Underwater Visual SLAM with Depth Uncertainty and Medium Modeling

Rui Liu Sheng Fan Wenguan Wang Yi Yang*
ReLER, CCAI, Zhejiang University
https://defaultrui.github.io/dus

Abstract

Underwater visual simultaneous localization and mapping (SLAM) faces critical challenges in light attenuation and degraded geometric consistency. Despite recent advances of visual SLAM in indoor and urban scenes, these approaches typically assume a clear medium and neglect medium-light interactions, leading to performance degradation in underwater environments. To overcome these limitations, we propose DUV-SLAM, a dense underwater visual SLAM framework that integrates uncertainty-aware geometry estimation with physics-inspired neural scattering modeling. Our method introduces two core innovations: i) depth uncertainty quantification derived from differentiable bundle adjustment, which propagates geometric confidence to guide mapping optimization; and ii) a neural-Gaussian hybrid representation that combines adaptive 3D Gaussians for underwater reconstruction with a neural field capturing wavelength-dependent medium properties, optimized using a combination of photometric, geometric, and distribution losses. Experiments on synthetic and real-world datasets demonstrate that DUV-SLAM achieves high-quality monocular reconstruction while maintaining real-time efficiency and robust tracking accuracy.

1. Introduction

Background. Autonomous underwater vehicles have garnered significant attention for applications such as ocean exploration and seabed mapping [38, 46]. Traditional localization systems are constrained by severe satellite signal loss and inertial drift. In contrast, underwater simultaneous localization and mapping (SLAM) utilizes onboard sensors, *e.g.*, sonars or cameras, to incrementally build a map of underwater environments (*mapping*) while concurrently estimating vehicle position (*tracking*) [32, 75]. Unlike sonar-based SLAM [8, 47], visual SLAM provides a lightweight, cost-effective solution that maintains original hydrodynamics and controllability of vehicles [42, 60]. Although visual SLAM has been extensively studied in indoor and urban environments [9, 13, 62, 63], its applications in underwater domains are relatively underexplored [23].

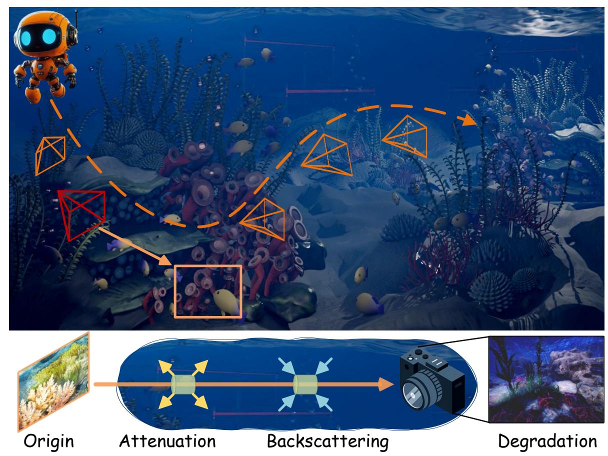

Figure 1. Due to light attenuation and backscattering in water medium, underwater visual SLAM introduces unique challenges for both *tracking* and *mapping* with degraded observations.

Motivation. Despite notable advancements in visual SLAM with recent Neural Radiance Fields (NeRF) [40] and 3D Gaussian Splatting (3DGS) [22], existing methods remain fragile in unbounded underwater environments due to several inherent challenges [25, 49] (see Fig. 1). ❶ Light attenuation and scattering induce wavelength-dependent color distortion and depth-dependent visibility degradation [1]. Such illumination fluctuations impair the reliability of feature correspondences, thereby destabilizing state estimation in *tracking*. Current NeRF- or 3DGS-based approaches [19, 39, 91], which rely on clear-medium priors, fail to accurately model these effects, leading to suboptimal rendering results for *mapping*. ❷ Furthermore, degraded image quality hinders reliable depth estimation during geometric initialization, amplifying outlier rates [10, 29]. This violation of geometric consistency introduces uncertainty in map reconstruction and impede optimization convergence [16]. These challenges cause cascading degradation in SLAM pipelines, perturbing bundle adjustment and ultimately corrupting scene representations [50, 51]. While recent underwater reconstruction techniques [26, 28, 66]

*Corresponding author: Yi Yang.

model medium-light interactions via offline computation, real-time state estimation and online mapping in large-scale underwater scenarios remain understudied [3, 48].

Method. In this work, we propose DUV-SLAM, a dense underwater visual SLAM framework that explicitly quantifies depth uncertainty and reconstructs underwater scenes as a hybrid 3DGS representation coupled with a neural scattering medium. The framework operates through two parallel threads, *i.e.*, *tracking* and *mapping*. In the *tracking* thread (§3.1), DUV-SLAM iteratively estimates camera poses and depths via a differentiable neural solver, leveraging recurrent dense optical flow for ordered images [68]. Depth uncertainty is derived from variance matrices in differentiable bundle adjustment, serving as a geometry prior for reliable underwater reconstruction. Meanwhile, DUV-SLAM builds a globally consistent 3D map using an adaptive 3DGS representation enhanced with differentiable medium properties (§3.2). The learnable parameters of the medium are conditioned on viewing angles per color channel, enforcing constancy along rays. The *mapping* process is jointly optimized by a photometric loss during rendering and a geometric loss weighted by the depth uncertainty. Moreover, due to sparse geometric cues from limited initial frames, which exacerbate convergence instability in noisy underwater scenes, a distribution loss is used to regularize the optimization. This loss penalizes deviations from tracked depth priors by aligning the α-blended depth histogram with uncertainty-derived confidence bounds, thereby ensuring geometric consistency.

Specifically, to address challenge ❶, our DUV-SLAM introduces a physics-inspired hybrid representation that models the entire 3D volume of underwater environments. This is achieved by integrating explicit 3DGS geometry with a separate neural scattering field to capture medium properties. For challenge ❷, depth variance is initially derived from the end-to-end *tracking* to quantify state uncertainty and then propagated into the *mapping* stage for confidence-aware metric reconstruction. Through uncertainty-guided optimization, DUV-SLAM effectively suppresses noise, achieving robust scene reconstruction.

Benchmarking. To the best of our knowledge, this work is the first to comprehensively benchmark performance degradation of existing visual SLAM methods in both real-world and simulated underwater environments [3, 48, 74], revealing critical limitations in tracking accuracy and mapping quality. We evaluate DUV-SLAM against existing monocular frameworks on both synthetic and real-world underwater datasets [3, 48, 74] (§4). Experimental results indicate that DUV-SLAM achieves high-quality reconstruction while maintaining competitive tracking accuracy and speed, *e.g.*, $\mathbf{2.40}$m ATE and $\mathbf{29.44}$ PSNR with $\mathbf{26.52}$ FPS *vs* 3.37m ATE and 24.16 PSNR with 23.09 FPS of Photo-SLAM [19] in real-world scenes [48]. This demonstrates the potential of our method in practical applications.

2. Related Work

In this section, we briefly review the most relevant studies about visual SLAM and underwater reconstruction due to limited space, while referring readers to comprehensive surveys [4, 8, 14, 45, 65, 71, 83, 86, 87] for more references.

Dense Visual SLAM. Early visual SLAM systems [5, 11, 41, 53, 70] rely on hand-crafted features and graph optimization, achieving remarkable real-time performance. Later approaches [67, 69] leverage deep neural networks for robust feature extraction, improving both localization and mapping accuracy. Dense map representations have been explored using point clouds [54], grids [35, 44, 58], and meshes [72] for navigation [12, 33, 34]. Recent advances in NeRF [40] and 3DGS [22] enable high-quality spatial representations with continuous surface modeling and reduced memory overhead. To optimize camera poses, several approaches [21, 39, 64, 78, 91] minimize dense photometric errors between rendered and observed images. To enhance tracking efficiency, decoupled systems [7, 17, 19, 37, 50, 84, 85] integrate additional feature-based or end-to-end *tracking* modules [5, 69] with NeRF- or 3DGS-based *mapping*.

Despite these advancements, modern visual SLAM techniques remain largely unexplored in underwater environments due to the complex optical conditions. In this work, our DUV-SLAM integrates physics-based medium modeling into 3DGS-based mapping and achieves accurate *tracking* and *mapping* via uncertainty-guided optimization. Furthermore, we systematically evaluate and benchmark existing approaches in underwater environments.

Underwater Reconstruction. With the increasing deployment of cameras, visual SLAM has emerged as an alternative to traditional sonar- and bathymetry-based approaches [25, 45]. However, underwater imaging is severely affected by light attenuation and scattering, leading to image distortion and degraded features [14, 76]. To recover color information for 3D reconstruction, deep learning-based methods [2, 57, 61] alleviate caustic effects based on a revised underwater image formation model [1, 30, 31]. NeRF-based underwater reconstruction [26, 59, 66, 81, 89] decomposes the NeRF rendering equations into direct and backscatter components for learning both scene geometry and medium parameters. Recently, there has been a surge of 3DGS-based studies released on arXiv. These concurrent studies [6, 28, 36, 73, 79, 88, 90] adopt 3DGS as an explicit representation, improving training efficiency and rendering performance on several underwater images.

The proposed DUV-SLAM performs online state estimation and concurrent underwater reconstruction. By aligning depth information obtained from iterative tracking and rendering, DUV-SLAM further optimizes both *tracking* and *mapping*. Moreover, depth uncertainty quantification is introduced to weigh this alignment process, effectively suppressing outliers and improving robustness.

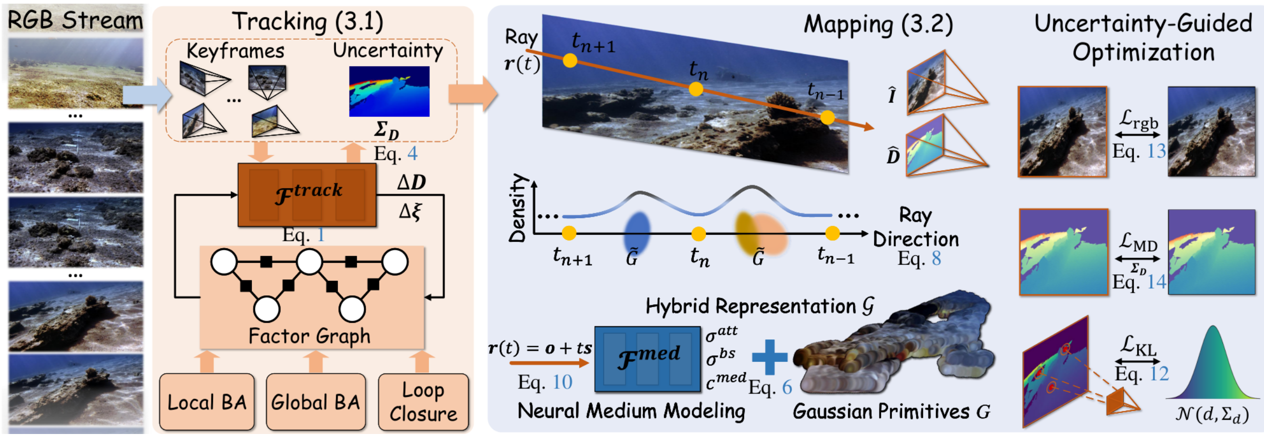

Figure 2. Overview of DUV-SLAM. Given an input RGB stream, the *tracking* module $\mathcal{F}^{track}$ iteratively estimates camera pose and depth using recurrent optical flow while maintaining a factor graph for BA and loop closure (§3.1). Depth uncertainty, derived from BA, is incorporated as a geometric prior into *mapping*. The underwater scene is represented by a set of Gaussian primitives $\mathcal{G}$ coupled with a neural medium model $\mathcal{F}^{med}$, both of which are jointly optimized via uncertainty-guided learning (§3.2). See §3 for more details.

Uncertainty in SLAM. Early study [70] adopts extended Kalman filters for mapping and introduces uncertainty modeling to enhance robustness. [80] leverages deep networks to jointly estimate depth, pose, and photometric uncertainty in visual odometry. Building upon this, neural SLAM methods [10, 27, 50, 51, 55] use online uncertainty learning for dense depth estimation, enabling adaptive loss reweighting in tracking and mapping via self-supervised training. Similarly, 3DGS-based frameworks [18, 29, 77] utilize depth uncertainty to selectively optimize Gaussian primitives, improving geometric stability in monocular systems. Beyond these domain-specific strategies, recent researches [24, 43, 82] explore out-of-domain uncertainty estimation using statistical measures for generalization.

Given the inherent noise for underwater SLAM, we incorporate depth uncertainty quantification to regularize the rendering loss and mitigate drift in tracking. This improves both reconstruction fidelity and pose estimation accuracy in challenging underwater environments.

3. Method

Overview. Given an input RGB sequence $\mathcal{I} = \{\boldsymbol{I}_k \in \mathbb{R}^{HW \times 3}\}_{k=0}^{K}$, our method jointly estimates the state trajectory $\mathcal{S} = \{\boldsymbol{P}_k, \boldsymbol{D}_k\}_{k=0}^{K}$ with camera poses $\boldsymbol{P}_k \in \mathrm{SE}(3)$ and depths $\boldsymbol{D}_k \in \mathbb{R}^{HW}$ (*tracking*), while incrementally reconstructing a 3D map $\mathcal{G}$ of underwater environments (*mapping*). During tracking $\mathcal{F}^{track}(\mathcal{I}) \rightarrow \{\mathcal{S}, \boldsymbol{\Sigma_D}\}$ (§3.1), the depth covariance $\boldsymbol{\Sigma_D}$ is introduced to describe the state uncertainty in noisy conditions. To model the complex process of underwater light propagation, we integrate learnable parameters of scattering medium into 3D Gaussian Splatting (3DGS)-based representation of $\mathcal{G}$ (§3.2), jointly optimizing scene geometry and attenuation coefficient during *mapping*. Crucially, the depth uncertainty $\boldsymbol{\Sigma_D}$ is served as a geometric prior to suppress noise and optimize reconstruction with geometric consistency, *i.e.*, $\mathcal{F}^{map}(\mathcal{I}, \mathcal{S}, \boldsymbol{\Sigma_D}) \rightarrow \mathcal{G}$ (Fig. 2).

3.1. Tracking with State Uncertainty

To enable robust pose estimation in degraded underwater visibility, we formulate $\mathcal{F}^{track}$ as a differentiable neural solver that integrates a pretrained recurrent optical flow network [68] into tracking. The solver maintains a hierarchical factor graph $(\mathcal{V}, \mathcal{E})$, where nodes $\mathcal{V}$ store camera poses and inverse depth of keyframes, and edges $\mathcal{E}$ represent optical flow between pairs. For each incoming frame, we compute mean optical flow magnitude across frames to trigger dynamic graph maintenance by inserting the new keyframe and pruning redundant frames. In addition, multi-scale bundle adjustment (BA) is performed over $(\mathcal{V}, \mathcal{E})$ through local BA, loop closure, and global BA.

BA. Local BA operates over a sliding window of overlapping keyframes, optimizing camera poses and inverse depths through a dense BA layer [69]. The Mahalanobis reprojection error is minimized for correspondences:

$$\underset{\boldsymbol{P}, \boldsymbol{D}}{\arg\min} \sum_{(j,k) \in \mathcal{E}} \left\| \boldsymbol{p}_{j,k}^{*} - \Pi_c\left(\boldsymbol{P}_k \circ \boldsymbol{P}_j^{-1} \circ \Pi_c^{-1}(\boldsymbol{p}_j, \boldsymbol{D}_j)\right) \right\|_{\boldsymbol{\Sigma}_{jk}}^{2}, \quad (1)$$

where $\boldsymbol{p}_{j,k}^{*} \in \mathbb{R}^{HW \times 2}$ denotes optical flow-corrected correspondences [68] between frames $(j, k) \in \mathcal{E}$. Π_c^{-1} back-projects 2D pixels $\boldsymbol{p}_j$ with the depth $\boldsymbol{D}_j$ to 3D points, while the camera intrinsic parameters Π_c projects transformed 3D points to frame k via relative pose $\boldsymbol{P}_k \circ \boldsymbol{P}_j^{-1}$. The diagonal covariance matrix $\boldsymbol{\Sigma}_{jk}$ encodes per-pixel optical flow confidence, downweighting noisy measurements underwater.

State Uncertainty Quantification. To balance computational efficiency with uncertainty awareness, we compute Jacobians $\boldsymbol{J}(\boldsymbol{\xi}, \boldsymbol{D})$ exclusively for active keyframes within the optimization window. The pose and depth are updated iteratively by lie algebra $\boldsymbol{\xi} \in \mathbb{R}^6$ and additive corrections:

$$\hat{\boldsymbol{P}} = \mathrm{Exp}(\Delta \boldsymbol{\xi}) \circ \boldsymbol{P}, \quad \hat{\boldsymbol{D}} = \Delta \boldsymbol{D} + \boldsymbol{D}. \quad (2)$$

The damped Gauss-Newton solver iteratively computes increments $(\Delta\boldsymbol{\xi}, \Delta\boldsymbol{D})$ using the block-sparse Hessian $\boldsymbol{H}$:

$$\begin{bmatrix} \boldsymbol{H}_{\xi\xi} & \boldsymbol{H}_{\xi D} \\ \boldsymbol{H}_{D\xi} & \boldsymbol{H}_{DD} \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\xi} \\ \Delta\boldsymbol{D} \end{bmatrix} = -\boldsymbol{J}^\top(\boldsymbol{\xi}, \boldsymbol{D}), \tag{3}$$

with pose blocks $\boldsymbol{H}_{\xi\xi}$, diagonal depth terms $\boldsymbol{H}_{DD}$, and cross-terms $\boldsymbol{H}_{\xi D} = \boldsymbol{H}_{D\xi}^\top$. State uncertainty $(\boldsymbol{\Sigma_P}, \boldsymbol{\Sigma_D})$ is derived via Schur complement marginalization:

$$\begin{aligned} \boldsymbol{\Sigma_P} &= \left[\boldsymbol{H}_{\xi\xi} - \boldsymbol{H}_{\xi D}\boldsymbol{H}_{DD}^{-1}\boldsymbol{H}_{D\xi}\right]^{-1}, \\ \boldsymbol{\Sigma_D} &= \boldsymbol{H}_{DD}^{-1} + \boldsymbol{H}_{DD}^{-\top}\boldsymbol{H}_{D\xi}\boldsymbol{\Sigma_P}\boldsymbol{H}_{\xi D}\boldsymbol{H}_{DD}^{-1}, \end{aligned} \tag{4}$$

where $\boldsymbol{\Sigma_P}$ leverages sparse Cholesky decomposition for efficient computation [15, 51, 69]. The depth uncertainty $\boldsymbol{\Sigma_D}$ serves as an adaptive geometric prior during map optimization (Eq. 13). This uncertainty propagation mechanism is critical for maintaining geometric consistency and reconstruction fidelity in low-visibility conditions.

Loop Closure and Global BA. To mitigate accumulated drift, our system integrates loop closure detection and online global BA. Loop candidates are identified by thresholding the mean optical flow magnitude between the active keyframe window and historical keyframes. For mapping consistency, we periodically execute global BA over the complete keyframe set via a sparse pose graph optimization. This hierarchical approach maintains local accuracy while ensuring global consistency in underwater conditions.

3.2. Mapping with Underwater Medium Modeling

The *mapping* module $\mathcal{F}^{map}$ processes input frames $\mathcal{I} = \{\boldsymbol{I}_k\}_{k=0}^K$ with per-pixel depth uncertainty $\boldsymbol{\Sigma_D}$ from $\mathcal{F}^{track}$. We model the underwater scene as a hybrid 3D Gaussian Splatting (3DGS) representation $\mathcal{G} = \{G_i\}_{n=1}^N$ coupled with a neural scattering medium $\mathcal{F}^{med}$ that predicts attenuation coefficients. $\mathcal{G}$ and $\mathcal{F}^{med}$ are optimized jointly through a differentiable rendering pipeline.

Volume Rendering with Medium. To model light transport in underwater environments, we extend classical volume rendering [20, 40] by explicitly accounting for scattering medium rather than clear air. For a camera ray $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{s}$ with origin $\boldsymbol{o}$ and direction $\boldsymbol{s}$, the rendered color $C(\boldsymbol{r})$ integrates contributions from both scene instances and the participating medium:

$$\begin{aligned} C(\boldsymbol{r}) &= \int_0^\infty T(t)\big(\sigma^{ins}(t)c^{ins}(t) + \sigma^{med}(t)c^{med}(t)\big)dt, \\ T(t) &= \exp\Big(-\int_0^\infty \big(\sigma^{ins}(t) + \sigma^{med}(t)\big)dt\Big), \end{aligned} \tag{5}$$

where $c^{med}(t)/c^{ins}(t)$ and $\sigma^{med}(t)/\sigma^{ins}(t)$ are color and density of underwater instances and medium at position t. The transmittance $T(t)$ follows the Beer-Lambert law, attenuated exponentially by both scene and medium densities. t is the Euclidean distance along the ray, and is proportional to the estimated depth $\boldsymbol{D}$ along the camera's optical axis.

3DGS with Differentiable Medium. To achieve high-fidelity reconstruction, we represent underwater scenes through adaptive 3DGS [22] enhanced with learnable medium properties. Each 3D Gaussian primitive G is based on a Gaussian kernel defined by the covariance matrix $\boldsymbol{\Sigma}_G \in \mathbb{R}^{3\times3}$ and mean $\boldsymbol{\mu} \in \mathbb{R}^3$. During rendering, 3D Gaussians are projected onto 2D image space via perspective projection and corresponding Jacobian of the affine approximation. The resulting 2D Gaussians $\widetilde{G}$ are described by $\widetilde{\boldsymbol{\mu}} \in \mathbb{R}^2$ and $\widetilde{\boldsymbol{\Sigma}}_G \in \mathbb{R}^{2\times2}$. The pixel color $C(\boldsymbol{r})$ along ray integrates contributions from depth-ordered Gaussians $\{\widetilde{G}\}_{n=1}^N$ and the participating medium by α-blending. By subdividing the ray propagation path into N sampling intervals, Eq. 5 is discretized as:

$$C(\boldsymbol{r}) = \sum\nolimits_{n=1}^N C_n^G(\boldsymbol{r}) + C_\infty^{med}(\boldsymbol{r}), \tag{6}$$

where $C_n^G(\boldsymbol{r})$ denotes the color contribution from the Gaussian field over the n-th interval, and $C_\infty^{med}(\boldsymbol{r})$ accounts for the background medium beyond the last Gaussian $\widetilde{G}_N$. Assuming that the instance color c_n^{ins} and the blending coefficients α_n^{ins} remain constant within the interval $[t_{n-1}, t_n]$, the per-interval contribution can be expressed as:

$$C_n^G(\boldsymbol{r}) = T_n(t)c_n^{ins}\alpha_n^{ins} + \int_{t_{n-1}}^{t_n} T_n(t)\sigma^{med}c^{med}dt, \tag{7}$$

where α_n^{ins} is the product of the primitive's opacity and density, and $T_n(t)$ is the transmittance between $\widetilde{G}_{n-1}$ and $\widetilde{G}_n$. Specifically, $T_n(t)$ is decomposed as:

$$T_n(t) = T_n^{ins} \circ T_n^{med} = \prod\nolimits_{i=1}^{n-1}(1-\alpha_i^{ins}) \circ \exp(-\sigma^{med}t). \tag{8}$$

where the instance term T_n^{ins} captures the occlusion effect from preceding Gaussian primitives, and the medium term T_n^{med} accounts for the attenuation. For the first term in Eq. 7, we fix $T_n(t) = T_n(t_n)$ as a constant, while $T_n(t)$ is integrated over the interval for the second term. Owing to the unbounded nature of underwater environments, the background medium contribution is modeled as:

$$C_\infty^{med}(\boldsymbol{r}) = \int_{t_N}^\infty T_n(t)\sigma^{med}c^{med}dt. \tag{9}$$

Underwater-Adaptive Attenuation Model. Although general volume rendering models (*e.g.*, for haze or smoke) use a uniform medium attenuation term $\exp(-\sigma^{med}t)$ as in Eq. 5. Recent work [1] on underwater image formation indicates that underwater light propagation differs fundamentally from that in the atmosphere. In particular, the direct signal and the backscattered signal attenuate at different rates. The rendering equations Eq. 6 and 7 currently employ an identical attenuation coefficient σ^{med} for both the instance and medium components. To account for the distinct underwater properties, we introduce separate parameters σ^{att} and σ^{bs} into $T_n(t)$, *i.e.*, $\sigma^{med} = \sigma^{att}$ for the instance term and $\sigma^{med} = \sigma^{bs}$ for the medium term. Based on a wavelength-dependent optical model [1], we disentangle these effects through a spectral neural medium model $\mathcal{F}^{med}$, which takes the position-encoded viewing direction $\boldsymbol{s}$ as input and yields constant values along each ray $\boldsymbol{r}$:

$$(c^{med}, \sigma^{att}, \sigma^{bs}) = \mathcal{F}^{med}(t, \boldsymbol{s}). \tag{10}$$

While conventional underwater image formation [1] relies on raw sensor data, our method captures medium properties with differentiable photometric consistency, enabling effective learning directly from processed RGB images. This physics-informed modeling enhances wavelength-aware rendering quality in underwater conditions.

Uncertainty-Guided Optimization. To enhance reconstruction fidelity in noisy underwater conditions, we propose a probabilistic depth supervision mechanism that leverages tracking-derived uncertainty $\boldsymbol{\Sigma_D}$ (Eq. 4). The rendered depth $\hat{\boldsymbol{D}}(\boldsymbol{r})$ for pixel ray $\boldsymbol{r}$ is computed as the expected termination distance via α-blending:

$$\boldsymbol{d}(\boldsymbol{r}) = \frac{\sum_{n=1}^{N} T_n(t)\alpha_n^{ins} d_n}{\sum_{n=1}^{N} T_n(t)\alpha_n^{ins}}, \tag{11}$$

where d_n denotes the distance to the mean $\widetilde{\boldsymbol{\mu}}$ of Gaussian $\widetilde{G}_n$ along the camera's optical axis. Inverse depth is employed by default to improve numerical stability and convergence. We model the tracked depth $\boldsymbol{D}$ as a normal distribution $\mathcal{N}(t; d, \Sigma_d)$ per pixel and formulate depth consistency through KL divergence minimization:

$$\mathcal{L}_{\text{KL}} = \text{KL}\big[\mathcal{N}(\boldsymbol{D}(\boldsymbol{r}), \boldsymbol{\Sigma_D}(\boldsymbol{r})) \| \sum_{n=1}^{N} \tilde{T}_n \tilde{\alpha}_n \delta(d_n - \hat{D}(\boldsymbol{r}))\big], \tag{12}$$

where $\tilde{T}_n(t)$ and $\tilde{\alpha}_n$ are the normalized weights of transmittance and blending. This aligns the rendered depth histogram with the uncertainty-aware estimates from tracking.

The mapping objective $\mathcal{L}_{\text{map}}$ combines a photometric loss $\mathcal{L}_{\text{rgb}}$ and a geometric loss $\mathcal{L}_{\text{depth}}$:

$$\begin{aligned} \mathcal{L}_{\text{rgb}} &= \mathcal{L}_1(\hat{\boldsymbol{I}}, \boldsymbol{I}) + \mathcal{L}_{\text{DSSIM}}(\hat{\boldsymbol{I}}, \boldsymbol{I}), \\ \mathcal{L}_{\text{depth}} &= \mathcal{L}_{\text{MD}}(\hat{\boldsymbol{D}}, \boldsymbol{D}) + \mathcal{L}_{\text{RGB-C}}(\hat{\boldsymbol{D}}) + \mathcal{L}_{\text{KL}}, \end{aligned} \tag{13}$$

where $\mathcal{L}_{\text{RGB-C}}(\hat{\boldsymbol{D}})$ enforces cross-channel depth consistency and $\mathcal{L}_{\text{MD}}$ denotes the Mahalanobis distance weighted by tracking uncertainty $\boldsymbol{\Sigma_D}$:

$$\mathcal{L}_{\text{MD}}(\hat{\boldsymbol{D}}, \boldsymbol{D}_{track}) = \|\hat{\boldsymbol{D}} - \boldsymbol{D}\|^2_{\boldsymbol{\Sigma_D}}. \tag{14}$$

3.3. Implementation Details

Network Architecture. Our pipeline consists of parallel *tracking* and *mapping* threads. The *tracking* thread continuously minimizes the BA reprojection error within an active window of keyframes. The downsampled depth $\boldsymbol{D}$ enables efficient update followed by a learned upsampling operation [68] to recover full-resolution depth for *mapping*. The *mapping* thread optimizes all keyframes received from the *tracking* thread. For medium encoding, we use two linear layers with 64 features and a softplus activation, followed by three branches of dense layers: one with a sigmoid activation for predicting c^{med}, and two with softplus activations for σ^{att} and σ^{bs}. We modify the rasterization used in original 3DGS to incorporate depth rendering, introducing an additional gradient computation branch in the backward process for depth loss. Jointly optimization of 3DGS and water medium parameters is achieved by minimizing a combination of photometric loss and geometric loss (Eq. 13).

Gradient Scaling of Gaussian Primitives. In standard 3DGS, adaptive Gaussian densification enhances representation capability based on the 2D color gradient $\frac{\partial \mathcal{L}_{\text{rgb}}}{\partial C(\boldsymbol{r})}$. However, in underwater scenes, the presence of the attenuation factor $\exp(-\sigma^{med} t)$ causes gradients to diminish with distance t, limiting Gaussian splitting and cloning in distant regions, resulting in sparse distributions. To mitigate this effect, we adopt a refined densification strategy [52, 73] as $\frac{1}{\text{detach}(\exp(-\sigma^{med} t))} \cdot \frac{\partial \mathcal{L}_{\text{rgb}}}{\partial C_n(\boldsymbol{r})}$, where detach$(\cdot)$ denotes stopping gradient of its argument. At each densification and pruning step, the moving averages in the Adam optimizer of medium encoding are reset to maintain stability.

Reproducibility. Experiments are conducted on a desktop with an Intel i9-14900K and an NVIDIA RTX 4090 GPU. 3DGS rasterisation and gradients are computed by CUDA, while the remaining pipeline is developed in PyTorch.

4. Experiment

Datasets. We evaluate DUV-SLAM on one real-world dataset, FLSea [48], and two synthetic datasets, MIMIR-UW [3] and TartanAir-Ocean [74]. **i**) FLSea comprises real-world underwater scenes from two distinct regions, the Mediterranean Sea's Canyons and the Red Sea. The Canyons includes four sequences captured at shallow depths (4-7m), totaling over 8K frames. The Red Sea contains eight sequences with more than 13K frames range from 3-8 m. **ii**) MIMIR-UW is a synthetic dataset generated using Unreal Engine with four diverse environments, *i.e.*, 'SeaFloor', 'SeaFloor Algae', 'OceanFloor', and 'SandPipe'. It consists of about 31K frames with average path length of 283 m. **iii**) TartanAir-Ocean is a subset of the large-scale TartanAir benchmark, collected in photo-realistic simulation environments with varying lighting conditions. It includes 12 and 10 motion sequences in 'Easy' and 'Hard' splits, respectively, with over 7K and 6K frames, respectively (more details can be found in Appendix).

Baselines. We benchmark DUV-SLAM against seven representative visual SLAM approaches. ORB-SLAM3 [5] is a feature-based SLAM system relying on lightweight geometric features, while DROID-SLAM [69] is a deep learning-based SLAM system leveraging optical flow es-

Methods	Metrics	Canyons				Red Sea							
		UC	Fla	HC	TC	NP	LP	DP	PP	CTL	CPL	BDL	SP
ORB-SLAM3 [5]	ATE [cm] ↓	8.36	7.56	7.81	2.83	4.87	5.35	9.58	7.03	3.93	4.69	3.95	4.06
DROID-SLAM [69]	ATE [cm] ↓	7.56	5.32	2.39	2.75	3.16	4.88	5.87	8.55	3.20	2.85	6.43	4.02
NICE-SLAM [91]	ATE [cm] ↓	10.08	7.50	6.66	4.85	6.26	7.37	7.44	11.15	5.17	4.38	5.36	5.21
	PSNR ↑	11.61	12.83	13.45	12.74	14.32	11.01	11.33	13.45	11.99	16.48	10.71	9.78
	SSIM ↑	0.40	0.43	0.41	0.64	0.38	0.41	0.24	0.29	0.35	0.25	0.34	0.27
	LPIPS ↓	0.51	0.59	0.45	0.47	0.39	0.59	0.60	0.67	0.64	0.53	0.57	0.68
UncLe-SLAM [55]	ATE [cm] ↓	9.92	8.13	6.70	4.98	7.28	8.29	8.31	25.09	5.17	4.46	5.40	5.16
	PSNR ↑	14.80	15.67	17.75	14.20	18.51	8.66	10.50	15.09	18.66	14.55	16.13	17.20
	SSIM ↑	0.43	0.40	0.35	0.44	0.52	0.40	0.49	0.51	0.32	0.39	0.47	0.37
	LPIPS ↓	0.49	0.41	0.39	0.45	0.43	0.37	0.47	0.42	0.38	0.44	0.39	0.47
MonoGS [39]	ATE [cm] ↓	14.71	12.53	21.27	23.13	25.16	27.93	19.49	32.97	30.07	22.20	25.98	16.15
	PSNR ↑	18.54	21.66	13.10	24.15	28.79	13.08	22.30	19.26	17.64	20.37	18.39	19.76
	SSIM ↑	0.47	0.51	0.46	0.58	0.56	0.28	0.49	0.45	0.42	0.50	0.47	0.52
	LPIPS ↓	0.55	0.37	0.34	0.31	0.40	0.47	0.36	0.29	0.44	0.27	0.39	0.36
Photo-SLAM [19]	ATE [cm] ↓	9.33	7.09	6.39	3.37	4.52	4.62	9.86	6.81	3.57	4.57	3.77	5.08
	PSNR ↑	17.67	23.87	18.93	24.16	18.56	21.83	26.78	19.57	23.37	21.46	20.15	23.08
	SSIM ↑	0.46	0.60	0.53	0.68	0.54	0.69	0.72	0.59	0.64	0.71	0.67	0.73
	LPIPS ↓	0.63	0.42	0.37	0.30	0.48	0.39	0.28	0.41	0.39	0.29	0.34	0.30
Splat-SLAM [56]	ATE [cm] ↓	7.44	5.97	2.56	2.78	4.74	5.27	5.85	8.01	3.62	3.09	5.18	4.37
	PSNR ↑	21.93	23.22	23.19	20.02	22.75	19.04	21.46	20.56	24.86	27.43	19.63	18.83
	SSIM ↑	0.55	0.64	0.60	0.54	0.70	0.57	0.61	0.63	0.69	0.73	0.54	0.56
	LPIPS ↓	0.36	0.34	0.42	0.44	0.34	0.49	0.51	0.46	0.39	0.31	0.43	0.50
DUV-SLAM (ours)	ATE [cm] ↓	**6.73**	6.42	**2.56**	**2.40**	**3.07**	**4.50**	**5.71**	8.73	**2.07**	3.44	**3.95**	**4.02**
	PSNR ↑	**22.73**	**25.71**	**24.55**	**29.44**	**25.81**	**26.09**	**27.38**	**25.48**	**30.03**	**28.37**	**26.52**	**27.67**
	SSIM ↑	**0.68**	**0.74**	**0.62**	0.61	**0.62**	**0.61**	**0.78**	**0.74**	**0.79**	0.68	**0.78**	**0.69**
	LPIPS ↓	**0.30**	**0.19**	**0.22**	**0.28**	**0.36**	**0.28**	**0.29**	**0.19**	**0.17**	0.32	**0.15**	**0.24**

Table 1. Quantitative results of different monocular methods on FLSea [48] (§4.1). Best results are highlighted as **first** and second.

timation. NICE-SLAM [91] and MonoGS [39] represent dense SLAM paradigms based on NeRF and 3DGS, respectively. UncLe-SLAM [55] introduces an uncertainty learning module into a neural SLAM framework. We report its performance in Table 1, 2, and 3 to evaluate this module, despite its reliance on RGB-D inputs. Photo-SLAM [19] and Splat-SLAM [56] integrate the visual odometry of ORB-SLAM3 and DROID-SLAM, respectively, into a decoupled 3DGS-based *mapping* process. All baseline methods are reproduced using their publicly available implementations.

Evaluation Metrics. To evaluate *tracking* accuracy, the root mean square error of absolute trajectory error (ATE) [63] is used for the estimated trajectory. For *mapping*, we employ standard photometric quality metrics by comparing rendered keyframe images against camera images. Specifically, Peak Signal-to-Noise Ratio (PSNR) quantifies noise levels, Structural Similarity Index Measure (SSIM) measures structural distortion, and Learned Perceptual Image Patch Similarity (LPIPS) evaluates perceptual quality. For evaluating depth uncertainty, we use the Area Under Sparsification Error (AUSE) for Mean Absolute Error (MAE) [29], which captures the correlation between estimated uncertainty and true MAE in each rendered view. All reported results are averaged over five runs.

4.1. Quantitative Results

Performance on FLSea [48]. Table 1 presents a quantitative evaluation on Canyons and Rea Sea splits, including four and eight sequences, respectively. In terms of rendering, DUV-SLAM outperforms all 3DGS-based baselines (*i.e.*, MonoGS [39], Photo-SLAM [19], and Splat-SLAM [56]) across most sequences, benefiting from medium modeling and uncertainty-aware geometric optimization. The results of UncLe-SLAM [55] indicate that degradation in underwater environments remains significant, even with depth information as an additional input. Moreover, the tracking performance of DUV-SLAM is comparable to Photo-SLAM [19] with robust tracking modules [5], *e.g.*, $\mathbf{6.73}$cm *vs* 9.33cm ATE on 'UC' split.

Performance on MIMIR-UW [3]. Table 2 reports performance comparison of our method against other competitors across four underwater environments. Compared to other baselines, our method achieves top rendering quality on most scenes, including $\mathbf{28.68}$ PSNR and $\mathbf{0.77}$ SSIM on Sand Pipe 't0-d'. Although our approach demonstrates obvious improvements over baseline methods, the ATE exceeding 10 meters remains non-negligible, suggesting the need for further research on long-sequence tracking, *e.g.*, the 't1-l' split of Ocean Floor with more than 6K frames.

Performance on TartanAir-Ocean [74]. As shown in Table 3, DUV-SLAM demonstrates promising reconstruction performance in both 'Easy' and 'Hard' splits. Notably, in the more challenging 'Hard' split, our approach outperforms all 3DGS-based baselines, *e.g.*, $\mathbf{27.57}$ PSNR *vs* 24.13 of Splat-SLAM [56], demonstrating its effectiveness in complex underwater environments.

Uncertainty Quantification Evaluation. To assess the

Methods	Metrics	SeaFloor			SeaFloor Algae			Ocean Floor			Sand Pipe	
		t0	t1	t2	t0	t1	t2	t0-d	t0-l	t1-l	t0-d	t0-l
ORB-SLAM3 [5]	ATE [m] ↓	3.67	8.78	5.18	2.99	1.15	9.16	5.78	8.37	23.66	20.08	6.85
DROID-SLAM [69]	ATE [m] ↓	4.90	6.93	12.10	2.47	3.76	8.19	5.79	9.03	17.66	15.47	6.34
NICE-SLAM [91]	ATE [m] ↓	9.22	10.52	11.10	13.14	10.49	11.21	10.46	16.45	21.14	20.19	23.06
	PSNR ↑	16.01	15.24	15.60	14.34	14.98	13.97	24.65	20.93	19.98	24.54	20.11
	SSIM ↑	0.36	0.40	0.25	0.33	0.32	0.27	0.88	0.78	0.79	0.76	0.68
	LPIPS ↓	0.60	0.65	0.58	0.59	0.58	0.61	0.17	0.32	0.28	0.20	0.32
UncLe-SLAM [55]	ATE [m] ↓	9.35	11.51	9.66	8.47	10.64	7.31	15.01	9.09	7.93	10.24	8.18
	PSNR ↑	18.45	13.93	15.59	15.87	9.59	15.94	21.51	14.51	16.67	18.58	18.24
	SSIM ↑	0.50	0.47	0.45	0.54	0.55	0.60	0.59	0.56	0.40	0.35	0.52
	LPIPS ↓	0.61	0.59	0.56	0.65	0.49	0.43	0.39	0.63	0.67	0.60	0.58
MonoGS [39]	ATE [m] ↓	8.97	9.87	14.08	7.64	7.95	12.37	9.44	12.90	27.94	17.95	9.26
	PSNR ↑	17.13	21.46	19.00	20.52	24.80	17.61	20.91	17.50	20.22	23.06	19.81
	SSIM ↑	0.40	0.48	0.42	0.38	0.53	0.41	0.46	0.43	0.52	0.55	0.44
	LPIPS ↓	0.41	0.37	0.44	0.32	0.40	0.32	0.36	0.42	0.45	0.32	0.39
Photo-SLAM [19]	ATE [m] ↓	3.86	9.06	5.74	3.29	1.83	9.71	6.67	9.26	19.56	18.24	6.90
	PSNR ↑	20.84	22.53	21.01	20.67	24.46	25.49	21.64	22.26	30.12	25.68	23.42
	SSIM ↑	0.49	0.47	0.51	0.52	0.47	0.58	0.55	0.58	0.66	0.59	0.53
	LPIPS ↓	0.35	0.40	0.43	0.27	0.40	0.33	0.34	0.30	0.27	0.30	0.33
Splat-SLAM [56]	ATE [m] ↓	5.63	7.63	9.22	8.25	4.97	7.55	6.83	9.14	20.08	12.72	7.21
	PSNR ↑	24.95	20.14	19.86	25.19	24.67	23.20	22.19	18.06	19.96	24.39	21.76
	SSIM ↑	0.62	0.45	0.49	0.66	0.54	0.65	0.62	0.54	0.55	0.67	0.51
	LPIPS ↓	0.22	0.45	0.39	0.26	0.37	0.25	0.29	0.41	0.31	0.26	0.36
DUV-SLAM (ours)	ATE [m] ↓	4.70	**6.90**	9.21	2.03	3.79	**8.07**	**5.15**	**7.02**	**15.59**	15.03	**6.28**
	PSNR ↑	**25.11**	**23.69**	**25.14**	23.67	**25.84**	**27.15**	**28.04**	**25.56**	**24.70**	**28.68**	21.96
	SSIM ↑	**0.69**	**0.73**	**0.79**	**0.70**	**0.66**	**0.65**	**0.74**	**0.80**	**0.67**	**0.77**	**0.67**
	LPIPS ↓	**0.21**	**0.13**	**0.17**	**0.20**	**0.18**	**0.21**	**0.18**	**0.20**	**0.19**	**0.14**	**0.22**

Table 2. Quantitative results of different monocular methods on MIMIR-UW [3] (§4.1). Best results are highlighted as **first** and second.

Methods	Easy				Hard			
	ATE	PSNR	SSIM	LPIPS	ATE	PSNR	SSIM	LPIPS
ORB-SLAM3 [5]	2.03	–	–	–	2.10	–	–	–
DROID-SLAM [69]	1.15	–	–	–	0.87	–	–	–
NICE-SLAM [91]	7.44	20.61	0.70	0.49	5.15	19.08	0.65	0.46
UncLe-SLAM [55]	3.07	11.13	0.24	0.59	2.20	3.20	0.21	0.77
MonoGS [39]	9.49	21.68	0.61	0.40	11.22	19.79	0.51	0.45
Photo-SLAM [19]	2.70	22.19	0.65	0.36	2.60	20.37	0.59	0.40
Splat-SLAM [56]	1.52	23.54	0.59	0.39	1.06	24.13	0.57	0.36
DUV-SLAM (ours)	**1.26**	**28.86**	**0.78**	**0.19**	**0.83**	**27.57**	**0.67**	**0.25**

Table 3. Quantitative results of different monocular methods on Tartanair-Ocean [74] (§4.1). Best results are highlighted.

Methods	MIMIR-SeaFloor			Tartanair-Ocean		Average
	t0	t1	t2	Easy	Hard	
UncLe-SLAM [55]	0.22	0.29	0.26	0.18	0.30	0.25
DUV-SLAM (ours)	**0.17**	**0.20**	**0.23**	0.19	**0.25**	**0.21**

Table 4. Evaluation on depth uncertainty quantification on MIMIR-SeaFloor [3] and Tartanair-Ocean [74] (§4.1).

quality of depth uncertainty, we first reorder the pixel-level depth values of each frame based on the predicted uncertainty, and then compute the difference between the reordered MAE list and the ground-truth values. We report the normalized results for each sequence of underwater scenes in Table 4. Our method yields a lower average AUSE compared to UncLe-SLAM [55] ($\mathbf{0.21}$ *vs* 0.25), indicating better uncertainty quantification capability.

4.2. Qualitative Results

Rendering Visualization. Figure 3 presents visual comparisons of DUV-SLAM against other NeRF- or 3DGS-based baselines in underwater scenes of FLSea [48]. Benefiting from geometric priors derived from depth uncertainty and medium modeling, our approach recovers more details that are oversmoothed in competing methods. The improvement in PSNR further demonstrates the high fidelity of our renderings. DUV-SLAM preserves fine textures and edge details even in low-texture regions. These qualitative results validate the effectiveness of our approach and highlight its potential for robust underwater scene reconstruction.

4.3. Diagnostic Experiment

Medium Parameters. We first validate the design of learnable medium parameters σ^{att} and σ^{bs}. Based on the underwater image formation model [1], we introduce separate parameters for direct attenuation and backscatter in Eq. 10, ensuring $\sigma^{att} \neq \sigma^{bs}$. The results in Table 5 indicate that using a single coefficient ($\sigma^{att} = \sigma^{bs}$) fails to accurately model underwater medium-light interactions, leading to a decline in rendering performance across multiple metrics. In the Tartanair-Hard [74] split, PSNR drops from $\mathbf{27.57}$ to 26.93, highlighting the necessity of distinct attenuation and backscattering modeling for underwater resonstruction.

Uncertainty-Guided Optimization. We next investigate the impact of uncertainty-guided optimization on the geo-

Figure 3. Qualitative results on FLSea [48]. Previous visual SLAM methods lack medium modeling, leading to suboptimal reconstructions. Our method achieves better rendering quality with the highest PSNR and effectively preserves fine details, even in distant regions (§4.2).

Medium Parameters	FLSea-Canyons-Fla			Tartanair-Hard		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
w/o Medium Modeling	23.62	0.66	0.29	24.08	0.61	0.33
$\sigma^{att} = \sigma^{bs}$	25.13	0.70	0.18	26.93	0.69	0.29
$\sigma^{att} \neq \sigma^{bs}$	**25.71**	**0.74**	**0.19**	**27.57**	0.67	**0.25**

Table 5. Ablation study of different designs for medium parameters on FLSea [48] and Tartanair-Ocean [74] (§4.3).

$\mathcal{L}_{\text{depth}}$			FLSea-Canyons-Fla			Tartanair-Hard		
$\mathcal{L}_{\text{MD}}$	$\mathcal{L}_{\text{RGB-C}}$	$\mathcal{L}_{\text{KL}}$	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
–	–	–	23.77	0.61	0.34	24.78	0.60	0.28
–	✓	✓	24.83	0.69	0.27	25.60	0.62	0.29
✓	✓	–	25.01	0.69	0.23	26.70	0.64	0.25
✓	✓	✓	**25.71**	**0.74**	**0.19**	**27.57**	**0.67**	**0.25**

Table 6. Ablation study of uncertainty-guided optimization on FLSea [48] and Tartanair-Ocean [74] (§4.3).

metric loss $\mathcal{L}_{\text{depth}}$ in *mapping* (Eq. 13). Depth uncertainty derived from *tracking* is served as a geometric prior to regularize the Mahalanobis distance loss $\mathcal{L}_{\text{MD}}$ (Eq. 14) and minimize the depth distribution loss $\mathcal{L}_{\text{KL}}$ (Eq. 12). In Table 6, the absence of $\mathcal{L}_{\text{MD}}$ results in a noticeable performance drop, with PSNR decreasing from $\mathbf{27.57}$ to 25.60. In addition, we observe that $\mathcal{L}_{\text{MD}}$ and $\mathcal{L}_{\text{KL}}$ are complementary, jointly contributing to improved geometric optimization.

Runtime and Memory Analysis. Table 7 presents the runtime and memory consumption evaluated on FLSea [48] using an NVIDIA RTX 4090 GPU. Our DUV-SLAM operates at an average speed of $\mathbf{23.09}$ FPS with a peak memory usage of $\mathbf{9.43}$ GB. While 3DGS-based baselines support the GPU-accelerated rasterization, they incur slightly higher per-iteration latency due to low-light fitting requirements. Despite introducing learnable medium parameters, DUV-SLAM converges efficiently, demonstrating its potential for real-world underwater applications.

Different Tracking Modules. Table 8 shows the results of map rendering using different tracking modules. To evaluate the effectiveness of our overall design, we incorporate

Metrics	MonoGS [39]	Photo-SLAM [19]	Splat-SLAM [56]	Ours
GPU [GB]	14.57	8.21	17.66	9.43
Average FPS	20.17	23.09	16.33	**26.52**

Table 7. Runtime and memory usage of NeRF- and 3DGS-based baselines on 'TC' split of FLSea [48] (§4.3).

Medium Modeling	FLSea-Canyons		MIMIR-Sand		Tartanair-Hard	
	ATE	PSNR	ATE	PSNR	ATE	PSNR
MonoGS* [39]	22.03	25.61	18.27	23.89	11.09	22.51
Photo-SLAM* [19]	3.64	26.07	17.63	26.11	2.77	23.36
DUV-SLAM (ours)	**2.40**	**29.44**	**15.03**	**28.68**	**0.83**	**27.57**

Table 8. Medium modeling with different tracker on FLSea-Canyons-TC [48], MIMIR-Sand Pipe-t0-d [3], and Tartanair-Hard [74]. '*' denotes incorporating medium modeling (§4.3).

medium modeling (Eq. 7) into various baseline methods. By directly optimizing 3D Gaussians with medium properties for *tracking*, MonoGS* [39] achieves promising rendering performance, obtaining a PSNR of 25.61 on FLSea [48]. Although Photo-SLAM* [19] leverages well-crafted ORB features [5] for real-time *tracking*, our method surpasses it in both trajectory estimation ($\mathbf{2.40}$cm *vs* 3.64cm ATE) and rendering ($\mathbf{29.44}$ *vs* 26.07 PNSR) due to geometric priors.

5. Conclusion

In this work, we propose DUV-SLAM, a dense underwater visual SLAM framework that addresses the inherent challenges of unbounded underwater environments. By quantifying depth uncertainty and incorporating a physics-inspired hybrid 3DGS representation, our method enables robust real-time state estimation and high-fidelity scene reconstruction. Extensive evaluations on synthetic and real-world datasets demonstrate that DUV-SLAM outperforms existing baselines in both tracking accuracy and rendering quality with competitive tracking speed and memory efficiency. Future work will focus on extending the framework to more diverse underwater scenes and longer sequences.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (No. 62372405), Fundamental Research Funds for the Central Universities (226-2025-00057), and CIE-Tencent Robotics X Rhino-Bird Focused Research Program.

References

[1] Derya Akkaynak and Tali Treibitz. A revised underwater image formation model. In *CVPR*, 2018. 1, 2, 4, 5, 7

[2] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *CVPR*, 2019. 2

[3] Olaya Álvarez-Tuñón, Hemanth Kanner, Luiza Ribeiro Marnet, Huy Xuan Pham, Jonas le Fevre Sejersen, Yury Brodskiy, and Erdal Kayacan. Mimir-uw: A multipurpose synthetic dataset for underwater navigation and inspection. In *IROS*, 2023. 2, 5, 6, 7, 8

[4] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE TRO*, 32(6):1309–1332, 2016. 2

[5] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE TRO*, 37(6):1874–1890, 2021. 2, 5, 6, 7, 8

[6] Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting. *arXiv preprint arXiv:2401.03890*, 2024. 2

[7] Chi-Ming Chung, Yang-Che Tseng, Ya-Ching Hsu, Xiang-Qian Shi, Yun-Hung Hua, Jia-Fong Yeh, Wen-Chin Chen, Yi-Ting Chen, and Winston H Hsu. Orbeez-slam: A real-time monocular visual slam with orb features and nerf-realized mapping. In *ICRA*, 2023. 2

[8] Yang Cong, Changjun Gu, Tao Zhang, and Yajun Gao. Underwater robot sensing technology: A survey. *Fundamental Research*, 1(3):337–345, 2021. 1, 2

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 1

[10] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, 2022. 1, 3

[11] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *ECCV*, 2014. 2

[12] Jianzhe Gao, Rui Liu, and Wenguan Wang. 3d gaussian map with open-set semantic grouping for vision-language navigation. In *ICCV*, 2025. 2

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1

[14] Salma P González-Sabbagh and Antonio Robles-Kelly. A survey on underwater computer vision. *ACM Computing Surveys*, 55(13s):1–39, 2023. 2

[15] Jie Han, Wenxue Liu, Yusheng Zheng, Arash Khalatbarisoltani, Yalian Yang, and Xiaosong Hu. Health-conscious predictive energy management strategy with hybrid speed predictor for plug-in hybrid electric vehicles: Investigating the impact of battery electro-thermal-aging models. *Applied Energy*, 352:121986, 2023. 4

[16] Akshay Hinduja, Bing-Jui Ho, and Michael Kaess. Degeneracy-aware factors with applications to underwater slam. In *IROS*, 2019. 1

[17] Christian Homeyer, Leon Begiristain, and Christoph Schnörr. Droid-splat: Combining end-to-end slam with 3d gaussian splatting. *arXiv preprint arXiv:2411.17660*, 2024. 2

[18] Jiarui Hu, Xianhao Chen, Boyin Feng, Guanglin Li, Liangjing Yang, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Cg-slam: Efficient dense rgb-d slam in a consistent uncertainty-aware 3d gaussian field. In *ECCV*, 2024. 3

[19] Huajian Huang, Longwei Li, Hui Cheng, and Sai-Kit Yeung. Photo-slam: Real-time simultaneous localization and photorealistic mapping for monocular stereo and rgb-d cameras. In *CVPR*, 2024. 1, 2, 6, 7, 8

[20] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984. 4

[21] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In *CVPR*, 2024. 2

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):139–1, 2023. 1, 2, 4

[23] Ayoung Kim and Ryan M Eustice. Real-time visual slam for autonomous underwater hull inspection using visual saliency. *IEEE TRO*, 29(3):719–733, 2013. 1

[24] Vladimir R Kostic, Giacomo Turri, Pietro Novelli, Karim Lounici, Massimiliano Pontil, et al. Neural conditional probability for uncertainty quantification. In *NeurIPS*, 2024. 3

[25] Jungwoo Lee and Younggun Cho. Mesh-based photorealistic and real-time 3d mapping for robust visual perception of autonomous underwater vehicle. *arXiv preprint arXiv:2404.18395*, 2024. 1, 2

[26] Deborah Levy, Amit Peleg, Naama Pearl, Dan Rosenbaum, Derya Akkaynak, Simon Korman, and Tali Treibitz. Seathru-nerf: Neural radiance fields in scattering media. In *CVPR*, 2023. 1, 2

[27] Deng Li, Aming Wu, Yaowei Wang, and Yahong Han. Prompt-driven dynamic object-centric learning for single domain generalization. In *CVPR*, 2024. 3

[28] Huapeng Li, Wenxuan Song, Tianao Xu, Alexandre Elsig, and Jonas Kulhanek. Watersplatting: Fast underwater 3d scene reconstruction using gaussian splatting. *arXiv preprint arXiv:2408.08206*, 2024. 1, 2

[29] Ruiqi Li and Yiu-ming Cheung. Variational multi-scale representation for estimating uncertainty in 3d gaussian splatting. In *NeurIPS*, 2024. 1, 3, 6

[30] Xingyuan Li, Jinyuan Liu, Zhixin Chen, Yang Zou, Long Ma, Xin Fan, and Risheng Liu. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *ECCV*, 2024. 2

[31] Xingyuan Li, Zirui Wang, Yang Zou, Zhixin Chen, Jun Ma, Zhiying Jiang, Long Ma, and Jinyuan Liu. Difiisr: A diffusion model with gradient guidance for infrared image super-resolution. In *CVPR*, 2025. 2

[32] Yu Ling, Ye Li, Teng Ma, Zheng Cong, Shuo Xu, and Zhihui Li. Active bathymetric slam for autonomous underwater exploration. *Applied Ocean Research*, 130:103439, 2023. 1

[33] Rui Liu, Xiaohan Wang, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation. In *ICCV*, 2023. 2

[34] Rui Liu, Wenguan Wang, and Yi Yang. Vision-language navigation with energy-based policy. In *NeurIPS*, 2024. 2

[35] Rui Liu, Wenguan Wang, and Yi Yang. Volumetric environment representation for vision-language navigation. In *CVPR*, 2024. 2

[36] Shaohua Liu, Junzhe Lu, Zuoya Gu, Jiajun Li, and Yue Deng. Aquatic-gs: A hybrid 3d representation for underwater scenes. *arXiv preprint arXiv:2411.00239*, 2024. 2

[37] Yunxuan Mao, Xuan Yu, Zhuqing Zhang, Kai Wang, Yue Wang, Rong Xiong, and Yiyi Liao. Ngel-slam: Neural implicit representation-based global consistent low-latency slam system. In *ICRA*, 2024. 2

[38] Ivan Masmitja, Mario Martin, Tom O'Reilly, Brian Kieft, Narcís Palomeras, Joan Navarro, and Kakani Katija. Dynamic robotic tracking of underwater targets using reinforcement learning. *Science Robotics*, 8(80):eade7811, 2023. 1

[39] Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian splatting slam. In *CVPR*, 2024. 1, 2, 6, 7, 8

[40] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2, 4

[41] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011. 2

[42] Paul M Newman, John J Leonard, and Richard J Rikoski. Towards constant-time slam on an autonomous underwater vehicle using synthetic aperture sonar. In *International Symposium of Robotics Research*, 2005. 1

[43] Hai-Vy Nguyen, Fabrice Gamboa, Reda Chhaibi, Sixin Zhang, Serge Gratton, and Thierry Giaccone. Combining statistical depth and fermat distance for uncertainty quantification. In *NeurIPS*, 2024. 3

[44] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM TOG*, 32(6):1–11, 2013. 2

[45] Liam Paull, Sajad Saeedi, Mae Seto, and Howard Li. Auv navigation and localization: A review. *IEEE Journal of Oceanic Engineering*, 39(1):131–149, 2013. 2

[46] Giacomo Picardi, Mrudul Chellapurath, Saverio Iacoponi, Sergio Stefanni, Cecilia Laschi, and Marcello Calisti. Bioinspired underwater legged robot for seabed exploration with low environmental disturbance. *Science Robotics*, 5(42): eaaz1012, 2020. 1

[47] Sharmin Rahman, Alberto Quattrini Li, and Ioannis Rekleitis. Svin2: An underwater slam system using sonar, visual, inertial, and depth sensor. In *IROS*, 2019. 1

[48] Yelena Randall and Tali Treibitz. Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets. *arXiv preprint arXiv:2302.12772*, 2023. 2, 5, 6, 7, 8

[49] David Ribas, Pere Ridao, Juan Domingo Tardós, and José Neira. Underwater slam in man-made structured environments. *Journal of Field Robotics*, 25(11-12):898–921, 2008. 1

[50] Antoni Rosinol, John J Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. In *IROS*, 2023. 1, 2, 3

[51] Antoni Rosinol, John J Leonard, and Luca Carlone. Probabilistic volumetric fusion for dense monocular slam. In *WACV*, 2023. 1, 3, 4

[52] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. Revising densification in gaussian splatting. In *ECCV*, 2024. 5

[53] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *CVPR*, 2013. 2

[54] Erik Sandström, Yue Li, Luc Van Gool, and Martin R Oswald. Point-slam: Dense neural point cloud-based slam. In *ICCV*, 2023. 2

[55] Erik Sandström, Kevin Ta, Luc Van Gool, and Martin R Oswald. Uncle-slam: Uncertainty learning for dense neural slam. In *ICCV Workshop*, 2023. 3, 6, 7

[56] Erik Sandström, Keisuke Tateno, Michael Oechsle, Michael Niemeyer, Luc Van Gool, Martin R Oswald, and Federico Tombari. Splat-slam: Globally optimized rgb-only slam with 3d gaussians. *arXiv preprint arXiv:2405.16544*, 2024. 6, 7, 8

[57] Jonathan Sauder and Devis Tuia. Self-supervised underwater caustics removal and descattering via deep monocular slam. In *ECCV*, 2024. 2

[58] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *CVPR*, 2019. 2

[59] Advaith Venkatramanan Sethuraman, Manikandasriram Srinivasan Ramanagopal, and Katherine A Skinner. Waternerf: Neural radiance fields for underwater scenes. In *OCEANS Gulf Coast*, 2023. 2

[60] Florian Shkurti, Ioannis Rekleitis, Milena Scaccia, and Gregory Dudek. State estimation of an underwater robot using visual and inertial information. In *IROS*, 2011. 1

[61] Katherine A Skinner, Eduardo Iscar, and Matthew Johnson-Roberson. Automatic color correction for 3d reconstruction of underwater scenes. In *ICRA*, 2017. 2

[62] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital

replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 1

[63] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IROS*, 2012. 1, 6

[64] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *ICCV*, 2021. 2

[65] Yucheng Suo, Fan Ma, Linchao Zhu, Tianyi Wang, Fengyun Rao, and Yi Yang. From trial to triumph: Advancing long video understanding via visual context sample scaling and self-reward alignment. *arXiv preprint arXiv:2503.20472*, 2025. 2

[66] Yunkai Tang, Chengxuan Zhu, Renjie Wan, Chao Xu, and Boxin Shi. Neural underwater scene representation. In *CVPR*, 2024. 1, 2

[67] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *CVPR*, 2017. 2

[68] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 2, 3, 5

[69] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In *NeurIPS*, 2021. 2, 3, 4, 5, 6, 7

[70] Sebastian Thrun, Michael Montemerlo, Daphne Koller, Ben Wegbreit, Juan Nieto, and Eduardo Nebot. Fastslam: An efficient solution to the simultaneous localization and mapping problem with unknown data association. *Journal of Machine Learning Research*, 4(3):380–407, 2004. 2, 3

[71] Fabio Tosi, Youmin Zhang, Ziren Gong, Erik Sandström, Stefano Mattoccia, Martin R Oswald, and Matteo Poggi. How nerfs and 3d gaussian splatting are reshaping slam: a survey. *arXiv preprint arXiv:2402.13255*, 2024. 2

[72] Ignacio Vizzo, Tiziano Guadagnino, Benedikt Mersch, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. Kiss-icp: In defense of point-to-point icp–simple, accurate, and robust registration if done the right way. *IEEE RAL*, 8(2):1029–1036, 2023. 2

[73] Haoran Wang, Nantheera Anantrasirichai, Fan Zhang, and David Bull. Uw-gs: Distractor-aware 3d gaussian splatting for enhanced underwater scene reconstruction. *arXiv preprint arXiv:2410.01517*, 2024. 2, 5

[74] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *IROS*, 2020. 2, 5, 6, 7, 8

[75] Stefan B Williams, Paul Newman, Gamini Dissanayake, and Hugh Durrant-Whyte. Autonomous underwater simultaneous localisation and map building. In *ICRA*, 2000. 1

[76] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. In *ICCV*, 2021. 2

[77] Ziheng Xu, Qingfeng Li, Chen Chen, Xuefeng Liu, and Jianwei Niu. Glc-slam: Gaussian splatting slam with efficient loop closure. *arXiv preprint arXiv:2409.10982*, 2024. 3

[78] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *CVPR*, 2024. 2

[79] Daniel Yang, John J Leonard, and Yogesh Girdhar. Seasplat: Representing underwater scenes with 3d gaussian splatting and a physically grounded image formation model. *arXiv preprint arXiv:2409.17345*, 2024. 2

[80] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *CVPR*, 2020. 3

[81] Xiangpeng Yang, Linchao Zhu, Xiaohan Wang, and Yi Yang. Dgl: Dynamic global-local prompt tuning for text-video retrieval. In *AAAI*, 2024. 2

[82] Xihang Yue, Yi Yang, and Linchao Zhu. Holistic physics solver: Learning pdes in a unified spectral-physical space. In *ICML*, 2025. 3

[83] Bingbing Zhang, Daxiong Ji, Shuo Liu, Xinke Zhu, and Wen Xu. Autonomous underwater vehicle navigation: a review. *Ocean Engineering*, 273:113861, 2023. 2

[84] Youmin Zhang, Fabio Tosi, Stefano Mattoccia, and Matteo Poggi. Go-slam: Global optimization for consistent 3d instant reconstruction. In *ICCV*, 2023. 2

[85] Wancai Zheng, Xinyi Yu, Jintao Rong, Linlin Ou, Yan Wei, and Libo Zhou. Gsorb-slam: Gaussian splatting slam benefits from orb features and transmittance information. *arXiv preprint arXiv:2410.11356*, 2024. 2

[86] Dewei Zhou, You Li, Fan Ma, Zongxin Yang, and Yi Yang. Migc++: Advanced multi-instance generation controller for image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[87] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *CVPR*, 2024. 2

[88] Dewei Zhou, Mingwei Li, Zongxin Yang, and Yi Yang. Dreamrenderer: Taming multi-instance attribute control in large-scale text-to-image models. In *ICCV*, 2025. 2

[89] Jingchun Zhou, Tianyu Liang, Dehuan Zhang, and Zongxin He. Waterhe-nerf: Water-ray tracing neural radiance fields for underwater scene reconstruction. *arXiv preprint arXiv:2312.06946*, 2024. 2

[90] Zhenglin Zhou, Xiaobo Xia, Fan Ma, Hehe Fan, Yi Yang, and Tat-Seng Chua. Dreamdpo: Aligning text-to-3d generation with human preferences via direct preference optimization. In *ICML*, 2025. 2

[91] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *CVPR*, 2022. 1, 2, 6, 7