frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 09 June 2025
DOI 10.3389/fmars.2025.1582126

OPEN ACCESS

EDITED BY
Yakun Ju,
University of Leicester, United Kingdom

REVIEWED BY
Fickrie Muhammad,
Bandung Institute of Technology, Indonesia
Haiyang Qiu,
Guangzhou Maritime College, China

*CORRESPONDENCE
Zhibin Yu
yuzhibin@ouc.edu.cn

RECEIVED 24 February 2025
ACCEPTED 19 May 2025
PUBLISHED 09 June 2025

CITATION
Wang Z, Yu Z and Zheng B (2025) YOLO-NeRFSLAM: underwater object detection for the visual NeRF-SLAM. *Front. Mar. Sci.* 12:1582126. doi: 10.3389/fmars.2025.1582126

YOLO-NeRFSLAM: underwater object detection for the visual NeRF-SLAM

Zhe Wang[1], Zhibin Yu[1,2]* and Bing Zheng[1,2]

[1]Faculty of Information Science and Engineering, Ocean University of China, Qingdao, Shandong, China, [2]Key Laboratory of Ocean Observation and Information of Hainan Province, Sanya Oceanographic Institution, Ocean University of China, Sanya, Hainan, China

Accurate and reliable dense mapping is crucial for understanding and utilizing the marine environment in applications such as ecological monitoring, archaeological exploration, and autonomous underwater navigation. However, the underwater environment is highly dynamic: fish and floating debris frequently appear in the field of view, causing traditional SLAM to be easily disturbed during localization and mapping. In addition, common depth sensors and depth estimation techniques based on deep learning tend to be impractical or significantly less accurate underwater, failing to meet the demands of dense reconstruction. This paper proposes a new underwater SLAM framework that combines neural radiance fields (NeRF) with a dynamic masking module to address these issues. Through a Marine Motion Fusion (MMF) strategy—leveraging YOLO to detect known marine organisms and integrating optical flow for pixel-level motion analysis—we effectively screen out all dynamic objects, thus maintaining stable camera pose estimation and pixel-level dense reconstruction even without relying on depth data. Further, to cope with severe light attenuation and the dynamic nature of underwater scenes, we introduce specialized loss functions, enabling the reconstruction of underwater environments with realistic appearance and geometric detail even under high turbidity conditions. Experimental results show that our method significantly reduces localization drift caused by moving entities, improves dense mapping accuracy, and achieves favorable runtime efficiency in multiple real underwater video datasets, demonstrating both its potential and advanced capabilities in dynamic underwater settings.

KEYWORDS

visual SLAM, NeRF-SLAM, underwater SLAM, object detection, novel view reconstruction

1 Introduction

Underwater exploration and mapping play a pivotal role in marine ecological studies, underwater archaeology, and autonomous navigation. Achieving high-quality 3D reconstruction and object detection is essential for these applications, yet conventional vision-based SLAM systems—successful in terrestrial environments—encounter serious

limitations underwater. A shortage of reliable depth information and the complexities of dense reconstruction pose major challenges. Although improvements in underwater SLAM have enhanced feature detection, they have not thoroughly addressed dynamic interference from marine life or realized a truly pixel-level mapping framework.

Underwater conditions present three fundamental difficulties. First, they are inherently dynamic: fish, plankton, and floating debris constantly drift through the field of view, making feature matching and pose estimation prone to drift. Second, accurate depth data are typically absent or unreliable: scattering and non-uniform lighting frequently degrade depth sensors and offline estimation methods, undermining the feasibility of dense 3D reconstruction. Third, optical attenuation and color distortion—caused by absorption and scattering in seawater—further diminish image quality, reducing the fidelity of dense reconstructions. Consequently, underwater SLAM must simultaneously address dynamic interference and the lack of accurate depth measurements.

To solve dynamic environments, while some approaches incorporate object detection to exclude moving elements—such as DynaSLAM Bescos et al. (2018) or CNN-SLAM Tateno et al. (2017)—these methods are often trained on terrestrial imagery and fail to account for the diverse marine organisms and debris, as well as the distinctive underwater optical properties. Moreover, most semantic SLAM pipelines rely on feature-level or bounding-box-based strategies, lacking an efficient pixel-level reconstruction solution. To adapt to underwater conditions, we propose a Marine Motion Fusion (MMF) module that combines YOLObased detection for known marine species with optical flow to capture unrecognized motion, excluding all dynamic regions in both the SLAM frontend and NeRF reconstruction. This ensures that only truly static parts of the scene are accurately processed. Also, current semantic SLAM methods do not fully meet the need for high-fidelity mapping in underwater settings. We further employed NeRF for picture level reconstruction. Research has shown that combining NeRF with SLAM can improve both mapping and rendering, as in iMAP Sucar et al. (2021) or NICE-SLAM Zhu et al. (2022), yet these frameworks frequently rely on depth data to bolster reconstruction quality and often struggle with dynamic disturbances or severe optical degradation. Other single-view NeRF-SLAM variants, such as ORBEEZ-SLAM Chung et al. (2023), also face difficulties in underwater scenarios, yielding subpar reconstructions and inadequate handling of moving objects. Although these techniques are effective in controlled indoor environments, they largely assume stable depth inputs and near-static scenes, rendering them ill-suited for real-world underwater domains. To achieve high-accuracy, pixel-level reconstruction in an environment where depth sensors or reliable depth estimations are generally impractical, we leverage neural radiance fields (NeRF) as the SLAM mapping backend. We further introduce specialized loss functions—including light attenuation and color consistency—to address seawater-specific optical issues, such as brightness falloff and color shifts caused by scattering. Our experiments on multiple real underwater datasets demonstrate that the proposed framework outperforms conventional underwater SLAM approaches and other solutions that rely on inaccurate depth estimates or fail to manage dynamic entities. We additionally evaluate its runtime feasibility, showing it can provide high-precision localization and mapping without imposing excessive computational overhead.

In summary, this study integrates SLAM, underwater-specific NeRF design, dynamic object detection, and optical flow into a single system tailored to underwater applications. The result is a novel approach that delivers both dense pixel-level reconstruction and robust detection—an essential capability for tasks demanding detailed spatial mapping and real-time assessment of marine biodiversity. The principal contributions of this work are:

- Optical Flow based dynamic Masking coupling dynamic masking with depth-free dense mapping via the Marine Motion Fusion module, thereby excluding moving objects during SLAM and NeRF processes more efficiently.
- Depth-free high-Quality NeRF reconstruction for underwater environment: We employed underwater-specific NeRF reconstruction constraints to account for optical attenuation and color distortion without depth inputs for dense mapping.
- Comprehensive Experimental Evaluation: We conducted extensive evaluations on real-world underwater video sequences, confirming significant improvements in trajectory accuracy, dense reconstruction quality, and runtime efficiency, thus providing a practical solution for resource-limited underwater platforms.

To evaluate the performance of our framework, we conducted extensive experiments under various underwater conditions, demonstrating its superior performance over traditional SLAM systems in mapping precision and reconstruction quality. Our results highlight the effectiveness of combining NeRF, SLAM, and the MMF module, marking significant progress in underwater scene understanding and dense mapping.

2 Related work

2.1 Related work on traditional SLAM

Filter-based SLAM approaches, such as EKF-SLAM Bailey et al. (2006) and Particle Filter SLAM Thrun (2002), estimate robot pose and map probabilistically but suffer from high computational costs and linearization issues. In contrast, optimization-based methods, notably Graph SLAM Grisetti et al. (2010), reframe SLAM as a pose graph optimization, enabling more accurate and scalable solutions. Visual SLAM has gained prominence due to the richness of camera data, with well-known systems like PTAM Klein and Murray (2007), LSD-SLAM Engel et al. (2014), and ORB-SLAM Mur-Artal et al. (2015); Mur-Artal and Tardós (2017) each offering various balances of direct vs. feature-based techniques. However, these methods generally assume static scenes and often degrade in underwater environments characterized by low visibility, light absorption, and scattering.

Existing deep-learning SLAM methods improve robustness by detecting and masking moving elements. For instance, CNN-SLAM Tateno et al. (2017) and DynaSLAM Bescos et al. (2018) use deep learning to identify dynamic regions, while YOLO Redmon (2016); Redmon and Farhadi (2018) has been integrated with SLAM Bescos et al. (2021) for real-time object detection. Yet, these solutions typically assume land-based imagery and struggle underwater, where domain mismatches result in poor detection accuracy, partial bounding boxes, and overlooked marine life.

2.2 Related work on underwater SLAM methods

Underwater environments pose unique challenges for SLAM, including unpredictable lighting, turbidity, and dynamic marine life. Early work by Eustice et al. Eustice et al. (2006) employed advanced registration to improve underwater image matching, while Negahdaripour and Firoozfam Negahdaripour and Firoozfam (2006) introduced specialized motion estimation in scattering media. Modifications to ORB-SLAM2 for underwater scenarios include applying color corrections and enhancements Liu et al. (2023), as well as leveraging stereo setups and artificial illumination Pizarro et al. (2009). Dual-SLAM Huang et al. (2020) further maintains two parallel threads for robust tracking and refined mapping but can encounter difficulties when rapid environmental changes or marine life movements obscure visual features.

Recent research increasingly focuses on *image preprocessing and feature extraction* tailored for underwater conditions. Our previous work, ULL-SLAM Xin et al. (2023), incorporated low-light enhancement into SLAM's front end, bolstering feature detection under insufficient illumination. Similarly, Zheng et al. Zheng et al. (2023) proposed real-time GAN-based image enhancement for monocular SLAM in turbid waters. While such techniques mitigate visibility issues, they offer only partial solutions when confronting *highly dynamic* scenes, where moving objects can still disrupt feature matching and pose estimation.

Although these methods advance underwater SLAM by addressing poor illumination and color distortion, major gaps persist. Handling the *lack of reliable depth information* and *managing dynamic interference* remain significant barriers to stable pose estimation and dense mapping in complex underwater domains. This work aims to fill these gaps by introducing a framework that integrates dynamic masking and depth-free dense reconstruction methods specifically designed for underwater conditions.

2.3 Related work on Neural Radiance Fields and NeRF-SLAM

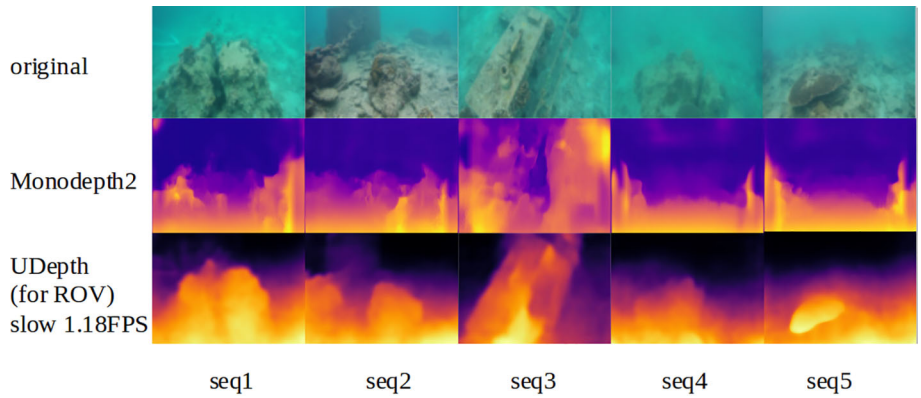
Neural Radiance Fields (NeRF) Mildenhall et al. (2020) have revolutionized scene representation by using neural networks to model the volumetric density and view-dependent emitted radiance at any 3D point in a scene. NeRF optimizes a continuous 5D function (spatial coordinates and viewing directions) to produce photorealistic novel views from input images. While NeRF achieves impressive results, it requires dense sampling and is computationally intensive, often taking hours or days to train on a single scene.

Efforts to improve efficiency include Fast NeRF Garbin et al. (2021) and PlenOctrees Yu et al. (2021), which accelerate rendering times but still face challenges in dynamic scenes and complex environments. Instant Neural Graphics Primitives (instant-ngp) Müller et al. (2022) introduces a multi-resolution hash encoding to achieve real-time rendering and training speeds, significantly reducing computational requirements. NeRF requires highly accurate camera poses, typically obtained using ground-truth systems like COLMAP Schönberger et al. (2016); Schönberger and Frahm (2016). The absence of reliable ground truth pose estimation methods in underwater settings makes this a significant challenge. Furthermore, NeRF assumes high-quality, static input images, which are rarely available in underwater environments characterized by low visibility and dynamic disturbances.

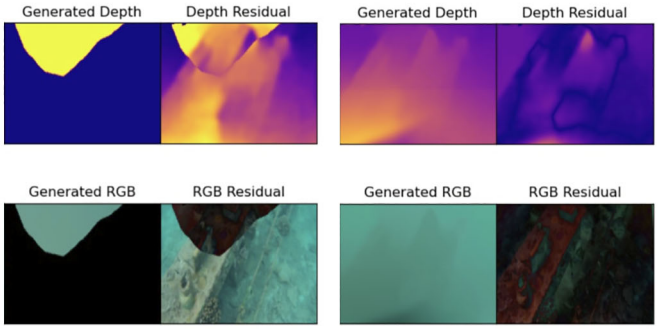
Meanwhile, combining NeRF with SLAM has shown promise for unifying tracking and rendering. iMAP Sucar et al. (2021) and NICE-SLAM Zhu et al. (2022) and VoxGraph Reijgwart et al. (2019) embed neural scene representations into SLAM for real-time dense mapping demonstrate progress in dense mapping for terrestrial environments, yet heavily rely on depth sensors that underperform underwater due to scattering and turbidity. And BARF Lin et al. (2021) jointly optimizes camera parameters and radiance fields. However, these methods typically assume static conditions or include partial depth cues that are unreliable in underwater domains. NeRF-SLAM Rosinol et al. (2023) still depends on depth sensors, limiting its applicability to monocular underwater cameras. Other works like Orbeez-SLAM Chung et al. (2023) merges ORB-SLAM2 with NeRF for monocular dense mapping but struggles underwater when facing sparse features and heavy distortions. As shown in Figure 1, general-purpose models like Monodepth2 Godard et al. (2019) (middle row) struggle to generalize to underwater conditions, producing inconsistent and inaccurate depth maps in areas with low texture or dynamic elements. Specialized networks like UDepth Yu et al. (2023) remain too slow (processing 602 frames in 8.5 minutes) for real-time SLAM. These shortcomings highlight the pressing need for a *depth-free* dense mapping framework that can handle dynamic scenes in underwater environments.

3 Materials and methods

This study introduces a novel integration of YOLO for dynamic object detection, NeRF for dense 3D reconstruction, and SLAM for localization and mapping. This system is meticulously designed for underwater environments, addressing challenges like dynamic interference, low visibility, and occlusions. The system comprises four primary modules: YOLO for real-time object detection and



(a) Comparison of depth maps generated by Monodepth2 and UDepth.



(b) RGB-D methods reconstruction results (depth from UDepth).

FIGURE 1
Underwater depth generation methods and RGB-D reconstruction result. **(a)** Comparison of depth maps generated by Monodepth2 Godard et al. (2019) and a dedicated underwater depth model (UDepth Yu et al. (2023)) across five underwater sequences. The top row shows the original underwater images (seq1 to seq5). The middle row displays the depth maps predicted by Monodepth2 Godard et al. (2019), a general-purpose monocular depth estimation model trained on terrestrial datasets. The bottom row illustrates the depth maps produced by UDepth Yu et al. (2023), a model specifically designed for underwater environments. **(b)** RGB-D reconstruction result using UDepth depth generation method.

masking, a SLAM frontend for feature extraction and tracking, NeRF for photorealistic scene reconstruction. By unifying these components, the system achieves precise localization, robust mapping, and accurate object detection, enabling diverse underwater applications such as marine habitat exploration and underwater archaeology.

3.1 System architecture

The overall system architecture, shown in Figure 2, demonstrates the tight integration of underwater object detection with optical flow, SLAM, and NeRF modules. The underwater object detection with an optical flow module identifies dynamic objects, such as marine life, creating masks to prevent interference with SLAM localization. Simultaneously, NeRF processes the filtered data to generate a dense, high resolution 3D reconstruction of the underwater scene. This architecture supports real-time object detection, dynamic-aware mapping, and environment understanding, making it particularly suitable for underwater environments that require both adaptability and precision. Integrating YOLO for real-time object detection ensures that moving objects do not corrupt SLAM's feature tracking process, which is critical in dynamic underwater environments. This module prevents localization drift and enables real-time object recognition, providing valuable information about the underwater scene. The output from the SLAM backend is further enhanced by NeRF with specially designed underwater loss, which reconstructs the environment with high spatial resolution, enabling photorealistic visualizations as shown in Figure 3. By effectively integrating these components, our system achieves robust localization and accurate mapping while simultaneously generating high-fidelity, photorealistic 3D reconstructions of underwater scenes.

3.2 Marine Motion Fusion module

Underwater environments are inherently dynamic, with fish, plankton, and floating debris (see Figure 4) frequently disrupting

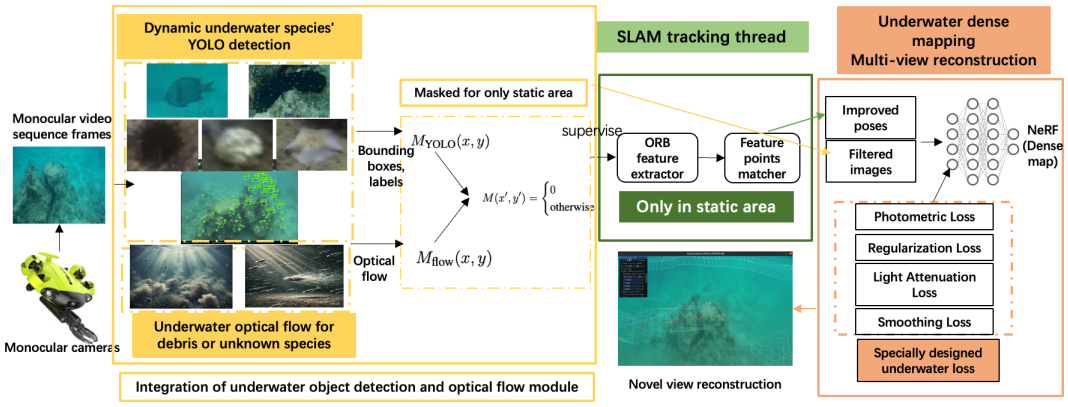


FIGURE 2
System architecture of the proposed YOLO-NeRF-SLAM framework. The yellow modules represent the integration of YOLO detection and optical flow for handling dynamic objects, unknown species, and floating debris in the underwater environment, ensuring only static areas are used for feature matching. The green module highlights the SLAM tracking thread, where ORB features are extracted and matched to improve pose estimation accuracy. The orange module represents the underwater dense mapping and multi-view reconstruction process using NeRF, optimized with specialized loss functions.

SLAM's assumption of static scenes. When confronted by such motion, traditional algorithms suffer from incorrect feature associations and drift. Although many YOLO-SLAM methods work on land, they struggle underwater due to low visibility, light distortion, and intensive marine movement, thus relying solely on YOLO detection is insufficient. Our approach addresses these obstacles by integrating Optical Flow analysis with YOLO, forming a hybrid pipeline that effectively handles both known categories (e.g., fish, holothurians, echinus) and unidentified or subtle background motion. This strategy ensures that only truly static areas contribute to SLAM feature extraction and pose estimation, mitigating erroneous associations and improving localization in dynamic underwater settings.

3.2.1 YOLO detection and exclusion

YOLO treats object detection as a regression problem, predicting bounding boxes and class probabilities in one forward pass. Let the detection results for each input frame be represented in Equation 1:

$$B = \left\{ \left(x_{\min}^{(i)}, \; y_{\min}^{(i)}, \; x_{\max}^{(i)}, \; y_{\max}^{(i)}, \; c^{(i)}, \; s^{(i)} \right) \right\}_{i=1}^{N}, \tag{1}$$

where $x_{\min}^{(i)}$, $y_{\min}^{(i)}$, $x_{\max}^{(i)}$, $y_{\max}^{(i)}$ indicate the top-left and bottom-right coordinates of the i-th bounding box, $c^{(i)}$ is the detected object class (e.g., fish, echinus), and $s^{(i)}$ is the confidence score.

For each input RGB image frame I, the YOLO outputs are converted into a preliminary binary mask $M_{\text{YOLO}}(x, y)$ that identifies known dynamic objects as shown in Equation 2:

$$M_{\text{YOLO}}(x, y) = \begin{cases} 0, \text{ if } (x, y) \text{ lies inside any bounding box in } B, \\ 1, \text{ otherwise}. \end{cases} \tag{2}$$

Any feature points within a YOLO detection box are considered dynamic and thus ignored by SLAM's feature extraction. This step significantly reduces mismatches caused by clearly recognized moving objects. It also ensures that SLAM focuses exclusively on static features, reducing drift and enhancing localization robustness.

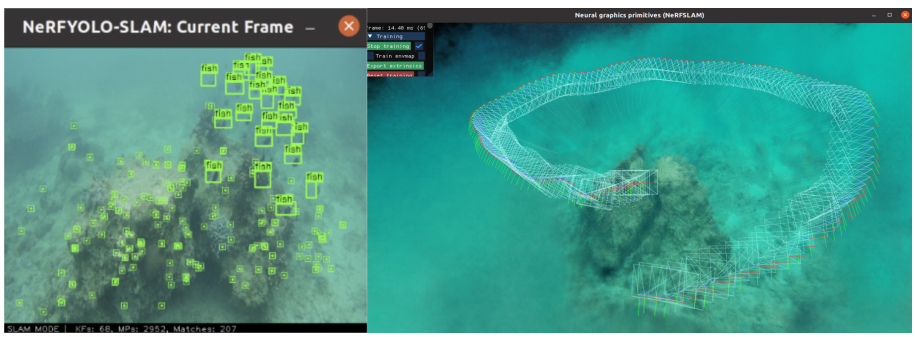


FIGURE 3
An example of the proposed YOLO-NeRF-SLAM system in action. The left panel shows the dynamic object detection module using YOLO, where marine species are detected and marked with bounding boxes. The right panel illustrates the dense 3D reconstruction process powered by NeRF, visualizing the camera trajectory and the reconstructed underwater scene.

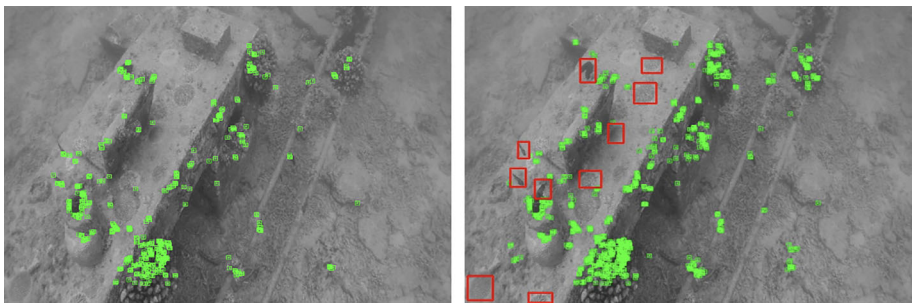


FIGURE 4
The filtered feature extraction comparison. The left image is the unfiltered feature extraction process. And the right image is the filtered feature extraction process.

3.2.2 Optical flow

While YOLO excels at detecting predefined object classes, it cannot capture all dynamic entities, such as non-classified marine debris or subtle background motion caused by water currents. To address this gap, we incorporate the Lucas-Kanade Optical Flow Method Lucas and Kanade (1981), which analyzes pixel-level motion and identify general dynamic regions. The integration of optical flow complements YOLO by expanding the scope of dynamic masking, ensuring that all significant motion, whether class-specific or not, is effectively excluded from SLAM computations.

Optical flow is a computer vision technique that estimates pixel motion between consecutive frames, generating a dense motion field to identify dynamic regions even where there are no distinct features by analyzing apparent motion patterns in underwater sequences. This is especially valuable in scenes with sandy seabeds or open water, where feature-based methods often fail, and in scenarios featuring unpredictable motion from marine life, both of which demand pixel-level motion analysis.

The motion vector $v(x,y)$of optical flow is calculated in Equation 3 as:

$$v(x,y) = \sqrt{(u(x,y))^2 + (v(x,y))^2}, \tag{3}$$

where $u(x,y)$ and $v(x,y)$ are the horizontal and vertical components of optical flow, respectively. A threshold τ is applied to classify pixels with significant motion as shown in Equation 4:

$$M_{\text{flow}}(x,y) = \begin{cases} 1, \text{ if } v(x,y) > \tau, \\ 0, \text{ otherwise}. \end{cases} \tag{4}$$

Any feature lying in these high-motion areas is also excluded, extending the dynamic filtering beyond YOLO's known classes.

3.2.3 Fusion of YOLO and optical flow

Our approach combines underwater object detection capabilities with optical flow motion analysis to achieve precise dynamic exclusion. We combine bounding box detection with pixel-level motion to form a comprehensive exclusion strategy, as shown in Equation 5:

$$M(x,y) = \begin{cases} 0, \text{ if } (x,y) \in \text{YOLO regions or high} - \text{motion areas (optical flow)}, \\ 1, \text{ otherwise}. \end{cases} \tag{5}$$

Here, $(x,y) \in$ YOLO regions indicates that the features within the bounding box detected by YOLOv5, and $(x,y) \in M_{\text{flow}}(x,y)$ indicates that the features are classified as a high-motion area by optical flow. In practice, rather than explicitly masking images, *we simply do not use* any feature points for SLAM if:

$$M(x,y) = 0.$$

Hence, our SLAM pipeline omits key points within bounding boxes or high-motion areas. Integrating optical flow with YOLOv5 enables the system to handle both known object classes and unclassified motion, enhancing robustness by capturing subtle movements, improving precision by minimizing dynamic noises. Restricting SLAM computations to genuinely static elements reduces the noise and drift often encountered in conventional systems when faced with fish, drifting debris, or other moving targets. In addition, isolating the static background prevents artifacts in NeRF's 3D reconstructions, mitigating localization drift and mapping inconsistencies. Consequently, our YOLO-based detection and dynamic masking strategy proves resilient even under highly dynamic underwater conditions.

3.3 Dynamic-aware SLAM frontend for pose estimation

Once dynamic regions are excluded by the YOLO+Optical Flow pipeline, the SLAM frontend focuses on extracting features from the remaining static areas. This dynamic-aware strategy is crucial for underwater scenarios, where motion interference from marine life can easily degrade pose estimation.

3.3.1 Feature extraction and matching with mask

We adopt ORB (Oriented FAST and Rotated BRIEF) features in static regions only. Let $M(x,y)$ be the binary mask indicating whether pixel (x,y) is static ($M = 1$) or belongs to a detected dynamic region ($M = 0$). The FAST detector is then applied only

where $M(x, y) = 1$, ignoring any feature points inside YOLO bounding boxes or high-motion optical flow areas.

Each ORB feature $\mathbf{f}_i = (\mathbf{p}_i, \mathbf{d}_i)$ consists of: $\mathbf{p}_i = (x_i, y_i)$: the 2D keypoint location, valid only if $M(x_i, y_i) = 1$, $\mathbf{d}_i$: the corresponding BRIEF descriptor, typically 256 bits.

To match features between consecutive frames, we compute the Hamming distance (Equation 6):

$$C(\mathbf{f}_i, \quad \mathbf{f}_j) \; = \; \sum_{l=1}^{L} \left| \quad d_{il} - d_j \quad _l \right|, \tag{6}$$

and retain descriptor pairs with the smallest distance (below a threshold). Critically, any feature fi falling in dynamic regions (where $M(x_i, y_i) = 0$) is discarded at this stage. From Figure 4, we can see that all he moving stuffs like fishes have been excluded from the feature extraction part. Thus, only features from static areas contribute to matching.

3.3.2 Pose estimation with RANSAC and masked residuals

Let Ω_{static} denote the set of matched keypoint pairs that survive the dynamic mask in frames k and $k + 1$. We aim to find the relative pose (Equation 7) $\mathbf{T}_{k, \quad k+1} \in \text{SE}(3)$, composed of a rotation $\boldsymbol{R} \in \text{SO}(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$:

$$\mathbf{T}_{k, \quad k+1} \; = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0^\top & 1 \end{bmatrix} \; . \tag{7}$$

3.3.3 Masked reprojection error

For each match $i \in \Omega_{\text{static}}$, let $\mathbf{p}_i^{(k+1)}$ be the observed 2D location in frame $k + 1$, and $\mathbf{P}_i^{(k)}$ the corresponding 3D point in frame k's coordinate system. We minimize the total reprojection error:

$$\mathcal{E} = \sum_{i \in \Omega_{\text{static}}} \| \, \mathbf{p}_i^{(k+1)} - \; \pi(\mathbf{T}_{k,\, k+1} \, \mathbf{P}_i^{(k)}) \, \|^2, \tag{8}$$

where $\pi(\cdot)$ projects a 3D point onto the image plane via the intrinsic matrix $\mathbf{K}$. Since Ω_{static} only contains key points passing the mask test ($M = 1$), dynamic outliers are excluded from the start.

We use RANSAC (Random Sample Consensus) to robustly find $\mathbf{T}_{k,\, k+1}$: a minimal subset of matches is sampled to estimate an initial pose, and the number of inliers with low reprojection error is counted. Iterating over multiple samples yields the best-fitting transformation. Finally, a nonlinear refinement (e.g., Levenberg-Marquardt) over all inliers in Ω_{static} further minimizes Equation 8. Removing dynamic points beforehand makes the RANSAC procedure less prone to spurious matches, thus improving pose accuracy in underwater applications.

3.4 NeRF-based reconstruction for underwater scenes

NeRF (Neural Radiance Fields)Mildenhall et al. (2020) replaces the traditional SLAM backend mapping to achieve dense and high-quality 3D reconstruction in this framework. Unlike conventional SLAM, which often generates sparse or semi-dense maps, NeRF parametrizes the scene as a volumetric radiance field, allowing for a more accurate representation of fine details. Instant-NGP Müller et al. (2022) is utilized as the NeRF baseline to ensure real-time performance. This is essential in scenarios where computational resources are limited or real-time feedback is necessary.

For each 3D point $\mathbf{x} = (x, y, z)$, NeRF models the scene by learning a volumetric density $\sigma(\mathbf{x})$ and an RGB color $\mathbf{c}(\mathbf{x}, \mathbf{d})$, where $\boldsymbol{d}$ represents the viewing direction. The observed pixel color $\mathbf{C}(\mathbf{r})$ along a ray $r(t)$ is computed using the following volumetric rendering Equation 9:

$$\mathbf{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d}) \, dt, \tag{9}$$

where $T(t)$ denotes the transmittance along the ray up to point t, defined as Equation 10:

$$T(t) = \exp\left(- \int_{t_n}^{t} \sigma(\mathbf{r}(s)) \, ds \right), \tag{10}$$

representing the probability that the ray is not occluded up to t. The term $\sigma(\mathbf{r}(t))$ represents the density at a point along the ray, and $\mathbf{c}(\mathbf{r}(t), \mathbf{d})$ gives the color at that point as a function of the viewing direction $\boldsymbol{d}$.

The volumetric rendering process ensures the scene is reconstructed with fine details, even in dynamic underwater environments. By integrating this into our SLAM pipeline, the system achieves real-time performance and significantly enhances the quality of the reconstructed scene compared to traditional SLAM methods.

Once the volumetric rendering formula is incorporated, we optimize the NeRF parameters through gradient-based learning. The following section introduces four core losses used during training, two of which (*Light Attenuation Loss* and *Smoothing Loss*) are especially adapted to underwater conditions.

3.5 Loss functions for underwater NeRF

The training of our underwater NeRF model involves a combination of standard and specialized loss functions designed to address the unique challenges of underwater environments. While the photometric loss Mildenhall et al. (2020) and regularization loss Krogh and Hertz (1991) are common in NeRF-based methods, our framework introduces two additional loss functions: a light attenuation loss to compensate for color distortion caused by wavelength-dependent absorption in water and a smoothing loss to mitigate noise and discontinuities in the reconstructed geometry. This combination ensures high-fidelity underwater 3D reconstructions with improved realism and robustness.

3.5.1 Base loss

Photometric Loss We adopt a standard photometric (reconstruction) loss by comparing the rendered color C_i^{pred} against the ground-truth color C_i^{gt} for each static pixel i:

$$\mathcal{L}_{\text{photo}} = \frac{1}{N}\sum_{i=1}^{N} \| C_i^{\text{pred}} - C_i^{\text{gt}} \|^2, \tag{11}$$

where N in the Equation 11 is the number of static pixels used in training. This loss encourages the network to synthesize realistic colors consistent with actual underwater imagery Mildenhall et al. (2020).

To prevent overfitting and encourage a smoother density field, we impose a simple L_2 penalty on the NeRF parameters θ:

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{reg}} \quad \| \theta \|_2^2, \tag{12}$$

where λ_{reg} is a hyperparameter controlling regularization strength Krogh and Hertz (1991). Equation 12 helps stabilize training, especially when only a limited number of high-quality underwater images are available.

3.5.2 Light attenuation loss

A key challenge in underwater environments is wavelength-dependent light absorption, leading to color casts and brightness decay over distance. To address this issue, we incorporate a light attenuation loss inspired by underwater image formation models Akkaynak and Treibitz (2018):

For each 3D sample along the ray, we postulate that color $\mathbf{c}_{\text{base}}(\mathbf{x}, \mathbf{d})$ is further modulated by an exponential factor:

$$\mathbf{c}_{\text{water}}(\mathbf{x}, \mathbf{d}) = \mathbf{c}_{\text{base}}(\mathbf{x}, \mathbf{d}) \odot \exp(-\alpha\, d(\mathbf{x})), \tag{13}$$

where $d(\mathbf{x})$ in the Equation 13 is the distance from the camera (or light source) to $\boldsymbol{x}$, and $\alpha = [\alpha_r, \alpha_g, \alpha_b]^\top$ is a vector of learnable absorption coefficients for each color channel.

Upon rendering, the resulting pixel color $\mathbf{C}^{\text{pred}}(\alpha)$ depends on α. We compare it to the observed color $\mathbf{C}^{\text{gt}}$ as shown in Equation 14:

$$\mathcal{L}_{\text{atten}} = \frac{1}{N}\sum_{i=1}^{N} \| \mathbf{C}_i^{\text{pred}}(\alpha) - \mathbf{C}_i^{\text{gt}} \|^2 . \tag{14}$$

By minimizing $\mathcal{L}_{\text{atten}}$, the network learns appropriate attenuation coefficients to correct underwater color distortion. This term is particularly beneficial for scenes with large distance variations or strong wavelength dependent absorption.

3.5.3 Smoothing loss

Even after accounting for attenuation, the reconstructed geometry or radiance fields may exhibit spurious noise and sharp discontinuities—especially in underwater scenes with uneven illumination. A Smoothing Loss as shown in Equation 15 penalizes abrupt changes in the density or surface geometry, encouraging more natural surfaces:

$$\mathcal{L}_{\text{smoothing}} = \frac{1}{M}\sum_{j=1}^{M} \| \nabla\sigma(\mathbf{x}_j) \|^2, \tag{15}$$

where $\nabla\sigma(\mathbf{x}_j)$ denotes the gradient of density with respect to spatial coordinates at sample point $\mathbf{x}_j$. This approach—similar to surface regularization used in mesh or implicit-surface reconstructions—helps eliminate high-frequency artifacts and yields more coherent underwater surfaces.

4 Experiments and results

In this section, we compare our proposed method against several state-of-the-art SLAM systems in challenging underwater scenarios. We choose ORB-SLAM3 Campos et al. (2021) as our primary baseline due to its versatility in monocular, stereo, and RGB-D setups with robust feature-based tracking. We also evaluate the original ORB-SLAM2 Mur-Artal and Tardós (2017), ORBEEZ-SLAM Chung et al. (2023), Dual-SLAM Huang et al. (2020), iMap Sucar et al. (2021), and NICE-SLAM Zhu et al. (2022).

- ORB-SLAM2 Mur-Artal and Tardós (2017) is a classic feature-based SLAM that supports monocular, stereo, and RGB-D inputs. It remains a popular choice for both academic benchmarks and real-world applications.
- ORB-SLAM3 Campos et al. (2021) extends ORB-SLAM2 by integrating inertial data and a refined system architecture, delivering robust performance across multiple sensor modalities.
- ORBEEZ-SLAM Chung et al. (2023) combines ORB-SLAM2 for camera tracking with InstantNGP Müller et al. (2022) for mapping, enabling real-time, high-quality 3D reconstruction on land.
- Dual-SLAM Huang et al. (2020) operates by maintaining two parallel SLAM threads—one focusing on accuracy, the other on robustness.
- iMap Sucar et al. (2021) is an implicit mapping approach that incrementally builds a scene representation in real-time using a neural SDF (Signed Distance Function). Although primarily used for RGB-D datasets on indoor and static scenarios, we include it here for completeness.
- NICE-SLAM Zhu et al. (2022) expands on neural implicit mapping by jointly optimizing camera poses and a grid-based latent code for the scene. Similar to iMap, used for RGB-D datasets.

4.1 Implementation details

We use YOLOv5 as our basic object detection model, which is typically implemented in Python based on the PyTorch framework. At the same time, the SLAM system is a visual SLAM framework written in C++. To integrate the two and enable real-time semantic mapping, this paper adopts LibTorch Imambi et al. (2021) as the solution, leveraging it to combine the object detection module with the SLAM system efficiently.

To adapt YOLO for underwater object detection, we trained it on the RUOD Fu et al. (2023) dataset, which contains 14,000 underwater images and 74,903 annotated objects, covering 10 categories of underwater targets. This dataset includes the

primary object categories relevant to underwater research: holothurian, echinus, starfish, scallop, and fish.

Our training and inference processes were conducted on an NVIDIA RTX 3090 GPU, providing the computational power necessary to handle the tasks of real-time SLAM.

4.2 Datasets

URPC Dataset. We adopt the URPC dataset Liu et al. (2021), which provides underwater imagery with multiple underwater marine life (see Figure 5). We follow the same data structures as in the paper ULLSLAM Xin et al. (2023). According to ULL-SLAM Xin et al. (2023), the URPC dataset's camera intrinsics and distortion parameters were obtained through COLMAP Schönberger and Frahm (2016); Schönberger et al. (2016) as pseudo-ground truth. COLMAP jointly optimizes camera poses and intrinsic parameters by matching sparse features across underwater image sequences. While COLMAP's model does not explicitly model underwater refraction, the implicit compensation through reprojection error minimization partially accounts for refractive effects in shallow-water scenarios with stable optical conditions. Specifically, COLMAP applies the Brown–Conrady photogrammetric model to approximate lens distortion, which includes radial distortion (k_1,k_2) and tangential distortion (p_1,p_2) coefficients. COLMAP's sparse bundle adjustment then minimizes the reprojection error, implicitly absorbing refractive effects into the optimized parameters Jordt-Sedlazeck and Koch (2013). And underwater calibration can approximate refraction via changes in focal length and distortion Jordt-Sedlazeck and Koch (2013).

Self-Collected Dataset With Pseudo Ground-truth. Additionally, we capture five underwater video segments using a FIFISH ROV with a 4K camera around Wuzhizhou Island in Sanya, Hainan Province, China, at diving depths of 6–10m. The recorded segments with multiple marine life (see Figure 5) contain 1226, 2506, 1250, 3669, and 1204 frames, respectively. We continue to use the previous method of obtaining ground truth poses for underwater SLAM datasets Ferrera et al. (2019); Rahman et al. (2022), employing COLMAP Schönberger and Frahm (2016); Schönberger et al. (2016) to obtain ground-truth camera trajectories under challenging underwater conditions. As in the URPC dataset, the COLMAP pipeline here also adopts the Brown–Conrady distortion model, with intrinsic parameters and distortion coefficients estimated directly from image sequences. Although this method does not explicitly model the underwater refraction physical process, as Telem and Filin (2010) states, the empirical distortion parameters effectively compensate for the refraction effect in shallow water environments.

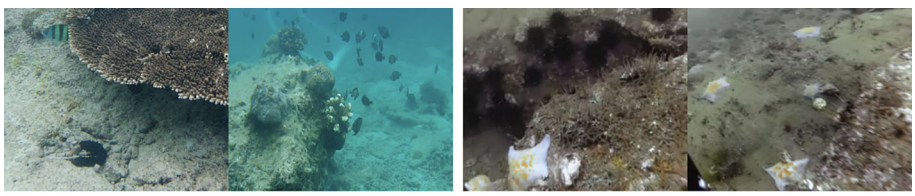
EASI Tank Dataset With Real Ground Truth. To further test our model, we select a public underwater SLAM dataset named EASI Yang et al. (2023). This dataset was collected in a controlled turbidity water tank (1.8m length, 1.1m depth) equipped with high-contrast feature patterns at the bottom to enhance SLAM feature detection. Ground truth was acquired via four overhead Vicon tracking cameras covering a 3×3m area: a dual-sided rigid frame held an underwater GoPro camera for image capture, while aerial-side retroreflective markers enabled synchronized 6-DOF pose recording by the Vicon system with 0.1mm positional accuracy. The intrinsic parameters of the underwater camera used here is provided in Yang et al. (2023), we directly used their provided radial distortion (k_1,k_2) and tangential distortion (p_1,p_2) coefficients to solve the approximates refractive effects.

4.3 Evaluation metrics

To evaluate SLAM performance, we utilize three key metrics: (1) absolute trajectory error (ATE), (2) relative pose error (RPE), and (3) initialization efficiency. The ATE measures the deviation between the ground truth camera trajectory and the trajectory estimated by SLAM. RPE evaluates the accuracy of relative motion estimation between consecutive frames. The initialization efficiency evaluates how quickly SLAM initialization occurs, represented by the number of frames required for successful initialization. To evaluate the reconstruction mapping result, we test the rendered image quality using the Peak signal-to-noise ratio (PSNR) and structural Similarity Index (SSIM).

4.4 SLAM performance analysis

This section evaluates the performance of various SLAM systems on two datasets: the URPC dataset and a self-collected dataset. For each dataset, we compare different SLAM methods based on two key metrics: Absolute Trajectory Error (ATE) and Relative Pose Error (RPE).



(a) Self-collected dataset example (b) URPC dataset example

FIGURE 5
Schematic illustrations of marine organisms in the dataset. **(a)** is example image of self-collected dataset. **(b)** is example image of URPC dataset.

4.4.1 URPC dataset

Monocular SLAM quantitative result(ATE, RPE) From Table 1, it is clear that our method achieves the best results on the URPC dataset, with the lowest values for both ATE and RPE. Our method shows its superiority by significantly reducing errors compared to the other methods, making it the most accurate solution.

Monocular SLAM trajectory result The Figure 6 compares the performance of several SLAM methods. (a) The trajectory comparison shows that while all methods follow the general path of the ground truth (GT), our method exhibits the closest match, with smaller deviations than methods like ORB-SLAM3. (b) Fitting results indicate that our method performs better in aligning with the *X*, *Y*, and *Z* axes and maintains more consistent rotational estimates (roll, pitch, and yaw) compared to the others. (c) The boxplot of Absolute Trajectory Error (ATE) highlights Our's superior performance, with the smallest median error and fewer large deviations, confirming its higher accuracy and stability over the other methods.

TABLE 1 Quantization errors of four different monocular SLAM systems on URPC dataset.

Method	ATE↓	RPE↓
ORB-SLAM2 Mur-Artal and Tardós (2017)	2.472832	0.047186
Dual-SLAM Huang et al. (2020)	2.383535	0.061062
ORB-SLAM3 Campos et al. (2021)	2.858885	0.050863
ORBEEZ-SLAM Chung et al. (2023)	2.425408	0.044877
Ours	**2.069632**	**0.044870**

↓ means the lower the number the better result. The bold text indicates the best performance of the method under the corresponding evaluation index.

4.4.2 Self-collected dataset

For the self-collected dataset, we compare the same four monocular methods and also include three RGB-D SLAM methods. The RGB-D methods, which use depth information from the UDepth Yu et al. (2023), are expected to provide more accurate results compared to monocular SLAM.

Monocular SLAM quantitative result(ATE, RPE, Initialization) From Table 2, the results demonstrate that our method consistently achieves superior localization accuracy and robustness. This improvement is attributed to the integration of YOLO for dynamic object detection and optical flow for masking, which ensures that feature matching is limited to static regions, reducing ambiguities caused by moving objects.

Monocular SLAM trajectory result Figure 7 illustrates the estimated trajectories for all five monocular SLAM systems. While ORB-SLAM2 performs well in static regions, it suffers from significant drift in dynamic environments. ORBEEZ-SLAM offers moderate improvements but struggles with fast-moving objects. In contrast, our system maintains stable and accurate trajectories under all tested conditions, demonstrating robustness and accuracy in challenging underwater conditions with low visibility and dynamic interference.

Quantitative result compared with RGB-D SLAM From Table 3, the comparison confirms that our monocular SLAM method outperforms the RGB-D SLAM methods in some sequences, demonstrating its ability to achieve higher accuracy

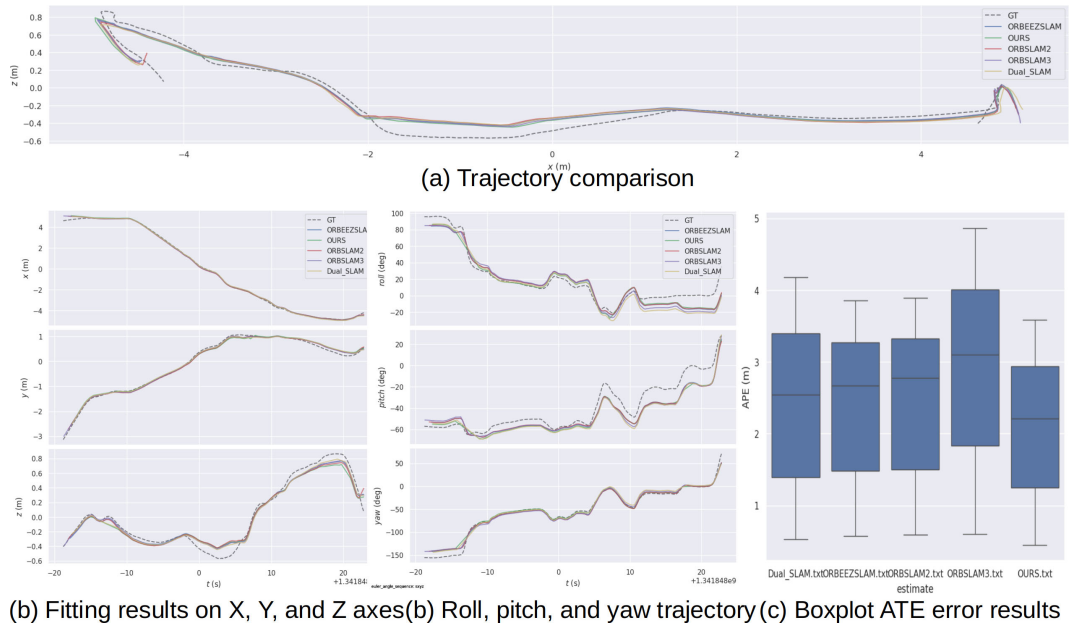


FIGURE 6
The estimated trajectory performance of different SLAM systems on URPC dataset. **(a)** Trajectory comparison between ground truth in different SLAM methods. **(b)** Fitting results on the X, Y, and Z axes as well as roll, pitch, and yaw trajectories. **(c)** Boxplot of Absolute Trajectory Error (ATE) for different methods.

TABLE 2 Our method and four monocular SLAM systems performed in five segments of real underwater challenge environments provided by our self-captured videos.

Video clips	Method	ATE↓	RPE↓	Initialization↓
seg1	ORB-SLAM2 Mur-Artal and Tardós (2017)	2.892000	0.014883	19
	ORB-SLAM3 Campos et al. (2021)	2.628203	0.110978	15
	Dual-SLAM Huang et al. (2020)	2.441731	0.217128	15
	ORBEEZ-SLAM Chung et al. (2023)	2.361609	0.045725	22
	Ours	1.965533	0.036166	13
seg2	ORB-SLAM2 Mur-Artal and Tardós (2017)	3.496794	0.430973	17
	ORB-SLAM3 Campos et al. (2021)	2.857591	0.056270	11
	Dual-SLAM Huang et al. (2020)	2.559526	0.039886	10
	ORBEEZ-SLAM Chung et al. (2023)	3.248732	0.038621	11
	Ours	3.228299	0.232200	8
seg3	ORB-SLAM2 Mur-Artal and Tardós (2017)	3.394621	0.696246	21
	ORB-SLAM3 Campos et al. (2021)	3.350094	0.014883	21
	Dual-SLAM Huang et al. (2020)	3.523472	0.379043	16
	ORBEEZ-SLAM Chung et al. (2023)	2.908638	0.160419	24
	Ours	2.492974	0.148776	16
seg4	ORB-SLAM2 Mur-Artal and Tardós (2017)	3.196740	0.179387	19
	ORB-SLAM3 Campos et al. (2021)	2.981525	0.122864	16
	ORBEEZ-SLAM Chung et al. (2023)	2.617896	0.066356	14
	Dual-SLAM Huang et al. (2020)	2.887165	0.261579	17
	Ours	2.547515	0.064097	14
seg5	ORB-SLAM2 Mur-Artal and Tardós (2017)	2.807136	0.077025	10
	ORB-SLAM3 Campos et al. (2021)	1.729453	0.064448	12
	ORBEEZ-SLAM Chung et al. (2023)	2.745240	0.048803	14
	Dual-SLAM Huang et al. (2020)	2.963576	0.207393	19
	Ours	1.269742	0.083663	11

Under the evaluation index of the SLAM system, our method can achieve greater results in challenging dynamic underwater environments compared with other systems.
↓ means the lower the number the better result.

and robustness even without depth information. This highlights the effectiveness of our approach in challenging scenarios where RGB-D methods typically rely on depth maps, which our method handles with similar or better precision.

4.4.3 EASI Tank Dataset with real ground truth

Since the previous two datasets relied on pseudo-ground truth to evaluate SLAM performance, we further validate our model's effectiveness by introducing the EASI dataset—an underwater dataset with real ground truth—for more rigorous verification.

Monocular SLAM quantitative result(ATE) From Table 4, it is clear that the absolute trajectory error (ATE) comparison reveals significant limitations in both ORB-SLAM2 and ORBEEZ-SLAM. This performance constraint stems from their shared architectural foundation - ORBEEZ-SLAM directly inherits ORB-SLAM2 as its baseline framework. While ORBEEZ-SLAM attempts to improve upon the original system through various modifications, its fundamental SLAM pipeline remains constrained by ORBSLAM2's inherent design choices. The quantitative results demonstrate that neither approach achieves satisfactory accuracy in our underwater evaluation scenarios, suggesting that the baseline architecture itself may be ill-suited for challenging aquatic environments. This shared limitation highlights the need for more specialized underwater SLAM architectures rather than incremental improvements to existing land-based systems. Our method shows its superiority by significantly reducing errors compared to the other methods, making it the most accurate solution.

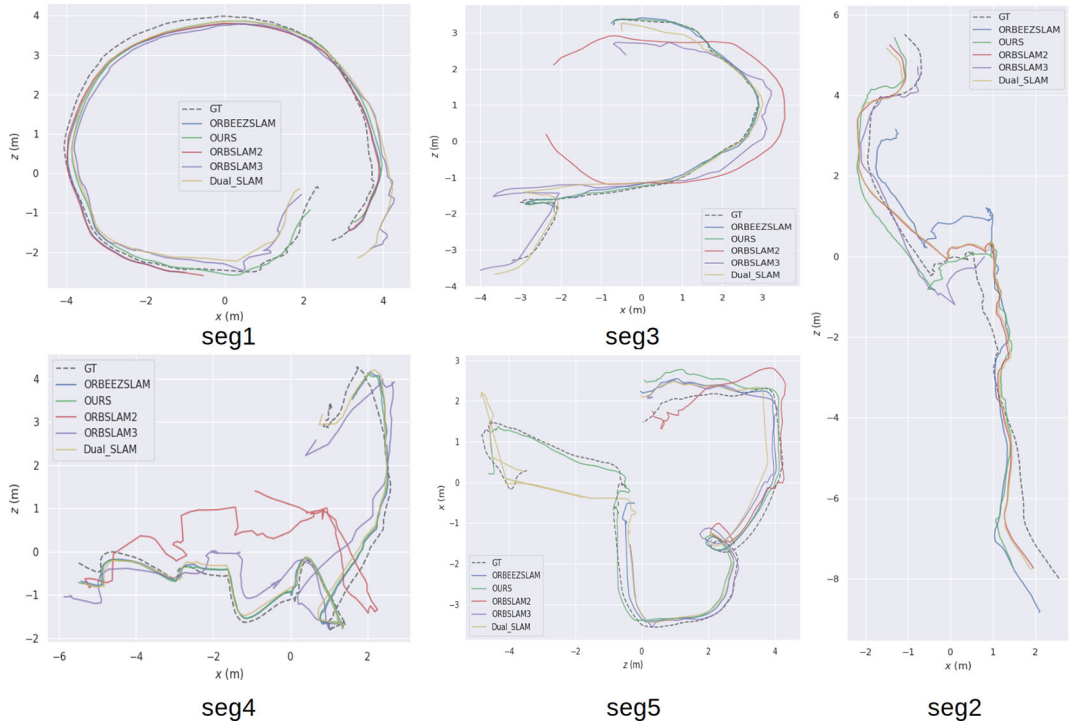


FIGURE 7
The estimated trajectory performance of different SLAM systems across five real-life underwater video segments. The dashed black line represents the ground truth (GT) trajectory, while the red, green, blue, purple, and yellow lines represent ORB-SLAM2, our proposed method, ORBEEZ-SLAM, ORBSLAM3, and Dual-SLAM, respectively.

TABLE 3 Our method and three other RGB-D SLAM systems performed in five segments of real underwater challenge environments provided by our self-captured videos.

Video clips	Method	ATE↓	RPE↓
seg1	i-MAP Sucar et al. (2021)+UDepth Yu et al. (2023)	2.297465	0.057245
	NICE-SLAM Zhu et al. (2022)+UDepth Yu et al. (2023)	2.248826	0.049274
	ORBEEZ-SLAM Chung et al. (2023)+UDepth Yu et al. (2023)	1.963923	0.037347
	Ours	1.965533	0.036166
seg2	i-MAP Sucar et al. (2021) +UDepth Yu et al. (2023)	3.927464	0.328764
	NICE-SLAM Zhu et al. (2022)+UDepth Yu et al. (2023)	3.582692	0.285217
	ORBEEZ-SLAM+UDepth Yu et al. (2023)	3.319824	0.221976
	Ours	3.228299	0.232200
seg3	i-MAP Sucar et al. (2021)+UDepth Yu et al. (2023)	2.932657	0.187153
	NICE-SLAM Zhu et al. (2022)+UDepth Yu et al. (2023)	2.738374	0.183642
	ORBEEZ-SLAM+UDepth Yu et al. (2023)	2.452714	0.148923
	Ours	2.492974	0.148776
seg4	i-MAP Sucar et al. (2021)+UDepth Yu et al. (2023)	2.648826	0.089274
	NICE-SLAM Zhu et al. (2022)+UDepth Yu et al. (2023)	2.578376	0.084736
	ORBEEZ-SLAM+UDepth Yu et al. (2023)	2.562716	0.071289
	Ours	2.547515	0.064097
seg5	i-MAP Sucar et al. (2021)+UDepth Yu et al. (2023)	1.473563	0.108463
	NICE-SLAM Zhu et al. (2022)+UDepth Yu et al. (2023)	1.753552	0.174548
	ORBEEZ-SLAM+UDepth Yu et al. (2023)	1.249617	0.092353
	Ours	1.249742	0.083663

↓ means the lower the number the better result.

TABLE 4 Our method and three monocular SLAM systems performed in five segments of tank experiments provided by EASI dataset.

Video clips	Method	ATE↓
seg1	ORB-SLAM2 Mur-Artal and Tardós (2017)	2.826594
	ORB-SLAM3 Campos et al. (2021)	0.019573
	ORBEEZ-SLAM Chung et al. (2023)	2.732843
	Ours	0.019274
seg2	ORB-SLAM2 Mur-Artal and Tardós (2017)	2.318397
	ORB-SLAM3 Campos et al. (2021)	0.018925
	ORBEEZ-SLAM Chung et al. (2023)	2.467274
	Ours	0.018736
seg3	ORB-SLAM2 Mur-Artal and Tardós (2017)	1.928345
	ORB-SLAM3 Campos et al. (2021)	0.029384
	ORBEEZ-SLAM Chung et al. (2023)	1.293847
	Ours	0.019868
seg4	ORB-SLAM2 Mur-Artal and Tardós (2017)	2.493547
	ORB-SLAM3 Campos et al. (2021)	0.017291
	ORBEEZ-SLAM Chung et al. (2023)	2.837644
	Ours	0.018283
seg5	ORB-SLAM2 Mur-Artal and Tardós (2017)	failed
	ORB-SLAM3 Campos et al. (2021)	0.035837
	ORBEEZ-SLAM Chung et al. (2023)	failed
	Ours	0.023938

Under the evaluation index of the SLAM system, our method can achieve greater results in challenging dynamic underwater environments compared with other systems.
↓ means the lower the number the better result.

Monocular SLAM trajectory result The Figure 8 compares the performance of four SLAM methods on seg 1. (a) The trajectory comparison shows that ORB-SLAM2 and Orbeez-SLAM does not follow the basic line of the general ground truth while ORB-SLAM3 and our method follow the general path of the ground truth (GT), our method exhibits the closest match, with smaller deviations than methods like ORB-SLAM3. (b) Fitting results indicate that our method performs better in aligning with the X, Y, and Z axes compared to the others.

In conclusion, our proposed method demonstrates superior performance in challenging underwater environments by effectively addressing dynamic interference and low visibility. Integrating YOLO, optical flow enables robust localization, setting a new benchmark for underwater SLAM systems.

4.5 Novel view reconstruction results

In this section, we generate novel view images from several dense SLAM approach's 3D reconstruction to visually assess their performance under dynamic underwater conditions. Other dense approaches tend to blur finer textures or exhibit color shifts around moving objects, whereas our YOLO-NeRFSLAM consistently preserves scene detail and color fidelity.

Figure 9 illustrates the mapping outcomes of several dense mapping results from SLAM systems and colmap systems in underwater scenarios, highlighting their respective dense or semi-dense reconstruction traits. Colmap relies on sparse feature points, resulting in point clouds that capture only partial scene details.

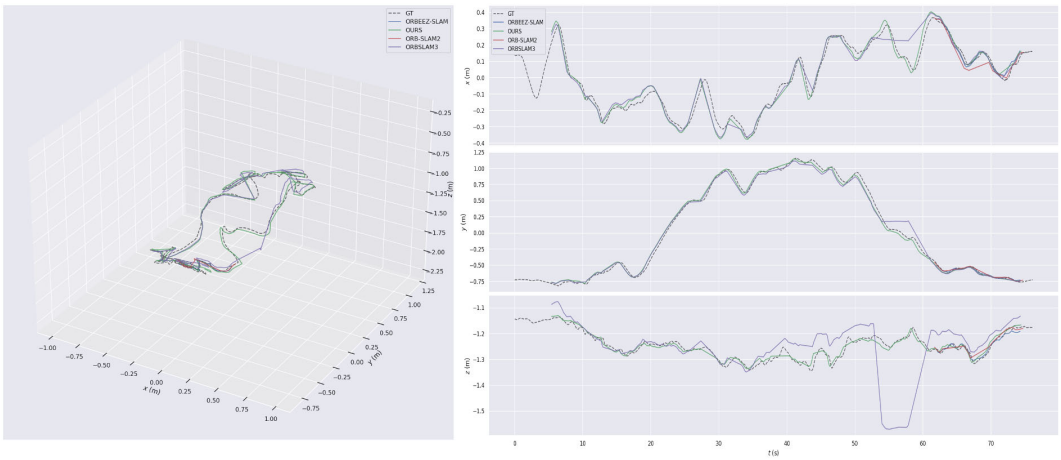
In contrast, methods such as iMAP Sucar et al. (2021) and NICE-SLAM Zhu et al. (2022) attempt to fuse neural scene representations into SLAM pipelines, using estimated depth and color maps to produce denser reconstructions. However, as both were originally designed for near-static indoor environments, they frequently yield depth errors or visible residual artifacts under high scattering or dynamic disturbances common in underwater settings. Similarly, ORBEEZ-SLAM Chung et al. (2023) combines monocular ORB-SLAM2 with NeRF for a form of dense mapping, yet in the presence of drifting debris or fish, its results often appear blurred or incomplete due to insufficient handling of dynamic interference.

By contrast, our framework leverages the Marine Motion Fusion (MMF) module to robustly exclude most underwater dynamic elements at the SLAM frontend, then employs a custom NeRF-based reconstruction approach specifically tuned for underwater conditions. Operating without dedicated depth sensors, our system still captures significantly richer details of seabed structures and marine objects (see the rightmost column in Figure 10), even in low-texture or scattering-heavy regions. The introduced dynamic masking further mitigates the artifacts and distortions that typically arise from fish activity or uneven optical attenuation, leading to a more coherent and accurate dense reconstruction. Overall, our approach merges effective dynamic exclusion with an underwater-specific NeRF pipeline, offering a more comprehensive solution for high-fidelity mapping in challenging marine environments.

Novel view reconstruction's numerical result is a key metric for evaluating SLAM systems, as it directly reflects the quality and fidelity of the generated 3D scene. We compared our method with ORBEEZSLAM Chung et al. (2023) to evaluate its performance. Metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) were used to assess reconstruction quality, as shown in Table 5.

Our method outperforms ORBEEZ-SLAM and the baseline across all evaluated scenes, achieving higher PSNR and SSIM values while maintaining lower LPIPS values. For instance, in segment 2, our system achieved a PSNR of 23.40, SSIM of 0.79, and LPIPS of 0.19, compared to ORBEEZ-SLAM's 20.62, 0.71, and 0.18. The inclusion of specialized loss functions plays a crucial role in preserving fine details and structural accuracy under challenging underwater conditions.

Figure 10 visually compares reconstruction results across five scenes. The first column shows results from ORBEEZ-SLAM, which struggles with dynamic interference and produces blurred reconstructions with noticeable artifacts. The second column



(a) EASI seg1 trajectory comparison (b) Fitting result of EASI seg1 on X, Y, and Z axes

FIGURE 8
The estimated trajectory performance of different SLAM systems on EASI dataset seg 1. **(a)** EASI seg1 trajectory comparison. **(b)** Fitting result of EASI seg1 on X, Y, and Z axes.

shows results from our system without the specialized loss functions, highlighting improved structural fidelity but lacking fine details in complex regions. The third column demonstrates the full capabilities of our method, with sharper edges, enhanced textures, and higher structural consistency, particularly in dynamic and low-light conditions.

4.6 Real-time analysis

In this section, we compare the runtime performance of our method with state-of-the-art NeRF-based SLAM models on our dataset. The runtime is evaluated for frames per second (fps), where higher values indicate better real-time performance. Despite incorporating additional components such as target detection and optical flow for enhanced accuracy, our method remains real-time, showing competitive performance when compared to existing models. Table 6 summarizes the results for five different segments of the dataset, demonstrating that our method performs efficiently even with these additional features.

4.7 Ablation study

To thoroughly evaluate the contributions of various components in our proposed framework, we conducted an ablation study on both the SLAM and NeRF modules using three underwater scenes. The study focuses on assessing the impact of dynamic object masking, optical flow integration, and specialized loss functions under different configurations. For the SLAM module, we tested three configurations: without dynamic object masking, with YOLO-based masking only, and with the full system integrating both YOLO and optical flow. The results, summarized in Table 7, reveal that dynamic masking significantly improves localization accuracy, as indicated by lower ATE and RPE values across all scenes. For example, in segment 1, the ATE decreased from 2.372 without masking to 2.212 with YOLO masking and further to 1.966 when optical flow was added. Similarly, optical flow integration provided additional robustness by addressing non-classified dynamic interference, such as floating debris and subtle water currents, ensuring more reliable feature matching and localization. The integration of YOLO and optical flow also

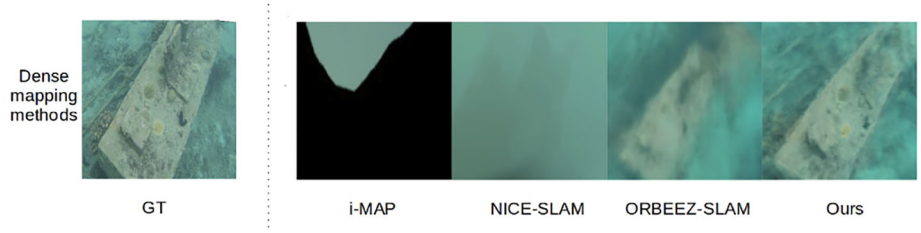


FIGURE 9
Visually comparison of different SLAM systems for underwater environment reconstruction.

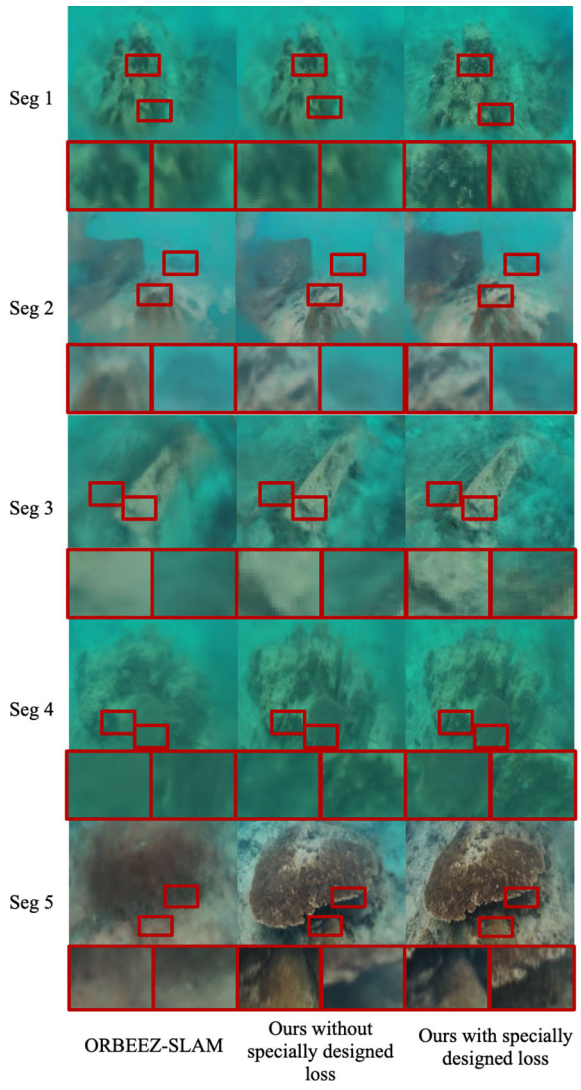


FIGURE 10
Novel view reconstruction comparison across five real underwater scenes.

TABLE 5 Results on our dataset.

Scene	Method	PSNR↑	SSIM↑	LPIPS↓
seg1	Orbeez-SLAM Chung et al. (2023)	22.85	0.73	0.22
	Ours	**24.79**	**0.77**	**0.21**
seg2	Orbeez-SLAM Chung et al. (2023)	20.62	0.71	**0.18**
	Ours	**23.40**	**0.79**	**0.19**
seg3	Orbeez-SLAM Chung et al. (2023)	24.15	0.83	0.20
	Ours	**26.85**	**0.86**	**0.20**
seg4	Orbeez-SLAM Chung et al. (2023)	19.71	0.72	0.19
	Ours	**22.74**	**0.75**	**0.17**
seg5	Orbeez-SLAM Chung et al. (2023)	21.39	0.79	0.17
	Ours	**23.41**	**0.82**	**0.16**

Our model outperforms ORBEEZ-SLAM across all scenes, achieving the best PSNR, SSIM, and LPIPS values. PSNR↑/SSIM↑/LPIPS↓.
The bold text indicates the best performance of the method under the corresponding evaluation index.

reduced initialization time, demonstrating the system's efficiency in handling complex underwater environments.

For the NeRF module, we evaluated the effect of dynamic object masking and specialized loss functions, including light attenuation, smoothing, and regularization losses, which were added on top of the common photometric loss. As shown in Table 8, dynamic masking alone substantially improved reconstruction quality by excluding interference from dynamic objects, leading to higher PSNR and SSIM values. For instance, in segment 2, the PSNR increased from 20.94 in the baseline configuration to 21.80 with masking. Incorporating the specialized loss functions further enhanced reconstruction fidelity, achieving a PSNR of 23.40 and an SSIM of 0.79 in the same segment. These results demonstrate that dynamic masking reduces interference in mapping, while the specialized losses address underwater-specific challenges, such as light attenuation and turbidity, resulting in sharper details and better structural consistency.

Overall, the ablation study highlights the effectiveness of the proposed enhancements in both SLAM and NeRF modules. Dynamic object masking through YOLO significantly improves SLAM performance by reducing ATE and RPE, while optical flow integration ensures robustness in handling non-classified motion. In the NeRF module, masking dynamic regions improves reconstruction quality, and specialized loss functions further refine structural fidelity, allowing the system to adapt to the complexities of underwater environments. Across all tested scenes, the full system configuration consistently outperformed the baselines, setting a new benchmark for underwater SLAM and 3D reconstruction tasks.

4.8 Result analysis and future work

Result Analysis. Our model demonstrates robust performance and high accuracy in dynamic and multi-illumination environments. Compared to traditional ORB-SLAM, our method is more resilient in dynamic scenes, effectively handling noise from moving objects. Compared with Instant-NGP, our model offers superior map quality and consistency, producing high-fidelity 3D reconstructions suited for applications requiring detailed mapping, such as underwater archaeology and ecological monitoring.

The ablation study confirms the effectiveness of the YOLO module and photometric consistency loss. These modules improve the system's adaptability and maintain stable, high-accuracy localization and mapping in complex environments.

TABLE 6 Runtime comparison.

Method	Setting	Segments				
		seg1	seg2	seg3	seg4	seg5
Orbeez-SLAM Chung et al. (2023)	w/o Depth	18.932	17.575	19.746	16.347	22.192
	w/Depth	1.111	1.106	1.113	1.101	1.120
Ours	w/o Depth	17.492	16.274	18.398	16.074	20.375

We compare performance with and without depth generation.
Frame per second [fps] (↑) when running on our dataset.

TABLE 7 Ablation study results for the SLAM module across five scenes.

Scene	No Masking		YOLO Only		YOLO + Optical Flow	
	ATE↓	RPE↓	ATE↓	RPE↓	ATE↓	RPE↓
seg1	2.372	0.041	2.212	0.048	**1.966**	**0.036**
seg2	3.354	0.347	3.213	0.302	**3.228**	**0.232**
seg3	2.732	0.169	2.434	0.141	**2.493**	**0.149**
seg4	2.856	0.949	2.457	0.787	**2.548**	**0.054**
seg5	1.945	0.943	1.763	0.075	**1.269**	**0.083**

Metrics include ATE (m) ↓, RPE (m) ↓, and initialization time (s) ↓.
The bold values represent the best result.

TABLE 8 Ablation study results for the NeRF module across five scenes.

Scene	Baseline with base loss		+ Masking		+ Masking + Light Attenuation Loss		+ Masking + Smoothing Loss		+ Masking + Full Loss	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
seg1	22.15	0.71	23.32	0.76	24.79	0.77	24.42	0.77	**24.79**	**0.77**
seg2	20.94	0.74	21.80	0.77	22.79	0.77	23.06	0.77	**23.40**	**0.79**
seg3	24.48	0.79	25.95	0.81	26.14	0.84	26.55	0.84	**26.85**	**0.86**
seg4	21.86	0.71	21.97	0.71	22.29	0.73	22.44	0.75	**22.71**	**0.75**
seg5	22.64	0.77	23.24	0.79	23.31	0.81	23.39	0.81	**23.41**	**0.82**

Metrics include PSNR ↑ (higher is better) and SSIM ↑ (higher is better).
The bold values represent the best result.

Future Work. While our current method achieves accurate dense reconstruction in real-world underwater scenes, future work will explore the integration of explicit refraction models to further improve geometric consistency under complex optical conditions. This enhancement may extend the applicability of our system to more challenging underwater environments, such as deep-sea scenarios or areas with severe light distortion.

5 Conclusion

This paper proposes a SLAM framework that combines underwater object detection, optical flow analysis, and a NeRF-based approach specifically tailored for the challenges of dynamic interference and low visibility in underwater environments. By employing YOLO to detect moving objects and integrating optical flow to capture unclassified motion, the system precisely excludes all dynamic areas during SLAM computations, focusing feature extraction and pose estimation solely on static backgrounds. Leveraging a customized NeRF for pixel-level 3D reconstruction in underwater scenarios, this framework not only overcomes the limitations of conventional SLAM in turbid waters but also achieves high-precision dense mapping underwater. Experimental results on multiple real-world underwater datasets demonstrate notable advantages in reducing localization drift, enhancing trajectory stability, and recovering underwater scene details, even in

settings with high dynamics or sparse texture. Overall, this research offers a scalable and robust solution for underwater localization and mapping, laying a solid foundation for further advances in underwater robotics and environmental monitoring.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical approval was not required for the study involving animals in accordance with the local legislation and institutional requirements because our research did not involve any direct experimental procedures or interventions with fish or other marine species. We merely gathered underwater imagery in a passive, observational manner, without handling, restraining, or otherwise impacting the animals' natural behaviors. Consequently, no contact-based experimentation was performed, and no fish or invertebrates were harmed or subjected to manipulation during this study. The work focuses on computer vision and SLAM/image processing techniques, rather than any form of experimental handling, and thus does not constitute animal research.

Author contributions

ZW: Data curation, Formal Analysis, Methodology, Writing – original draft, Writing – review & editing. ZY: Conceptualization, Funding acquisition, Supervision, Writing – review & editing. BZ: Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the National Natural Science Foundation of China (Grant No. 62171419), Hainan Province Science and Technology Special Fund of China (Grant No. ZDYF2022SHFZ318), National Key Research and Development Program of China (Grant No. 2022YFD2401304).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Akkaynak, D., and Treibitz, T. (2018). "A revised underwater image formation model," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. Salt Lake City, UT, USA: IEEE 6723–6732. doi: 10.1109/CVPR.2018.00703

Bailey, T., Nieto, J., Guivant, J., Stevens, M., and Nebot, E. (2006). "Consistency of the EKF-SLAM algorithm," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IEEE)*. Beijing, China: IEEE 3562–3568. doi: 10.1109/IROS.2006.281644

Bescos, B., Campos, C., Tardós, J. D., and Neira, J. (2021). DynaSLAM II: Tightly-coupled multi-object tracking and slam. *IEEE robotics automation Lett.* 6, 5191–5198. doi: 10.1109/LRA.2021.3068640

Bescos, B., Fácil, J., Civera, J., and Neira, J. (2018). DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robotics Automation Lett.* 3, 4076–4083. doi: 10.1109/LSP.2016.

Campos, C., Elvira, R., Rodríguez, J. J. G., Montiel, J. M., and Tardós, J. D. (2021). ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM. *IEEE Trans. robotics* 37, 1874–1890. doi: 10.1109/TRO.2021.3075644

Chung, C.-M., Tseng, Y.-C., Hsu, Y.-C., Shi, X.-Q., Hua, Y.-H., Yeh, J.-F., et al. (2023). "Orbeez-SLAM: A real-time monocular visual slam with orb features and nerf-realized mapping," in *2023 IEEE International Conference on Robotics and Automation (ICRA) (IEEE)*. London, United Kingdom: IEEE 9400–9406. doi: 10.1109/ICRA48891.2023.10160950

Engel, J., Schöps, T., and Cremers, D. (2014). "LSD-SLAM: Large-scale direct monocular SLAM," in *European conference on computer vision (Springer)*. Zurich, Switzerland: Springer, Cham 834–849.

Eustice, R. M., Singh, H., Leonard, J. J., and Walter, M. R. (2006). Visually mapping the rms titanic: Conservative covariance estimates for slam information filters. *Int. J. robotics Res.* 25, 1223–1242. doi: 10.1177/0278364906072512

Ferrera, M., Creuze, V., Moras, J., and Trouvé-Peloux, P. (2019). Aqualoc: An underwater dataset for visual–inertial–pressure localization. *Int. J. Robotics Res.* 38, 1549–1559. doi: 10.1177/0278364919883346

Fu, C., Liu, R., Fan, X., Chen, P., Fu, H., Yuan, W., et al. (2023). Rethinking general underwater object detection: Datasets, challenges, and solutions. *Neurocomputing* 517, 243–256. doi: 10.1016/j.neucom.2022.10.039

Garbin, S. J., Kowalski, M., Johnson, M., Shotton, J., and Valentin, J. (2021). "FastNeRF: High-fidelity neural rendering at 200fps," in *Proceedings of the IEEE/CVF international conference on computer vision*. Montreal, QC, Canada: IEEE 14346–14355. doi: 10.1109/ICCV48922.2021.01408

Godard, C., Mac Aodha, O., Firman, M., and Brostow, G. J. (2019). "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*. Seoul, Korea (South): IEEE 3828–3838.

Grisetti, G., Kummerle, R., Stachniss, C., and Burgard, W. (2010). A tutorial on graph-based slam. *IEEE Intelligent Transportation Syst. Magazine* 2, 31–43. doi: 10.1109/MITS.2010.939925

Huang, H., Lin, W.-Y., Liu, S., Zhang, D., and Yeung, S.-K. (2020). "Dual-SLAM: A framework for robust single camera navigation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE)*. Las Vegas, NV, USA: IEEE 4942–4949. doi: 10.1109/IROS45743.2020.9341513

Imambi, S., Prakash, K. B., and Kanagachidambaresan, G. (2021). "Pytorch," in *Programming with TensorFlow: solution for edge computing applications*, Springer, Cham 87–104.

Jordt-Sedlazeck, A., and Koch, R. (2013). "Refractive structure-from-motion on underwater images," in *Proceedings of the IEEE international Conference on Computer Vision*. Sydney, NSW, Australia: IEEE 57–64. doi: 10.1109/ICCV.2013.14

Klein, G., and Murray, D. (2007). "Parallel tracking and mapping for small ar workspaces," in *2007 6th IEEE and ACM international symposium on mixed and augmented reality (IEEE)*. Nara, Japan: IEEE 225–234. doi: 10.1109/ISMAR.2007.4538852

Krogh, A., and Hertz, J. (1991). A simple weight decay can improve generalization. *Adv. Neural Inf. Process. Syst.* 4, 950–957.

Lin, C.-H., Ma, W.-C., Torralba, A., and Lucey, S. (2021). "BaRF: Bundle-adjusting neural radiance fields," in *Proceedings of the IEEE/CVF international conference on computer vision*. Montreal, QC, Canada: IEEE 5741–5751. doi: 10.1109/ICCV48922.2021.00569

Liu, C., Li, H., Wang, S., Zhu, M., Wang, D., Fan, X., et al. (2021). "A dataset and benchmark of underwater object detection for robot picking," in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (IEEE)*. Shenzhen, China: IEEE 1–6. doi: 10.1109/ICMEW53276.2021.9455997

Liu, Y., Wang, Y., Xie, C., Guan, Z., Zhu, J., and Qin, J. (2023). "An improved image enhancement method for underwater robot SLAM," in *2023 IEEE International Conference on Mechatronics and Automation*. Harbin, Heilongjiang, China: IEEE 2366–2371. doi: 10.1109/ICMA57826.2023.10215743

Lucas, B. D., and Kanade, T. (1981). "An iterative image registration technique with an application to stereo vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, Vancouver, BC, Canada: Morgan Kaufmann Publishers Inc. Vol. 2. 674–679.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020). "NeRF: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision*. Glasgow, Scotland, United Kingdom: Springer, Cham 405–421. doi: 10.1007/978-3-030-58452-8_24

Müller, T., Evans, A., Schied, C., and Keller, A. (2022). Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graphics (TOG)* 41, 1–15. doi: 10.1145/3528223.3530127

Mur-Artal, R., Montiel, J., and Tardós, J. (2015). ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans. Robotics* 31, 113–127. doi: 10.1109/TRO.2015.2463671

Mur-Artal, R., and Tardós, J. D. (2017). ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Trans. Robotics* 33, 1255–1262. doi: 10.1109/TRO.2017.2705103

Negahdaripour, S., and Firoozfam, P. (2006). An rov stereovision system for ship-hull inspection. *IEEE J. oceanic Eng.* 31, 551–564. doi: 10.1109/JOE.2005.851391

Pizarro, O., Eustice, R. M., and Singh, H. (2009). Large area 3-d reconstructions from underwater optical surveys. *IEEE J. Oceanic Eng.* 34, 150–169. doi: 10.1109/JOE.2009.2016071

Rahman, S., Quattrini Li, A., and Rekleitis, I. (2022). Svin2: A multi-sensor fusion-based underwater slam system. *Int. J. Robotics Res.* 41, 1022–1042. doi: 10.1177/02783649221110259

Redmon, J. (2016). "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. Las Vegas, NV, USA: IEEE. doi: 10.1109/CVPR.2016.91

Redmon, J., and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Reijgwart, V., Millane, A., Oleynikova, H., Siegwart, R., Cadena, C., and Nieto, J. (2019). Voxgraph: Globally consistent, volumetric mapping using signed distance function submaps. *IEEE Robotics Automation Lett.* 5, 227–234. doi: 10.1109/LSP.2016.

Rosinol, A., Leonard, J. J., and Carlone, L. (2023). "NeRF-SLAM: Real-time dense monocular SLAM with neural radiance fields," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE)*. Detroit, MI, USA: IEEE 3437–3444. 10.1109/IROS55552.2023.10341922

Schönberger, J. L., and Frahm, J.-M. (2016). "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE. doi: 10.1109/CVPR.2016.445

Schönberger, J. L., Zheng, E., Pollefeys, M., and Frahm, J.-M. (2016). "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*. Amsterdam, The Netherlands: Springer, Cham. doi: 10.1007/978-3-319-46487-9_31

Sucar, E., Liu, S., Ortiz, J., and Davison, A. J. (2021). "iMAP: Implicit mapping and positioning in real-time," in *IEEE International Conference on Computer Vision*. Montreal, QC, Canada: IEEE 6229–6238. doi: 10.1109/ICCV48922.2021.00617

Tateno, K., Tombari, F., Laina, I., and Navab, N. (2017). "CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. Honolulu, HI, USA: IEEE 6243–6252. doi: 10.1109/CVPR.2017.695

Telem, G., and Filin, S. (2010). Photogrammetric modeling of underwater environments. *ISPRS J. photogrammetry Remote Sens.* 65, 433–444. doi: 10.1016/j.isprsjprs.2010.05.004

Thrun, S. (2002). Particle filters in robotics. *UAI (Citeseer)* 2, 511–518.

Xin, Z., Wang, Z., Yu, Z., and Zheng, B. (2023). ULL-SLAM: underwater low-light enhancement for the front-end of visual SLAM. *Front. Mar. Sci.* 10, 1133881. doi: 10.3389/fmars.2023.1133881

Yang, J., Gong, M., Nair, G., Lee, J. H., Monty, J., and Pu, Y. (2023). "Knowledge distillation for feature extraction in underwater VSLAM," in *2023 IEEE International Conference on Robotics and Automation (ICRA) (IEEE)*. London, United Kingdom: IEEE 5163–5169. doi: 10.1109/ICRA48891.2023.10161047

Yu, A., Li, R., Tancik, M., Li, H., Ng, R., and Kanazawa, A. (2021). "Plenoctrees for real-time rendering of neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal, QC, Canada: IEEE 5752–5761. doi: 10.1109/ICCV48922.2021.00570

Yu, B., Wu, J., and Islam, M. J. (2023). "UDepth: Fast monocular depth estimation for visually-guided underwater robots," in *2023 IEEE International Conference on Robotics and Automation (ICRA) (IEEE)*. London, United Kingdom: IEEE 3116–3123. doi: 10.1109/ICRA48891.2023.10161471

Zheng, Z., Xin, Z., Yu, Z., and Yeung, S.-K. (2023). Real-time GAN-based image enhancement for robust underwater monocular slam. *Front. Mar. Sci.* 10, 1161399. doi: 10.3389/fmars.2023.1161399

Zhu, Z., Peng, S., Larsson, V., Xu, W., Bao, H., Cui, Z., et al. (2022). "Nice-SLAM: Neural implicit scalable encoding for SLAM," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. New Orleans, LA, USA: IEEE 12786–12796. doi: 10.1109/CVPR52688.2022.01245