Enhanced DUSt3R for Underwater 3D Reconstruction in Shallow Water Environments

Tsuyoshi Shimano[1], Takashi Fuse[1]

[1] Department of Civil Engineering, The University of Tokyo, Tokyo, Japan - (shimano@trip.t.u-tokyo.ac.jp, fuse@civil.t.u-tokyo.ac.jp)

Keywords: Underwater Photogrammetry, 3D Reconstruction, Shallow Water, DUSt3R, Underwater Simulation.

Abstract

Shallow-water environments present significant challenges for underwater photogrammetry due to light caustics and the combined effects of absorption and scattering caused by water turbidty. These optical disturbances degrade image quality, disrupt feature matching, and ultimately reduce the reliability of 3D reconstruction using traditional SfM (Structure from Motion) pipeline. In this study, we focus on these two dominant factors and investigate a 3D reconstruction framework inspired by recent feed-forward architectures such as DUSt3R (Dense and Unconstrained Stereo 3D Reconstruction). To support this approach, we develop a synthetic data generation pipeline capable of simulating shallow-water visual conditions. Preliminary experiments indicate a possible trend for integrating physics-aware image formation with DUSt3R-type feed-forward reconstruction. However, several limitations remain: the current model does not yet achieve stable accuracy, real-world underwater validation has not been conducted, and computation costs remain high due to complex training procedures. Future work will focus on refining the network architecture, exploring DUSt3R-derived multi-view and high-fidelity extensions, accelerating computation, and validating the pipeline in real shallow-water environments. Additionally, integrating advanced rendering techniques may further improvethe refinement of 3D reconstruction.

1. Introduction

Underwater 3D reconstruction has become increasingly important for various applications including structure monitoring (Wu et al., 2023), archaeological documentation (Drap et al., 2015, Henderson et al., 2013), and environmental monitoring (Leon et al., 2015, Marre et al., 2020). In particular, shallow water areas are critical for infrastructure construction and maintenance because they strongly accelerate the corrosion of steel components (Melchers and Jeffrey, 2014). The increased corrosion rate arises from shallow-water characteristics such as elevated dissolved oxygen levels, intensified biofilm and microbial activity, frequent oxygen replenishment by wave and tide action, and wet–dry cycling near the tidal zone. These factors collectively enhance electrochemical reactions, making steel elements in shallow water substantially more vulnerable to corrosion than those in deeper environments.

In shallow-water environments, several factors severely degrade the reliability of image-based 3D reconstruction (Hu et al., 2023). The two dominant issues are: (1) dynamic light caustics and (2) depth-dependent absorption and scattering caused by water turbidity. These phenomena distort image appearance and significantly hinder conventional SfM and MVS pipelines. In particular, SfM relies on stable feature extraction and matching, but these optical disturbances reduce feature contrast and consistency, making robust correspondence estimation difficult.

Meanwhile, feed-foward 3D reconstruction frameworks, such as DUSt3R (Wang et al., 2024), have recently gained widespread attention. These models leverage large amounts of teacher data, including images and ground-truth depth, and can directly infer 3D geometry from image pairs without explicit feature extraction or matching. Although such approaches show strong potential for underwater applications, current models are trained almost exclusively on terrestrial datasets. Incorporating underwater imagery into the training process would enable these feed-forward architectures to better handle underwater-specific optical conditions and potentially achieve more accurate reconstruction in shallow-water environments.

In this paper, we present an enhanced DUSt3R-based framework tailored for underwater 3D reconstruction in shallow-water environments. To support this approach, we generate synthetic underwater datasets using Blender that incorporate caustics, attenuation, and scattering effects. We then adapt the DUSt3R architecture to better handle these underwater-specific optical conditions and train the model using the generated dataset. Finally, we evaluate its reconstruction performance in comparison with the baseline model.

This research contributes to underwater photogrammetry by introducing a new approach for 3D reconstruction specifically designed for shallow-water environments. We develop a synthetic dataset using Blender that incorporates underwater optical effects to better represent shallow-water imaging conditions. Building on this dataset, we adapt the DUSt3R feed-forward reconstruction framework to handle underwater-specific image degradation by integrating physics-aware modeling into the training process. Finally, we present preliminary experiments that demonstrate the potential of this approach and highlight both the challenges and prospects of applying feed-forward 3D reconstruction to underwater scenes.

2. Related Work

2.1 Challenges with Shallow Water Environments

Underwater photogrammetry in shallow-water environments faces several challenges, particularly the effects of light caustics, the absorption and scattering caused by water turbidity, and violations of the collinearity condition induced by ray refraction at the camera housing. In this study, we focus on former two dominant factors because they have the greatest impact on 3D reconstruction in shallow water. Previous research

has repeatedly highlighted their importance, although most existing studies address them primarily from an image restoration perspective rather than within a 3D reconstruction framework.

While camera-housing refraction is also an important consideration in real underwater imaging, especially for accurate reconstruction, previous studies have proposed various calibration methods. In simple aapproaches, the effects of refraction are absorbed into the distortion parameters of a pinhole camera model (Sedlazeck and Koch, 2012, Shortis, 2019). However, this approximation has been shown to be insufficient in complex conditions. Therefore, more recent studies introduce explicit refractive models, providing both theoretical foundations and practical calibration methods (Agrawal et al., 2012, Jordt-Sedlazeck and Koch, 2012, Sun et al., 2024).

2.1.1 Caustics In general optics, caustics refer to the envelopes formed when light rays are reflected or refracted by curved surfaces or objects. In shallow-water environments, caustics denote the intensity patterns of light created on the seabed as sunlight is reflected and refracted by the water surface. The presence of caustics causes large variations in brightness along their contours, which increases the number of false feature detections. Moreover, since caustic patterns are formed by surface waves, they change dynamically over time. Therefore, in monocular imaging, the caustic patterns differ from frame to frame, making image acquisition and analysis more challenging. In practical 3D reconstruction studies, image capture is often performed during early morning hours when the sun is low to avoid the occurrence of caustics (Casella et al., 2024).

Several studies have been conducted on image restoration techniques aimed at removing caustics, primarily using statistical or image-processing-based methods (Gracias et al., 2008, Trabes and Jordan, 2015, Tripathy et al., 2020). However, with the recent advances in deep learning, more sophisticated approaches have been proposed. For instance, a two-stage convolutional neural network has been introduced, where the first stage detects caustic regions and the second stage performs image restoration (Forbes et al., 2018). In addition, under the assumption that SfM can be applied, a self-supervised approach has been proposed that uses depth maps obtained from SfM to restore caustic-free images (Sauder and Tuia, 2024). Nevertheless, these methods mainly focus on image restoration itself and do not directly address the entire 3D reconstruction process.

2.1.2 Light Absorption and Scattering Unlike in the atmosphere, water absorbs light, which leads to image degradation. In particular, light in the red wavelength region is strongly absorbed, causing underwater images to exhibit a dominant blue or green tone (Hu et al., 2023). To address these issues, various image processing techniques—such as contrast equalization and smoothing—have been proposed and compared (Mangeruga et al., 2018, Jian et al., 2021). Similar to the case of caustics, recent studies have increasingly adopted deep learning–based approaches (Cong et al., 2024). Among them, some methods assume that depth maps are available through SfM and perform physics-based image restoration under physical constraints, following the optical imaging equations (Sauder and Tuia, 2024).

Furthermore, several recent approaches incorporate explicit underwater image formation models into 3D neural representations. SeaThru-NeRF (Levy et al., 2023) extends NeRF (Mildenhall et al., 2021) by embedding wavelength-dependent light attenuation into the volumetric rendering process, enabling physically grounded color restoration during 3D reconstruction. Similarly, AquaGS (Shi et al., 2024) integrates underwater imaging physics into Gaussian Splatting (Kerbl et al., 2023), allowing the model to jointly reason about geometry, scene radiance, and depth-dependent light attenuation. These methods demonstrate that combining neural 3D representations with physics-based underwater imaging models can significantly improve color fidelity in shallow-water environments. However, because they still rely on iterative volume rendering or optimization-based pipelines, their computational demands remain substantially higher than feed-forward models.

2.2 Feed-forward 3D Reconstruction

The classical pipeline for 3D reconstruction typically consists of Structure from Motion (SfM) and Multi-View Stereo (MVS). These methods have an inherently iterative structure. Specifically, SfM extracts and matches local features across images and then performs Bundle Adjustment (BA)—an iterative optimization process—to jointly refine the camera poses and a sparse 3D point cloud. MVS then utilizes the camera poses estimated by SfM to establish dense pixel-wise correspondences, estimate per-pixel depth, and fuse them into a dense point cloud or mesh model. However, due to their inherently iterative nature, these methods face systematic challenges in terms of robustness, efficiency, and usability. Because of their sequential workflow, errors can easily propagate between stages, making them vulnerable to challenging scenarios such as texture-less regions.

With the advancement of deep learning, new approaches have been developed to avoid such error propagation by directly inferring globally consistent camera poses and dense geometry from image sets (Zhang et al., 2025). These methods typically consist of a powerful encoder for extracting image features, a Transformer-based global matching module for establishing dense pixel-wise correspondences, and a joint decoding head that estimates the relative camera poses and 3D structure from these correspondences. In essence, they replace the traditional multi-stage workflow with a single, unified deep network that encapsulates the entire 3D reconstruction process.

Since the introduction of DUSt3R (Wang et al., 2024), a new research paradigm known as Feed-forward 3D Reconstruction has emerged, with DUSt3R recognized as its pioneering work. Following DUSt3R, numerous studies such as MASt3R (Leroy et al., 2024), VGGT (Wang et al., 2025), and Fast3R (Yang et al., 2025) have further improved correspondence accuracy, multi-view consistency, and real-time capability.

DUSt3R features a fully feed-forward architecture that enables accurate 3D reconstruction from an unposed image pair in a single forward pass, without requiring any prior knowledge of camera poses or intrinsic parameters. Given a pair of images, a shared encoder extracts deep visual features, which are then processed by a cross-attention Transformer serving as a decoder to establish dense pixel-wise correspondences. From these correspondences, a regression head jointly estimates the relative camera poses and dense 3D geometry, forming a simple yet effective pipeline. During training, the model minimizes the point map regression error, ensuring that corresponding 3D points are accurately aligned. A confidence map is also used to weight the loss, reflecting the reliability of each estimated point. Because DUSt3R operates on an image-pair basis, it is flexible across diverse imaging conditions and achieves a good balance between generality and computational efficiency without requiring large-scale input data. Furthermore, its concise and

elegant architecture has served as a foundational baseline for subsequent feed-forward models, guiding both design and comparison in later studies.

3. Methodology

In this study, we aim to improve the accuracy of underwater 3D reconstruction in shallow-water environments by developing an integrated framework consisting of (1) the generation of a physically plausible synthetic underwater dataset, and (2) an enhanced version of DUSt3R that accounts for underwater optical properties.

In shallow-water environments, although image acquisition with cameras is relatively easy, obtaining depth maps or accurate 3D geometry is extremely challenging. Consequently, “ground-truth” depth maps that can serve as supervision for learning-based methods are generally unavailable. Although several studies have proposed methods for synthesizing underwater images and depth maps (Zwilgmeyer et al., 2021, Lv et al., 2025), no existing dataset explicitly models shallow-water optical properties, particularly caustics and turbidity (absorption and scattering). To address this limitation, we generate a synthetic underwater dataset using Blender, where underwater optical models and environmental conditions are reproduced. The dataset consists of rendered RGB images, depth maps, and camera parameters. Furthermore, caustics and water turbidity (absorption and scattering) are randomized during rendering, enabling the generation of diverse samples that reflect the visual variability observed in real shallow-water conditions.

Next, using this synthetic underwater dataset, we train an extended version of DUSt3R that explicitly incorporates underwater optical characteristics and estimates depth maps from image pairs. The original DUSt3R is designed for terrestrial environments and does not account for underwater-specific degradations such as caustics and turbidity. In this work, we integrate a Lambert–Beer–based attenuation model together with frequency-domain features, enabling more accurate depth estimation under underwater imaging conditions.

This chapter provides a detailed description of (1) the construction of the synthetic underwater dataset, and (2) the architecture and training strategy of the improved DUSt3R model.

3.1 Synthetic Underwater Dataset Generation

In this study, we use the BlendedMVS dataset (Yao et al., 2020)—also employed in the training of DUSt3R—as the source dataset for constructing synthetic ground-truth training data. BlendedMVS contains a wide variety of outdoor real-world scenes captured in terrestrial environments, and for each scene it provides a corresponding mesh model in .obj format. These .obj models can be easily imported into Blender, making the dataset well suited for synthesizing underwater-like renderings. This enables us to partially bridge the domain gap in imaging conditions between air and water. However, since BlendedMVS does not include any real underwater scenes, differences in scene content between terrestrial and underwater environments remain.

For all 112 scenes included in BlendedMVS, we performed underwater-style rendering using Blender. For each scene, two underwater turbidity conditions (defined by scattering and absorption properties) were prepared. For each turbidity condition, 24 camera viewpoints were randomly selected such that their indices formed a continuous sequence, and rendering was conducted from each of those viewpoints. If a scene originally contained fewer than 24 available camera views, all available viewpoints were used. During rendering from each camera view, caustics and other shallow-water optical degradations were randomized, ensuring a diverse set of environmental conditions. This rendering pipeline is designed from a sim-to-real perspective (Zhao et al., 2020) using domain randomization, aiming to reduce the domain gap between synthetic data and real shallow-water environments by incorporating physically motivated and randomized optical degradations such as caustics and turbidity.

3.1.1 Simulation of Caustics Caustics, caused by refraction and reflection of sunlight at the water surface, ideally require precise simulation using ray tracing or physically based rendering. However, such methods impose a high computational cost and are impractical for large-scale dataset generation.

Therefore, following workflows commonly used in POSER (Menna et al., 2024) and similar tools, we adopted a simplified caustics simulation using Blender’s Voronoi texture node. Specifically, a white Voronoi pattern was overlaid on the surface of the target mesh, and combined with Add, Multiply, and Color Mix nodes in a randomized manner. The Voronoi node was configured with Dimensions = 3D and Feature = Smooth F1. Although this approach lacks full physical accuracy, it effectively reproduces the spatial and temporal randomness of caustics typically observed in shallow-water environments.

3.1.2 Simulation of Underwater Turbidity Underwater turbidity arises from two main optical phenomena: scattering and absorption.

In this study, these effects were simulated by applying Blender’s Principled Volume node to the World environment. The scattering and absorption colors were randomized within a blue–green range, replicating the characteristic coloration of shallow-water environments. In addition, the density parameter of the volume was randomized to generate varying levels of turbidity, ranging from relatively clear water to highly degraded visibility.

Through this process, we constructed a dataset that encompasses a wide range of underwater optical conditions with varying degrees of turbidity and color shift. This diversity allows the enhanced DUSt3R model to be trained under multiple illumination and visibility conditions characteristic of real shallow-water scenes.

3.2 Adaptation of DUSt3R for Underwater Reconstruction

In this study, we extend the feed-forward 3D reconstruction model DUSt3R to underwater environments and construct an enhanced version tailored to the optical characteristics of shallow water. Fig. 1 illustrates the overall architecture of the proposed model.

3.2.1 DUSt3R We first provide a brief overview of DUSt3R, which serves as the baseline of our work. DUSt3R is a feed-forward 3D reconstruction model that directly predicts a 3D point map $X^{1,1}, X^{2,1} \in \mathbb{R}^{W\times H\times 3}$ and its corresponding confidence map $C^1, C^2 \in \mathbb{R}^{W\times H}$ from two uncalibrated input images $I^1, I^2 \in \mathbb{R}^{W\times H\times 3}$ through a single

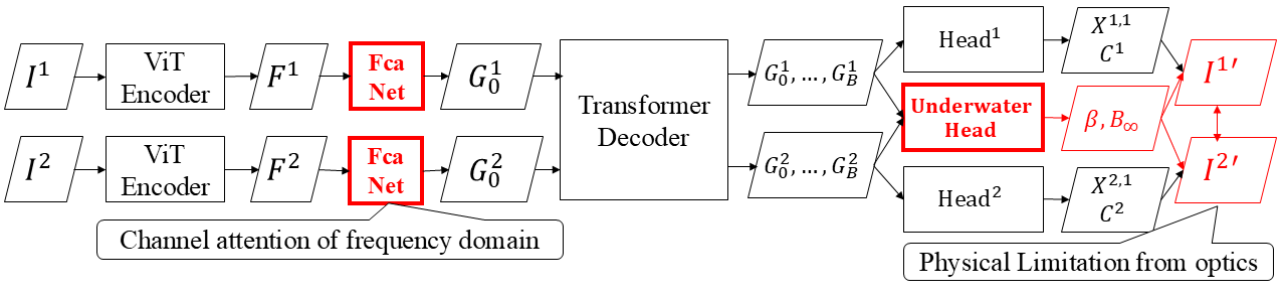


Figure 1. Overall architecture of the enhanced DUSt3R.

forward pass. Each element of the point map and confidence map corresponds to a single pixel. The two images I^1, I^2 arefirst processed by a shared-weight vision transformer encoder to extract visual features F^1, F^2. These encoded features are then fed into a Transformer decoder $\left(G_0^1 = F^1,\ G_0^2 = F^2\right)$, which incorporates cross-attention layers to model the geometric and photometric relationships between the two views. Through this cross-view interaction $\left(G_{i+1}^1 = \mathrm{Decode}(G_i^1, G_i^2),\ G_{i+1}^2 = \mathrm{Decode}(G_i^2, G_i^1)\right)$, the decoder produces refined, correspondence-aware features $\left(G^1 = \left\{G_0^1, \ldots, G_B^1\right\},\ G^2 = \left\{G_0^2, \ldots, G_B^2\right\}\right)$. A regression head at the final stage jointly estimates the 3D point map based on these correspondences. Importantly, both predict point maps $X^{1,1}, X^{2,1}$ are expressed in the coordinate frame of the first view, enabling direct pixel-wise comparison and straightforward multi-view fusion. During training, the point-map regression error is employed as the main loss, and the confidence map is used to weight the regression residuals so that unreliable regions contribute less to the optimization.

3.2.2 Handling Caustics Caustics in shallow-water environments arise from the refraction and reflection of sunlight by surface waves, producing periodic or quasi-periodic band-like patterns with strong local brightness variations. Although visually prominent, these patterns are unrelated to the geometry of underwater structures and may introduce incorrect correspondences or instability in geometric prediction when directly used for feature extraction.

To mitigate this issue, we integrate a Frequency Channel Attention (FcaNet) (Qin et al., 2021) module into the encoder of DUSt3R. Given an input feature map $X \in \mathbb{R}^{H\times W\times C}$, FcaNet performs multi-spectral channel analysis based on 2D Discrete Cosine Transform (DCT). Specifically, for each channel $X_c \in \mathbb{R}^{H\times W}$, a set of predefined 2D-DCT basis functions $B_{u_k,v_k}(h,w)$ is applied to extract K frequency components.

$$f_{c,k} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X_c(h,w)\, B_{u_k,v_k}(h,w),\ k = 1,\ldots,K \quad (1)$$

Here, $f_{c,k}$ denotes the k-th frequency coefficient of channel c, and (u_k, v_k) are the corresponding frequency indices. These coefficients from a frequency descriptor $F = [f_{c,1}, \ldots, f_{c,K}]$, which is fed into a lightweight MLP to obtain a channnel-wise attention weight $a_c \in (0,1)$. Finally, each channnel is rescaled as $\hat{X}_c = a_c X_c$. Through this process, FcaNet assigns larger weights to channels dominated by non-caustic geometric structures, while suppressing channels in which periodic high-frequency caustic patterns dominate. As a result, caustics-induced artifacts are attenuated, and structurally meaningful features are preserved. This frequency-aware reweighting is lightweight, differentiable, and fully compatible with DUSt3R's feed-forward encoder.

In this study, FcaNet is inserted directly after the encoder blocks of DUSt3R. Once, the two images are pocessed by the shared-weight encoder, their features F^1, F^2 are refined by the the shared-weight FcaNet.

$$G_0^1 = \mathrm{FcaNet}(F^1), \quad G_0^2 = \mathrm{FcaNet}(F^2) \quad (2)$$

This opertion replaces the original encoder output with frequency-enhanced tokens before they are fed into the Transformer decoder. FcaNet can be seamlessly inserted into the original DUSt3R encoder without altering the overall feed-forward pipeline, thereby improving robustness to caustics while preserving inference efficiency.

3.2.3 Handling Turbidity Underwater turbidity arises from light absorption and scattering, causing distance-dependent contrast reduction and color shifts. If such depth-dependent attenuation is not considered, even identical 3D structures may appear differently across viewpoint pairs, leading to degraded correspondence estimation and reduced accuracy in point-map regression.

To address this problem, we introduce a lightweight regression module, referred to as the Underwater Head, applied to the output features of the Transformer decoder. This module predicts underwater optical parameters for each image pair, including the attenuation coefficients and background light components defined in the Lambert–Beer attenuation model. These predicted parameter maps are then incorporated into the loss formulation (Section 3.3) to jointly evaluate consistency between geometric prediction and photometric restoration.

The Underwater Head operates in parallel with the existing regression head for point-map estimation, enabling the model to infer both 3D geometry and underwater optical parameters from the same feature representation. This design allows our enhanced DUSt3R to explicitly model underwater optical degradation while maintaining a fully feed-forward reconstruction pipeline.

The relationship between the optical parameters and the restored image formation process is formulated using the Lambert–Beer law, and the details of its integration into the loss function are described in Section 3.3.

3.3 Loss Function

In this study, we refine the loss functions of DUSt3R by incorporating underwater environment parameters predicted by the Underwater Head. We first describe the original DUSt3R loss

formulation, then introduce the Lambert–Beer law that forms the basis of our physical modeling, and finally present the proposed loss function.

3.3.1 Loss Function in DUSt3R The core error term in DUSt3R is the 3D regression loss. It is defined as the Euclidean distance between the predicted 3D point and its ground-truth correspondence for all valid pixels where the correspondences are available:

$$L_{\text{regr}} = \left\| \frac{1}{z} X_i^{v,1} - \frac{1}{\bar{z}} \bar{X}_i^{v,1} \right\| \tag{3}$$

Here, $X_i^{v,1}$ and $\bar{X}_i^{v,1}$ denote the predicted and ground-truth 3D points, respectively, and z and $\bar{z}$ are scale factors used to normalize their magnitudes. Following DUSt3R, these scale factors are estimated using a normalization function: $z = \text{norm}(X^{1,1}, X^{2,1}), \bar{z} = \text{norm}(\bar{X}^{1,1}, \bar{X}^{2,1})$ which computes a consistent global scale from a pair of point maps. This normalization allows the regression error to be evaluated in a scale-invariant manner.

DUSt3R further extends this formulation to a confidence-aware loss, in which the 3D regression error is weighted by the predicted confidence map:

$$L_{\text{conf}} = \sum_{v \in 1,2} \sum_{i \in \mathcal{D}^v} C_i^{v,1} L_{\text{regr}}(v,i) - \alpha \log C_i^{v,1} \tag{4}$$

A regularization term is added to prevent the confidence values from collapsing uniformly toward zero.

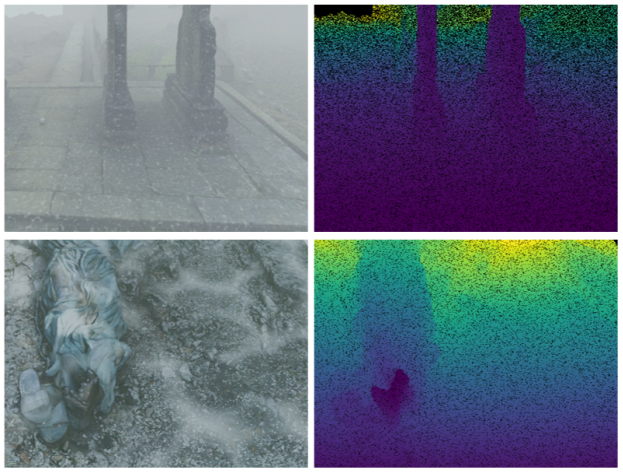
3.3.2 Derivation of the Physical Model An underwater image can be modeled as the sum of direct transmission from the object and scattered background light generated by particles suspended in the water. According to the Lambert–Beer law, light is exponentially attenuated as it travels through a participating medium (Carlevaris-Bianco et al., 2010).

$$I(x) = I'(x)\exp(-\beta d) + B_\infty \{1 - \exp(-\beta d)\} \tag{5}$$

Here, I denotes the observed color value, I' is the true scene radiance, d is the optical path length, β is the attenuation coefficient, and B_∞ represents the background light. It is widely known that absorption and scattering vary with wavelength; therefore, the attenuation coefficients differ for each spectral band (Chiang and Chen, 2011, Akkaynak and Treibitz, 2018). In this work, we model the attenuation using separate coefficients for each RGB channel.

3.3.3 Proposed Loss Function Design Using the underwater environment parameters predicted by the underwater head, we introduce a new loss term called the Lambert–Beer consistency loss. The underwater head predicts the wavelength-dependent attenuation coefficients and background colors for each RGB channel. In addition, the optical path length from each viewpoint is obtained from the 3D point maps estimated by DUSt3R.

Based on the Lambert–Beer formulation, the latent clean radiance can be estimated as:



(a) Rendering images (left) and corresponding depth maps (right).

(b) Additional rendering examples across different scenes.

Figure 2. Examples of the synthetic underwater dataset.

$$\hat{I}'(x) = \frac{I(x) - B_\infty \{1 - \exp(-\beta d)\}}{\exp(-\beta d)} \tag{6}$$

To enforce photometric consistency between the two viewpoints, we construct a soft matching matrix:

$$M_{ij} = \exp\left(-\frac{\left\| X_i^{1,1} - X_j^{2,1} \right\|^2}{2\sigma^2} \right) \tag{7}$$

which assigns higher weights to geometrically closer 3D point pairs in a fully differentiable manner. Using these weights, the Lambert–Beer Consistency Loss is defined as:

$$L_{\text{LB}} = \sum_{i \in \mathcal{D}^1} \sum_{j \in \mathcal{D}^2} M_{ij} \left\| \hat{I}_i' - \hat{I}_j' \right\|^2 \tag{8}$$

Finally, the overall loss function of the proposed model is given by:

$$L_{\text{total}} = L_{\text{conf}} + \lambda_{\text{LB}} L_{\text{LB}} \tag{9}$$

This combined loss jointly optimizes geometric consistency, confidence prediction, and physical underwater consistency, enabling more robust reconstruction under various underwater conditions.

4. Experimental Results and Discussion

This section first presents the characteristics and outcomes of the synthetic dataset generated using the proposed data generation pipeline. We then describe the training conditions of the improved DUSt3R model. Finally, we report both quantitative

Table 1. Multi-view depth performance comparison.

Method	rel ↓	τ ↑
Dust3R (baseline)	**0.106**	**0.279**
Enhanced Dust3R	0.117	0.196
Improvement	-0.284	-0.038

and qualitative results of multi-view depth estimation obtained using the trained model and discuss the observed trends.

4.1 Synthetic Dataset Overview

The synthetic dataset constructed in this study consists of 5,316 RGB images, the corresponding depth maps, and relative camera parameters for each image. Fig. 2 illustrates representative samples of the synthetic underwater dataset. Fig. 2(a) presents the generated RGB images and their associated depth maps for representative scenes, showing a gate on a cobble-stone surface and a statue on the ground. Fig. 2(b) shows additional rendering examples across different scenes. From left to right, the scenes depict a pedestal, a doll, and houses. The first two scenes exhibit pronounced caustic patterns, while the rightmost scene demonstrates strong turbidity effects.

From the figure, it can be observed that underwater-specific visual effects—such as locally enhanced illumination patterns caused by caustics and depth-dependent light attenuation—are realistically reproduced. The generated depth maps generally exhibit smooth and consistent structures, indicating that the dataset offers sufficiently reliable ground truth for supervised learning. However, a small number of regions show missing or invalid depth values due to rendering limitations. These artifacts were excluded during training, but they highlight areas where further refinement of the rendering pipeline is needed.

4.2 Training Details

We initialized the model using the 512-pixel-resolution pretrained DUSt3R model. To reduce computational cost and improve stability during training, all encoder and decoder layers except the final layer were frozen. The final layer was left trainable to allow adjustment to the newly introduced components, such as the FcaNet module and the underwater head.

Modules that do not exist in the original DUSt3R architecture—namely the FcaNet frequency attention module and the underwater head—were randomly initialized. This partial freezing strategy preserves the robustness of the base model while enabling effective learning of underwater-specific physical effects.

We further evaluated the sensitivity to the loss weight factor λ_{LB} for the Lambert–Beer consistency loss, which controls the balance between geometric regression and photometric consistency, by testing values of $\{0.01, 0.1, 1.0\}$. All configurations resulted in similar performance, and we report the results for $\lambda_{\mathrm{LB}} = 0.1$ in the following sections. The model was trained for 10 epochs using the AdamW optimizer with a learning rate of $1e-4$ and a batch size of 1, which was the maximum allowed by our GPU memory constraints.

4.3 3D Reconstruction Performance

Multi-view depth estimation was performed using the improved DUSt3R model. Specifically, 8 scenes were selected from the 112 total scenes, and for each scene, depth values were obtained

Figure 3. An example of input image (left), predicted depth map by the original DUSt3R (middle), and predicted depth map by the enhanced DUSt3R (right).

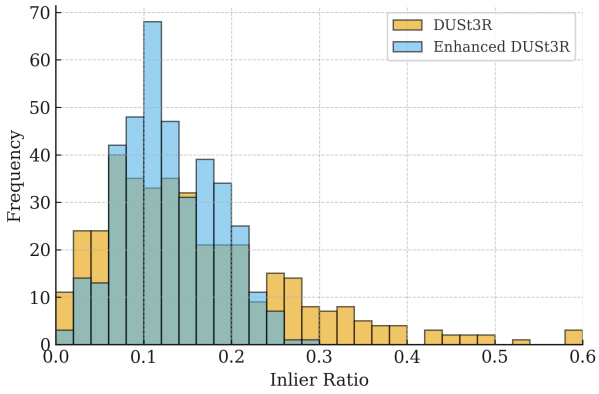


Figure 4. Comparison of Inlier Ratio (τ).

by extracting the z-coordinates from the point maps computed from 24 images. The resulting depth maps were merged using confidence-based weighted integration.

Quantitative evaluation was conducted using Absolute Relative Error (rel) and Inlier Ratio (τ) with a threshold of 1.03. The comparison results are summarized in Table 1. As shown, both metrics deteriorated when using the improved DUSt3R.

Examples of the estimated depth maps are shown in Fig. 3. In the input image, a bell-shaped object is placed on a planar surface on the left side, while the upper-right region corresponds to the background with a large depth value. The prediction by the original DUSt3R clearly distinguishes the bell-shaped object, the supporting planar surface, and the background, exhibiting well-defined depth discontinuities between these regions. In the predicted depth map by the improved model, the bell-shaped object visible in the image is not properly reconstructed, and the model fails to capture its actual geometry. The predicted depth becomes unrealistically deep toward the upper-right region, resulting only in a vague estimation of the object’s boundary rather than a faithful recovery of its shape. Furthermore, the depth map exhibits periodic stripe-like patterns, indicating instability in the geometric prediction.

Additionally, the histogram of inlier ratios (Fig. 4) shows that scenes with originally low inlier ratios benefited from the enhanced model, exhibiting an increase in correct correspondences. In contrast, scenes that previously achieved high inlier ratios tended to show reduced performance, indicating that the improvement is not uniform across all scenes.

The degradation in performance can be attributed to several factors. First, FcaNet enhances feature responses in the frequency domain; however, this may have unintentionally introduced periodic artifacts in the predicted depth maps, resulting in

grid-like patterns. Second, if the underwater attenuation model is not properly learned, the network may overfit to color consistency rather than geometric accuracy, causing the predicted depth to deviate from physically plausible values. Third, from a loss function design perspective, the inclusion of weak or unreliable correspondences in the photometric consistency loss due to the use of a soft matching matrix may have degraded the geometric estimation. Since such correspondences can relate unrelated pixels across images, they may introduce incorrect supervision signals, leading to inaccurate depth predictions.

Moreover, the use of image pairs may impose inherent limitations under shallow-water environments. In particular, when the appearance of the scene varies significantly between views due to caustics, light absorption and scattering, pairwise matching alone may be insufficient to resolve ambiguous correspondences. This lack of redundancy across multiple views may contribute to instability in geometric reconstruction. Finally, due to GPU memory constraints and the overall training time required, the batch size was restricted to 1, and the number of training epochs was limited to 10. These constraints likely led to unstable gradient estimation, and premature termination before convergence. Consequently, the model may not have reached a stable optimum, leading to the observed accuracy deterioration.

5. Conclusion

In this study, we explored a data-driven approach to underwater 3D reconstruction under shallow-water conditions, where strong caustics and wavelength-dependent attenuation significantly degrade image-based geometric estimation. Our primary contribution is the design of a DUSt3R-based architecture that explicitly incorporates underwater optical effects—namely light caustics, absorption, and scattering—into the depth estimation process To support this model, we constructed a physically informed synthetic data generation pipeline that produces paired RGB images, depth maps, and camera parameters while realistically simulating shallow-water light transport. In addition, we examined a frequency-domain enhancement strategy using FcaNet and investigated its potential effects on underwater geometry reconstruction.

Although the proposed model did not outperform the baseline DUSt3R and exhibited limited depth estimation accuracy, this work highlights several important challenges inherent to underwater 3D reconstruction. In particular, the combination of complex optical distortions and the current learning framework introduces ambiguity in correspondence estimation. The use of soft matching in the photometric consistency loss may allow weak or unreliable correspondences to influence optimization, while the pairwise formulation may lack sufficient redundancy to resolve such ambiguities under severe appearance variations. Furthermore, the model suffered from insufficient convergence due to computational constraints, and the intended effect of the underwater optical constraints was not fully realized. In addition, the applicability of the proposed approach to real underwater imagery remains to be demonstrated.

Future work will focus on addressing these limitations. First, improving the training strategy, including stabilization with larger batch sizes and longer training schedules, will be essential for achieving reliable performance. Second, more robust correspondence modeling, such as improved weighting strategies or filtering of unreliable matches, may help mitigate the impact of ambiguous correspondences. Third, extending the framework beyond pairwise matching to incorporate multi-view geometric constraints could improve reconstruction stability under complex optical conditions. Furthermore, applying the model to real shallow-water environments will require explicitly accounting for refractive effects at interfaces between different media, such as water, glass, and air, introduced by the camera housing. Finally, deeper integration of physically based rendering and underwater light transport simulation may provide more accurate supervision and enable more robust underwater 3D reconstruction.

References

Agrawal, A., Ramalingam, S., Taguchi, Y., Chari, V., 2012. A theory of multi-layer flat refractive geometry. *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 3346–3353.

Akkaynak, D., Treibitz, T., 2018. A revised underwater image formation model. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6723–6732.

Carlevaris-Bianco, N., Mohan, A., Eustice, R. M., 2010. Initial results in underwater single image dehazing. *Oceans 2010 Mts/IEEE Seattle*, IEEE, 1–8.

Casella, E., Scicchitano, G., Rovere, A., 2024. Accuracy and precision of shallow-water photogrammetry from the sea surface. *Remote Sensing*, 16(22), 4321.

Chiang, J. Y., Chen, Y.-C., 2011. Underwater image enhancement by wavelength compensation and dehazing. *IEEE transactions on image processing*, 21(4), 1756–1769.

Cong, X., Zhao, Y., Gui, J., Hou, J., Tao, D., 2024. A comprehensive survey on underwater image enhancement based on deep learning. *arXiv preprint arXiv:2405.19684*.

Drap, P., Merad, D., Hijazi, B., Gaoua, L., Nawaf, M. M., Saccone, M., Chemisky, B., Seinturier, J., Sourisseau, J.-C., Gambin, T. et al., 2015. Underwater photogrammetry and object modeling: a case study of Xlendi Wreck in Malta. *Sensors*, 15(12), 30351–30384.

Forbes, T., Goldsmith, M., Mudur, S., Poullis, C., 2018. DeepCaustics: Classification and removal of caustics from underwater imagery. *IEEE Journal of Oceanic Engineering*, 44(3), 728–738.

Gracias, N., Negahdaripour, S., Neumann, L., Prados, R., Garcia, R., 2008. A motion compensated filtering approach to remove sunlight flicker in shallow water images. *OCEANS 2008*, IEEE, 1–7.

Henderson, J., Pizarro, O., Johnson-Roberson, M., Mahon, I., 2013. Mapping submerged archaeological sites using stereo-vision photogrammetry. *International Journal of Nautical Archaeology*, 42(2), 243–256.

Hu, K., Wang, T., Shen, C., Weng, C., Zhou, F., Xia, M., Weng, L., 2023. Overview of underwater 3D reconstruction technology based on optical images. *Journal of Marine Science and Engineering*, 11(5), 949.

Jian, M., Liu, X., Luo, H., Lu, X., Yu, H., Dong, J., 2021. Underwater image processing and analysis: A review. *Signal Processing: Image Communication*, 91, 116088.

Jordt-Sedlazeck, A., Koch, R., 2012. Refractive calibration of underwater cameras. *European conference on computer vision*, Springer, 846–859.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 139–1.

Leon, J. X., Roelfsema, C. M., Saunders, M. I., Phinn, S. R., 2015. Measuring coral reef terrain roughness using ‘Structure-from-Motion’close-range photogrammetry. *Geomorphology*, 242, 21–28.

Leroy, V., Cabon, Y., Revaud, J., 2024. Grounding image matching in 3d with mast3r. *European Conference on Computer Vision*, Springer, 71–91.

Levy, D., Peleg, A., Pearl, N., Rosenbaum, D., Akkaynak, D., Korman, S., Treibitz, T., 2023. Seathru-nerf: Neural radiance fields in scattering media. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 56–65.

Lv, Q., Dong, J., Li, Y., Chen, S., Yu, H., Zhang, S., Wang, W., 2025. UWStereo: A large synthetic dataset for underwater stereo matching. *IEEE Transactions on Circuits and Systems for Video Technology*.

Mangeruga, M., Cozza, M., Bruno, F., 2018. Evaluation of underwater image enhancement algorithms under different environmental conditions. *Journal of Marine Science and Engineering*, 6(1), 10.

Marre, G., Deter, J., Holon, F., Boissery, P., Luque, S., 2020. Fine-scale automatic mapping of living Posidonia oceanica seagrass beds with underwater photogrammetry. *Marine Ecology Progress Series*, 643, 63–74.

Melchers, R. E., Jeffrey, R. J., 2014. Corrosion of steel piling in seawater harbours. *Proceedings of the Institution of Civil Engineers-Maritime Engineering*, 167number 4, Thomas Telford Ltd, 159–172.

Menna, F., McAvoy, S., Nocerino, E., Tanduo, B., Giuseffi, L., Calantropio, A., Chiabrando, F., Teppati Losè, L., Lingua, A. M., Sandin, S. et al., 2024. POSER: an oPen sOurce Simulation platform for tEaching and tRaining underwater photogrammetry. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 265–272.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99–106.

Qin, Z., Zhang, P., Wu, F., Li, X., 2021. Fcanet: Frequency channel attention networks. *Proceedings of the IEEE/CVF international conference on computer vision*, 783–792.

Sauder, J., Tuia, D., 2024. Self-supervised underwater caustics removal and descattering via deep monocular slam. *European Conference on Computer Vision*, Springer, 214–232.

Sedlazeck, A., Koch, R., 2012. Perspective and non-perspective camera models in underwater imaging–overview and error analysis. *Outdoor and Large-Scale Real-World Scene Analysis: 15th International Workshop on Theoretical Foundations of Computer Vision, Dagstuhl Castle, Germany, June 26-July 1, 2011. Revised Selected Papers*, Springer, 212–242.

Shi, J., Xu, J., He, J., Lin, Z., 2024. Aquags: Fast underwater scene reconstruction with sfm-free gaussian splatting. *2024 IEEE 10th International Conference on Underwater System Technology: Theory and Applications (USYS)*, IEEE, 1–6.

Shortis, M., 2019. Camera calibration techniques for accurate measurement underwater. *3D recording and interpretation for maritime archaeology*, 11–27.

Sun, Y., Zhou, T., Zhang, L., Chai, P., 2024. Underwater camera calibration based on double refraction. *Journal of Marine Science and Engineering*, 12(5), 842.

Trabes, E., Jordan, M. A., 2015. Self-tuning of a sunlight-deflickering filter for moving scenes underwater. *2015 XVI Workshop on Information Processing and Control (RPIC)*, IEEE, 1–6.

Tripathy, A., Singh, M., Roy, H., 2020. Clustering based approach for underwater sunlight flicker removal. *Global Oceans 2020: Singapore–US Gulf Coast*, IEEE, 1–7.

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D., 2025. Vggt: Visual geometry grounded transformer. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5294–5306.

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J., 2024. Dust3r: Geometric 3d vision made easy. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20697–20709.

Wu, T., Hou, S., Sun, W., Shi, J., Yang, F., Zhang, J., Wu, G., He, X., 2023. Visual measurement method for three-dimensional shape of underwater bridge piers considering multirefraction correction. *Automation in Construction*, 146, 104706.

Yang, J., Sax, A., Liang, K. J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M., 2025. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21924–21935.

Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L., 2020. BlendedMVS: A Large-scale Dataset for Generalized Multi-view Stereo Networks. *Computer Vision and Pattern Recognition (CVPR)*.

Zhang, W., Wu, Y., Li, S., Ma, W., Ma, X., Li, Q., Wang, Q., 2025. Review of Feed-forward 3D Reconstruction: From DUSt3R to VGGT. *arXiv preprint arXiv:2507.08448*.

Zhao, W., Queralta, J. P., Westerlund, T., 2020. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. *2020 IEEE symposium series on computational intelligence (SSCI)*, IEEE, 737–744.

Zwilgmeyer, P. G. O., Yip, M., Teigen, A. L., Mester, R., Stahl, A., 2021. The varos synthetic underwater data set: Towards realistic multi-sensor underwater data with ground truth. *Proceedings of the IEEE/CVF international conference on computer vision*, 3722–3730.