Multi-Modal Fusion for Underwater Localization through Hierarchical Uncertainty Awareness

Yuanbo Xue[1,2], *Student Member, IEEE,* Yaming Ou[3], *Graduate Student Member, IEEE,* Junfeng Fan[3,*], *Senior Member, IEEE,* Chao Zhou[3], *Senior Member, IEEE,* Bing Wang[1,2,*], *Member, IEEE,* Chih-yung Wen[1,2], *Member, IEEE*

***Abstract*—Traditional vehicle localization methods, such as visual simultaneous localization and mapping, often struggle underwater, especially under conditions like motion blur and low light. To address these challenges, this paper presents a novel multi-modal fusion system for underwater localization that integrates hierarchical uncertainty awareness (UA) into the acoustic/visual/inertial/pressure (AVIP) fusion framework, termed UA-AVIP, to achieve robust and accurate localization. The Doppler Velocity Log is introduced for reliable scale information and pressure sensor for water-depth measurements. A dual-framework approach is introduced to parallelize the fusion of acoustic and pressure data in an Extended Kalman Filter framework with the integration of visual and inertial measurements via a visual-inertial system. The final estimates are subsequently unified through factor graph optimization. Furthermore, a hierarchical UA strategy is developed to effectively manage uncertainties at multiple levels within the dual-framework fusion. At the measurement level, noise variance is adaptively estimated through spectrum analysis. At the module level, a deep learning-assisted sub-model enhances acoustic and pressure estimation. At the system level, an adaptive weighting strategy within the factor graph addresses dynamic uncertainties. Both self-collected and open-source datasets in the pool, lakes, and oceans under distinct scenarios are used to train and validate the effectiveness and superiority of the proposed UA-AVIP system. The results underscore the critical importance of UA in underwater localization. The source code is available at https://github.com/HKPolyU-UAV/UA-AVIP.**

***Index Terms*—Uncertainty awareness, Multi-Modal fusion, Deep learning-assisted estimation, Underwater localization, Underwater vehicles.**

I. INTRODUCTION

UNDERWATER vehicles serve as intelligent tools for marine development with their advanced environmental perception, robust payload capabilities, and high system stability [1], [2]. Equipped with sophisticated sensors, underwater vehicles facilitate both instrumental measurement and situational awareness in underwater environments [3]–[5]. In particular, the precise localization forms the foundation of reliable underwater exploration [6]. One of the most effective approach for achieving accurate localization is Simultaneous Localization and Mapping (SLAM) [7].

Relying on visual features, traditional visual SLAM and visual inertial navigation system (VINS) perform well for aerial and ground applications [8]. However, the underwater localization presents significantly greater challenges due to several unique environmental factors. Firstly, the light attenuation and suspended particles in underwater environments severely degrade the visual quality [9]. Therefore, visual estimations under constrained optical conditions require underwater adaptations [10]. Secondly, underwater environment variables, such as currents and tides, can result in unpredictable vehicle vibration [11]. The above dynamic conditions can lead to accumulated drifts in localization estimates based on the inertial navigation system [12]. Moreover, underwater vehicles lack access to the Global Positioning System, which further poses localization challenges [13].

Based on the above issues, developing a localization method suitable for harsh underwater environments is a crucial issue that urgently needs to be addressed in the field of underwater robotics. A promising solution for underwater localization is to fuse multi-modal information for optimal estimates [14]–[16]. Nevertheless, multi-modal fusion is subject to hierarchical uncertainties arising from measurement noise, odometry module estimation, joint optimization, etc. Effectively managing these uncertainties is crucial for achieving robust and reliable localization in underwater environments.

For multi-modal fusion-based localization, existing methods have integrated additional sensing modalities with visual/inertial (VI) estimation to enhance performance. For example, Ou et al. [6] developed a hybrid vision system based on an active vision device. Moreover, Joshi [7] and Rahman et al. [14] introduced range measurements from a scanning sonar and depth measurements from a Pressure Sensor (PS) to complement visual inertial odometry (VIO). Furthermore, Wang et al. [17] explored graph-based localization by fusing multi-source acoustic measurements, specifically a Doppler Velocity Log (DVL) and an imaging sonar. The initial pose was derived from dead reckoning (DR) by fusing DVL, Inertial Measurement Unit (IMU) and PS. While above methods have demonstrated promising results on self-collected datasets, certain challenges remain. First, environmental adaptability is limited. The work by [6] did not address the localization in natural environments, such as lakes or oceans. Second, robustness against modality-specific failure is often overlooked. Notably, [7] and [14] did not investigate scenarios involving limited visibility or experiments with visual failures, which are

[1]Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong. yuanbo.xue@connect.polyu.hk, {bingwang, cywen}@polyu.edu.hk
[2]Research Centre for Unmanned Autonomous System, The Hong Kong Polytechnic University, Hong Kong.
[3]Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. {ouyaming2021, junfeng.fan, chao.zhou}@ia.ac.cn
Yuanbo Xue and Yaming Ou contributed equally to this work.
Corresponding author: Bing Wang and Junfeng Fan.

common in underwater settings. Conversely, [17] employed sonar without visual measurements, the lack of visual integration may limit reliability when acoustic interference occurs.

For uncertainty awareness (UA) in underwater localization, some works have addressed noise estimation and outlier detection [18]–[22]. For instance, Zhang et al. [18] integrated noise estimation into kinetic control for subsea instrumentation. Similarly, Wang et al. [19] and Bian et al. [20] advanced filtering algorithms to mitigate measurement biases. However, [18]–[20] primarily modeled uncertainties at the measurement level alone. Specifically, [18] was restricted to short-range electromagnetic applications, while [19], [20] did not consider uncertainties in the odometry estimation or the joint optimization within multi-modal fusion. To explicitly model uncertainties of the graph system, [22] employed a probabilistically consistent approach for state estimation. Noises and bias states were incorporated as nodes in the factor graph. Although bias trajectories alongside the vehicle state were jointly estimated, this method relied on the still-water assumption, rendering flow-derived velocity measurements invalid in real dynamic environments with ocean currents.

As demonstrated by above works, both multi-modal fusion and uncertainty-aware method have been shown to substantially improve localization performance. Nevertheless, few works investigate localization uncertainties, especially from a hierarchical perspective. To address this gap, our objective is to fully integrate the hierarchical UA into the Acoustic/Visual/Inertial/Pressure (AVIP) fusion system, say UA-AVIP, to achieve underwater localization in diverse scenarios with reduced localization error. Compared with previous studies, the main contributions of this work are as follows.

1) A novel multi-modal fusion system, termed UA-AVIP, is proposed for underwater localization. Unlike the methods in [6], [7], [14], [17], the proposed UA-AVIP introduces a dual-framework architecture which incorporates DVL, stereo camera, IMU, and PS. The system employs the Extended Kalman Filter (EKF) to tightly integrate acoustic and pressure measurements, while leveraging VINS to handle visual and inertial data. These two complementary frameworks are seamlessly coupled through the factor graph optimization (FGO).

2) An uncertainty-aware method that operates on different levels is designed, which enables more accurate localization. Compared to methods in [18]–[22], which only manage measurement noises, the proposed method achieves hierarchical awareness. Specifically, at the measurement level, the noise variance is dynamically estimated through spectrum analysis. At the module level, deep learning-based refinement is implemented for acoustic and pressure estimation. Meanwhile, we employ an adaptive weighting strategy at the system level to facilitate robust modal fusion within the FGO.

3) Extensive real-world experiments conducted in the pool, lakes, and open oceans validate the effectiveness and superiority of the proposed underwater localization system. In contrast to prior work that relied primarily on simulations or controlled indoor environments (e.g., [6], [18], [22]), the present system undergoes a comprehensive evaluation in diverse and challenging conditions. The results demonstrate the robustness and high precision when UA is incorporated.

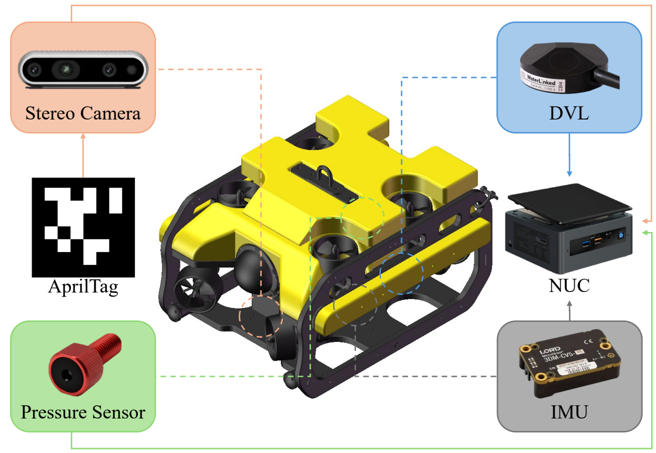


Fig. 1: Mechatronic design and sensors of the underwater multi-modal system.

II. Preliminaries and Problem Statement

Given an underwater vehicle with UA, the dual-framework system provides the state estimation for localization. Acoustic and pressure measurements are fused to estimate vehicle poses based on the EKF framework while augmenting the VINS framework in scenarios with sparse visual features. An hierarchical method is constructed to account for uncertainties at various levels. This section provides fundamental concepts and states the problem described.

A. Notations

To evaluate the real-world performance, an underwater multi-modal system is designed, as shown in Fig. 1. The sensor suite comprises the DVL, the stereo camera, the IMU, and the PS, which are denoted as $\mathcal{A}$, $\mathcal{V}$, $\mathcal{I}$, and $\mathcal{P}$, respectively.

The state of the underwater vehicle is defined as

$$\mathbf{x} = \left[{}_{\mathcal{W}}\mathbf{p}_{\mathcal{I}}^{\top}, {}_{\mathcal{W}}\mathbf{q}_{\mathcal{I}}^{\top}, {}_{\mathcal{W}}\mathbf{v}_{\mathcal{I}}^{\top}, \mathbf{b}_g^{\top}, \mathbf{b}_a^{\top} \right]^{\top} \in \mathbb{R}^3 \times \mathbb{SO}(3) \times \mathbb{R}^9, \quad (1)$$

where ${}_{\mathcal{W}}\mathbf{p}_{\mathcal{I}}$, ${}_{\mathcal{W}}\mathbf{q}_{\mathcal{I}}$ and ${}_{\mathcal{W}}\mathbf{v}_{\mathcal{I}}$ represent the position, orientation and velocity in the inertial frame $\mathcal{I}$ with respect to the world frame $\mathcal{W}$, respectively. $\mathbf{b}_g$ and $\mathbf{b}_a$ are the biases of the gyroscope and the accelerometer. $\mathbb{SO}(3)$ denotes the Special Orthogonal Group. The transformation from $\mathcal{I}$ to $\mathcal{W}$ is

$${}_{\mathcal{W}}\mathbf{T}_{\mathcal{I}} = \begin{bmatrix} {}_{\mathcal{W}}\mathbf{R}_{\mathcal{I}} & {}_{\mathcal{W}}\mathbf{p}_{\mathcal{W},\mathcal{I}} \\ \mathbf{0} & 1 \end{bmatrix} \in \mathbb{SE}(3), \quad (2)$$

where ${}_{\mathcal{W}}\mathbf{R}_{\mathcal{I}} \in \mathbb{SO}(3)$ and ${}_{\mathcal{W}}\mathbf{p}_{\mathcal{W},\mathcal{I}} \in \mathbb{R}^3$. $\mathbb{SE}(3)$ is the Special Euclidean Group. ${}_{\mathcal{W}}\mathbf{R}_{\mathcal{I}}$ denotes the rotation matrix and ${}_{\mathcal{W}}\mathbf{p}_{\mathcal{W},\mathcal{I}}$ presents the translation vector.

For acoustic measurements, the high-precision ${}_{\mathcal{W}}\mathbf{v}_{\mathcal{A}} \in \mathbb{R}^3$ is obtained from the Doppler effect. ${}_{\mathcal{W}}\mathbf{p}_{\mathcal{A}}$ is estimated by fusing ${}_{\mathcal{W}}\mathbf{v}_{\mathcal{A}}$ with onboard gyroscope measurements via DR

$$\begin{aligned} \mathbf{p}_{\mathcal{A}} &= \left[\mathbf{p}_{\mathcal{A}_x}, \mathbf{p}_{\mathcal{A}_y}, \mathbf{p}_{\mathcal{A}_z} \right]^{\top} \\ &= \left[\int \mathbf{v}_{\mathcal{A}_x}\, dt, \int \mathbf{v}_{\mathcal{A}_y}\, dt, \int \mathbf{v}_{\mathcal{A}_z}\, dt \right]^{\top}. \end{aligned} \quad (3)$$

$\mathbf{q}_{\mathcal{A}}$ can be computed given the initial orientation ${}_{\mathcal{W}}\mathbf{q}_{\mathcal{I}}(t_0)$, the angular velocity $\boldsymbol{\omega}_{\mathcal{I}}$, and the quaternion multiplication $\otimes$

$$\mathbf{q}_{\mathcal{A}} = {}_{\mathcal{W}}\mathbf{q}_{\mathcal{I}} \otimes {}_{\mathcal{I}}\mathbf{q}_{\mathcal{A}}, \tag{4}$$

where

$$\begin{aligned} {}_{\mathcal{W}}\mathbf{q}_{\mathcal{I}}(t) &= \Delta\mathbf{q} \otimes {}_{\mathcal{W}}\mathbf{q}_{\mathcal{I}}(t_0), \\ \Delta\mathbf{q} &= \left[\cos\left(\frac{\Lambda}{2}\right), \mathbf{n}\sin\left(\frac{\Lambda}{2}\right)\right]), \\ \Lambda &= |\boldsymbol{\omega}_{\mathcal{I}}|\,\Delta t, \quad \mathbf{n} = \frac{\boldsymbol{\omega}_{\mathcal{I}}}{|\boldsymbol{\omega}_{\mathcal{I}}|}. \end{aligned} \tag{5}$$

B. Nonlinear Filtering and Graph Optimization

1) Nonlinear Filtering: The EKF framework delineates a nonlinear extension of Kalman filtering in which the state transition and observation models are defined by [23]

$$\begin{aligned} \boldsymbol{x}_k &= f(\boldsymbol{x}_{k-1}, \mathbf{0}) + \boldsymbol{W}_{k-1}, \\ \boldsymbol{z}_k &= h(\boldsymbol{x}_k) + \boldsymbol{V}_k, \end{aligned} \tag{6}$$

where $f(\cdot)$ is the state transition function and $h(\cdot)$ is the observation function. $\boldsymbol{x}_k$ and $\boldsymbol{z}_k$ represent the EKF state vector and the measurement at the k-th sample point, respectively. $\boldsymbol{W}_{k-1}$ is the process noise and $\boldsymbol{V}_k$ is the measurement noise. $\boldsymbol{W}_k$ and $\boldsymbol{V}_k$ are assumed to be zero-mean Gaussian noises with covariance $\boldsymbol{Q}_k$ and $\boldsymbol{R}_k$, respectively.

The vehicle state is predicted and updated iteratively. In the prediction step, the state estimate is predicted as

$$\hat{\boldsymbol{x}}_{k|k-1} = f(\hat{\boldsymbol{x}}_{k-1|k-1}, \mathbf{0}), \tag{7}$$

where $\hat{\boldsymbol{x}}_{k|k-1}$ is the estimate of $\boldsymbol{x}_k$ given $\boldsymbol{z}_{1:k-1} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{k-1}\}$. The predicted covariance estimate is

$$\boldsymbol{P}_{k|k-1} = \boldsymbol{F}_k \boldsymbol{P}_{k-1|k-1} \boldsymbol{F}_k^{\top} + \boldsymbol{Q}_{k-1}, \tag{8}$$

where

$$\boldsymbol{F}_k = \left.\frac{\partial f}{\partial \boldsymbol{x}}\right|_{\hat{\boldsymbol{x}}_{k-1|k-1}, \mathbf{0}}. \tag{9}$$

In the update step, the process can be described as

$$\begin{aligned} \tilde{\boldsymbol{y}}_k &= \boldsymbol{z}_k - h(\hat{\boldsymbol{x}}_{k|k-1}), \\ \boldsymbol{S}_k &= \boldsymbol{H}_k \boldsymbol{P}_{k|k-1} \boldsymbol{H}_k^{\top} + \boldsymbol{R}_k, \\ \boldsymbol{K}_k &= \boldsymbol{P}_{k|k-1} \boldsymbol{H}_k^{\top} \boldsymbol{S}_k^{-1}, \\ \hat{\boldsymbol{x}}_{k|k} &= \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k \tilde{\boldsymbol{y}}_k, \\ \boldsymbol{P}_{k|k} &= (\boldsymbol{I} - \boldsymbol{K}_k \boldsymbol{H}_k) \boldsymbol{P}_{k|k-1}, \\ \boldsymbol{H}_k &= \left.\frac{\partial h}{\partial \boldsymbol{x}}\right|_{\hat{\boldsymbol{x}}_{k|k-1}}, \end{aligned} \tag{10}$$

where $\tilde{\boldsymbol{y}}_k$ is the measurement residual, $\boldsymbol{S}_k$ is the residual covariance, $\boldsymbol{K}_k$ is the Kalman gain, $\hat{\boldsymbol{x}}_{k|k}$ and $\boldsymbol{P}_{k|k}$ are the update estimates of state and covariance.

2) Graph Optimization: The factor graph $G(\mathcal{H}, \mathcal{X}, \mathcal{E})$ is a probabilistic model with the factor node $\eta_i \in \mathcal{H}$, the variable node $\mathbf{x}_j \in \mathcal{X}$, and the edge $\varepsilon_{ij} \in \mathcal{E}$ that connects η_i to $\mathbf{x}_j$ [24]. In the sliding window, state variables of multiple frames are presented by Eq. (1) with K being the window size

$$\mathcal{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K] \in (\mathbb{R}^3 \times \mathbb{SO}(3) \times \mathbb{R}^9)^K. \tag{11}$$

To estimate the unknown $\mathcal{X}$, assume that all factors follow

$$\begin{aligned} &\eta_i(\mathcal{X}_i) \propto \exp\left\{-\frac{1}{2}\left\|h_i(\mathcal{X}_i) - \boldsymbol{z}_i\right\|^2_{\boldsymbol{\Sigma}_i}\right\}, \\ &\boldsymbol{\Sigma}_i \in \mathbb{S}^{d_i}_{++}, \quad d_i = \dim(\boldsymbol{z}_i) \in \mathbb{Z}^+. \end{aligned} \tag{12}$$

Then, a nonlinear least-squares optimization problem can be formulated as

$$\hat{\mathcal{X}} = \arg\min_{\mathcal{X}} \sum_i \left\|h_i(\mathcal{X}_i) - \boldsymbol{z}_i\right\|^2_{\boldsymbol{\Sigma}_i}. \tag{13}$$

C. Problem Statement

Given an underwater vehicle with onboard sensing, i.e., $\mathcal{A}, \mathcal{V}, \mathcal{I}, \mathcal{P}$, we aim to develop a graph-based state estimation system to provide the reliable estimation of $\mathbf{x}$. The proposed system is designed as shown in Fig. 2. A dual-framework mechanism is introduced for multi-modal fusion (see Section III-A), and a hierarchical approach is proposed to solve the localization uncertainties (see Section III-B).

Specifically, the visual and inertial modalities function as a conventional VINS framework, while the acoustic and pressure modalities serve as an acoustic/pressure (AP) estimator fusing within the EKF framework. The proposed hierarchical approach explicitly distinguishes uncertainties at three levels:

- Time-varying noise and non-stationary signals introduce underwater uncertainties at the measurement level.
- AP estimator exhibits module-level uncertainties, such as accumulated errors and unsteady estimations [25], [26].
- Factor reliability within the graph quantifies system-level uncertainties during dual-framework integration.

III. Methodology

This section outlines the proposed state estimation system, with a particular focus on UA for underwater applications. A dual-framework mechanism is first presented to parallelize multi-modal data and is subsequently optimized through a factor graph. To enhance the fusion reliability, we describe a hierarchical uncertainty-aware strategy that addresses measurement, module, and system uncertainties using the dynamic noise estimation, the deep learning-based refinement, and the adaptive weighting strategies, respectively.

A. Dual-Framework Multi-Modal Fusion

1) EKF-based AP Integration: The EKF is employed for the AP estimator due to the highly linear mapping between the state vector and the sensor observations. Compared to advanced filters for high nonlinearities and rotational singularities, the EKF provides a computationally efficient solution for first-order linearization without the overhead of sigma-point sampling or complex manifold projections [27], [28]. The framework is initialized when the DVL observes a velocity relative to the sea bottom (see Eq. (3)). Updates are performed at each DVL timestamp. The temporally closest pressure measurement is identified and incorporated, as the PS frequency is much higher than that of DVL (see Section IV-A).

The AP state vector in Eq. (6) is defined as $\boldsymbol{x}_{\mathcal{AP}} = [x, y, z, \phi, \theta, \psi, v_x, v_y, v_z]^{\top} \in \mathbb{R}^9$. Here, ϕ, θ, and ψ are

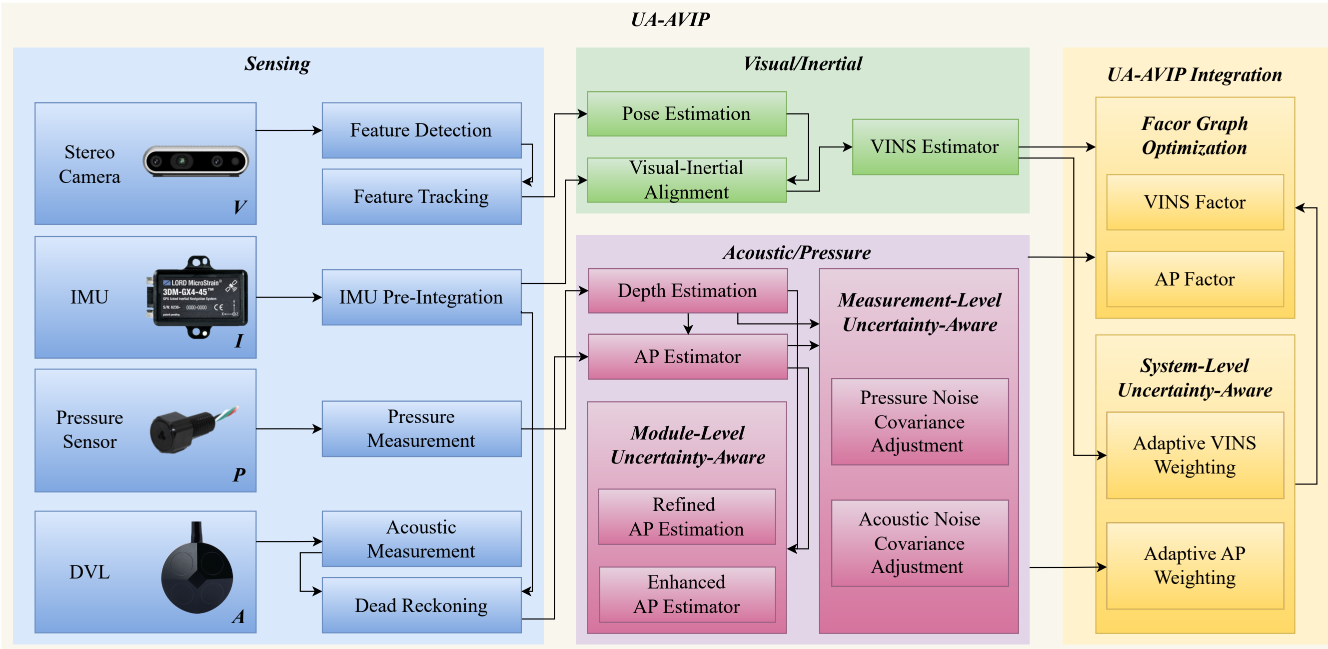


Fig. 2: The proposed system architecture is equipped with DVL, stereo camera, IMU, and PS. AP estimator integrates acoustic and pressure measurements with the IMU based on the EKF framework; The camera combines the IMU to establish the VINS.

the roll, pitch, and yaw angles, respectively. Meanwhile, the measurement vector $\boldsymbol{z}_{\mathcal{AP}} \in \mathbb{R}^9$ is given by

$$\begin{aligned} \boldsymbol{z}_{\mathcal{AP}} &= \left[{}_{\mathcal{W}}\mathbf{p}_{\mathcal{AP}}^{\top}, {}_{\mathcal{W}}\mathbf{q}_{\mathcal{AP}}^{\top}, {}_{\mathcal{W}}\mathbf{v}_{\mathcal{AP}}^{\top}\right]^{\top} \\ &= \left[{}_{\mathcal{W}}p_{\mathcal{A}_x}, {}_{\mathcal{W}}p_{\mathcal{A}_y}, {}_{\mathcal{W}}p_{\mathcal{P}_z}, \theta_{\mathcal{A}_\phi}, \theta_{\mathcal{A}_\theta}, \theta_{\mathcal{A}_\psi}, v_{\mathcal{A}_x}, v_{\mathcal{A}_y}, v_{\mathcal{A}_z}\right]^{\top}. \end{aligned} \tag{14}$$

Given Eqs. (6-10), we can conclude that $\boldsymbol{H}_{\mathcal{AP}_k} = \frac{\partial h}{\partial \boldsymbol{x}} = \boldsymbol{I}_9$. We assume

$$\begin{aligned} &\boldsymbol{V}_{\mathcal{AP}_k} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{R}_{\mathcal{AP}_k}\right), \qquad (15) \\ &\boldsymbol{R}_{\mathcal{AP}_k} = \mathrm{diag}(\sigma_{\mathcal{A}_x}^2, \sigma_{\mathcal{A}_y}^2, \sigma_{\mathcal{P}_z}^2, \sigma_{\mathcal{A}_\phi}^2, \sigma_{\mathcal{A}_\theta}^2, \sigma_{\mathcal{A}_\psi}^2, \sigma_{\mathcal{A}_{v_x}}^2, \sigma_{\mathcal{A}_{v_y}}^2, \sigma_{\mathcal{A}_{v_z}}^2). \end{aligned}$$

Thus, the state update can be described as

$$\begin{aligned} \hat{\boldsymbol{x}}_{\mathcal{AP}_{k|k}} =& \hat{\boldsymbol{x}}_{\mathcal{AP}_{k|k-1}} \\ &+ \boldsymbol{P}_{\mathcal{AP}_{k|k-1}} \boldsymbol{S}_{\mathcal{AP}_k}^{-1} \left(\boldsymbol{z}_{\mathcal{AP}_k} - \hat{\boldsymbol{x}}_{\mathcal{AP}_{k|k-1}}\right). \end{aligned} \tag{16}$$

Therefore, EKF is used to obtain AP odometry by fusing DVL measurements and PS information.

Remark 1. The EKF-based AP measurement model remains full-rank (9), which preserves the filter observability. For the depth measurement, the prediction is triggered by $v_{\mathcal{A}_z}$, while the correction is driven by ${}_{\mathcal{W}}p_{\mathcal{P}_z}$ during the update.

2) VINS-Fusion-based VI Integration: For the VIO framework, the VINS-Fusion is employed to tightly couple the pre-integrated IMU measurements and visual observations provided by [8]. The system processes camera images for feature tracking and IMU data for motion propagation.

IMU measurements are preintegrated between consecutive keyframes and the raw measurement model is

$$\begin{aligned} \tilde{\mathbf{a}}_t &= \mathbf{a}_t + {}_{\mathcal{I}}\mathbf{R}_{\mathcal{W}}\mathbf{g}^{\mathcal{W}} + \mathbf{b}_a + \mathbf{n}_a, \\ \tilde{\boldsymbol{\omega}}_t &= \boldsymbol{\omega}_t + \mathbf{b}_g + \mathbf{n}_g, \end{aligned} \tag{17}$$

where $\tilde{\mathbf{a}}_t$ and $\tilde{\boldsymbol{\omega}}_t$ are measured acceleration and angular velocity, $\mathbf{a}_t$ and $\boldsymbol{\omega}_t$ are real vehicle acceleration and angular velocity, $\mathbf{g}^{\mathcal{W}}$ is gravity, $\mathbf{n}_a$ and $\mathbf{n}_g$ are Gaussian noises [29]. The pre-integration from keyframe b_i to b_j yields

$$\begin{aligned} \boldsymbol{\alpha}_{b_j}^{b_i} &= \iint_{t\in[t_i,t_j]} {}_{b_i}\mathbf{R}_t \left(\hat{\mathbf{a}}_t - \mathbf{b}_{a_t}\right) \mathrm{d}t^2, \\ \boldsymbol{\beta}_{b_j}^{b_i} &= \int_{t\in[t_i,t_j]} {}_{b_i}\mathbf{R}_t \left(\hat{\mathbf{a}}_t - \mathbf{b}_{a_t}\right) \mathrm{d}t, \\ \boldsymbol{\gamma}_{b_j}^{b_i} &= \int_{t\in[t_i,t_j]} \frac{1}{2}\boldsymbol{\Omega}\left(\hat{\boldsymbol{\omega}}_t - \mathbf{b}_{g_t}\right) {}_{b_i}\boldsymbol{\gamma}_t \mathrm{d}t, \end{aligned} \tag{18}$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are increments in position, velocity and rotation, and $\boldsymbol{\Omega}(\cdot)$ is the quaternion kinematics matrix. The discrete implementation uses the midpoint Euler integration, with covariance propagation

$$\boldsymbol{P}_{b_j}^{b_i} = \boldsymbol{F}_{\mathcal{VI}} \boldsymbol{P}_{b_j}^{b_i} \boldsymbol{F}_{\mathcal{VI}}^{\top} + \boldsymbol{G}_{\mathcal{VI}} \boldsymbol{Q}_{\mathcal{VI}} \boldsymbol{G}_{\mathcal{VI}}^{\top}, \tag{19}$$

where $\boldsymbol{F}_{\mathcal{VI}}$ and $\boldsymbol{G}_{\mathcal{VI}}$ are Jacobian matrices of the dynamics and noise, and $\boldsymbol{Q}_{\mathcal{VI}}$ is the noise covariance of VIO framework.

Visual features are tracked using optical flow across frames. For a feature first observed in frame c_i at normalized coordinates $\mathbf{u}_i = [u_i, v_i]^{\top}$, the reprojection in frame $c_j = [u_j, v_j, 1]^{\top}$ is

$$\begin{aligned} c_j =& \frac{1}{z_j^c} \mathbf{K}_{\mathcal{VI}} \mathbf{T}_{\mathcal{VI}_j} \mathbf{R}_{\mathcal{W}} \Big(\frac{{}_{\mathcal{W}}\mathbf{R}_{\mathcal{I}_i}\, {}_{\mathcal{V}}\mathbf{T}_{\mathcal{I}} c_j}{\lambda} \\ &+ {}_{\mathcal{W}}\mathbf{p}_{\mathcal{I}_i} - {}_{\mathcal{W}}\mathbf{p}_{\mathcal{I}_j} \Big), \end{aligned} \tag{20}$$

where $\mathbf{K}_{\mathcal{V}}$ is the camera intrinsic matrix, ${}_{\mathcal{I}}\mathbf{T}_{\mathcal{V}}$ and ${}_{\mathcal{V}}\mathbf{T}_{\mathcal{I}}$ are extrinsics, λ is the inverse depth, and z_j^c is the depth in the camera frame. Therefore, the VINS estimator is achieved by minimizing the cost function of the maximum a posteriori [30]

$$\mathcal{J}_{\mathcal{VI}}(\mathcal{X}_{\mathcal{VI}}) = \|\mathbf{r}_p - \mathbf{H}_p\mathcal{X}_{\mathcal{VI}}\|^2 + \sum_{k\in\mathcal{B}} \left\|\mathbf{r}_{\mathcal{B}}\left(\hat{\mathbf{z}}^{b_k}_{b_{k+1}}, \mathcal{X}_{\mathcal{VI}}\right)\right\|^2_{\boldsymbol{P}^{b_k}_{b_{k+1}}} + \sum_{(l,j)\in\mathcal{C}} \rho\left(\left\|\mathbf{r}_{\mathcal{C}}\left(\hat{\mathbf{z}}^{c_j}_l, \mathcal{X}_{\mathcal{VI}}\right)\right\|^2_{\boldsymbol{P}^{c_j}_l}\right). \tag{21}$$

Among them, $\rho(s) = s$ if $s \leq 1$; otherwise, $\rho(s) = 2\sqrt{s} - 1$. $\{\mathbf{r}_p, \mathbf{H}_p\}$ denotes the prior information from marginalization, $\mathbf{r}_{\mathcal{B}}\left(\hat{\mathbf{z}}^{b_k}_{b_{k+1}}, \mathcal{X}\right)$ represents the IMU residual, and $\mathbf{r}_{\mathcal{C}}\left(\hat{\mathbf{z}}^{c_j}_l, \mathcal{X}\right)$ indicates the visual residual.

3) Factor Graph-based Dual-Framework Optimization: Corresponding to the dual-framework estimators, i.e., the AP estimator and the VINS estimator, the sliding window-based graph optimization is utilized to incorporate the AP factor and the VINS factor. Table I details the algorithmic parameters. Similar to Eq. (13), the objective is to minimize cost $\mathcal{J}(\mathcal{X})$

$$\mathcal{J}(\mathcal{X}) = \sum_{k=1}^{K}\left(\left\|\mathbf{r}^k_{\mathcal{AP}}\right\|^2_{\boldsymbol{\Sigma}^k_{\mathcal{AP}}} + \left\|\mathbf{r}^k_{\mathcal{VI}}\right\|^2_{\boldsymbol{\Sigma}^k_{\mathcal{VI}}}\right), \tag{22}$$

where

$$\begin{aligned}\left\|\mathbf{r}^k_{\mathcal{AP}}\right\|^2_{\boldsymbol{\Sigma}^k_{\mathcal{AP}}} &\triangleq (\mathbf{r}^k_{\mathcal{AP}})^\top(\boldsymbol{\Sigma}^k_{\mathcal{AP}})^{-1}\mathbf{r}^k_{\mathcal{AP}}, \\ \left\|\mathbf{r}^k_{\mathcal{VI}}\right\|^2_{\boldsymbol{\Sigma}^k_{\mathcal{VI}}} &\triangleq (\mathbf{r}^k_{\mathcal{VI}})^\top(\boldsymbol{\Sigma}^k_{\mathcal{VI}})^{-1}\mathbf{r}^k_{\mathcal{VI}},\end{aligned} \tag{23}$$

with factor residuals $\mathbf{r}^k_{\mathcal{AP}}$ and $\mathbf{r}^k_{\mathcal{VI}}$ given by

$$\begin{aligned}\mathbf{r}^k_{\mathcal{AP}} &= \begin{bmatrix}\mathbf{r}_{p,\mathcal{AP}} \\ \mathbf{r}_{q,\mathcal{AP}} \\ \mathbf{r}_{v,\mathcal{AP}}\end{bmatrix} = \begin{bmatrix}{}_{\mathcal{W}}\mathbf{p}^k_{\mathcal{I}} - {}_{\mathcal{W}}\mathbf{p}^k_{\mathcal{AP}} \\ 2\cdot\log\left({}_{\mathcal{W}}\mathbf{q}^k_{\mathcal{I}} \otimes {}_{\mathcal{W}}\mathbf{q}^k_{\mathcal{AP}}{}^{-1}\right) \\ {}_{\mathcal{W}}\mathbf{v}^k_{\mathcal{I}} - {}_{\mathcal{W}}\mathbf{v}^k_{\mathcal{AP}}\end{bmatrix} \in \mathbb{R}^9, \\ \mathbf{r}^k_{\mathcal{VI}} &= \begin{bmatrix}\mathbf{r}_{p,\mathcal{VI}} \\ \mathbf{r}_{q,\mathcal{VI}} \\ \mathbf{r}_{g,\mathcal{VI}} \\ \mathbf{r}_{a,\mathcal{VI}}\end{bmatrix} = \begin{bmatrix}{}_{\mathcal{W}}\mathbf{p}^k_{\mathcal{I}} - {}_{\mathcal{W}}\mathbf{p}^k_{\mathcal{VI}} \\ 2\cdot\log\left({}_{\mathcal{W}}\mathbf{q}^k_{\mathcal{I}} \otimes {}_{\mathcal{W}}\mathbf{q}^k_{\mathcal{VI}}{}^{-1}\right) \\ \mathbf{b}^k_g - \mathbf{b}^k_{g,\mathcal{VI}} \\ \mathbf{b}^k_a - \mathbf{b}^k_{a,\mathcal{VI}}\end{bmatrix} \in \mathbb{R}^{12}.\end{aligned} \tag{24}$$

$\log(\cdot)$ is the logarithm map that converts a quaternion to a minimal rotation vector representation.

For the AP estimation $\boldsymbol{z}_{\mathcal{AP}}$, the state error $\delta\hat{\mathbf{x}}_{\mathcal{AP}}$ is defined as the minimal representation of the perturbation that lives in the tangent space of the state manifold

$$\delta\hat{\mathbf{x}}^k_{\mathcal{AP}} = \left[\delta_{\mathcal{W}}\mathbf{p}^\top_{\mathcal{AP}}, \delta_{\mathcal{W}}\boldsymbol{\theta}^\top_{\mathcal{AP}}, \delta_{\mathcal{W}}\mathbf{v}^\top_{\mathcal{AP}}, \delta\mathbf{b}^\top_g, \delta\mathbf{b}^\top_a\right]^\top, \tag{25}$$

where $\boldsymbol{\theta}_{\mathcal{AP}} = \left[\boldsymbol{\theta}_{\mathcal{AP}_\phi}, \boldsymbol{\theta}_{\mathcal{AP}_\theta}, \boldsymbol{\theta}_{\mathcal{AP}_\psi}\right]^\top \in \mathbb{R}^3$ is the equivalent rotation associated with $\mathbf{q}_{\mathcal{AP}}$. Assume $\mathbf{r}^k_{\mathcal{AP}} \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}^k_{\mathcal{AP}}\right)$ and the covariance matrix of the AP factor is formulated as

$$\boldsymbol{\Sigma}^k_{\mathcal{AP}} = \boldsymbol{J}^k_{\mathcal{AP}} \cdot \boldsymbol{Q}^k_{\mathcal{AP}}\left(\delta\hat{\mathbf{x}}^{k+1}_{\mathcal{AP}} \mid \mathbf{x}^k_{\mathcal{AP}}, \boldsymbol{z}^k_{\mathcal{AP}}\right) \cdot \boldsymbol{J}^k_{\mathcal{AP}}{}^\top, \tag{26}$$

with the Jacobian matrix being

$$\boldsymbol{J}^k_{\mathcal{AP}} = \frac{\partial\mathbf{r}^k_{\mathcal{AP}}}{\partial\mathbf{x}^k} \tag{27}$$

$$= \begin{bmatrix}\boldsymbol{I}_3 & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \frac{\partial\mathbf{r}_{q,\mathcal{AP}}}{\partial_{\mathcal{W}}\mathbf{q}_{\mathcal{I}}} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} \\ \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \boldsymbol{I}_3 & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3}\end{bmatrix} \in \mathbb{R}^{9\times15}. \tag{28}$$

TABLE I: Algorithm and Fusion Parameters for UA-AVIP

Category	Parameter	Value
EKF	Initial State Covariance	diag(0.1, 0.1, ...)
	Initial DVL Noise	0.1 m/s
	Initial PS Noise	0.01 m
	Initial Process Noise	0.01
VINS	Max Tracking Features	250
	Min Distance between Features	15 pixels
	Accelerometer Noise Density	$0.00741\ \mathrm{m/s^2/\sqrt{Hz}}$
	Gyroscope Noise Density	$0.00105\ ^\circ\mathrm{/s/\sqrt{Hz}}$
	Accelerometer Walk Noise	$0.00006\ \mathrm{m/s^2/\sqrt{Hz}}$
	Gyroscope Walk Noise	$0.00004\ ^\circ\mathrm{/s/\sqrt{Hz}}$
FGO	Sliding Window Size	30
	Max Iterations	10

Similarly, for the VINS estimation $\boldsymbol{z}_{\mathcal{VI}} = [{}_{\mathcal{W}}\mathbf{p}^\top_{\mathcal{VI}}, {}_{\mathcal{W}}\mathbf{q}^\top_{\mathcal{VI}}, \mathbf{b}^\top_{g,\mathcal{VI}}, \mathbf{b}^\top_{a,\mathcal{VI}}]^\top$ given $\mathbf{r}^k_{\mathcal{VI}} \sim \mathcal{N}(0, \boldsymbol{\Sigma}^k_{\mathcal{VI}})$, the covariance matrix of the VINS factor is

$$\boldsymbol{\Sigma}^k_{\mathcal{VI}} = \boldsymbol{J}^k_{\mathcal{VI}} \cdot \boldsymbol{Q}^k_{\mathcal{VI}}\left(\delta\hat{\mathbf{x}}^{k+1}_{\mathcal{VI}} \mid \mathbf{x}^k_{\mathcal{VI}}, \boldsymbol{z}^k_{\mathcal{VI}}\right) \cdot \boldsymbol{J}^k_{\mathcal{VI}}{}^\top, \tag{29}$$

with the Jacobian matrix being $\boldsymbol{J}^k_{\mathcal{VI}} = \frac{\partial\mathbf{r}^k_{\mathcal{VI}}}{\partial\delta\hat{\mathbf{x}}^k_{\mathcal{VI}}} \in \mathbb{R}^{12\times15}$.

B. Hierarchical Uncertainty Awareness

1) STFT-based Noise Estimation for Measurement-Level Uncertainty: For EKF-based AP integration, $\boldsymbol{Q}$ quantifies the temporal evolution of measurement-level uncertainty [31]. Due to the time-varying noises and non-stationary signals in underwater environments, the Short-Time Fourier Transform (STFT) provides time-frequency analysis, which enables the detection and quantification of real-time noise conditions [32]. In this approach, sensor signals are divided into short and overlapping windows. A Fourier transform is performed on each segment to capture dynamic spectral characteristics [33].

For position measurements, at time t and frequency ν, the STFT can be expressed as

$$\mathbf{X}_p(t,\nu) = \int_{-\infty}^{\infty} \mathbf{p}_{\mathcal{AP}}(\tau)w(\tau - t)e^{-j2\pi\nu\tau}\mathrm{d}\tau, \tag{30}$$

where τ is the integration variable representing the time within the signal, $w(\tau - t)$ represents a window function centered around t. For orientation measurements $\mathbf{q}_{\mathcal{AP}}$ and equivalent rotation $\boldsymbol{\theta}_{\mathcal{AP}}$, the STFT can be computed as follows

$$\mathbf{X}_o(t,\nu) = \int_{-\infty}^{\infty} \boldsymbol{\theta}_{\mathcal{AP}}(\tau)w(\tau - t)e^{-j2\pi\nu\tau}\mathrm{d}\tau. \tag{31}$$

Based on the STFT results, the power spectral density could be calculated to describe the frequency content of the noise

$$\begin{aligned}\mathbf{D}_p(\nu) &= \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} |\mathbf{X}_p(t,\nu)|^2\mathrm{d}t, \\ \mathbf{D}_o(\nu) &= \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} |\mathbf{X}_o(t,\nu)|^2\mathrm{d}t.\end{aligned} \tag{32}$$

where $|\cdot|^2$ denotes the element-wise squared magnitude. Then, the noise variances in the frequency domain can be estimated.

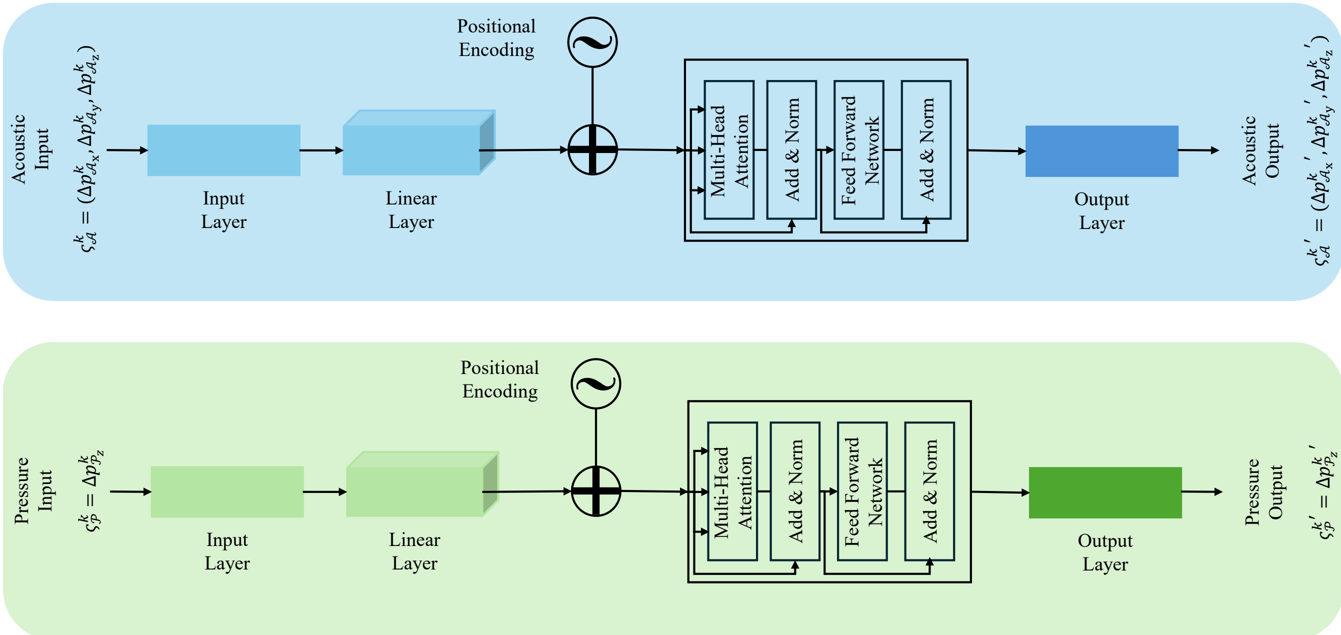


Fig. 3: The Transformer architecture for acoustic and pressure refinements consisting of the input layer, the linear layer, the Transformer encoder and the output layer. Two separate Transformer models are designed for acoustic and pressure estimations.

Assume that sensor measurements follow a white-noise process. The noise variances can be approximated as

$$\boldsymbol{\sigma}_p^2 = \int_{-\infty}^{\infty} \mathbf{D}_p(\nu)\mathrm{d}\nu, \qquad \boldsymbol{\sigma}_o^2 = \int_{-\infty}^{\infty} \mathbf{D}_o(\nu)\mathrm{d}\nu, \tag{33}$$

where $\boldsymbol{\sigma}_p^2 = \left[\sigma_{p,1}^2, \sigma_{p,2}^2, \sigma_{p,3}^2\right]^\top$ and similarly for $\boldsymbol{\sigma}_o^2$. Assume that the noises affecting position and orientation are uncorrelated. Specifically, the variances are incorporated into the diagonal elements of the process noise covariance matrix

$$\boldsymbol{Q} = \mathrm{diag}(\sigma_{p,1}^2, \sigma_{p,2}^2, \sigma_{p,3}^2, \sigma_{o,1}^2, \sigma_{o,2}^2, \sigma_{o,3}^2). \tag{34}$$

For practical implementation, discrete sensor signals are processed using the STFT with a Hann window of 16 samples for the DVL and 64 samples for the PS, with a 50% overlap. Prior to the STFT, each signal segment is demeaned to prevent constant offsets or slow baseline drifts from biasing the spectral power estimates. Given the sampling frequencies of 9 Hz for the DVL and 67 Hz for the PS, these parameters are empirically selected to provide an optimal balance between time-domain responsiveness to dynamic underwater noise and sufficient frequency-domain resolution for variance estimation.

2) Transformer-based Localization for Module-Level Uncertainty: Two Transformer-based models are employed to refine the acoustic and pressure position estimates [34]. The model leverages the capabilities of Transformer architectures to capture temporal patterns and overcome estimator uncertainties. As shown in Fig. 3, the displacements of the DVL DR estimation and the depth displacements estimated by the PS, whenever observed, are used as input. The inputs are processed to generate refined and corrected displacements as the model output. The input at the k-th sample point is defined as

$$\boldsymbol{\varsigma}_{\mathcal{A}}^k = \Delta \mathbf{p}_{\mathcal{A}}^k \in \mathbb{R}^{M\times 3}, \quad \boldsymbol{\varsigma}_{\mathcal{P}}^k = \Delta \mathbf{p}_{\mathcal{P}_z}^k \in \mathbb{R}^{N\times 1}, \tag{35}$$

where $\Delta \mathbf{p}$ is the displacement. M, N are the sequence lengths.

During training, the ground truth displacement labels the acoustic model in three directions, denoted as $\mathbf{y}_{\mathcal{A}}^{\mathrm{gt}} \in \mathbb{R}^{M\times 3}$, and provides supervision for the pressure model in the z-direction, denoted as $\mathbf{y}_{\mathcal{P}}^{\mathrm{gt}} \in \mathbb{R}^{N\times 1}$.

The Transformer architecture comprises an input linear layer, a multi-layer Transformer encoder, and an output linear layer. The input linear layer projects the input data into a higher-dimensional space for processing, as

$$\mathbf{z}_0 = \mathbf{W}_i \boldsymbol{\varsigma}^k + \mathbf{b}_i. \tag{36}$$

$\mathbf{W}_i$ is the learnable weight matrix of the input layer and $\mathbf{b}_i$ denotes the learnable bias of the input projection layer. For acoustics, $\mathbf{W}_i^{\mathcal{A}} \in \mathbb{R}^{3\times d_m}$ with d_m being the hidden dimension of the Transformer model. Similarly, $\mathbf{W}_i^{\mathcal{P}} \in \mathbb{R}^{1\times d_m}$. The Transformer encoder in this work consists of three layers, each with four attention heads. It is responsible for capturing temporal dependencies within the input sequence

$$\mathbf{z}_l = \mathcal{T}_l\left(\mathbf{z}_{l-1}\right), \tag{37}$$

where $\mathcal{T}_l(\cdot)$ denotes the l-th Transformer encoder layer. The output linear layer then maps encoded features back to the original displacement space as

$$\hat{\mathbf{y}} = \mathbf{W}_o \mathbf{z}_L + \mathbf{b}_o, \quad L = 3, \tag{38}$$

where $\mathbf{z}_L$ is the output from the last Transformer layer, $\mathbf{W}_o$ and $\mathbf{b}_o$ are the weights and biases of the output layer.

To optimize the model parameters, the mean squared error loss function is used to quantify the average squared difference

$$MSE_m = \begin{cases} \frac{1}{M}\sum_{i=1}^{M}\left\|\hat{\mathbf{y}}_{\mathcal{A},i} - \mathbf{y}^{\mathrm{gt}}_{\mathcal{A},i}\right\|_2^2 & \text{if } m = \mathcal{A}, \\ \frac{1}{N}\sum_{j=1}^{N}\left\|\hat{\mathbf{y}}_{\mathcal{P},j} - \mathbf{y}^{\mathrm{gt}}_{\mathcal{P},j}\right\|_2^2 & \text{if } m = \mathcal{P}. \end{cases} \tag{39}$$

During inference, the refined estimations serve as enhanced AP estimators and can subsequently be integrated into the EKF framework for the adaptive AP fusion.

3) Adaptive Weighting Strategy for System-Level Uncertainty: In the FGO framework, an adaptive weighting strategy is employed to modulate the estimator contributions based on their statistical uncertainties. This approach leverages the chi-square distribution to dynamically adjust the weightings at the system level. The chi-square threshold is computed as

$$\chi^2_{c,df} = Q^{-1}_{\chi^2}(c, df), \tag{40}$$

where c is the desired confidence level, df is the degrees of freedom for measurements, and $Q^{-1}_{\chi^2}(\cdot)$ denotes the inverse cumulative distribution function for the chi-square distribution.

Then, we determine the adaptive standard deviation σ, as

$$\sigma = \sigma_0 \cdot \max\left(1, \sqrt{\frac{\chi^2}{\chi^2_{c,df}}}\right), \tag{41}$$

where σ_0 is the nominal standard deviation, $\chi^2 = \mathbf{r}^\top(\boldsymbol{\Sigma})^{-1}\mathbf{r} = \mathbf{r}^\top(\sigma^2\boldsymbol{I})^{-1}\mathbf{r}$ is the chi-square statistic of $\mathbf{r}$ in Eq. (22). It ensures higher uncertainty of the estimator contributes less to the dual-framework optimization.

In the event of total visual dropout, such as the feature tracking failure or initialization errors, $\mathbf{r}_{\mathcal{VI}}$ is excluded from $\mathcal{J}(\mathcal{X})$ in Eq. (22), ensuring the FGO relies on the remaining $\mathbf{r}_{\mathcal{AP}}$ and $\boldsymbol{\Sigma}_{\mathcal{AP}}$ until visual tracking is restored. The pseudocode of the adaptive weighting strategy is illustrated in Algorithm 1.

IV. Experiments

This section begins with the experimental setup and the introduction of the remotely operated vehicle (ROV) platform. Then we describe the training dataset and analyze the training results. Thereafter, datasets captured in real-world scenarios including a lab pool, lakes, and oceans, are utilized to evaluate the performance of the proposed method. Specifically, two ablation studies are conducted to highlight the contribution of different UA levels and different sensor configurations. Finally, quantitative metrics are presented regarding the running time and memory of individual modules on the embedded platform.

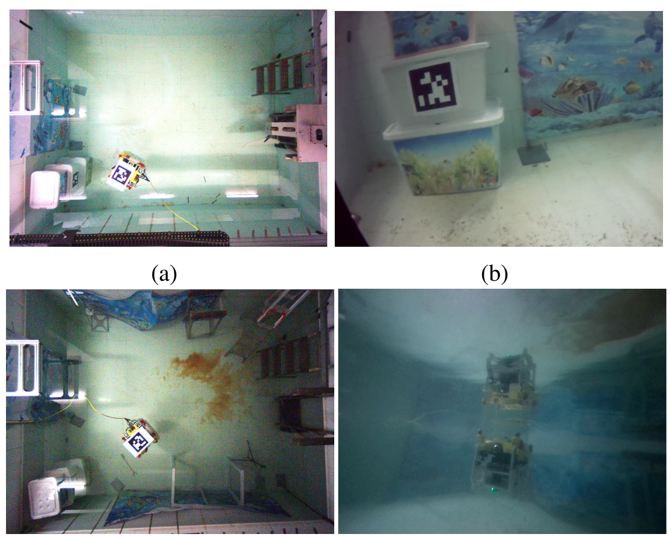
A. Experimental Setup and ROV Platform

To validate the effectiveness and superiority of the proposed system in the lab pool, an ROV, findROV H20 pro, integrated with a stereo depth camera, a DVL, an external IMU, and a PS, is utilized as our experimental platform. Table II specifies the specifications of the onboard sensing, the ROV, and the processor. To account for the scattering and absorption in underwater environments, the intrinsic parameters of the stereo camera were calibrated underwater for practical applications. The laboratory pool measures $5\ \mathrm{m}\times 4\ \mathrm{m}\times 2\ \mathrm{m}$ and is shown in Fig. 4. Several storage boxes and stepladders are placed within the pool for feature extraction. To provide the ground truth, we set a global camera (1292×946 pixel, 17 Hz) above the pool, following [6]. The global camera detects and tracks the AprilTag which is rigidly attached to the top of the ROV. The AprilTag remains above the water surface for all experiments to ensure reliable tracking. This setup provides a reliable and

Algorithm 1 Adaptive Weighted FGO-based AVIP Integration

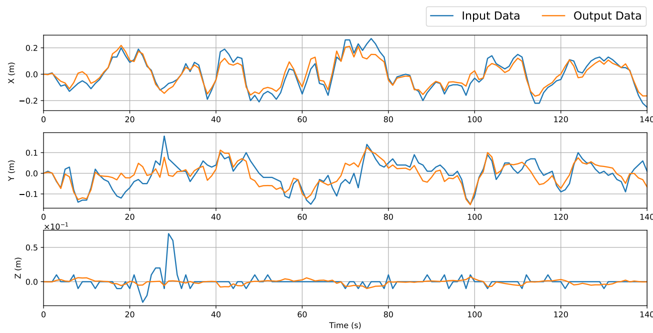
Input: Estimations from AP and VINS factor $\boldsymbol{z}_{\mathcal{AP}}$, $\boldsymbol{z}_{\mathcal{VI}}$
Variance estimates for each factor $\sigma_{\mathcal{AP}}$, $\sigma_{\mathcal{VI}}$
Confidence level $c_{\mathcal{AP}}$, $c_{\mathcal{VI}}$
Degrees of freedom $df_{\mathcal{AP}}$, $df_{\mathcal{VI}}$
Output: Optimized state estimates $\mathcal{X}$

Initialize factor graph $G \leftarrow \emptyset$
for each estimation k **do**
  $\chi^2_{\mathcal{AP}} \leftarrow \mathbf{r}^\top_{\mathcal{AP}}(\boldsymbol{\Sigma}_{\mathcal{AP}})^{-1}\mathbf{r}_{\mathcal{AP}}$
  $\sigma^k_{\mathcal{AP}} \leftarrow \sigma_{\mathcal{AP}_0} \cdot \max\left(1, \sqrt{\chi^2_{\mathcal{AP}}/Q^{-1}_{\chi^2}(c_{\mathcal{AP}}, df_{\mathcal{AP}})}\right)$
  $\boldsymbol{\Sigma}^k_{\mathcal{AP}} \leftarrow {\sigma^k_{\mathcal{AP}}}^2\boldsymbol{I}$
  $G \leftarrow G \cup \{\mathbf{r}^k_{\mathcal{AP}}, \boldsymbol{\Sigma}^k_{\mathcal{AP}}\}$
  if estimation is from VINS **then**
    $\chi^2_{\mathcal{VI}} \leftarrow \mathbf{r}^\top_{\mathcal{VI}}(\boldsymbol{\Sigma}_{\mathcal{VI}})^{-1}\mathbf{r}_{\mathcal{VI}}$
    $\sigma^k_{\mathcal{VI}} \leftarrow \sigma_{\mathcal{VI}_0} \cdot \max\left(1, \sqrt{\chi^2_{\mathcal{VI}}/Q^{-1}_{\chi^2}(c_{\mathcal{VI}}, df_{\mathcal{VI}})}\right)$
    $\boldsymbol{\Sigma}^k_{\mathcal{VI}} \leftarrow {\sigma^k_{\mathcal{VI}}}^2\boldsymbol{I}$
    $G \leftarrow G \cup \{\mathbf{r}^k_{\mathcal{VI}}, \boldsymbol{\Sigma}^k_{\mathcal{VI}}\}$
  else
    continue
  end if
end for
$\arg\min_{\mathcal{X}} \sum_{k=1}^{K}\left(\left\|\mathbf{r}^k_{\mathcal{AP}}(\mathcal{X})\right\|^2_{\boldsymbol{\Sigma}^k_{\mathcal{AP}}} + \left\|\mathbf{r}^k_{\mathcal{VI}}(\mathcal{X})\right\|^2_{\boldsymbol{\Sigma}^k_{\mathcal{VI}}}\right)$
return $\mathcal{X}$

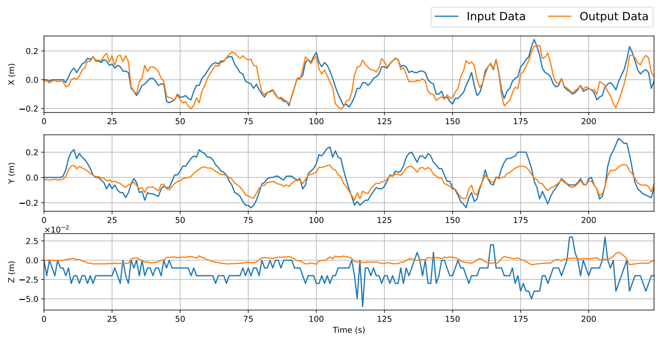


(a) (b) (c) (d)

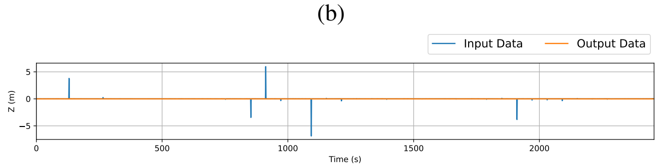
Fig. 4: Experimental setups in the laboratory pool environment. (a) Overhead view of Pool 01. (b) Underwater view of Pool 01. (c) Overhead view of Pool 02 and Pool 03. (d) Underwater view of Pool 02 and Pool 03.



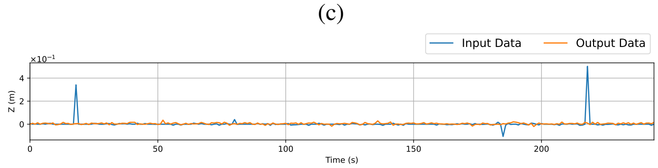
(a)



(b)



(c)



(d)

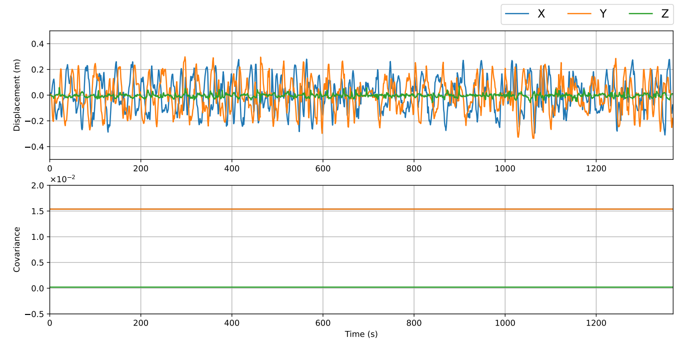
Fig. 5: Results of Transformer training. (a) Comparison between raw and refined DVL DR for Pool 01. (b) Comparison between raw and refined DVL DR for Pool 02. (c) Comparison between raw and refined PS for Pool 01. (d) Comparison between raw and refined PS for Pool 02.

TABLE II: Hardware Specifications for UA-AVIP

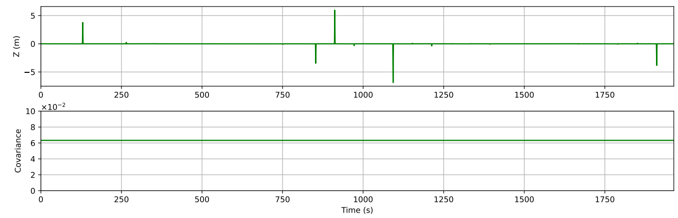
Sensor	Model	Specification
DVL	Water Linked A50	9 Hz
IMU	3DM-CV5-AHRS	67 Hz
PS	Blue Robotics Bar30	62 Hz
Camera	RealSense D435i	30 Hz 640 × 480 pixel
Vehicle	FindROV H20 pro	8 thrusters
Processor	Intel NUC with Core i5	1.6 GHz 15.5 GB RAM

accurate benchmark for performance evaluation [35], [36].

Table III describes the datasets obtained from different sources. Pool 01-03 are datasets we collected in the aforementioned lab pool. Lake/Ocean 01-06 are publicly available datasets captured from natural lakes and oceans [38].



(a)



(b)

Fig. 6: Evaluation of datasets for training. (a) Input data and corresponding covariance of DVL DR. (b) Input data and corresponding covariance of PS.

TABLE III: Experimental Datasets for Performance Evaluation

Datasets	Path Length (m)	Max Speed (m/s)	Min Speed (m/s)
Pool 01	24.8289	0.7805	0.0229
Pool 02	12.4366	0.4686	0.0031
Pool 03	11.1678	0.7235	0.0317
Lake/Ocean 01	423.4343	14.7068	0.0014
Lake/Ocean 02	289.5209	16.0242	0.0000
Lake/Ocean 03	207.9788	9.1532	0.0317
Lake/Ocean 04	108.0718	11.1258	0.2358
Lake/Ocean 05	69.9877	8.5199	0.0416
Lake/Ocean 06	11.5323	5.7902	0.0000

TABLE IV: Competing Methods for Underwater Localization

Method	Sensor				UA	Learning
	DVL	Camera	IMU	PS		
DVL DR	✓	×	✓	×	×	×
EKF [37]	✓	✓	✓	✓	×	×
FGO [21]	✓	✓	✓	✓	×	×
SVIn2 [14]	×	✓	✓	✓	×	×
VINS-Fusion [8]	×	✓	✓	×	×	×
AVIP	✓	✓	✓	✓	×	×
UA-AVIP	✓	✓	✓	✓	✓	✓

B. Dataset Collection and Neural Network Training

As detailed in Section III-B2, the Transformer was utilized to refine the estimation captured from the DVL and the PS. The model was implemented using Python on a Linux-based operating system and ROS Noetic. Training was conducted on a desktop PC equipped with a 32-core i9-13900K processor. The datasets for training were obtained by manually operating the ROV in the pool exclusively under the scenario in Fig. 4(c).

To improve generalization and mitigate overfitting, training sequences with aggressive maneuvering were augmented, exposing the deep learning model to dynamic conditions. In

TABLE V: Performance Evaluation on Self-Collected Real-World Underwater Datasets (Lowest in **Bold**)

Method			DVL DR	EKF [37]	FGO [21]	SVIn2 [14]	VINS-Fusion [6], [8]	AVIP	UA-AVIP
Pool 01	ATE (m)	RMSE	2.8906	1.3053	0.5499	0.4165	0.1743	0.0928	**0.0439**
		MAE	2.8892	1.1492	0.5338	0.3836	0.1475	0.0789	**0.0381**
	RTE (m)	RMSE	0.1842	0.0645	0.0297	0.6112	0.0720	0.0400	**0.0200**
		MAE	0.1195	0.0332	0.0214	0.4151	0.0370	0.0211	**0.0107**
Pool 02	ATE (m)	RMSE	0.4433	0.3968	0.2702	0.0834	0.2237	0.0769	**0.0398**
		MAE	0.3944	0.3855	0.2642	0.0738	0.1779	0.0694	**0.0355**
	RTE (m)	RMSE	0.0384	0.0563	0.0743	0.2576	0.0711	0.0788	**0.0106**
		MAE	0.0294	0.0370	0.0484	0.1341	0.0492	0.0514	**0.0071**
Pool 03	ATE (m)	RMSE	2.5678	1.1675	1.1066	0.2348	0.3130	0.2640	**0.1751**
		MAE	2.5651	1.0124	1.0592	0.2093	0.2919	0.2446	**0.1661**
	RTE (m)	RMSE	0.1258	0.0943	0.1232	0.0965	0.0942	**0.0887**	0.0899
		MAE	0.0944	0.0618	0.0727	0.0948	0.0617	0.0610	**0.0569**

TABLE VI: Performance Evaluation on Open-Source Real-World Underwater Datasets (Lowest in **Bold**)

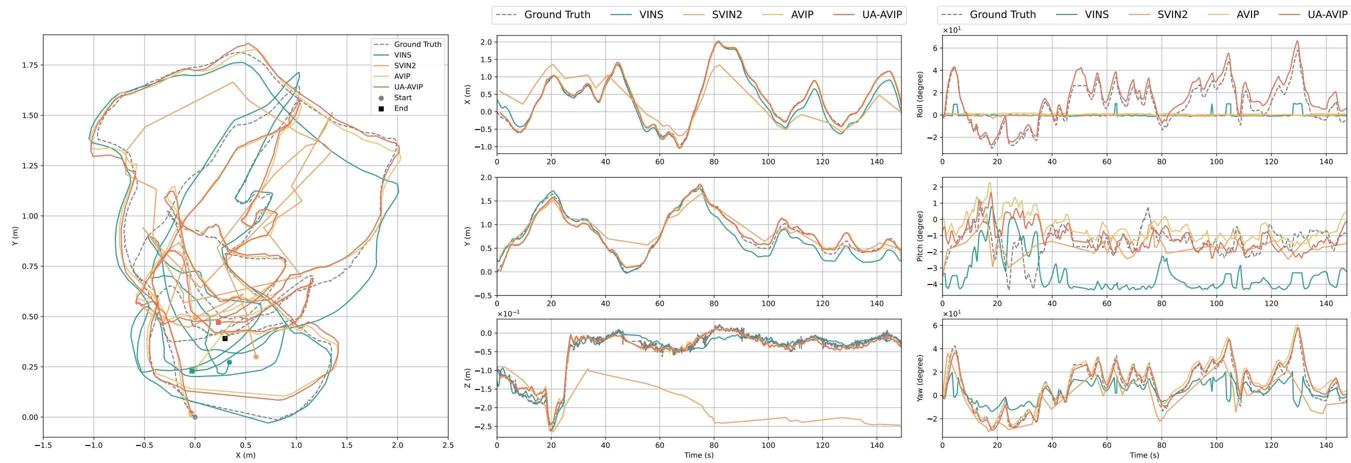
Method			DVL DR	EKF [37]	FGO [21]	SVIn2 [14]	VINS-Fusion [6], [8]	AVIP	UA-AVIP
Lake/Ocean 01	ATE (m)	RMSE	9.8261	9.8332	9.7886	N/A	N/A	3.3336	**2.9138**
		MAE	9.4723	9.4791	9.4299	N/A	N/A	2.6268	**2.5144**
	RTE (m)	RMSE	0.3341	0.1613	0.1612	N/A	N/A	0.1629	**0.1202**
		MAE	0.1953	0.0685	0.0684	N/A	N/A	0.0686	**0.0398**
Lake/Ocean 02	ATE (m)	RMSE	7.7925	7.4563	7.7863	N/A	N/A	6.9088	**2.1574**
		MAE	7.5628	7.2529	7.5560	N/A	N/A	6.7269	**1.7614**
	RTE (m)	RMSE	0.4762	0.3374	0.1672	N/A	N/A	0.2151	**0.1374**
		MAE	0.3617	0.3584	0.0935	N/A	N/A	0.1098	**0.0584**
Lake/Ocean 03	ATE (m)	RMSE	5.9316	5.5805	5.9291	N/A	N/A	4.7927	**1.2138**
		MAE	5.2556	4.8798	5.2533	N/A	N/A	4.2505	**1.1857**
	RTE (m)	RMSE	0.1442	0.1827	0.3010	N/A	N/A	0.1246	**0.0738**
		MAE	0.0986	0.1423	0.2659	N/A	N/A	0.0993	**0.0463**
Lake/Ocean 04	ATE (m)	RMSE	1.8976	1.8979	1.7801	N/A	N/A	1.0070	**0.9905**
		MAE	1.6537	1.6539	1.5583	N/A	N/A	0.9340	**0.9186**
	RTE (m)	RMSE	0.1787	0.1786	0.1778	N/A	N/A	0.3918	**0.1270**
		MAE	0.0678	0.0677	0.0677	N/A	N/A	0.2225	**0.0287**
Lake/Ocean 05	ATE (m)	RMSE	4.2149	4.0927	7.3381	N/A	N/A	4.0859	**1.2447**
		MAE	3.8163	3.7126	6.7256	N/A	N/A	3.7571	**1.1726**
	RTE (m)	RMSE	0.2040	0.1059	0.2466	N/A	N/A	0.1901	**0.1033**
		MAE	0.1184	0.0682	0.1905	N/A	N/A	0.1146	**0.0671**
Lake/Ocean 06	ATE (m)	RMSE	1.4960	1.4951	1.4959	N/A	N/A	0.6140	**0.2000**
		MAE	1.4313	1.4304	1.4312	N/A	N/A	0.5530	**0.1879**
	RTE (m)	RMSE	0.0212	0.0212	0.0216	N/A	N/A	0.0230	**0.0211**
		MAE	0.0177	0.0175	0.0176	N/A	N/A	0.0133	**0.0077**

addition, we included dropout layers and zero-mean Gaussian noises during the training to reduce feature co-adaptation. The employed Transformer-based model is lightweight, with a hidden dimension of 64, a feed-forward dimension of 256, 4 attention heads, and 3 encoder layers. The network was trained using the Adam optimizer with a learning rate of 0.001, a dropout rate of 0.1, a batch size of 16, and 50 training epochs. The training dataset consists of 3826 samples, each containing displacement measurements paired with the corresponding ground truth displacements. Consequently, the input dimension was defined as 3 for the DVL DR refinement with an inference frequency of approximately 1 Hz and 1 for the PS at an inference frequency of approximately 17 Hz. In addition, the training and evaluation data were separated at the sequence level to avoid temporal leakage between adjacent samples. Fig. 5 displays the results of the trained model for two representative sequences, Pool 01 and Pool 02. Specifically, Figs. 5(a) and 5(b) show that the learning-based method helps align acoustic estimations with learned temporal patterns. Figs. 5(c) and 5(d) show that the model effectively rejects outliers from the PS. Then, the refined DVL and PS are intended for fusion into the EKF-based AP integration.

Remark 2. Fig. 6 shows the displacements and covariances of the collected datasets for training, from which we can clearly find some outliers. Outliers in the raw data of the PS can significantly affect the performance of data fusion, which is the motivation for employing the Transformer in this study.

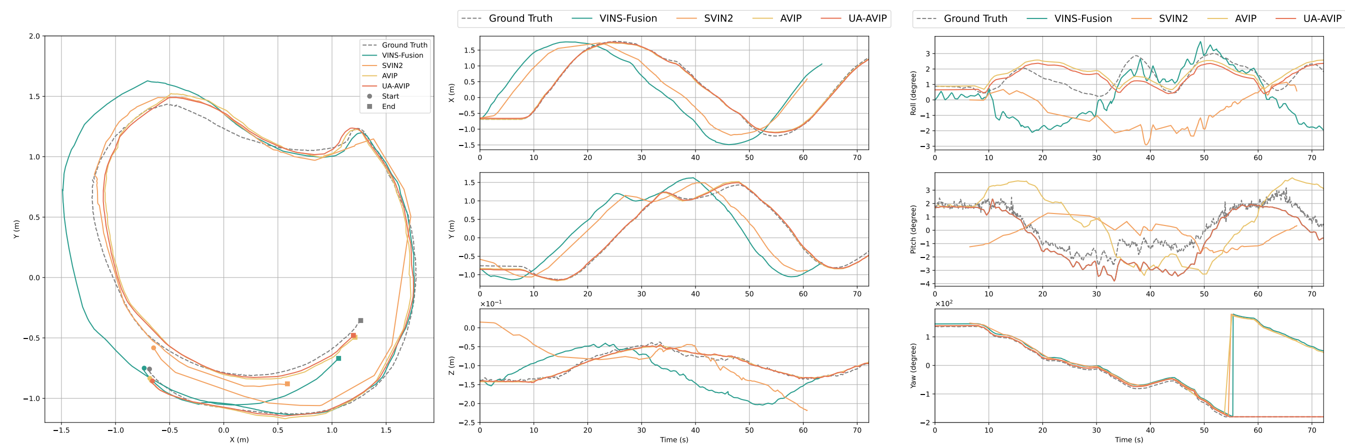
C. Performance Evaluation and Comparison

To evaluate the performance across different conditions, we conduct a comparative study using the nine datasets in Table III. The performance of the UA-AVIP is compared against five practical and state-of-the-art (SOTA) methods in Table IV, including the onboard DVL DR, a baseline EKF [37], a baseline FGO [21], an underwater SLAM system SVIn2 [14], and the multi-modal state estimator VINS-Fusion [8]. Additionally, the AVIP system without the UA technique is also evaluated. The root mean square error (RMSE) and the mean absolute error (MAE) of the absolute trajectory error (ATE) and the relative trajectory error (RTE) are computed separately [39]. ATE



(a) (b) (c)

Fig. 7: Performance evaluation on Pool 01. (a) Trajectory comparison between our proposed and SOTA systems. (b) Position estimation of our proposed and SOTA systems. (c) Orientation estimation of our proposed and SOTA systems.



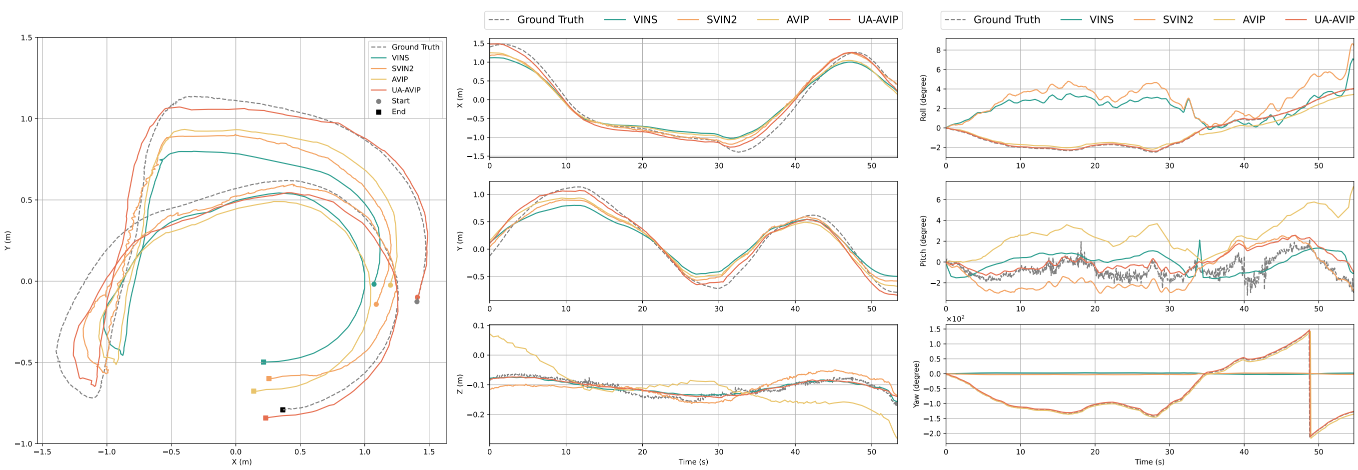
(a) (b) (c)

Fig. 8: Performance evaluation on Pool 02. (a) Trajectory comparison between our proposed and SOTA systems. (b) Position estimation of our proposed and SOTA systems. (c) Orientation estimation of our proposed and SOTA systems.

is employed to evaluate the global consistency of estimated trajectories, while RTE quantifies the local accuracy [40].

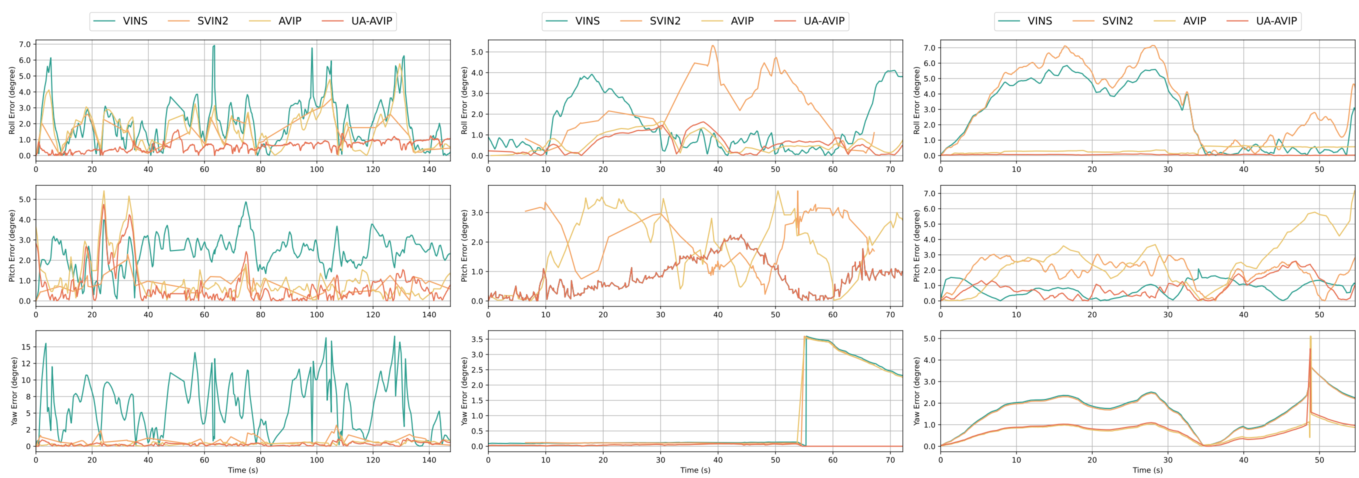
1) Performance Evaluation on Datasets Pool 01-03: For experiments in the lab pool, we develop two distinct real-world scenarios, as depicted in Figs. 4(a) and 4(c). Table V presents the quantitative performance of the evaluated localization methods. The proposed UA-AVIP, which incorporates UA, achieves the lowest ATE in all cases. It demonstrates the superior accuracy in trajectory estimation. The AVIP system (Ours w/o UA) exhibits a reduction in ATE and outperforms existing SOTA methods. The results underscore the effectiveness of the proposed dual-framework multi-modal fusion. Among the evaluated methods, DVL DR exhibits the highest ATE, followed by EKF and FGO. While SVIn2 and VINS-Fusion are competitive in pool sequences, UA-AVIP achieves RMSE ATE reductions ranging from 25.43% to 74.81%. Furthermore, UA-AVIP exhibits the lowest RTE across nearly all metrics and maintains the local consistency. Compared to SVIn2 which exhibits substantial local drifts, UA-AVIP effectively suppresses RMSE RTE from 6.84% to 96.73%.

Qualitative estimations are depicted in Figs. 7-9. Figs. 7(a), 8(a), and 9(a) present the trajectories estimated on the pool sequences. Among them, the red curve represents the trajectory obtained from the proposed UA-AVIP system, which aligns with the ground truth trajectory. As observed in Figs. 7(b), 8(b), and 9(b), SVIn2 exhibits a larger position error in the z-direction compared to the x and y directions. On the other hand, the position errors of VINS-Fusion accumulate over time across all directions. The results illustrated in Figs. 7(c), 8(c), and 9(c) indicate that UA-AVIP provides robust orientation estimation. Specifically, for the yaw angle, Fig. 10 demonstrates our estimation aligns closely with the ground truth. Within the proposed system, yaw drift is actively managed through multi-modal constraints with complementary orientation corrections, alongside hierarchical UA, which ensures orientation noise estimation and adaptive



(a) (b) (c)

Fig. 9: Performance evaluation on Pool 03. (a) Trajectory comparison between our proposed and SOTA systems. (b) Position estimation of our proposed and SOTA systems. (c) Orientation estimation of our proposed and SOTA systems.



(a) (b) (c)

Fig. 10: Orientation evaluation over time on Datasets Pool 01-03. (a) Pool 01. (b) Pool 02. (c) Pool 03.

factor weighting. Notably, the yaw estimation exhibits higher accuracy than the roll and pitch angles. This discrepancy may be attributed to a minor physical misalignment between the AprilTag and the vehicle frame during installation. Despite this systematic bias, UA-AVIP precisely tracks dynamic motion trends. As shown in Figs. 7(c) and 10(a), even roll angles vary greatly, UA-AVIP maintains high consistency with the ground truth and outperforms competing methods in capturing rapid orientation changes.

2) Performance Evaluation on Datasets Lake/Ocean 01-03: Table VI reports the performance of the methods evaluated in six open-source lake and ocean datasets. In particular, both SVIn2 and VINS-Fusion fail to generate valid trajectories in the lake/ocean scenarios due to the lack of visual measurements in natural sequences. In contrast, the proposed UA-AVIP demonstrates robust applicability with the lowest localization error across all sequences. Additionally, AVIP significantly improves localization performance, particularly in the challenging Lake/Ocean 01 sequence, achieving an RMSE ATE of 3.3336 m and an MAE RTE of 0.0686 m over a trajectory length of 423.4343 m. Traditional approaches, such as DR, EKF, and FGO, exhibit higher RMSE values ranging from 1.4951 m to 9.8332 m for ATE and from 0.0212 m to 0.4762 m for RTE. For example, on the Lake/Ocean 03 sequence, UA-AVIP achieves an RMSE ATE of 1.2138 m compared to 5.9291 m for FGO, representing an improvement of around 79.53%, and an MAE RTE of 0.0463 m against 0.1423 m for EKF, a 67.46% reduction in error. The results underscore the superiority of UA-AVIP in complex open-water environments, where existing visual SLAM systems fail and conventional methods are prone to significant drift.

D. Ablation Study

In UA-AVIP, multi-level uncertainty is first taken into account during the integration of acoustic, visual, inertial, and pressure data. To validate the performance of the proposed

TABLE VII: Ablation Study for Different Sensor Setups (Lowest in **Bold**)

Sensor Setup				Pool 01				Pool 02				Pool 03			
				ATE (m)		RTE (m)		ATE (m)		RTE (m)		ATE (m)		RTE (m)	
IMU	Camera	DVL	PS	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
✓	✓	×	×	0.1743	0.1475	0.0720	0.0370	0.2237	0.1779	0.0711	0.0492	0.3130	0.2919	0.0942	0.0617
✓	✓	×	✓	0.1398	0.1190	0.2280	0.0201	0.1391	0.1091	0.0719	0.0463	0.1918	0.1783	0.0988	0.0623
✓	×	✓	✓	0.0451	0.0390	0.0698	0.0480	0.0400	0.0357	0.1166	0.0880	0.1995	0.1854	0.2432	0.1894
✓	×	✓	×	0.0732	0.0635	0.0548	0.0252	0.1114	0.0999	0.1730	0.1306	0.2091	0.1950	0.0995	0.0607
✓	✓	✓	×	0.0465	0.0409	0.0399	0.0211	0.1087	0.0966	0.0664	0.0432	0.2078	0.1955	0.0958	0.0657
✓	✓	✓	✓	**0.0439**	**0.0381**	**0.0200**	**0.0107**	**0.0398**	**0.0355**	**0.0106**	**0.0071**	**0.1751**	**0.1661**	**0.0899**	**0.0569**

TABLE VIII: Ablation Study for Different Levels of Uncertainty Awareness (Lowest in **Bold**)

Uncertainty Awareness			Pool 01				Pool 02				Pool 03			
M.	Mod.	Sys.	ATE (m)		RTE (m)		ATE (m)		RTE (m)		ATE (m)		RTE (m)	
Level	Level	Level	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
×	×	×	0.0928	0.0789	0.0400	0.0211	0.0769	0.0694	0.0788	0.0514	0.2640	0.2446	0.0887	0.0610
✓	×	×	0.0696	0.0596	0.0316	0.0127	0.0721	0.0647	0.0714	0.0502	0.1998	0.1872	0.0833	0.0606
×	✓	✓	0.0677	0.0576	0.0294	0.0178	0.0423	0.0377	0.0356	0.0126	0.2114	0.1969	0.0855	0.0639
✓	×	✓	0.0662	0.0568	0.0290	0.0171	0.0592	0.0530	0.0443	0.0184	0.1990	0.1872	**0.0814**	**0.0519**
✓	✓	×	0.0646	0.0569	0.0209	0.0213	0.0520	0.0404	0.0375	0.0163	0.1988	0.1865	0.0959	0.0659
✓	✓	✓	**0.0439**	**0.0381**	**0.0200**	**0.0107**	**0.0398**	**0.0355**	**0.0106**	**0.0071**	**0.1751**	**0.1661**	0.0899	0.0569

Abbreviations: M. Level (Measurement Level), Mod. Level (Module Level), Sys. Level (System Level).

TABLE IX: Ablation Study for Transformer Refinements

Transformer	Pool 01		Pool 02		Pool 03	
	ATE (m)		ATE (m)		ATE (m)	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
OFF	0.0662	0.0568	0.0592	0.0530	0.1990	0.1872
ON	0.0439	0.0381	0.0398	0.0355	0.1751	0.1661

algorithm, three ablation studies are performed on (1) the sensor configurations, (2) the hierarchical uncertainties and (3) the Transformer contribution. Specifically, the first compares the contributions of different sensor setups, while the second involves the removal of certain UA sub-models. The third separates the contribution of the refinement module from other UA components. Noted that Lake/Ocean datasets only include DVL and IMU measurements, which disable VI estimations. Therefore, ablation studies are performed using Pool datasets.

1) Localization with Different Sensor Setups: We study how different sensor setups affect the system performance. To isolate the effect from uncertainty, in this section, all levels of uncertainty-aware modules are fully integrated into the localization system. Table VII presents a comparative analysis of localization accuracy with and without the integration of visual odometry, acoustic odometry, and pressure information.

The results indicate that the combination of all sensors yields the lowest error across all sequences. This suggests that the integration of multiple modalities significantly enhances the localization accuracy. The inclusion of the DVL, particularly when paired with the IMU, substantially reduces localization errors compared to the absence of the DVL in most cases. Furthermore, the integration of the water-pressure results in a modest enhancement in performance.

2) Localization with Different Levels of Uncertainty Awareness: We investigate the impact of different levels of uncertainty on pose estimation accuracy using the proposed method. Specifically, sub-models at the measurement, module, and system levels are independently evaluated.

For quantitative analysis, the RMSE and MAE of the ATE are computed. As shown in Table VIII, each level independently contributes to the performance improvements, whereas the combination of all uncertainty levels results in the lowest ATE. It shows that incorporating hierarchical UA is crucial for underwater localization. In addition, there is a notable reduction in errors when measurement and module-level awareness are both active. It indicates that the Fourier transform and deep learning enhance performance when combined.

3) Effectiveness of the Transformer-Based Refinement: To explicitly validate the contribution of the module-level UA, we conduct a targeted control comparison to isolate the Transformer-based AP refinement. In this ablation study, the localization performance is compared with the refinement module enabled (ON) and disabled (OFF). When disabled, the proposed system relies purely on raw DVL DR and pressure measurements without deep learning refinement.

The comparison demonstrates a consistent accuracy enhancement when the Transformer refinement is active, shown in Table IX. For instance, in the longer Pool 01, the RMSE ATE reduces from 0.0662 m to 0.0439 m, and the MAE ATE from 0.0568 m to 0.0381 m. Similar error reductions are observed in Pool 02-03. The experimental results demonstrate that by providing the FGO with refined AP estimations, the localization errors during the global optimization can be significantly reduced, particularly for the long-duration mission.

E. Computation Cost Analysis

The lab experiments were performed on an embedded computer equipped with an Intel i5-8259U CPU @ 1.6 GHz, 15.5 GB RAM. To evaluate computational efficiency, we recorded the average running time, frequency, and memory consumption of different modules using datasets Pool 01-03.

Table X shows that the AP estimation operates at $4-5$ Hz, which is synchronized with the DVL DR frequency. Since the VI estimation operates at a higher frequency, the AP estimations are interpolated to temporally align with the camera.

TABLE X: Computation Cost of Our Proposed Method

UA Module	Latency (ms)	Memory (MB)
STFT-based Noise Estimation	1.46	122.21
Transformation-based Localization	825.95	510.88
Adaptive Factor Weighting	0.97	89.76
Localization Module	**Frequency (Hz)**	**Memory (MB)**
AP Estimation	4.53	61.01
VI Estimation	14.77	91.21
Dual-Framework (Pose Output)	14.97	137.04

Consequently, the FGO is triggered at the VINS update rate. The dual-framework mechanism integrates interpolated AP factor alongside VINS factor, performing joint optimization whenever the EKF framework receives AP updates. This graph-based optimization achieves a pose output frequency of 14.97 Hz and a memory usage of 137.04 MB, ensuring the real-time localization for closed-loop navigation and control.

For the UA module, the measurement and system-level modules indicate lightweight computations, serving as local and global adjustment steps, respectively. In contrast, the module-level processing requires more time and memory due to the incorporation of a Transformer-based architecture, resulting in an end-to-end latency of 825.95 ms on the CPU-only platform. To prevent this latency, the proposed system propagates the vehicle state in real-time while waiting for the UA module. Once computed, UA updates are applied as a posteriori corrections. With the availability of additional GPU resources in the future, the running time and memory usage are expected to decrease substantially.

V. Conclusion

This work focuses on the multi-modal fused underwater localization accounting for hierarchical uncertainties. A robust and accurate state estimation system named UA-AVIP is proposed by integrating acoustic, visual, inertial, and pressure information. The uncertainties in sensor measurements, odometry estimates, and nonlinear optimization are investigated in the proposed UA-AVIP. We conducted extensive experiments in the pool, lakes, and oceans. Qualitative and quantitative results prove that UA-AVIP outperforms other SOTA systems with more precise localization in complex underwater environments. Future work will focus on collecting more complex real-world data and developing advanced filters under minimally controlled conditions. Furthermore, based on the reliable state estimates, we also aim to introduce learning-based path planning algorithms for obstacle avoidance in unstructured terrains that bridge localization and planning.

Acknowledgments

This work was jointly supported by the Research Grants Council of Hong Kong (25206524), the Young Scientists Fund of the National Natural Science Foundation of China (42301520), and the Platform Project of Unmanned Autonomous Systems Research Centre (P0049516).

References

[1] Y. Fei and B. Xu, "A novel toa-based adaptive multi-auvs cooperative localization method under unknown nlos/los errors and outliers," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–10, 2025.

[2] Y. Ding, K. Li, H. Mei, S. Liu, and G. Hou, "Watermono: Teacher-guided anomaly masking and enhancement boosting for robust underwater self-supervised monocular depth estimation," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–14, 2025.

[3] F. Ding, R. Wang, T. Zhang, G. Zheng, Z. Wu, and S. Wang, "Real-time trajectory planning and tracking control of bionic underwater robot in dynamic environment," *Cyborg and Bionic Systems*, vol. 5, p. 0112, 2024.

[4] H. Huang, S. Zhang, D. Wang, K.-V. Ling, F. Liu, and X. He, "A novel bayesian-based adaptive algorithm applied to unobservable sensor measurement information loss for underwater navigation," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023.

[5] H. Huang, Z. Liu, L. Zhang, D. Wang, and J. Wang, "Via-slam: An underwater visual–inertial–acoustic slam with integrated dvl," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–11, 2025.

[6] Y. Ou, J. Fan, C. Zhou, P. Zhang, and Z.-G. Hou, "Hybrid-vins: Underwater tightly coupled hybrid visual inertial dense slam for auv," *IEEE Transactions on Industrial Electronics*, pp. 1–11, 2024.

[7] B. Joshi, "Robust underwater state estimation and mapping," Ph.D. dissertation, University of South Carolina, 2022.

[8] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE transactions on robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[9] S. Xu, J. S. Willners, Z. Hong, K. Zhang, Y. R. Petillot, and S. Wang, "Observability-aware active extrinsic calibration of multiple sensors," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2091–2097.

[10] H. Huang, M. Zhang, L. Zhang, D. Wang, and B. Wang, "A novel sonar image preprocessing method for auv positioning based on underwater slam," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–14, 2025.

[11] S. Xu, T. Luczynski, J. S. Willners, Z. Hong, K. Zhang, Y. R. Petillot, and S. Wang, "Underwater visual acoustic slam with extrinsic calibration," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 7647–7652.

[12] Y. Hou, L. Cheng, Q. Wang, J. Li, and Y. Ke, "An underwater multisource fusion anomaly detection navigation algorithm based on factor graph and lstm," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–11, 2025.

[13] J. Zhang, J. Song, S. Yang, H. Guo, and A. Song, "Absolute localization of underwater pipeline defects based on multimodal information fusion and active perception," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–13, 2025.

[14] S. Rahman, A. Quattrini Li, and I. Rekleitis, "Svin2: A multi-sensor fusion-based underwater slam system," *The International Journal of Robotics Research*, vol. 41, no. 11-12, pp. 1022–1042, 2022.

[15] Q. Liang, W. Xiao, J. Long, and D. Zhang, "A multimodal robust recognition method for grasping objects with robot flexible grippers," *IEEE/ASME Transactions on Mechatronics*, pp. 1–12, 2024.

[16] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, vol. 91, pp. 424–444, 2023.

[17] J. Wang, F. Chen, Y. Huang, J. McConnell, T. Shan, and B. Englot, "Virtual maps for autonomous exploration of cluttered underwater environments," *IEEE Journal of Oceanic Engineering*, vol. 47, no. 4, pp. 916–935, 2022.

[18] J. Zhang, X. Xiang, W. Li, and Q. Zhang, "Adaptive neural control of flight-style auv for subsea cable tracking under electromagnetic localization guidance," *IEEE/ASME Transactions on Mechatronics*, vol. 28, no. 5, pp. 2976–2987, 2023.

[19] X. Wang, B. Xu, and Y. Guo, "Minimum error entropy robust delay filter for multi-auv cooperative localization," *IEEE/ASME Transactions on Mechatronics*, pp. 1–11, 2024.

[20] Y. Bian, Z. Li, G. Wang, H. Qin, M. Hu, X. Qin, and R. Ding, "A unidirectional trend imm method for sins/dvl/usbl navigation system under the time-varying noise environment," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–14, 2025.

[21] L. Yan, S. Chang, X. Wang, L. Zhang, and J. Liu, "A dual-stage coverage path planning method for bathymetric survey using an auv in graph-based slam framework considering positioning uncertainty," *Ocean Engineering*, vol. 312, p. 119252, 2024.

[22] Y. Ling, T. Ma, Y. Li, E. Tian, X. Sun, S. Xu, and C. Qi, "Underwater tightly coupled state estimation method using imu and artificial lateral line systems," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–9, 2024.

[23] X. Zhang, B. He, and S. Gao, "An integrated navigation method for small-sized auv in shallow-sea applications," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 3, pp. 2878–2890, 2022.

[24] S. Bai, W. Wen, L.-T. Hsu, and P. Yang, "Factor graph optimization-based smartphone imu-only indoor slam with multi-hypothesis turning behavior loop closures," *IEEE Transactions on Aerospace and Electronic Systems*, 2024.

[25] S. Gu, L. Zhang, S. Guo, L. Zheng, R. An, T. Jiang, and A. Xiong, "Communication and cooperation for spherical underwater robots by using acoustic transmission," *IEEE/ASME Transactions on Mechatronics*, vol. 28, no. 1, pp. 292–301, 2022.

[26] F. Qiu and W. Zhang, "Auv-aided joint time synchronization and localization of underwater target with propagation speed uncertainties," *Ocean Engineering*, vol. 283, p. 115060, 2023.

[27] L. Sun, Y. Wang, X. Hui, X. Ma, X. Bai, and M. Tan, "Underwater robots and key technologies for operation control," *Cyborg and Bionic Systems*, vol. 5, p. 0089, 2024.

[28] D. He, W. Xu, and F. Zhang, "Symbolic representation and toolkit development of iterated error-state extended kalman filters on manifolds," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 12, pp. 12 533–12 544, 2023.

[29] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual–inertial odometry," *IEEE Transactions on Robotics*, vol. 33, no. 1, pp. 1–21, 2016.

[30] W. Zhou, Y. Pan, J. Liu, T. Wang, and H. Zha, "Visual-inertial-wheel odometry with wheel-aided maximum-a-posteriori initialization for ground robots," *IEEE Robotics and Automation Letters*, vol. 9, no. 5, pp. 4814–4821, 2024.

[31] Q. Wang, K. Liu, and Z. Cao, "System noise variance matrix adaptive kalman filter method for auv ins/dvl navigation system," *Ocean Engineering*, vol. 267, p. 113269, 2023.

[32] C. Chen, C. X. Lu, B. Wang, N. Trigoni, and A. Markham, "Dynanet: Neural kalman dynamical model for motion estimation and prediction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 12, pp. 5479–5491, 2021.

[33] X. Liang, Y. Wang, and H. Gu, "An anti-noise feature extraction and improved harris hawks optimization for on-load tap changer mechanical fault diagnosis," *IEEE Sensors Journal*, 2024.

[34] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[35] J. Kallwies, B. Forkel, and H.-J. Wuensche, "Determining and improving the localization accuracy of apriltag detection," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8288–8294.

[36] C. Schempp and S. Schulz, "High-precision absolute pose sensing for parallel mechanisms," *Sensors*, vol. 22, no. 5, p. 1995, 2022.

[37] P. Zhang, Z. Wu, J. Wang, S. Kong, M. Tan, and J. Yu, "An open-source, fiducial-based, underwater stereo visual-inertial localization method with refraction correction," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4331–4336.

[38] C. Wang, C. Cheng, D. Yang, G. Pan, and F. Zhang, "Underwater auv navigation dataset in natural scenarios," *Electronics*, vol. 12, no. 18, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/18/3788

[39] M. Grupp, "evo: Python package for the evaluation of odometry and slam." https://github.com/MichaelGrupp/evo, 2017.

[40] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 573–580.

Yuanbo Xue received the master's degree focusing in emergency systemic design and simulation from Beihang University in 2023. Since the same year, she has been studying for a doctoral degree in the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University.

Her current research is in the field of multimodal localization for underwater applications.

Yaming Ou received the B.E. degree in automation from Southeast University, Nanjing, China, in 2021. He is currently pursuing the Ph.D. degree in control theory and control engineering with the Institute of Automation, Chinese Academy of Sciences (IACAS), Beijing, China.

His research interests include underwater 3D vision, SLAM, multi-sensor fusion, and autonomous robot navigation.

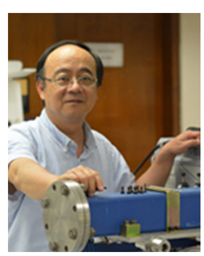
Junfeng Fan (Member, IEEE) received the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences (IACAS), Beijing, China, in 2019.

He is currently an Associate Professor of Control Theory and Control Engineering with the State Key Laboratory of Management and Control for Complex Systems, IACAS, Beijing. His research interests include robot vision and underwater robots.

Chao Zhou received the B.S. degree in automation from Southeast University, Nanjing, China, in 2003, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences (IACAS), Beijing, China, in 2008.

He is currently a Professor of Control Theory and Control Engineering with the State Key Laboratory of Management and Control for Complex Systems, IACAS, Beijing. His research interests include underwater robots and bionic robots.

Bing Wang is an Assistant Professor in Robotics and Autonomous Systems at the Faculty of Engineering, The Hong Kong Polytechnic University. He obtained his DPhil degree in 2022 from the Department of Computer Science at the University of Oxford. His research is at the forefront of spatial intelligence, a dynamic field focused on advancing human-level 3D spatial perception and world understanding for mobile robotics. The primary objective is to enhance the reliability, intelligence, and security of intelligent machines in real-world environments.



Chih-yung Wen received the B.S. degree from the Department of Mechanical Engineering, National Taiwan University, in 1986, and the M.S. and Ph.D. degrees from the Department of Aeronautics, California Institute of Technology, in 1989 and 1994, respectively. In 2012, he joined the Department of Mechanical Engineering, The Hong Kong Polytechnic University, as a professor. In 2019, he became the Interim Head of the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University. His current research interests include aircraft control and reinforcement learning control.