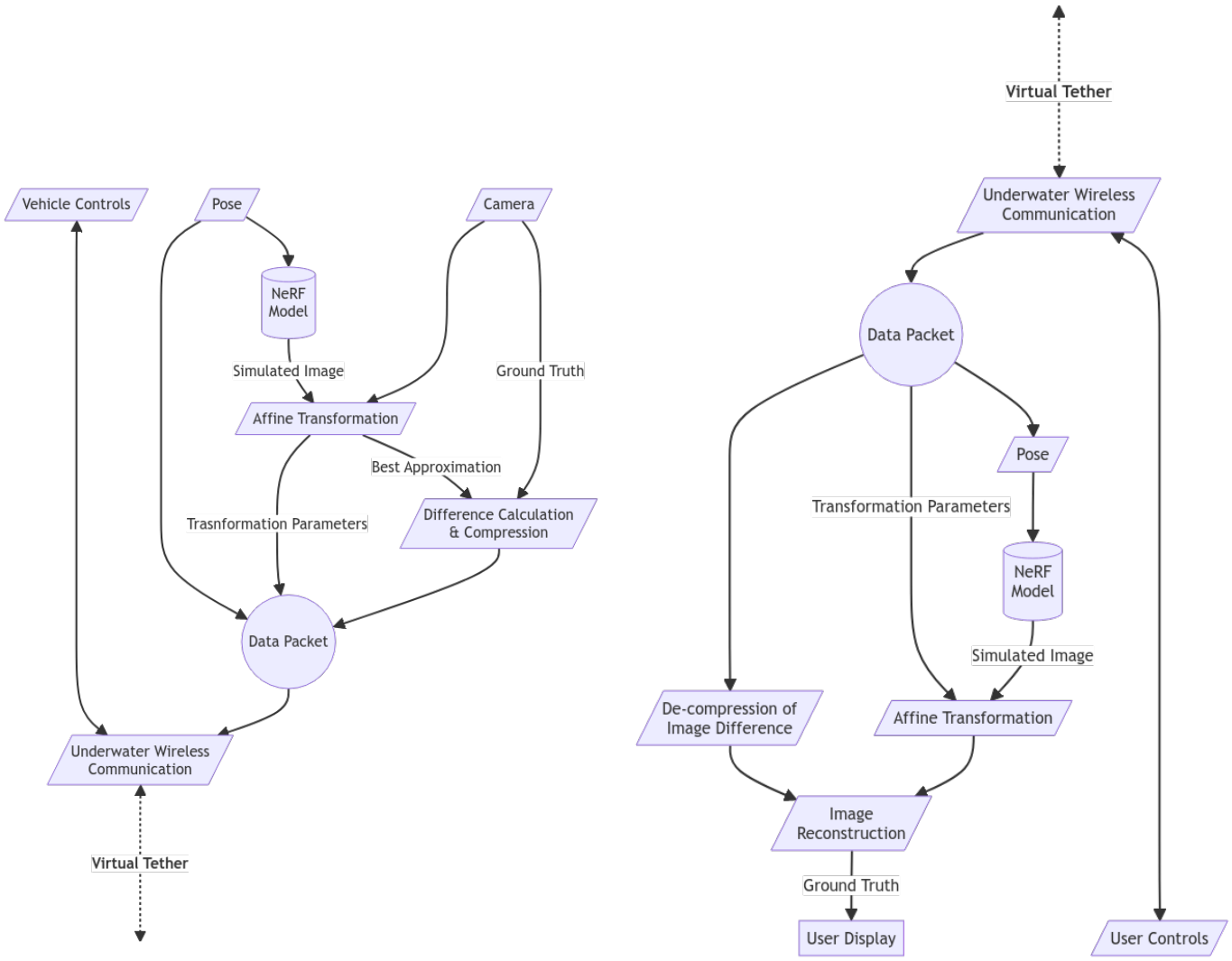
II. Virtual Tethering Architecture

We are interested in making the intervention missions completely wireless and thus replace the physical tether with an virtual alternative. One straightforward approach is to utilize the data collected during inspection runs and train a machine learning models for successive missions that compresses the incoming data stream. For instance, the utilization of autoencoder based compression for scientific data has demonstrated compression ratios of up to 50% [12]. However, even with this degree of compression, it remains infeasible for acoustic modems to provide a high FPS video stream. Consequently, this simple architecture, utilizing an environment-specific compression and decompression model, can provide a virtual tether with a large operational area but with lower FPS.

Broadly, a physical tether is used for transmitting live camera images, pose information and vehicle control status for ROV operators to maneuver the vehicle underwater. The bandwidth required for transmitting pose information and control commands is low, and thus, these can be transmitted using an acoustic link. However, the camera images require significant compression to achieve high FPS and make fully functional wireless ROV control feasible.

Interestingly, pose information and camera images are correlated. This correlation means that for a particular pose, there will be a corresponding camera view. One way to exploit this correlation is to make use of the data collected during the first inspection run and maintain a bag-of-poses with correlated views. Then, whenever the vehicle is in a particular pose during an intervention mission, we can query this bag-of-poses and render the equivalent camera view. However, such an approach will have two major challenges. First, during the inspection mission, we will have to visit each and every pose in the area to build a comprehensive bag-of-poses, which is infeasible. Second, there may be dynamic components in the view, such as schools of fish or new obstacles, that were not present during the intervention mission and thus render an incorrect view without the dynamic components.

We overcome the first challenge by learning a 3D model representation that can be queried to render an equivalent camera view using the pose information only. This representation is learned using the data collected during the inspection task, details about this model are discussed in Section II-A. During the intervention mission, we can use the vehicle pose



(a) (b)

Fig. 3: Flowchart diagrams depicting the implementation process within the vehicle and operator side. (a) illustrates the implementation flow within the vehicle before transmitting information via the virtual tether, while (b) depicts the implementation on the operator side. Vehicle and user control commands are transmitted using conventional underwater communication methods.

to render a view using the learned 3D representation and compute the difference between this rendered view and the live view captured using the camera. Calculating this difference in information can help in overcoming the second challenge of dynamic components, details about this difference calculation are discussed in Section II-B. Finally, we stack the difference and other information such as pose as discussed in Section II-C to successfully recover the live camera view by transmitting only the pose and the novel information instead of raw, live camera views. The level of compression offered by such an approach makes it feasible to transmit data over an acoustic link. Fig. 2 shows a graphical representation of our virtual tethering implementation.

A. NeRF model as a Database of Views

NeRF represents a scene in term of a continuous volumetric scene function through a fully-connected neural network [5]. This enables the learning of complex 3D representations using only a partial set of images and corresponding pose information. Once trained, NeRF models can generate novel views based solely on pose information. Essentially, a NeRF model captures the information of a complex 3D scene within the weights and biases of the neural network. This compact and easily queryable representation makes NeRFs well-suited for use as a database for rendering views.

A recent feasibility study demonstrated the applicability of NeRFs to underwater structures [13]. Building upon this study, we utilize data collected during the initial inspection mission

to train a NeRF model for learning a 3D representation of the underwater site. Subsequently, this trained NeRF model enables operators to render synthetic views solely based on received pose information from the vehicle. However, additional information is required for rendering a camera view from this synthetic view, as discussed in the following section.

B. Extracting Novel Information for Rendering Camera View

NeRF models offer the capability to generate synthetic representations of underwater structures. However, the live camera view may deviate from this synthetic representation due to two primary factors: the dynamic nature of underwater scenes, including fluctuations in lighting conditions, and inaccuracies in the vehicle's pose estimates. This discrepancy between views necessitates the extraction of novel information to reconstruct the live camera view from a given synthetic representation, with a critical requirement being that the representation remains compact enough for transmission via acoustic links.

A direct method for computing this novel information involves determining the pixel-wise differences between the live camera view and the synthetic representation. This approach is particularly effective in addressing the dynamic nature of underwater scenes. Ideally, the computed differences would manifest as a sparse matrix, making them well-suited for compression algorithms such as WebP [14], which allocate fewer bits to segments with low entropy. However, inaccuracies in pose estimates may result in slight shifts or tilts in the synthetic image, potentially diminishing the sparsity and increasing the size of the compressed image.

To mitigate the impact of noisy pose estimates, we compute an affine transformation of the synthetic view that effectively preserves the salient features observed in the live camera view while simultaneously minimizing the size of the resulting compressed image. Remarkably, the parameters of the affine transformation necessary for reconstructing the camera view are independent of the image dimensions. They rely solely on the specific transformations required such as translation or rotation, rendering it highly suitable for our intended purpose.

C. Transmitting and Reconstructing the View For Operator

For a fully functioning virtual tether, ensuring that operators can view live camera images while transmitting all required data via underwater wireless communication is imperative. We deploy two components to achieve this, one operating on the vehicle side and the other on the operator or top side. The flowchart for these individual components is depicted in Fig. 3.

The implementation process commences with training a NeRF model using the dataset collected during the initial inspection mission. Once trained, a copy of this model resides both on the vehicle and on the operator side. During successive missions, when a new camera and pose image are captured on the vehicle, our vehicle-side framework first utilizes the pose to generate the corresponding synthetic view using the NeRF model. Subsequently, it calculates the novel information and corresponding affine transformation parameters using the method outlined in Section II-B. The framework then encapsulates the novel information, transformation parameters, and vehicle pose into a data packet, which is transmitted via the underwater wireless link. It is noteworthy that training the NeRF model is necessary only at the outset of the process and may be repeated only if the size of data packets continually increases, suggesting significant overall changes in the underwater scene compared to the NeRF representation.

On the operator side, the framework receives the data packet containing vehicle pose, novel information, and transformation parameters via the underwater wireless link. Utilizing the copy of the trained NeRF model, the framework generates the synthetic view based on the received pose information. Subsequently, it applies the affine transformation to the synthetic view using the received transformation parameters. Finally, the framework decompresses and integrates the novel information into this transformed image to provide the operator with a live camera view. The flow of information and processes for both the vehicle-side and operator-side frameworks are depicted in flowchart diagrams in Fig. 3.

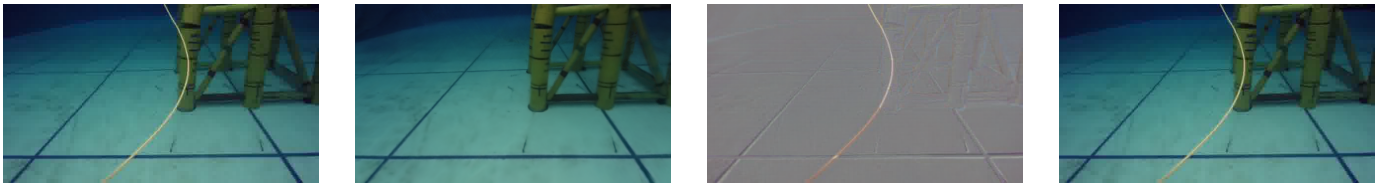
D. Vehicle Controls and Hybrid Underwater Communication

The bandwidth required for transmitting vehicle controls and other sensory information, such as battery levels, is generally low. In practice, such vehicle data can be transmitted using fewer bytes. We leverage on these well-established conventional methods to send vehicle data across, rendering our architecture capable of transmitting live camera views, vehicle poses, and control information using acoustic communication links only.

However, operators may occasionally require low-latency communication for downloading or uploading certain files to the ROV during inspection missions. Utilizing an acoustic link may not be feasible but an optical modem can provide this, given the vehicle is within a certain operational range. Consequently, our framework incorporates an optical modem alongside our acoustic modem for opportunistic low-latency and short-range communication. The framework seamlessly switches between acoustic and optical communication, enabling operators to take advantage of the auto-switch feature by navigating to areas where optical communication is viable to download large files such as mission logs. This hybrid underwater communication approach fulfills the final requirement of a physical tether, providing a viable virtual alternative to it.

III. Experiments & Results

We retrofitted a commercially available ROV with both an optical and an acoustic modem and used NVIDIA Orin to run the vehicle-side framework exclusively. The low-level controls on the vehicle were managed using ROS 2 [15] and associated packages. Additionally, we deployed a base station equipped with hydrophone array, an acoustic modem and an optical modem for communication and vehicle localization. On the operator side, we connected the base-station with a GPU-enabled system to run our operator-side framework.



(a) (b) (c) (d)

Fig. 4: Results from the field experiments showcasing the efficacy of our virtual tether architecture. (a) depicts the live camera view, while (b) presents the synthetic view generated using pose information and our trained NeRF model. The computed novel information is illustrated in (c), revealing the presence of the safety tether obstructing the camera view, which was not present during the initial data collection used for training the NeRF model. Finally, (d) displays the successfully reconstructed image achieved using the method outlined in Section II-C.

Implementation of the framework on both sides utilized the Julia programming language.

Our experiments took place in an ocean basin at TCOMS [1], where we employed our hybrid ROV to conduct underwater structure inspections, as depicted in Fig. 4a. The first inspection mission involved data collection to learn a NeRF representation, as detailed in Section II-A. Subsequent missions utilized our virtual tether architecture for data compression, resulting in significantly reduced throughput requirements. For instance, 720×320 pixel images were transmitted using only 16 to 35 kilobits after applying our compression techniques. During the experiment, we demonstrated video transmission at 1 to 2 FPS.

Fig. 4 provides insight into the process facilitated by our framework. The image displayed in Fig. 4a represents the live camera view, while the image in Fig. 4b is generated using the NeRF model. Subsequently, our vehicle-side framework computes novel information, depicted in Fig. 4c. The gray area in this image denotes segments with low entropy, demonstrating the compression potential as compared to the original camera image. Conversely, the non-gray portion represents novel information absent from the synthetic view, such as the safety tether obstructing the camera's perspective that was not present during the intervention mission. Finally, this novel information, along with other data points, is transmitted, resulting in the rendered image shown to the operator in Fig. 4d, employing the method outlined in Section II-C. It is evident that the final rendered image closely resembles the camera view. This experiment serves as proof-of-concept for our proposed virtual tether system and establishes a baseline for future optimization to achieve higher FPS.

IV. Conclusion

Our approach introduces a novel machine learning-based video compression architecture that leverages the repetitive nature of ROV missions. By transmitting only novel information in real-time alongside data necessary for reconstructing the vehicle's camera view, our architecture optimizes bandwidth usage, enabling effective low-bandwidth communication. Through our virtual tether framework, we achieve robust image reconstruction and provide satisfactory FPS for ROV operators to navigate the vehicle wirelessly. Utilizing NeRF models as a database, combined with our data manipulation techniques for image compression and transmission, demonstrates the efficacy of our framework. Field experiments validate the feasibility of our approach, showcasing that modern-day embedded computing capabilities have reached a practical point for implementing such techniques.

Acknowledgment

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

[2] R. Dunbar and A. Settery, "Video communications for the untethered submersible rover," in *Proceedings of the 1985 4th International Symposium on Unmanned Untethered Submersible Technology*, vol. 4. IEEE, 1985, pp. 140–149.

[3] N. Farr, A. Bowen, J. Ware, C. Pontbriand, and M. Tivey, "An integrated, underwater optical/acoustic communications system," in *OCEANS'10 IEEE SYDNEY*. IEEE, 2010, pp. 1–6.

[4] A. D. Bowen, M. V. Jakuba, N. E. Farr, J. Ware, C. Taylor, D. Gomez-Ibanez, C. R. Machado, and C. Pontbriand, "An un-tethered rov for routine access and intervention in the deep sea," in *2013 oceans-san diego*. IEEE, 2013, pp. 1–7.

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

[6] 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.

[7] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5855–5864.

[8] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[9] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

[10] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1983–1992.

[11] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.
[12] T. Liu, J. Wang, Q. Liu, S. Alibhai, T. Lu, and X. He, "High-ratio lossy compression: Exploring the autoencoder to compress scientific data," *IEEE Transactions on Big Data*, 2021.
[13] Y. M. Too, H. Vishnu, M. Chitre, B. Kalyan, L. Peng, and R. Mishra, "A feasibility study on novel view synthesis of underwater structures using neural radiance fields," in *2024 OCEANS-MTS/IEEE Singapore*. IEEE, 2024.
[14] "Reference on WebP compression techniques." [Online]. Available: https://developers.google.com/speed/webp/docs/compression
[15] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot operating system 2: Design, architecture, and uses in the wild," *Science Robotics*, vol. 7, no. 66, p. eabm6074, 2022. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.abm6074