MARVO: Marine-Adaptive Radiance-aware Visual Odometry

Anonymous CVPR submission

Paper ID 20972

Abstract

Underwater visual localization remains challenging due to wavelength-dependent attenuation, poor texture, and non-Gaussian sensor noise. We introduce MARVO, a physics-aware, learning-integrated Simultaneous Localization And Mapping framework that fuses underwater image formation modeling, differentiable matching, and reinforcement-learning optimization. At the front-end, we extend transformer-based feature matcher with a Physics-Aware Radiance Adapter that compensates for color-channel attenuation and contrast loss, yielding geometrically consistent feature correspondences under turbidity. These semi-dense matches are combined with inertial and pressure measurements inside a factor-graph localization backend, where we formulate a keyframe-based visual-inertial-barometric estimator using GTSAM library. Each keyframe introduces (i) Pre-integrated IMU motion factors, (ii) MARVO-derived visual pose factors, and (iii) barometric depth priors, giving a full-state MAP estimate in real time. Lastly, we introduce a Reinforcement-Learning-based Pose-Graph Optimizer that refines global trajectories beyond local minima of classical least-squares solvers by learning optimal retraction actions on SE(2). This work highlights the synergy between physics-guided feature adaptation, probabilistic multi-sensor fusion, and learning-based global optimization for robust underwater visual-inertial SLAM.

1. Introduction

Advances in feature matching [30], multi-sensor fusion [11] and factor-graph optimization [8] have empowered Visual Odometry (VO) and Simultaneous Localization and Mapping (SLAM) to achieve strong performance in terrestrial settings. Underwater environments, however, remain uniquely challenging. Light scattering, wavelength-dependent attenuation and strong non-Gaussian noise produce severe contrast loss, unstable features, and inconsistent long-horizon pose estimates. Classical VO and visual-inertial pipelines fail in such conditions because they neither correct the underlying physics of underwater image formation nor effectively fuse uncertain auxiliary sensors such as pressure and inertial measurements.

We propose the **MARVO**, a *Marine-Adaptive Radiance-Aware Visual Odometry* framework that couples physics-guided front-end perception with probabilistic multi-sensor fusion and offline learning-driven pose-graph refinement. The basic idea of MARVO is that robust underwater VO calls for both (i) perception modules that explicitly compensate for radiometric distortions, and (ii) back-end optimization that can escape the local minima typical of noisy, visually degraded trajectories.

At the front end, MARVO extends LoFTR [30] with a lightweight *Physics-Aware Radiance Adapter (PARA)*. PARA modulates intermediate descriptors using learned attenuation and visibility estimates, counteracting color-channel imbalance and restoring feature matchability before transformer attention. This radiance-aware formulation enables stable semi-dense correspondences in regions where standard LoFTR degrades.

These corrected visual factors are fused together with IMU preintegration [11] and barometric depth measures within a keyframe-based factor graph implemented in GTSAM [8]. Each keyframe contributes with PARA-enhanced visual pose constraints, motion factors and unary depth priors to create a consistent maximum-a-posteriori estimate. The system acts like a fixed-lag smoother, preserving real-time performance while maintaining global consistency across heterogeneous sensing modalities. To further improve long-term stability, MARVO applies a reinforcement-learning (RL) policy in order to refine the SE(2) pose graph; this is inspired by recent work in efficient pose-graph optimization [25]. Instead of solely relying on non-convex least-squares solvers which often get stuck in suboptimal minima, the learned policy proposes retraction actions to steer the trajectory towards globally consistent solutions. This learned refinement complements classical optimization and is particularly effective for turbid or visually sparse scenes.

MARVO integrates physics-informed perception, probabilistic fusion, and learning-based optimization into a reli-

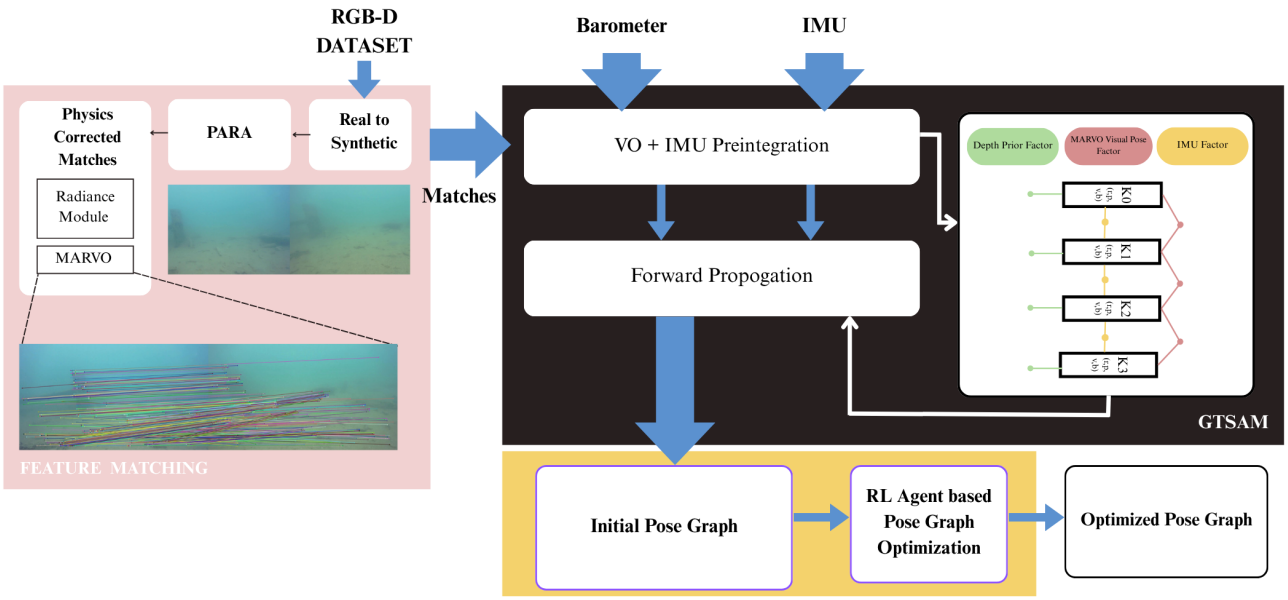

Figure 1. Overview of MARVO. PARA enhances LoFTR features using physically-informed radiance correction. Corrected visual factors are fused with IMU and barometric depth in a GTSAM factor graph to produce real-time VO. An offline reinforcement-learning agent performs pose-graph refinement to obtain globally consistent trajectories.

able underwater localization system. Its contributions are:

- **Physics-aware front end:** a differentiable radiance adapter that compensates wavelength-dependent attenuation within the LoFTR transformer pipeline.
- **Probabilistic multi-sensor fusion:** a visual–inertial–barometric estimator based on a GTSAM fixed-lag smoother with PARA-enhanced constraints.
- **Learning-based global optimization:** an RL-driven pose-graph optimizer that refines SE(2) trajectories beyond the limits of standard least-squares solvers.

Together, these elements provide a single powerful underwater VO system that is able to estimate geometrically consistent trajectories where traditional pipelines fail.

2. Related Works

MARVO leverages three fundamental research directions: feature matching, factor-graph visual-inertial odometry, and reinforcement-learning-based pose-graph optimization.

2.1. Feature Matching

Detector-based methods. Classical local features follow a paradigm of detect-describe-match. While handcrafted techniques such as SIFT [22], SURF [3], and ORB [29] enjoy robustness to viewpoint and illumination changes, they fail in texturepoor or turbidity degraded underwater imagery. Similarly, learned techniques that improve repeatability via data-driven keypoint detection, such as LIFT [34] and SuperPoint [9], are still limited by the fundamental sparsity of stable interest points in underwater scenes.

Detector-free methods. Dense matching methods avoid explicit keypoint detection and compute cost volumes or dense correlation fields directly. NCNet [28] and DRC-Net [20] impose neighborhood consensus on dense descriptors, although their convolutional receptive fields limit global reasoning. LoFTR [30] introduced an alternating mechanism of self- and cross-attention in establishing globally consistent semi-dense matches; since then, LoFTR has been widely adopted as a front-end for geometric vision tasks. However, descriptor quality degrades due to contrast loss, backscatter, and wavelength-dependent attenuation underwater, which motivated the physics-aware radiance adaptation in MARVO.

Transformers in geometric vision. Transformers have become foundational in correspondence and motion estimation due to their ability to model long-range relationships. Following ViT [10], attention-based architectures have been deployed in optical flow [31], correspondence estimation [30], and geometric reasoning. MARVO extends this paradigm with a Physics-Aware Radiance Adapter (PARA) that modulates LoFTR's intermediate features according to learned attenuation cues, restoring discriminability in underwater conditions while retaining global attention benefits.

CVPR #20972

2.2. Factor Graph-Based Visual{Inertial Odometry

Pose-graph and factor-graph estimation. Robotic state estimation is often framed as a sparse nonlinear least-squares problem over poses and landmarks, modeled probabilistically with factor graphs [8]. Solvers such as GTSAM [8], g2o [18], and iSAM2 [15] leverage sparsity to enable real-time inference and form the basis of many state-of-the-art VO and SLAM methods. Visual–inertial odometry and preintegration. Combining cameras with IMUs leads to drift-resistant trajectories, assuming that the system properly models high-rate inertial signals. IMU preintegration [11] allows for the efficient incorporation of continuous IMU measurements in fixed-lag smoothing frameworks by avoiding full re-integration during optimization. This has given rise to both highly accurate and computationally efficient VIO systems [19] suitable for real-time deployment. MARVO adheres to this paradigm, fusing PARA-enhanced visual constraints with IMU and barometric depth in a fixed-lag factor graph. In practical robotic systems, sensors run at different rates. Continuous-time fusion and multi-threaded architectures handle asynchronous data efficiently by interpolating trajectory states and marginalizing older poses [13, 23]. MARVO uses a similar strategy to enable real-time optimization with heterogeneous visual–inertial–pressure data.

2.3. Reinforcement-Learning-Based Pose-Graph Optimization

Classical PGO. Pose-graph optimization (PGO) solves for globally consistent trajectories by minimizing rotational and translational constraints over or, typically via Gauss-Newton or Levenberg-Marquardt with sparse factorizations. While effective, these solvers remain sensitive to poor initializations, unreliable loop closures, and strong noise-all common challenges in underwater environments. RL-based PGO. Recent work proposes reinforcement-learning-based optimizers that extend classical solvers by exploring the pose manifold beyond local gradients. RL-PGO [16] formulates planar PGO as a POMDP and learns retraction actions on $\mathrm{SE}(2)$ that escape suboptimal minima. Distributed extensions further leverage graph neural networks for multi-robot settings [17]. MARVO follows a similar philosophy, using a learned policy to refine pose-graph estimates and enhance global consistency in the presence of underwater visual degradation.

3. Dataset Augmentation

To generate realistic underwater-appearance training data, we utilise the *SyreaNet* synthesis module [33], which applies a physics-based underwater image formation model to in-air RGB-D datasets. We process RGB–depth pairs from ScanNet [6], TartanAir [32], and Hypersim [27], converting each into a synthetic underwater counterpart that simulates wavelength-dependent attenuation and depth-dependent scattering.

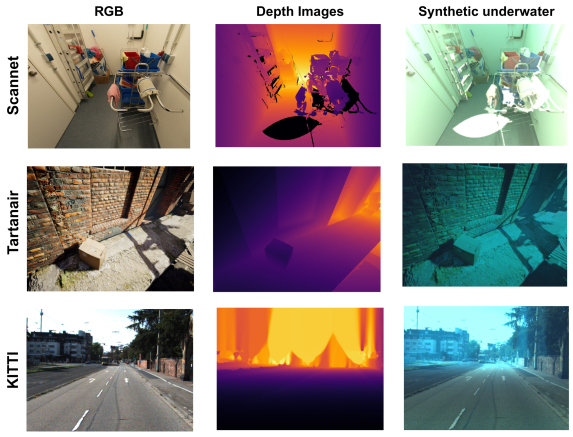

Figure 2. Examples of synthetic data used in MARVO. Each row shows an example from a different dataset (ScanNet, TartanAir, Hypersim): original RGB image, corresponding depth map, and the generated synthetic underwater image with simulated attenuation and scattering.

3.1. Physical Model

Following the revised underwater image formation model [1], the observed intensity for color channel $c \in \{R, G, B\}$ is

$$I_c(x) = J_s(x)\, W_c\, e^{-\beta_c^D z(x)} + B_c(x), \qquad (1)$$

where $J_s(x)$ is the in-air scene radiance, W_c is the diffuse downwelling term, β_c^D is the attenuation coefficient, $B_c(x)$ is the backscatter component, and $z(x)$ is the per-pixel range. This formulation decouples attenuation and backscatter, enabling physically faithful simulation of haze, color loss, and contrast degradation under varying water types.

3.2. Synthetic Generation

For every RGB–depth pair, SyreaNet samples β_c^D and $B_c(x)$ from empirical distributions calibrated on real underwater imagery, emulating different levels of turbidity and spectral absorption. Illumination variability is emulated by stochastically perturbing W_c. In total, we synthesize approximately 120k RGB-D pairs, around 40k per dataset. We resize all images to 640×480 and normalize them to zero-mean, unit variance before training. The resulting renderings preserve the geometric structure of the original datasets but embed realistic radiometric degradation.

3.3. Evaluation and Fine-Tuning

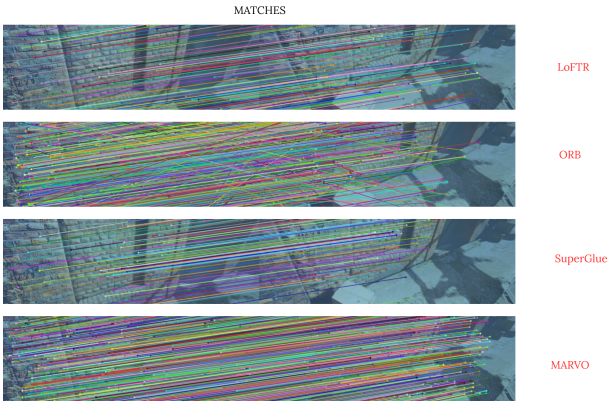

Figure 3. Qualitative feature matching comparison: MARVO produces denser and more geometrically stable correspondences than SuperGlue, ORB, and LoFTR under underwater conditions characterized by turbidity, color attenuation, and low texture. Conventional matchers degrade noticeably, while MARVO maintains semi-dense and spatially coherent matches through physics-aware radiance modulation.

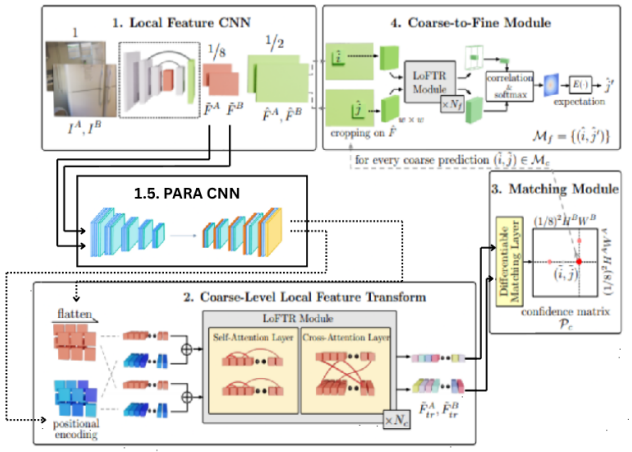

Figure 4. PARA architecture: The Physics-Aware Radiance Adapter takes coarse CNN features and predicts per-pixel attenuation and backscatter fields. These are used to generate a radiance correction mask $\Gamma(x)$, which normalizes intermediate descriptors before transformer matching. PARA compensates for wavelength-dependent attenuation, color imbalance, and contrast degradation common in underwater environments.

We asses the generalization of the front-end on KITTI [12] in a two-stage protocol: first, perform large-scale pretraining on synthetic underwater data; second, fine-tune on a small subset (10%) of real data. This helps in bridging the synthetic-to-real domain gap while preserving the physical diversity of the synthetic corpus.

We further benchmark the feature-matching pipeline on MegaDepth [21] using the standard image-pair splits and evaluation protocol. MegaDepth provides diverse Internet photo collections with SfM/MVS-derived geometry, allowing us to quantify improvements in match recall and pose accuracy under wide-baseline viewpoint and illumination changes.

4. Feature Matching

At the heart of visual odometry lies feature correspondence. MARVO extends the detector-free matcher LoFTR [30] by adding a physics-guided radiance module inspired from SyreaNet [33] for capturing the complex effects of attenuation, scattering, and color imbalance in underwater media. Combining them ensures geometric and photometric consistency even under wavelength-dependent distortions.

4.1. Detector-Free Transformer Backbone

MARVO follows LoFTR's detector-free paradigm and directly predicts semi-dense correspondences from input image pairs with no explicit keypoint detection. Given rectified images (I_A, I_B), convolutional encoders generate feature maps at $1/8$ resolution. The features are enriched with global context across both images by stacked self- and cross-attention layers. High-confidence correspondences are obtained from coarse correlation maps $\mathbf{C}_{\text{coarse}}$ using dual-softmax and mutual nearest-neighbor filtering and then refined to sub-pixel matches $\mathbf{C}_{\text{fine}}$ via differentiable correlation. This allows the detector-free architecture to perform robust matching in texture-poor underwater scenarios.

4.2. Physics-Aware Radiance Adapter (PARA)

To adapt transformer matching to underwater imaging, MARVO introduces a *Physics-Aware Radiance Adapter* (PARA) between the CNN encoder and transformer layers.

Let $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$ denote the encoder feature map at $1/8$ resolution and let $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ be the corresponding RGB image, both bilinearly downsampled to the same spatial size. PARA is implemented as a lightweight three-branch prediction head:

$$\mathbf{F}_{\text{shared}} = \phi_{\text{sh}}([\mathbf{F}, \text{BN}(\mathbf{I})]), \tag{2}$$

$$\hat{\boldsymbol{\beta}} \in \mathbb{R}^{H\times W\times 3} = \phi_{\beta}(\mathbf{F}_{\text{shared}}), \tag{3}$$

$$\hat{\mathbf{B}}_{\infty} \in \mathbb{R}^{H\times W\times 3} = \phi_{B}(\mathbf{F}_{\text{shared}}), \tag{4}$$

$$\hat{\mathbf{z}} \in \mathbb{R}^{H\times W\times 1} = \phi_{z}(\mathbf{F}_{\text{shared}}), \tag{5}$$

where ϕ_{sh} denotes two 3×3 convolutional layers with ReLU and batch normalization, and $\phi_{\beta}, \phi_{B}, \phi_{z}$ are 1×1 convolutional heads that produce per-pixel attenuation, backscatter, and a depth proxy respectively. All predictions are made at the feature resolution, so no extra decoder is required.

We model the underwater image formation for each color channel $c \in \{R, G, B\}$ as

$$I_c(x) = J_c(x)\, e^{-\beta_c(x) z(x)} + B^c_{\infty}(x)\big(1 - e^{-\beta_c(x) z(x)}\big), \tag{6}$$

where $J_c(x)$ is the in-air radiance, $\beta_c(x)$ is the attenuation coefficient, $B^c_\infty(x)$ is the asymptotic backscatter, and $z(x)$ is the range. During training on synthetic RGB–depth data, we supervise $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{B}}_\infty$ using the corresponding SyreaNet parameters and provide the ground-truth depth $z(x)$ to PARA. At test time, PARA relies only on the predicted proxy $\hat{\mathbf{z}}$.

To obtain a radiance-corrected estimate of $J_c(x)$ we invert Eq. (6) using the predicted fields:

$$\hat{J}_c(x) = \big(I_c(x) - \hat{B}^c_\infty(x)(1 - e^{-\hat{\beta}_c(x)\hat{z}(x)})\big) \cdot e^{\hat{\beta}_c(x)\hat{z}(x)}. \tag{7}$$

From $\hat{J}_c(x)$ we derive a scalar, channel-aggregated correction mask

$$\Gamma(x) = \frac{1}{3} \sum_{c\in\{R,G,B\}} \frac{\hat{J}_c(x)}{I_c(x) + \epsilon}, \tag{8}$$

with a small ϵ to avoid division by zero. The mask $\Gamma(x) \in \mathbb{R}^{H\times W\times 1}$ is then broadcast across channels and applied to the encoder features as

$$\tilde{\mathbf{F}}(x) = \mathrm{LN}\big(\Gamma(x) \odot \mathbf{F}(x)\big), \tag{9}$$

where $\odot$ denotes element-wise multiplication and $\mathrm{LN}(\cdot)$ is layer normalization. The transformer matcher in MARVO operates on $\tilde{\mathbf{F}}$ instead of $\mathbf{F}$, so all subsequent attention layers see features that have been explicitly corrected for wavelength-dependent attenuation and backscatter.

In practice, PARA adds fewer than 5% additional parameters relative to the LoFTR backbone and keeps the feature resolution unchanged, which preserves LoFTR's computational profile while significantly improving descriptor consistency in spectrally distorted regions.

4.3. Training Objectives

The front-end is trained using a combined geometric, photometric, and physics-based loss:

$$\mathcal{L} = \lambda_{\text{match}}\mathcal{L}_{\text{match}} + \lambda_{\text{photo}}\mathcal{L}_{\text{photo}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}}, \tag{10}$$

$$\mathcal{L}_{\text{match}} = \|\hat{\mathbf{P}} - \mathbf{P}^*\|_1, \tag{11}$$

$$\mathcal{L}_{\text{photo}} = 1 - \mathrm{SSIM}(I'_A, I'_B), \tag{12}$$

$$\mathcal{L}_{\text{phys}} = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\text{gt}}\|_1 + \|\hat{\mathbf{B}}_\infty - \mathbf{B}_{\infty,\text{gt}}\|_1. \tag{13}$$

Here $\boldsymbol{\beta}_{\text{gt}}$ and $\mathbf{B}_{\infty,\text{gt}}$ are the SyreaNet-derived supervision fields used in Eq. (6). $\mathcal{L}_{\text{match}}$ enforces geometric consistency of correspondences, $\mathcal{L}_{\text{photo}}$ encourages radiance-corrected view agreement through the PARA-normalized images I'_A, I'_B, and $\mathcal{L}_{\text{phys}}$ explicitly ties the predicted physical fields to the underlying underwater image formation model.

4.4. Synthetic-to-Real Adaptation

MARVO is trained on a combination of synthetic underwater data and real underwater images. A physics-based synthesis pipeline generates pairs $(I_{\text{air}}, I_{\text{uw}})$ across multiple turbidity and illumination conditions. Domain adaptation and consistency regularization encourage feature invariance across synthetic-real shifts, thus allowing for robust matching without environment-specific calibration.

4.5. Training Procedure

Our model is trained in two stages: (1) pre-training with 1.2×10^5 synthetic underwater image pairs from ScanNet [6], TartanAir [32] and Hypersim [27]; and (2) fine-tuning with $\sim$12k real underwater frames including 10% KITTI [12] and internal field data. The convolutional layers are partially frozen while fine-tuning transformer and PARA modules using real-world attenuation statistics. Training utilizes mixed-precision and multi-GPU parallelization on 4$\times$ NVIDIA A100 GPUs.

The developed physics-aware transformer recovers stable semi-dense matches even in spectrally distorted and feature-poor underwater regions, forming the primary constraints for downstream GTSAM-based state estimation and RL-PGO global optimization.

5. Pose Graph Localization

State estimation in MARVO is done by a lightweight factor graph that fuses: i) physics-aware visual constraints from PARA–LoFTR, ii) IMU preintegration, and iii) a barometric depth prior. A fixed-lag smoother is implemented for the back-end in GTSAM. Unlike typical VIO pipelines, MARVO involves two new components specific to underwater applications-a semi-dense visual factor derived from physics-corrected matches, and a unary depth factor that suppresses vertical drift.

5.1. State Representation

Every keyframe state $\mathbf{x}_i$ contains orientation, position, velocity, and IMU biases. Only keyframes remain in the optimization window for real-time operation.

5.2. Graph Factors

IMU Preintegration. We apply standard GTSAM preintegration to provide metric scale and short-term motion constraints.

Depth Prior (Underwater). A barometric pressure reading provides a explicit estimate of depth. We add a simple unary factor on $\mathbf{p}_i$ an inexpensive but highly effective constraint that eliminates the vertical drift common in monocular and low-visibility underwater VO.

MARVO Visual Factor. For every keyframe pair (i, j),

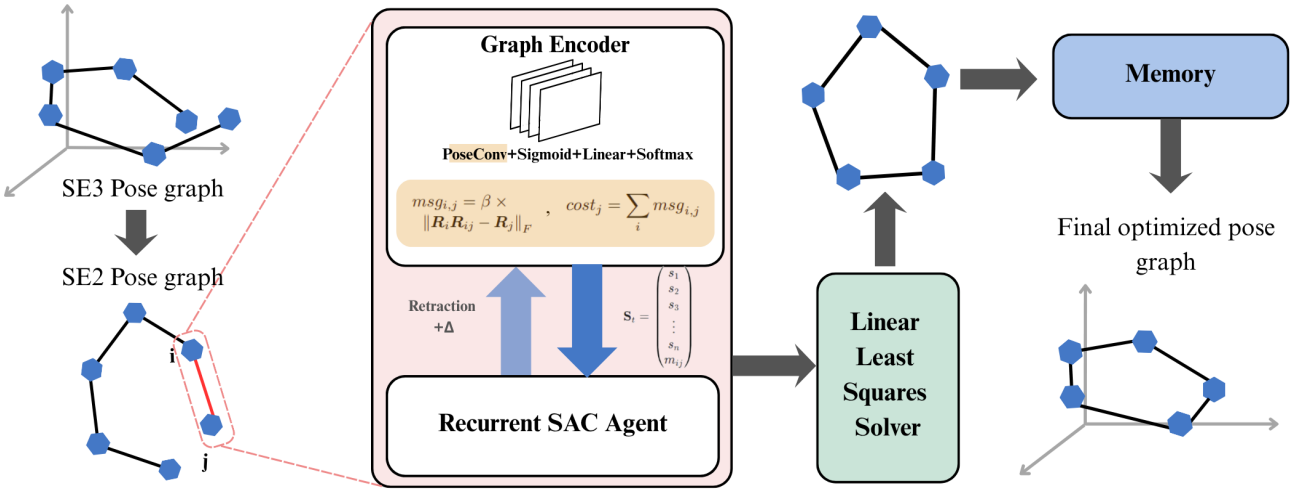

Figure 5. Reinforcement learning-based pose-graph optimization. A GNN encoder maps an initial pose graph to latent edge features, which condition a recurrent SAC agent. The agent iteratively applies retraction actions; a final linear least-squares step produces the optimized pose graph.

PARA–LoFTR provides semi-dense matches

$$\mathcal{M}_{ij} = \{(\mathbf{u}_k^i, \mathbf{u}_k^j)\}_{k=1}^N,$$

from which we estimate a relative pose via an essential matrix. A scale variable s_{ij} is co-optimized with the poses, improving robustness under partial stereo loss or poor geometric parallax.

Adaptive Covariance. MARVO weights each visual factor using physics-aware match confidences. The covariance scales inversely with inlier count and spatial coverage, allowing high-visibility frames to dominate while down-weighting degraded or spectrally imbalanced images automatically .

MARVO's localization back-end differs from conventional VIO in two aspects: (1) PARA–LoFTR supplies stable, semi-dense constraints even under turbidity and color attenuation, and (2) an underwater-specific depth factor provides drift-free vertical motion. These components produce a reliable initialization for the RL-based global refinement stage that follows.

6. Offline Agent-Initiated Pose Graph Optimisation

6.1. Foundations

Classical iterative solvers for pose-graph optimization, such as Gauss–Newton and Levenberg–Marquardt [8, 18], are efficient but prone to local minima under poor initialization or heavy noise. RL-PGO [16] addresses this by casting pose-graph optimization as a partially observable Markov decision process on $\mathrm{SE}(2)$ and learning a policy that applies continuous retraction actions to pose pairs. A message-passing GNN encodes edge residuals [2], enabling the agent to explore beyond local gradients and systematically bootstrap classical solvers on graphs of varying size.

6.2. Domain Adaptation for Autonomous Underwater Craft

AUVs and ROVs are often roll/pitch stabilized via ballast and thrusters, leaving yaw as the main rotational degree of freedom, while depth is accurately measured by a pressure sensor [11]. This motivates a dimensionality-reduced formulation: from the GTSAM-based visual–inertial–barometric frontend [7, 15] we extract the planar $\mathrm{SE}(2)$ states (x, y, θ), build a 2D pose graph on the horizontal plane, and keep roll, pitch, and depth fixed via barometric priors.

6.3. Architecture and Integration

Our optimizer follows a two-stage pipeline. First, the full $\mathrm{SE}(3)$ factor graph is optimized with a standard solver (iSAM2 or Levenberg–Marquardt) [8, 15] to obtain a consistent initial trajectory. We then project this trajectory to $\mathrm{SE}(2)$ and feed the planar graph to the RL agent for refinement. After convergence, the corrected (x, y, θ) poses are lifted back to $\mathrm{SE}(3)$ by reattaching the original roll, pitch, and depth estimates, yielding a globally consistent 3D trajectory.

Graph Encoding and Message Passing: We use a GNN encoder that aggregates orientation and translation residuals on each edge $(i, j) \in E$ [2, 16]. Message passing yields per-edge embeddings capturing patterns such as odometric chains and loop closures; these embeddings are pooled into the state fed to the policy.

Policy and Sequential Refinement: A recurrent Soft Actor–Critic (SAC) agent with LSTM history [14] selects an edge and outputs a retraction action in $\mathrm{SE}(2)$ at each step. Retractions are applied via the exponential map [4], enabling smooth, continuous pose updates while preserving manifold structure.

6.4. Key Innovation: Log-Weighted Orientation Cost

RL-PGO averages orientation errors uniformly over all edges [16], implicitly assuming equal importance. We instead weight each edge's rotational error by the magnitude of its associated translation, yielding the log-weighted orientation cost

$$OC_{\log} = \sqrt{\sum_{(i,j)\in E} w_{ij} \left\| R_i R_{ij} - R_j \right\|_F^2}, \tag{14}$$

with

$$w_{ij} = 1 + \beta \log\left(\frac{\|\mathbf{t}_{ij}\|}{\bar{t}} + \epsilon \right), \tag{15}$$

where $R_i, R_j, R_{ij} \in \mathrm{SO}(2)$ are absolute and relative rotations, $\mathbf{t}_{ij} \in \mathbb{R}^2$ is the measured translation, $\bar{t}$ is the mean translation magnitude over all edges, β controls the strength of the weighting, and ϵ is a small constant for numerical stability.

The logarithmic form remains monotonically increasing but sublinear in $\|\mathbf{t}_{ij}\|$: long-range constraints are emphasized without allowing a few very long, noisy edges to dominate the reward. Setting $\beta = 0$ recovers the uniform weighting of [16], while moderate β values better reflect the heterogeneous uncertainty patterns typical in underwater SLAM.

6.5. Post-Optimization Expansion and Final Refinement

After RL-based $\mathrm{SE}(2)$ refinement, we reattach roll, pitch, and depth from the initial GTSAM estimate (or barometric priors) to obtain a full $\mathrm{SE}(3)$ trajectory. A final short Levenberg–Marquardt refinement [18, 25] on the reconstructed 3D pose graph exploits the improved initialization to quickly converge to a high-quality local minimum, even in regimes where purely classical methods tend to stall [5].

7. Experiments

We evaluate MARVO across synthetic underwater benchmarks and real coastal field deployments, comparing against both classical and learning-based VO systems. Our experiments measure (1) correspondence quality under wavelength-dependent attenuation and (2) end-to-end odometry accuracy. Due to the absence of standardized underwater VO datasets with ground-truth poses and depth maps, our evaluation incorporates both physically rendered datasets and real sequences aligned to SfM-based ground truth. This limitation is inherent to underwater benchmarking: accurate supervision requires synchronized RGB, depth, and high-quality pose annotations, which are rarely available together in public datasets.

7.1. Evaluations

Feature Matching Accuracy. We first benchmark MARVO's radiance-aware correspondence module against SuperGlue and LoFTR [30]. Using RGB-D datasets rendered through a physically based underwater model (per-channel attenuation and scattering), we compute the pose estimation AUC at 5°, 10°, and 20°. MARVO achieves the highest accuracy across all thresholds, showing improved match stability under spectral degradation.

Table 1. Pose estimation AUC on synthetic underwater sequences.

Method	5°	10°	20°
SP + SuperGlue	25.4	42.2	59.7
LoFTR	42.9	59.5	68.2
MARVO (Ours)	**49.7**	**62.9**	**71.3**

Visual Odometry on Synthetic Underwater Dataset. We next evaluate end-to-end VO performance. All trajectories are scale-aligned using a similarity transform. MARVO substantially reduces ATE, angular drift, and relative pose error compared to classical baselines and modern VO pipelines.

Table 2. Synthetic underwater VO performance

Method	ATE (m) ↓	RPE (deg/m) ↓	Drift (%) ↓
ORB-SLAM3	6.45	1.38	5.9
LIBVISO2	5.12	1.14	5.1
MARVO (ours)	**2.47**	**0.61**	**2.2**

Note: Larger-scale multi-sequence evaluations are limited by the scarcity of datasets that provide jointly calibrated ground truth depth, images, and poses, all of which are required to generate physically realistic underwater renderings.

7.2. Real-World Field Data

We evaluate MARVO on real underwater imagery collected using a monocular camera, IMU, and depth-pressure sensors. Ground truth camera poses are obtained via COLMAP SfM and aligned using a 7-DoF similarity transform. These sequences contain severe turbidity, forward-scattering, and intermittent visibility. MARVO maintains significantly lower drift and angular error compared to classical VO

CVPR #20972

methods, validating its robustness in challenging real-world conditions.

Table 3. VO performance on real underwater field deployments (Scale Alligned)

Method	ATE (m) ↓	RPE (deg/m) ↓	Drift (%) ↓
ORB-SLAM3	4.12	0.92	3.8
LIBVISO2	3.47	0.85	3.1
MAST3R-SLAM [26]	2.52	0.58	2.2
VGGT-SLAM [24]	2.41	0.56	2.1
MARVO (ours)	**1.73**	**0.34**	**1.2**

7.3. Ablation Studies

We ablate MARVO to quantify the contribution of each module:

- **No PARA Module:** Removing physics-guided radiance correction degrades matchability and increases drift, confirming its necessity under spectral attenuation.
- **Replace Matcher with LoFTR:** Using vanilla LoFTR instead of PARA-enhanced descriptors produces significantly worse correspondence reliability.
- **Replace RL-PGO with Classical PGO:** Gauss–Newton optimization leads to higher drift, especially when loop closures are sparse.
- **No Physics-Based Radiance Norm:** Using PARA without physics-based normalization yields the largest AUC drop, indicating that physics supervision—not merely CNN modulation—is responsible for robustness.

Table 4. Ablation analysis of MARVO components.

Configuration	AUC @10° ↑	ATE (m) ↓	Drift (%) ↓
Full MARVO (ours)	**0.92**	**1.73**	**1.2**
No PARA Module	0.81	2.24	1.9
Replace Feature Matcher (LoFTR)	0.76	2.47	2.3
Replace RL-PGO w/ Classical PGO	0.84	2.05	1.7
No Physics-Based Radiance Norm	0.73	2.68	2.6

8. Conclusion

We introduced MARVO, a marine-adaptive visual odometry framework that integrates physics-guided feature correction, multi-sensor factor-graph estimation, and reinforcement-learned pose-graph optimization. By embedding a wavelength-dependent attenuation and backscatter model directly into the transformer correspondence pipeline, MARVO restores descriptor discriminability in visually degraded underwater scenes. Combined with an RL-enhanced global optimizer and visual–inertial–barometric frontend, MARVO produces stable, geometrically consistent trajectories across a wide spectrum of conditions.

Our evaluations demonstrate consistent improvements over SuperGlue, LoFTR [30], ORB-SLAM3, and LIBVISO2 across AUC, ATE, RPE, and drift metrics, both on synthetic underwater renderings and real-world field deployments. MARVO maintains tracking where classical detectors fail and significantly reduces global trajectory drift through RL-based refinement.

Limitations and Future Work. A major limitation is the lack of large-scale underwater VO datasets with ground-truth trajectories and depth maps. Physics-based rendering requires metric depth and camera calibration, and accurate odometry evaluation requires high-quality ground truth—two properties rarely available simultaneously. This limits the scope of quantitative benchmarks and necessitates hybrid evaluation using synthetic renderings and COLMAP-aligned real data.

Future extensions include joint 3D mapping (e.g., TSDF fusion or underwater-adapted MVS), learning full $\mathrm{SE}(3)$ global optimization with roll/pitch coupling, and integrating acoustic depth priors to handle extreme turbidity or complete visual dropout.

References

[1] Derya Akkaynak and Tali Treibitz. A revised underwater image formation model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6723–6732, 2018. 3

[2] Rana Azzam, Felix H Kong, Tarek Taha, and Yahya Zweiri. Pose-graph neural network classifier for global optimality prediction in 2d slam. *IEEE Access*, 9:80466–80477, 2021. 6

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 2

[4] José Luis Blanco-Claraco. A tutorial on $\mathbf{SE}(3)$ transformation parameterizations and on-manifold optimization. *arXiv preprint arXiv:2103.15980*, 2021. 7

[5] Giuseppe Calafiore, Luca Carlone, and Frank Dellaert. Pose graph optimization in the complex domain: Lagrangian duality, conditions for zero duality gap, and optimal solutions. *arXiv preprint arXiv:1505.03437*, 2015. 7

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 3, 5

[7] Frank Dellaert. Factor graphs and gtsam: A hands-on introduction. *Georgia Institute of Technology, Tech. Rep*, 2(4), 2012. 6

[8] Frank Dellaert, Michael Kaess, et al. Factor graphs for robot perception. *Foundations and Trends® in Robotics*, 6(1-2): 1–139, 2017. 1, 3, 6

[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[11] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. On-manifold preintegration for real-time visual–inertial odometry. *IEEE Transactions on Robotics*, 33 (1):1–21, 2016. 1, 3, 6

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013. 4, 5

[13] Patrick Geneva, Kevin Eckenhoff, and Guoquan Huang. Asynchronous multi-sensor fusion for 3d mapping and localization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5994–5999. IEEE, 2018. 3

[14] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018. 7

[15] Michael Kaess, Hordur Johannsson, Richard Roberts, Viorela Ila, John J Leonard, and Frank Dellaert. isam2: Incremental smoothing and mapping using the bayes tree. *The International Journal of Robotics Research*, 31(2):216–235, 2012. 3, 6

[16] Nikolaos Kourtzanidis and Sajad Saeedi. Rl-pgo: Reinforcement learning-based planar pose-graph optimization. *IEEE Control Systems Letters*, 7:3777–3782, 2023. 3, 6, 7

[17] Sai Krishna Ghanta and Ramviyas Parasuraman. Policies over poses: Reinforcement learning based distributed pose-graph optimization for multi-robot slam. *arXiv e-prints*, pages arXiv–2510, 2025. 3

[18] Rainer Kümmerle, Giorgio Grisetti, Hauke Strasdat, Kurt Konolige, and Wolfram Burgard. g 2 o: A general framework for graph optimization. In *2011 IEEE international conference on robotics and automation*, pages 3607–3613. IEEE, 2011. 3, 6, 7

[19] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual–inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34(3):314–334, 2015. 3

[20] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. *Advances in Neural Information Processing Systems*, 33:17346–17357, 2020. 2

[21] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 4

[22] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[23] Jiajun Lv, Xiaolei Lang, Jinhong Xu, Mengmeng Wang, Yong Liu, and Xingxing Zuo. Continuous-time fixed-lag smoothing for lidar-inertial-camera slam. *IEEE/ASME Transactions on Mechatronics*, 28(4):2259–2270, 2023. 3

[24] Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv preprint arXiv:2505.12549*, 2025. 8

[25] Gabriel Moreira, Manuel Marques, and Joao Paulo Costeira. Fast pose graph optimization via krylov-schur and cholesky factorization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1898–1906, 2021. 1, 7

[26] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16695–16705, 2025. 8

[27] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 3, 5

[28] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *Advances in neural information processing systems*, 31, 2018. 2

[29] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

[30] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 1, 2, 4, 7, 8

[31] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 2

[32] 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 *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020. 3, 5

[33] Junjie Wen, Jinqiang Cui, Zhenjun Zhao, Ruixin Yan, Zhi Gao, Lihua Dou, and Ben M Chen. Syreanet: A physically guided underwater image enhancement framework integrating synthetic and real images. *arXiv preprint arXiv:2302.08269*, 2023. 3, 4

[34] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European conference on computer vision*, pages 467–483. Springer, 2016. 2