U-VINDO: Underwater Visual-Inertial Odometry Enhanced with Robot Dynamics Predictions Powered by Port-Hamiltonian Neural ODE Networks

Yazan Maalla[1], Zein Alabedeen Barhoum[1], Maxim Popov[1], and Sergey Kolyubin[1]

***Abstract*— Reliable state estimation for autonomous underwater robots is challenging due to GPS denial, poor visibility, and complex hydrodynamic disturbances. Conventional visual–inertial odometry (VIO) ignores actuation dynamics, misinterpreting external forces as noise and causing drift. We introduce U-VINDO, a multi-sensor framework that integrates physically consistent robot dynamics with visual–inertial sensor fusion. A Port-Hamiltonian Neural ODE (PH-NODE) learns energy-conserving underwater dynamics, pre-integrated as a dynamics factor in a tightly coupled factor graph that jointly estimates trajectory and external forces. On real NTNU underwater datasets, U-VINDO consistently improves accuracy over VINS-Mono with only marginal overhead. In simulation with known force ground truth, it reduces average pose error by over 60% and achieves near-perfect force estimation (R=0.99). To our knowledge, this is the first integration of physically consistent neural ODEs into optimization-based SLAM, yielding interpretable force estimates for higher-level autonomy.**

Keywords: Underwater robotics, visual-inertial odometry, state estimation, SLAM, robot dynamics, factor graph optimization, external force estimation, autonomous underwater vehicles.

I. Introduction

Accurate localization is fundamental for autonomous robots in GPS-denied environments, where turbulent fluids, erratic currents, and limited visibility degrade sensor data. Conventional VIO and LIO fuse multi-sensor data to estimate trajectories, yet neglect the robot's dynamic model—a rich source of kinematic constraints. Early attempts relied on analytical rigid-body models that oversimplify nonlinear robot–environment interactions. Data-driven modeling can capture these nonlinearities but suffers from poor physical consistency and weak generalization. Hybrid strategies—enforcing physics while adapting residual effects via learning—offer a principled balance and can improve robustness in dynamic environments.

We propose U-VINDO which: (1) employs a Port-Hamiltonian Neural ODE (PH-NODE) [1] to learn a physically consistent underwater dynamics model; (2) fuses visual, inertial, and dynamics cues within a single factor-graph optimization; and (3) estimates external forces online to stabilize trajectory prediction under disturbances.

The main contributions are:

[1] Biomechatronics and Energy-Efficient Robotics Lab (BE2R Lab), ITMO University, Saint Petersburg, Russia. {yazanmaalla, s.kolyubin}@itmo.ru

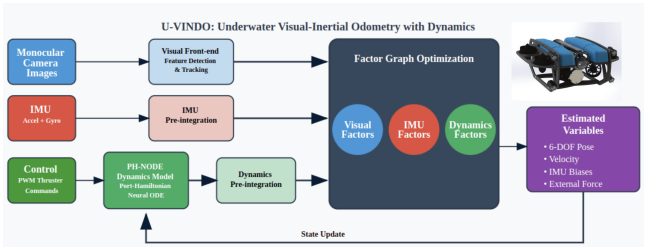


Fig. 1: U-VINDO pipeline: monocular features and IMU form visual and inertial factors; PWM commands drive the PH-NODE, preintegrated into a dynamics factor. A sliding-window factor graph jointly estimates pose, velocity, IMU biases, and external force.

- Dynamics-informed VIO: A unified factor-graph framework fusing visual, inertial, and learned dynamics constraints. The PH-NODE predicts control-induced motion, pre-integrated analogously to IMU factors, enabling explicit separation of commanded and disturbance-induced motion.
- Physically consistent dynamics in the factor graph: The PH-NODE enforces energy conservation while learning inertial, damping, and actuation parameters from data, mapping raw thruster PWM commands to the acceleration space.
- Online external force estimation: A dynamics residual compares PH-NODE predicted motion (the commanded motion following the learnt dynamics of the robot in the disturbance-free conditions) with visual–inertial observations, attributing discrepancies to an explicit external force variable jointly optimized with the trajectory.

II. Related Work

a) Dynamics-Aware SLAM.: Early works extended VIO with dynamics models for aerial robots. Pre-integrated dynamics factors for UAVs were introduced in [2], inspired by IMU preintegration [3]. VIMO [4] jointly optimized pose and external forces using analytical UAV dynamics, while VID-Fusion [5] improved force estimation at high computational cost. DIDO [6] combines a CNN de-bias module with a ResNet aerodynamic predictor and an EKF, and HDVIO [7] couples a temporal-convolution network with factor-graph optimization for high-rate aerodynamic force prediction. All of these target aerial platforms whose models do not transfer to the complex hydrodynamics of underwater vehicles.

b) Physics-Informed Learning for Dynamics.: PINNs [8] embed physical constraints into training losses, with UAV applications demonstrating improved state estimation [9]. Structure-preserving models include Hamiltonian NNs [10], Deep Lagrangian Networks [11], and PH-NODEs on $SE(3)$ [12], which capture dissipative forces without requiring known system parameters. The latter was recently applied to underwater robots in our prior work [1], achieving accurate long-term dynamics prediction — this is the model we build upon.

c) Underwater Dynamics-Enhanced Navigation.: Model-aided EKFs [13] fused simplified dynamics with current estimation but suffer from linearization errors. Pseudo-DVL methods [14] replaced Doppler log measurements using a simplified translational model, ignoring rotational hydrodynamics. Learning-based approaches [15] adapt dynamics online but lack physical consistency. Factor-graph SLAM [16] switches between VIO and a kinematic model for robustness in feature-poor environments, but ignoring hydrodynamic forces still leads to drift. DeepVL [17] learns velocity from proprioceptive inputs (IMU, thruster commands) fused in an EKF, maintaining odometry during visual blackout. Yet none of these integrate physically consistent learned dynamics into underwater SLAM using only a monocular camera and IMU, without extra sensors such as DVL or sonar.

III. Methodology

A. System Overview

Fig. 1 shows the full pipeline, extending VINS-Mono [18]. The front-end extracts image features and preintegrates IMU data. Control inputs feed the PH-NODE to generate a dynamics prior, and a dynamics residual enforces consistency between model-predicted and observed motion, enabling joint estimation of trajectory, biases, and external forces in a sliding-window factor graph.

The state at keyframe k augments the standard VIO state with an external disturbance term:

$$\mathbf{x}_k = \left[\mathbf{p}_k,\ \mathbf{q}_k,\ \mathbf{v}_k,\ \mathbf{b}_{g_k},\ \mathbf{b}_{a_k},\ f_{e_k}\right]^\top, \tag{1}$$

where $(\mathbf{p}_k, R(\mathbf{q}_k)) \in SE(3)$, $\mathbf{v}_k$ is body linear velocity, (b_{g_k}, b_{a_k}) are IMU biases, and $f_{e_k} \in \mathbb{R}^3$ is the external body-frame specific force to be estimated. The joint optimization cost is:

$$\mathcal{J} = e_{\text{visual}} + e_{\text{inertial}} + e_{\text{margin}} + e_{\text{dynamics}}. \tag{2}$$

The first three terms follow the standard VIO formulation [18], [3], [19]. The e_{dynamics} term is detailed in Section III-C.

B. Port-Hamiltonian Neural ODE

We adopt the PH-NODE from our prior work [1]. Let $\boldsymbol{\eta} \in SE(3)$ be the robot pose, $\boldsymbol{\xi} \in \mathbb{R}^6$ its body-frame twist, $\boldsymbol{\rho} = \mathbf{M}(\boldsymbol{\eta})\boldsymbol{\xi}$ the generalized momenta, and $\mathbf{u}$ the control input. The dynamics are:

$$\dot{\boldsymbol{\eta}} = \boldsymbol{\eta}\,\boldsymbol{\xi}^\wedge, \quad \dot{\boldsymbol{\xi}} = \mathbf{M}^{-1}(\boldsymbol{\eta})\,\dot{\boldsymbol{\rho}} + \dot{\mathbf{M}}^{-1}(\boldsymbol{\eta})\,\boldsymbol{\rho}, \tag{3}$$

where inertia $\mathbf{M}$, dissipation $\mathbf{D}$, potential V, and actuation map $\mathbf{g}$ are all neural networks parameterized to enforce positive definiteness via Cholesky decomposition. The first three components of $\dot{\boldsymbol{\xi}}$ yield the commanded linear acceleration $\mathbf{a}_d^b$ from control inputs alone, under nominal conditions with no external disturbances. For more details on the PH-NODE architecture and training, see [1].

C. Dynamics Preintegration and Residual

The dynamics factor enables external-force estimation by comparing motion predicted by the learned dynamics model with motion observed from visual–inertial odometry. The PH-NODE is trained on data collected under minimal external disturbances [20], so its output $\mathbf{a}_d^b$ reflects nominal actuation-driven motion. External interactions (currents, contacts) are represented as an additional body-frame specific acceleration f_e^b. The continuous-time kinematics separating commanded from disturbance motion are:

$$\dot{\mathbf{p}}_b^w = \mathbf{v}_b^w, \qquad \dot{\mathbf{v}}_b^w = \mathbf{R}_b^w\big(\mathbf{a}_d^b + f_e^b\big). \tag{4}$$

A zero-mean Gaussian prior $f_e^b \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_f)$ reflects that disturbances are typically small unless data suggest otherwise.

To handle high-rate control inputs, we preintegrate the dynamics between keyframes analogously to IMU preintegration [3]. Integrating over $[t_k, t_{k+1}]$ with $\Delta t = t_{k+1} - t_k$ and transforming to the body frame at t_k, the state-based increments are:

$$\boldsymbol{\beta}_k^{k+1} = \mathbf{R}_w^{b_k}\big(\mathbf{v}_{k+1} - \mathbf{v}_k\big) - f_{e_k}\,\Delta t, \tag{5}$$

$$\boldsymbol{\alpha}_k^{k+1} = \mathbf{R}_w^{b_k}\big(\mathbf{p}_{k+1} - \mathbf{p}_k - \mathbf{v}_k\,\Delta t\big) - \tfrac{1}{2} f_{e_k}\,\Delta t^2. \tag{6}$$

The PH-NODE provides the commanded acceleration $\mathbf{a}_d^b(t)$, which acts as a "virtual" measurement source. We preintegrate it in the b_k frame using first-order Euler updates, initializing $\hat{\boldsymbol{\alpha}}_{b_k}^{b_k} = \mathbf{0}$, $\hat{\boldsymbol{\beta}}_{b_k}^{b_k} = \mathbf{0}$, and propagating the orientation increment $\hat{\boldsymbol{\gamma}}$ from gyroscope measurements $\tilde{\boldsymbol{\omega}}$:

$$\hat{\boldsymbol{\gamma}}_{i+1}^{b_k} = \hat{\boldsymbol{\gamma}}_i^{b_k} \otimes \begin{bmatrix} 1 \\ \frac{1}{2}(\tilde{\boldsymbol{\omega}}_i - \mathbf{b}_{g_i})\delta t \end{bmatrix}, \tag{7}$$

$$\hat{\boldsymbol{\beta}}_{i+1}^{b_k} = \hat{\boldsymbol{\beta}}_i^{b_k} + R(\hat{\boldsymbol{\gamma}}_i^{b_k})\,\mathbf{a}_d^{b_i}\,\delta t, \tag{8}$$

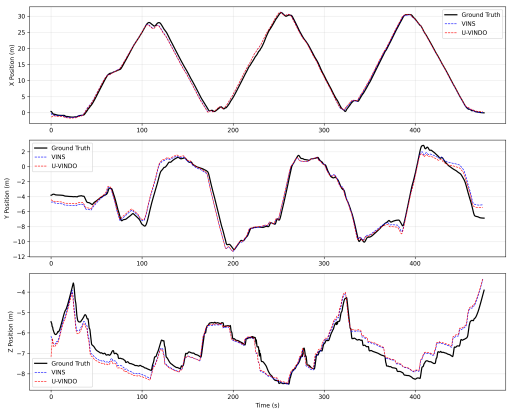
$$\hat{\boldsymbol{\alpha}}_{i+1}^{b_k} = \hat{\boldsymbol{\alpha}}_i^{b_k} + \hat{\boldsymbol{\beta}}_i^{b_k}\,\delta t + \tfrac{1}{2} R(\hat{\boldsymbol{\gamma}}_i^{b_k})\,\mathbf{a}_d^{b_i}\,\delta t^2. \tag{9}$$

The dynamics residual penalizes disagreement between state-predicted and model-predicted increments, together with a zero-mean prior on f_e:

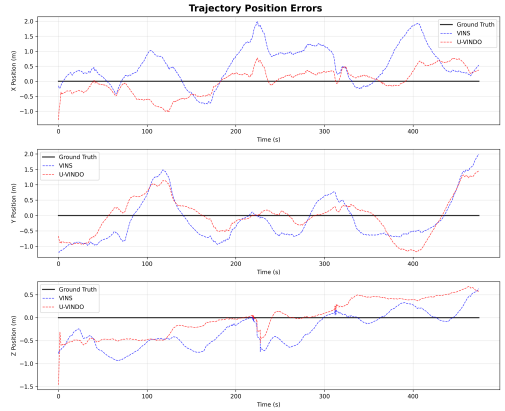
$$\mathbf{r}_d = \begin{bmatrix} \boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}} \\ \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \\ f_e \end{bmatrix}, \quad \mathbf{W}_d^k = \operatorname{diag}\big(\mathbf{P}_{\alpha\beta}^{-1},\ \boldsymbol{\Sigma}_f^{-1}\big), \tag{10}$$

where $\mathbf{P}_{\alpha\beta}$ is the preintegration covariance propagated via a linearized error-state model using an empirically calibrated dynamics covariance $\boldsymbol{\Sigma}_{\text{dyn}}$, estimated on a held-out validation set as the sample covariance of prediction residuals $\boldsymbol{\epsilon}_k = \mathbf{a}_{d,k} - \mathbf{a}_{gt,k}$.

The residual $\mathbf{r}_d$ is jointly optimized with all VIO states. If the trajectory already explains the observations, f_e stays



(a) Estimated trajectories.



(b) Position errors.

Fig. 2: NTNU `fjord_5`: VINS-Mono (blue, dashed) and U-VINDO (red, dashed) vs. ground truth (black, solid).

TABLE I: NTNU Dataset: ATE [m] and ARE [deg] (RMSE). **Bold** = best per sequence.

Sequence	Length [m]	APE VINS	APE U-VINDO	ARE VINS	ARE U-VINDO
fjord_1	156.6	**0.428**	0.445	1.233	**1.072**
fjord_2	287.3	**1.129**	1.142	**2.654**	3.067
fjord_3	183.0	**0.776**	0.789	**1.963**	2.055
fjord_4	229.9	**0.393**	0.490	**1.235**	1.341
fjord_5	239.5	0.851	**0.832**	1.863	**1.540**
fjord_6	387.4	1.245	**1.211**	2.872	**2.361**
mclab_1	133.8	0.594	**0.417**	1.842	**1.513**
mclab_2	128.8	0.367	**0.340**	1.984	**1.713**
Avg.	–	0.848	**0.833**	1.956	**1.833**

near zero due to the Gaussian prior. Persistent discrepancies between dynamics-predicted and observed motion drive f_e away from zero, indicating genuine external disturbances. Crucially, this prevents such disturbances from being absorbed as bias into visual–inertial residuals, preserving odometry accuracy even under significant environmental interactions.

TABLE II: Simulation results: trajectory accuracy (APE [m], ARE [deg], RMSE) and force estimation. **Bold** = best.

Sequence	APE VINS ↓	APE U-VINDO ↓	ARE VINS ↓	ARE U-VINDO ↓	Force RMSE [N/kg] ↓	Force Corr. ↑
Seq. 1	1.519	**0.844**	1.523	**0.716**	0.797	0.984
Seq. 2	1.183	**0.631**	0.877	**0.453**	0.292	0.995
Seq. 3	1.312	**0.579**	0.572	**0.387**	0.802	0.984
Seq. 4	1.261	**0.564**	1.090	**0.591**	0.781	0.981
Seq. 5	1.764	**0.362**	0.676	**0.385**	0.746	0.985
Seq. 6	2.292	**0.643**	1.612	**0.713**	0.765	0.984
Avg.	1.555	**0.604**	1.058	**0.541**	0.697	0.986

IV. Evaluation

A. Real-World Dataset

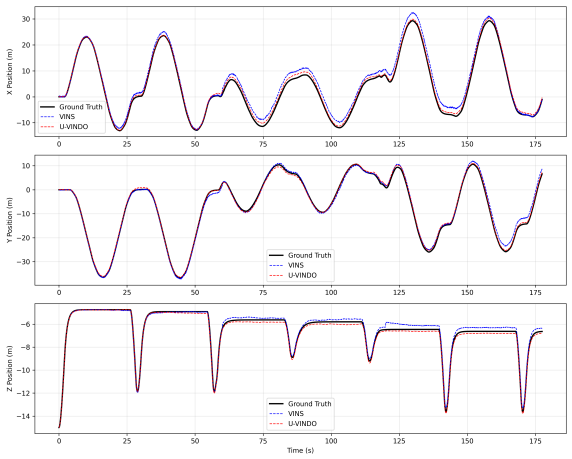
We evaluate on the NTNU underwater dataset [20], providing synchronized imagery, IMU, and thruster commands from a BlueROV2. Table I reports Absolute Pose Error (APE) and Absolute Rotation Error (ARE) for the baseline VINS-Mono [18] and for U-VINDO, and Fig. 2 illustrates the *fjord_5* sequence: panel a shows the estimated x, y, z trajectories against ground truth, and panel b shows the corresponding position errors. The comparison highlights that both methods track the ground truth closely. U-VINDO achieves slightly reduced drift. On average across the dataset, U-VINDO achieves consistent improvements: translation error improves by 1.77% and rotation error by 6.29%, with only 3.3% runtime overhead,demonstrating that the additional computational load is manageable within real-time constraints. These modest but consistent gains indicate that the dynamics prior can reduce drift and provide a stabilizing effect even in sequences without significant external perturbations. However, as the dataset does not include explicit force actuation or ground-truth external forces, it is unsuitable for quantitative validation of force estimation. To our knowledge, no public underwater dataset currently provides visual–inertial data, actuation signals, and ground-truth external forces simultaneously.

B. Simulation with Known Force Ground Truth

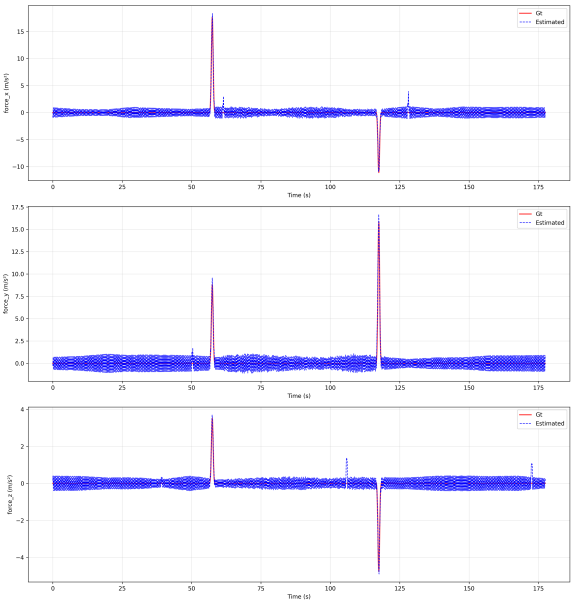
To quantify force estimation, we use HoloOcean [21] to generate six BlueROV2 sequences with full ground truth: synchronized images, IMU, PWM commands, and injected forces up to 20 N at random intervals. For this experiment, a PH-NODE pre-trained on the NTNU dataset is fine-tuned on simulator data to mitigate the sim-to-real gap.

Fig. 3a compares trajectories from VINS-Mono and U-VINDO: both follow ground truth closely, but U-VINDO shows consistently smaller deviations, especially during force injections (near 60 s and 120 s).

Table II shows U-VINDO reduces APE by >60% and ARE by ≈50% relative to the baseline, confirming that dynamics-informed constraints suppress drift under external disturbances. Force estimation (Fig. 3b) recovers both amplitude and timing of injected forces, achieving mean correlation $R = 0.986$ and average RMSE below 0.7 N/kg. These results indicate that the optimizer consistently assigns unmodeled dynamics to the explicit force variable rather than biasing pose or IMU residuals.



(a) Position vs time.



(b) Estimated force components.

Fig. 3: Simulation experiments with injected external forces in HoloOcean. (a) Position traces: U-VINDO remains closer to ground truth, particularly near injection intervals ($\approx$60 s and $\approx$120 s). (b) Recovered force components: estimated external forces closely follow ground truth injections.

V. Conclusion

We presented U-VINDO, the first underwater VIO system to integrate physically consistent neural ODE dynamics into a tightly coupled factor-graph optimizer with joint external force estimation. The dynamics factor stabilizes odometry in benign conditions and actively suppresses drift under disturbances, yielding interpretable force estimates useful for higher-level control. Future work will extend the framework to full 6-DoF torque modeling and online adaptive covariance for out-of-distribution regimes.

References

[1] Z. A. Barhoum and S. Kolyubin, "Physically consistent dynamic modeling of underwater robots for robust long-horizon motion prediction," *Journal of Instrument Engineering*, vol. 68, no. 11, pp. 983–995, 2025.

[2] A. Antonini, *Pre-Integrated Dynamics Factors and a Dynamical Agile Visual-Inertial Dataset for UAV Perception*. PhD thesis, Massachusetts Institute of Technology, 2018.

[3] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual–inertial odometry," *IEEE Transactions on Robotics*, vol. 33, pp. 1–21, 2015.

[4] B. Nisar, P. Foehn, D. Falanga, and D. Scaramuzza, "Vimo: Simultaneous visual inertial model-based odometry and force estimation," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2785–2792, 2019.

[5] Z. Ding, T. Yang, K. Zhang, C. Xu, and F. Gao, "Vid-fusion: Robust visual-inertial-dynamics odometry for accurate external force estimation," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14469–14475, 2021.

[6] K. Zhang, C. Jiang, J. Li, S. Yang, T. Ma, C. Xu, and F. Gao, "Dido: Deep inertial quadrotor dynamical odometry," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9083–9090, 2022.

[7] G. Cioffi, L. Bauersfeld, and D. Scaramuzza, "Hdvio: Improving localization and disturbance estimation with hybrid dynamics vio," *ArXiv*, vol. abs/2306.11429, 2023.

[8] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *J. Comput. Phys.*, vol. 378, pp. 686–707, 2019.

[9] D. Bianchi, N. Epicoco, M. D. Ferdinando, S. D. Gennaro, and P. Pepe, "Physics-informed neural networks for unmanned aerial vehicle system estimation," *Drones*, vol. 8, no. 716, 2024.

[10] S. Greydanus, M. Dzamba, and J. Yosinski, "Hamiltonian neural networks," 2019.

[11] M. Lutter, C. Ritter, and J. Peters, "Deep lagrangian networks: Using physics as model prior for deep learning," in *International Conference on Learning Representations (ICLR)*, 2019.

[12] T. Duong, A. Altawaitan, J. Stanley, and N. Atanasov, "Port-hamiltonian neural ODE networks on lie groups for robot dynamics learning and control," *IEEE Transactions on Robotics*, vol. 40, pp. 3695–3715, 2024.

[13] A. Martinez, L. Hernandez, H. Sahli, Y. Valeriano-Medina, M. Orozco-Monteagudo, and D. Garcia-Garcia, "Model-aided navigation with sea current estimation for an autonomous underwater vehicle," *International Journal of Advanced Robotic Systems*, vol. 12, p. 103, 2015.

[14] A. Karmozdi, M. Hashemi, H. Salarieh, and A. Alasty, "Implementation of translational motion dynamics for ins data fusion in dvl outage in underwater navigation," *IEEE Sensors Journal*, vol. 21, no. 5, pp. 6652–6662, 2021.

[15] B. Wehbe, M. Hildebrandt, and F. Kirchner, "A framework for on-line learning of underwater vehicles dynamic models," *2019 International Conference on Robotics and Automation (ICRA)*, pp. 7969–7975, 2019.

[16] B. Joshi, H. Damron, S. Rahman, and I. Rekleitis, "Sm/vio: Robust underwater state estimation - switching between model-based and visual inertial odometry," in *IEEE Conference on Robotics and Automation (ICRA)*, 2023.

[17] M. Singh and K. Alexis, "Deepvl: Dynamics and inertial measurements-based deep velocity learning for underwater odometry," 2025.

[18] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, pp. 1004–1020, 2017.

[19] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment – a modern synthesis," in *Vision Algorithms: Theory and Practice, International Workshop on Vision Algorithms*, vol. 1883 of *Lecture Notes in Computer Science*, pp. 298–372, Springer, 2000.

[20] M. Singh, M. Dharmadhikari, and K. Alexis, "An online self-calibrating refractive camera model with application to underwater odometry," 2023.

[21] E. Potokar, K. Lay, K. Norman, D. Benham, S. Ashford, R. Peirce, T. B. Neilsen, M. Kaess, and J. G. Mangelson, "Holoocean: A full-featured marine robotics simulator for perception and autonomy," *IEEE Journal of Oceanic Engineering*, vol. 49, no. 4, pp. 1322–1336, 2024.