Depth-Consistent Monocular Visual Trajectory Estimation for AUVs

Yangyang Wang, *Member*, *IEEE*, Xiaokai Liu, Dongbing Gu, *Senior Member*, *IEEE*, Jie Wang, *Senior Member*, *IEEE*, and Xianping Fu

***Abstract*—Visual trajectory estimation can endow autonomous underwater vehicles (AUVs) with environmental perception capabilities and has broad application prospects in the fields such as oceanographic surveys, underwater construction, and marine ranching. However, due to featureless images and depth ambiguity issues caused by underwater multiple mediums environments, monocular visual trajectory estimation in complex underwater environments remains a challenging problem. In this paper, we propose a monocular visual trajectory estimation method for AUVs, which can address the challenges of featureless and depth ambiguity by leveraging deep image representations and multi-view geometry. Specifically, we design a bidirectional optical flow consistency scheme that selects sparse correspondences from monocular dense predictions to deal with the featureless images, and then achieve AUV trajectory estimation through epipolar constraints. Furthermore, we propose an iterative depth-consistent method, which solves the problem of depth ambiguity by aligning geometrically triangulated depths to the scale-consistent deep depths. We also develop a low-cost, agile, and portable AUV picking system with real-time trajectory estimation capabilities, and carry out extensive experiments in the Yellow Sea to test its performance. The experimental results demonstrate the effectiveness of the proposed method.**

***Index Terms*—Trajectory estimation, localization, monocular vision, underwater, AUV.**

I. Introduction

In recent years, autonomous underwater vehicles (AUVs) are widely used for a variety of tasks, for instance, hydrographic survey, ship maintenance, and object picking [1], [2]. For complex ocean scenarios, localization function of AUVs is the foundation for accomplishing potential applications. Trajectory estimation capability could significantly improve the AUV autonomy while performing tasks in unknown scenarios [3], [4], [5]. The underwater positioning method based on acoustic sensors such as sonar and Doppler velocity log, has the advantages of high positioning accuracy and strong autonomy [6], [7]. However, this method requires expensive acoustic sensors, which limit their widespread application in civilian fields. The multi-node localization method based on underwater acoustic sensor networks has advantages such as wide detection range and strong adaptability [8]. However, this method needs to go through a cumbersome installation process before deployment and is not suitable for application in unknown marine scenarios. Moreover, sensor nodes are susceptible to passive motion caused by ocean currents or tides, resulting in deviation in the position of reference points, which affects the localization accuracy [9]. There is an urgent need for new trajectory estimation technologies that can support the autonomous operations of underwater vehicles or robots.

Simultaneous localization and mapping (SLAM) refers to the process of calculating the position and orientation of a vehicle or robot, with respect to its surroundings, while simultaneously mapping the environment. Visual SLAM has been successfully applied in indoor robot localization, drone localization, and self-driving vehicle localization [10], [11], [12]. It has been demonstrated to significantly enhance the intelligence and autonomy of robots [13], [14]. Compared with existing underwater positioning methods, it has the advantages of low cost, low energy consumption, and high resolution [15], [16], [17]. Current mobile platforms mainly use sensors such as depth cameras and multi-cameras to achieve visual SLAM. Depth cameras rely on structured light or Time-of-Flight principles to actively project light to measure the depth of the environment, but it has strict requirements for the environment and can only be used in indoor scenes. Calibration between multi-cameras is complex and consumes significant computational resources, making it difficult to deploy on small underwater embedded platforms with limited computing resources. Considering the above factors, monocular systems offer lower cost and power consumption and are easier to deploy in underwater environments.

Although monocular visual SLAM has broad application prospects in the field of AUV trajectory estimation, many problems still need to be solved urgently due to the complex underwater environment. For example, underwater environments exhibit phenomena such as featureless images, poor specificity of descriptors, and interference from moving objects, which pose challenges to the robustness of visual localization. The camera carried by AUVs is usually placed in a waterproof housing. When underwater cameras record images, light must traverse multiple mediums including water, underwater camera housing, lenses, camera sensor, etc., leading to alterations in

This work was supported in part by the National Natural Science Foundation of China under Grants 62471080 and 62271098, in part by the China Postdoctoral Science Foundation under Grant 2024M750294, in part by the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation under Grant GZB20240089, in part by the Natural Science Foundation of Liaoning Province under Grants 2024-BS-021 and 2024-MS-010, in part by the Science and Technology Program of Liaoning Province under Grant 2023JH26/10300010, and in part by the Educational Science Research Project of China Transportation Education Research Association under Grant JT2024YB104. (*Corresponding author: Jie Wang.*)

Yangyang Wang, Xiaokai Liu, Jie Wang and Xianping Fu are with the School of Information Science and Technology, Dalian Maritime University, Dalian 116026, P.R. China (e-mail: yyw@dlmu.edu.cn, xkliu@dlmu.edu.cn, wang_jie@dlmu.edu.cn, fxp@dlmu.edu.cn).

Dongbing Gu is with the School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ Colchester, U.K. (e-mail: dgu@essex.ac.uk).

the light's propagation path due to light refraction. The two most common types of radial distortion are pincushion distortion and barrel distortion. In addition, during the installation of underwater cameras, the lens and photosensitive plane were not strictly parallel, resulting in tangential distortion, as shown in Fig. 1. Distortion may cause changes in the position of features in the image, making it more difficult to extract and match underwater featureless images. Meanwhile, due to the lack of depth information, geometry-based monocular vision schemes often suffer from system drift issues, especially in large-scale environments. Loop closure detection is usually used to correct the error accumulation caused by system drift. However, there are few loop closure scenarios in marine ranching tasks, making it less suitable to correct system drift through loop closure detection. The above issues will affect the accuracy of underwater monocular visual trajectory estimation.

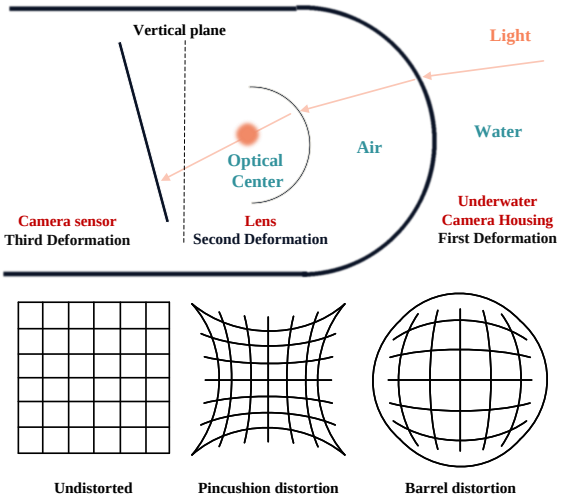


Fig. 1. Illustration of underwater camera imaging deformation. Light must traverse multiple mediums including water, underwater camera housing, lenses, camera sensor etc., leading to alterations in the light's propagation path. The two most common types of radial distortion are pincushion distortion and barrel distortion. Meanwhile, the camera sensor on the left side of the figure cannot be strictly parallel to the vertical plane, resulting in tangential distortion.

Deep neural networks have powerful capabilities in mining featureless image information and predicting scene depth values [18], [19], [20]. However, the underwater labelled datasets are limited, making it difficult to obtain reliable deep network models through supervised training. Therefore, we address the challenges in the field of underwater visual trajectory estimation through unsupervised deep learning methods. In this paper, we propose a depth-consistent monocular visual trajectory estimation method for AUVs, which combines self-supervised deep networks with multi-view geometry to achieve continuous trajectory estimation in underwater multiple medium environments. We have successfully implemented the proposed network on the NVIDIA Jetson TX2 platform. Experimental results show that the proposed method could produce a small error of 0.44 m in complex underwater scenarios and demonstrate a competitive performance. We summarize our contributions as follows:

- We propose a monocular visual trajectory estimation scheme, which integrates bidirectional optical flow consistency and epipolar constraints strategy to solve the trajectory estimation problem of AUVs in featureless environments.
- We propose an iterative depth-consistent method, which addresses the issue of depth ambiguity in the monocular vision by aligning geometrically triangulated depths to the scale-consistent deep depths.
- We develop a low-cost, agile, and portable underwater autonomous picking system with real-time trajectory estimation capabilities, and carry out extensive underwater experiments in the Yellow Sea to test its performance.

The paper is structured as follows. Section II gives a brief introduction to some related works. Section III presents an overview of the proposed depth-consistent monocular visual trajectory estimation method for AUVs. In Section IV, we elaborate the details of our proposed scheme. Section V describes our experimental results in the fire pool and the Yellow Sea. Finally, the conclusions and directions for future work are summarized in Section VI.

II. Related Works

AUV localization is the foundation for completing tasks such as marine resource exploration, aquaculture monitoring, and multi-AUV collaborative operations. To adapt to harsh underwater environments, most underwater localization algorithms are based on acoustic sensors, such as sonar, Doppler velocity log, long baseline, and ultra-short baseline [21], [22]. Dong et al. [23] proposed a sonar-inertial navigation system with an adaptive grouping optimization framework, which used the sonar-inertial constraint-based feature matching to increase the accuracy of data association. Zhang et al. [24] proposed a multiple model integrated navigation method for AUV surface missions to maintain the high-frequency output and robustness of the AUV navigation system. This scheme can be applied to the localization task of small AUVs in shallow-sea applications. The method of using acoustic sensors to collect data can provide good localization results. However, their high cost limits their widespread application.

In recent years, with the rapid development of underwater communication technologies, underwater acoustic sensing networks have been emergent and are attracting the interest of numerous industrial and academic researchers [25], [26], [27]. More specifically, Wang et al. [28] proposed an efficient localization scheme with velocity prediction, which solves the problems of excessive energy consumption and large localization errors caused by harsh underwater conditions like node mobility and huge ranging errors. Su et al. [29] proposed a mobile-beacon-based iterative localization mechanism, which achieves node hierarchically positioning of large-scale multihop underwater acoustic sensing networks with an aim at increasing the percentage of localized nodes and reducing the localization error in the network. In [30], the authors proposed a cooperative location-aware network, which includes surface buoys, AUVs, active sensor nodes and passive sensor nodes. This scheme can integrate the current estimation into the

localization of sensor nodes, which allows the co-design of modeling and optimization. The above works have achieved good results in localization. However, sensor nodes often have passive motions caused by ocean currents, which could also affect the accuracy of localization. In addition, the device requires tedious installation and debugging before use and is not suitable for application in unknown underwater scenarios.

Visual SLAM has attracted much attention due to its rich information content, wide applicability, low cost, and many other advantages [31], [32], [33]. Cui et al. [34] proposed a tailored incremental structure from motion framework for multi-camera systems, where the internal relative poses between cameras can not only be calibrated automatically but also serve as an additional constraint to improve the system robustness. Ye et al. [35] proposed a covisibility-based matching strategy that discovers covisible image pairs and iteratively extends the feature matches from the potential registration images. This method can process well-mixed data and unordered images. Zhan et al. [19] proposed a new odometry network called DF-VO, but the network does not apply for underwater applications. In [36], the authors proposed an optimization-based tightly coupled direct visual-inertial odometry, which fuses the visual and inertial measurements to provide real-time trajectory estimation. At the same time, the authors designed an iterative selection strategy to reject outliers during the data association and establish more visual constraints in the optimization. Afterwards, Luo et al. [37] proposed the monocular visual-inertial trajectory estimation with point-line fusion and backend adaptive optimization to improve the positioning accuracy and robustness of unmanned aerial vehicle navigation systems. The above works have shown good performance in both terrestrial and aerial scenarios. However, the complex and dynamic working environment renders these methods unsuitable for direct application in underwater environments.

As the demand for underwater resource development sharply rises, some scholars have begun to focus on the research of underwater SLAM methods. Wang et al. [38] solved the real-time dense 3D reconstruction problem for a resource-constrained autonomous underwater vehicle. This fully demonstrates the feasibility of using visual sensors to achieve SLAM in underwater environments. Nevertheless, there are still many problems that need to be solved urgently in the research of underwater visual SLAM. Fan et al. [39] proposed an approach named object level depth reconstruction network taking only RGB images as input for category-level 6D object pose estimation. However, due to the scattering properties of the water medium, RGB-D cameras are unable to obtain effective depth information in underwater environments.

Different from other existing studies that employ costly sonars, Doppler velocity log, or ultra-short baseline to acquire measurement data, our proposed scheme, as described in the following section, only includes a forward-looking monocular camera to effectively achieve underwater trajectory estimation. A key feature of our method is to achieve the visual trajectory estimation for a miniature underwater vehicle with a low-cost camera and limited computational resources. Therefore, a depth-consistent monocular visual trajectory estimation method for AUVs is proposed. Since our scheme is based on underwater vision, we would like to highlight the main differences with other papers. For example, there is no requirement for pre-prepared artificial markers, making it better suited for unfamiliar underwater scenarios.

III. System Overview

The architecture of the proposed depth-consistent monocular visual trajectory estimation method for AUVs is shown in Fig. 2. In this paper, we propose to use an optical flow network to match the corresponding feature pairs and a depth network to estimate the depth. Once matching pairs and depths are available, they are used to retrieve the pose via the epipolar geometry method for 2D-2D matching. In the learning stage, both optical flow and depth are used to construct the loss function. Specifically, we first corrected the distorted images caused by underwater multiple mediums environments and then fed a pair of underwater images into both networks simultaneously. Secondly, the optical flow network learns to generate forward optical flow and backward optical flow estimations. Due to interference from dynamic objects or occlusion, numerous outliers may occur. To address this issue, we propose a correspondence selection method based on bidirectional consistency to select reliable optical flows. Then, we use the epipolar geometry method for 2D-2D matching to estimate the trajectory. Thirdly, the monocular depth estimation network is based on the standard, fully convolutional, U-Net [40] to predict depth. Our proposed iterative depth-consistent method is able to solve the problem of depth ambiguity by aligning geometrically triangulated depths to the scale-consistent deep depths. The details of network architecture and its training will be introduced in Section IV.

IV. Monocular Visual Trajectory Estimation Method for AUVs

This section introduces the proposed depth-consistent monocular visual trajectory estimation method for AUVs. In general, our method addresses the challenges of featureless images and depth ambiguity by leveraging deep image representations and multi-view geometry. The detailed framework is shown in Fig. 2 and described in Algorithm 1. Table I shows the main notations used in this paper.

TABLE I
NOTATION DEFINITIONS

Name	Description
$\boldsymbol{L}_m^n$	The forward optical flow
$\boldsymbol{L}_n^m$	The backward optical flow
L_θ	The bidirectional prediction consistency of $\boldsymbol{L}_m^n$ and $\boldsymbol{L}_n^m$
L_ρ	The depth smoothness error of edge awareness
$\boldsymbol{K}$	The camera intrinsic matrix
$\boldsymbol{E}$	The Essential matrix
L_α	The photometric loss
L_β	The depth smoothness loss
L_χ	The depth consistency loss
L_{sum}	The overall objective function

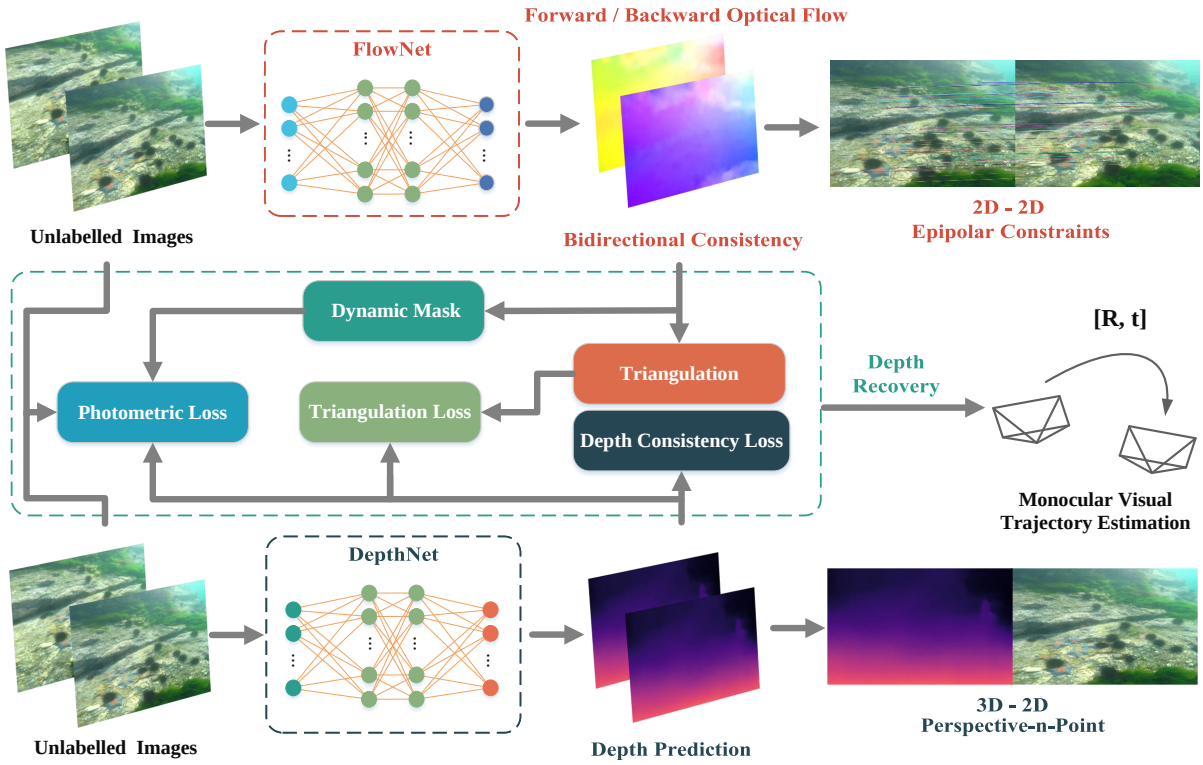


Fig. 2. The outline of the proposed algorithm.

Algorithm 1: Underwater Visual Trajectory Estimation

Input: Image sequence after correction: $\{\boldsymbol{I}_1, \boldsymbol{I}_2, \cdots, \boldsymbol{I}_n\}$
Output: Trajectory estimation: $\{\boldsymbol{T}_1, \boldsymbol{T}_2, \cdots, \boldsymbol{T}_n\}$

Initialize $\boldsymbol{T}_1 = \boldsymbol{I}$; $i = 2$
while $i \leq n$ **do**
  Calculate the forward and backward flow consistency
  Select the best 2D-2D matches through the consistency
  if 2D-2D matches $>$ threshold **then**
    Compute $\boldsymbol{E}$ through $(\boldsymbol{p}_m, \boldsymbol{p}_n)$ and solve $[\boldsymbol{R}, \boldsymbol{t}]$
    Triangulate $(\boldsymbol{p}_m, \boldsymbol{p}_n)$
    Estimate the scaling factor
    Get ${\boldsymbol{T}_i}^{i-1}$
  else
    Form the 3D-2D correspondences
    Solve $[\boldsymbol{R}, \boldsymbol{t}]$ by the Perspective-n-Point method
    Get ${\boldsymbol{T}_i}^{i-1}$
  end if
  $\boldsymbol{T}_i \leftarrow T_{i-1}{T_i}^{i-1}$
end while

A. *Distortion Correction*

Unlike on-the-ground working environments, the light must traverse multiple different transparent mediums including water, underwater camera housing, lenses, camera sensor, etc., leading to alterations in the light's propagation path due to light refraction. Correcting radial distortions requires consideration of calibration methods for multiple mediums refraction, ensuring geometric accuracy in captured images. Moreover, the lens cannot be installed strictly parallel to the photosensitive plane, resulting in tangential distortion. Therefore, distortion correction is essential, otherwise, it will seriously affect the accuracy of visual measurement.

We project the 3D point onto the normalized image plane and set their normalized coordinates to $[x, y]^T$. For radial distortion, we use three parameters k_1, k_2, and k_3 for correction. For tangential distortion, we use two parameters p_1 and p_2 for correction. Therefore, we describe the coordinate changes before and after distortion using the following formula:

$$\begin{cases} x_d = x(1+k_1r^2+k_2r^4+k_3r^6)+2p_1xy+p_2(r^2+2x^2) \\ y_d = y(1+k_1r^2+k_2r^4+k_3r^6)+p_1(r^2+2y^2)+2p_2xy \\ r^2 = x^2+y^2 \end{cases}, \quad (1)$$

where $[x, y]^T$ represents the coordinate of normalized plane point, $[x_d, y_d]^T$ represents the coordinate of distorted points, r represents the distance between any point X on the plane and the origin of the coordinate system, k_1, k_2, and k_3 represent radial distortion coefficient, p_1 and p_2 represent tangential distortion coefficient. In the underwater experimental scenario of this paper, we calculated k_1=-0.240329, k_2=0.0552096, k_3=0, p_1=-0.000150 and p_2=-0.000205 through offline calibration steps [41].

Between the pixel coordinate system and the imaging plane, we define the pixel coordinates as scaled by a factor of s_x on the u-axis, by a factor of s_y on the v-axis, and translated by $[c_x, c_y]^T$ from the origin. We project the corrected point onto the pixel plane through the camera intrinsic matrix to obtain

the correct position of the point on the image. The formula is written as follows:

$$\begin{cases} u = s_x f x_d + c_x \\ \\ v = s_y f y_d + c_y \end{cases}, \tag{2}$$

where $[u, v]^T$ represents the pixel coordinates, f represents the focal length.

B. Self-Supervised Optical Flow Estimation Network

The optical flow method is a scheme used to compute the motion information of objects from monocular consecutive images, which leverages the temporal changes of pixels in an image sequence and the correlation between adjacent frames to find the corresponding relationship between the previous and current frames.

We define a pair of images ($\boldsymbol{I}_m$, $\boldsymbol{I}_n$). In the optical flow network, convolution is first used to generate pyramid features that can extract information from different levels of the image. Then, the optical flow field is inferred from the high-level features $\mathcal{F}_m$ and $\mathcal{F}_n$ of the images $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$, and the estimated values of the optical flow field are gradually refined through iteration. For any sub-pixel displacement $\dot{\mathrm{x}}$, we interpolate $\mathcal{F}$ to obtain $\widetilde{\mathcal{F}}$ and its formula is written as follows:

$$\widetilde{\mathcal{F}}(\mathrm{x}) = \sum_{\mathrm{x}_s^m \in N(\mathrm{x}_s)} \mathcal{F}(\mathrm{x}_s^m)(1-|x_s - x_s^m|)(1-|y_s - y_s^m|), \tag{3}$$

where s represents the spatial resolution factor, $\mathrm{x}_s = \mathrm{x} + \dot{\mathrm{x}} = (x_s, y_s)^\top$ represents the source coordinates of the defined sampling point in the input feature map $\mathcal{F}$, $N(\mathrm{x}_s)$ represents the four adjacent pixels of x_s. Since it is possible to compute gradients efficiently, backpropagation of bilinear interpolation is allowed during network training.

During the process of descriptor matching, the point correspondences between images $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$ are established by calculating the correlation of high-level feature vectors in single pyramid features $\mathcal{F}_m$ and $\mathcal{F}_n$. The formula can be computed as follows:

$$c\,(\mathrm{x}, \mathrm{d}) = \mathcal{F}_m(\mathrm{x}) \cdot \mathcal{F}_n(\mathrm{x}+\mathrm{d}) \,/\, N, \tag{4}$$

where c represents the matching cost between point x in $\mathcal{F}_m$ and point $\mathrm{x}+\mathrm{d}$ in $\mathcal{F}_n$, $\mathrm{d} \in \mathbb{Z}$ represents the displacement vector of x, and N represents the length of the feature vector. The total cost value C is constructed by aggregating all matching costs c together.

In the descriptor matching unit M, the residual flow is inferred by filtering the matching cost values C. To match the spatial resolution of the pyramid features at the current level, the previous level of $\dot{\mathrm{x}}$ needs to be upsampled in spatial resolution (denoted by "$\uparrow s$") and amplitude (multiplied by scalar s) to $s\dot{\mathrm{x}}^{\uparrow s}$. The calculation process of the optical flow field $\dot{\mathrm{x}}_m$ is defined as follows:

$$\dot{\mathrm{x}}_m = \underbrace{M\big(C(\mathcal{F}_m, \widetilde{\mathcal{F}}_n; d)\big)}_{\Delta\dot{\mathrm{x}}_m} + s\dot{\mathrm{x}}^{\uparrow s}. \tag{5}$$

Since the cost capacity function in the descriptor matching unit is aggregated by measuring the correlation pixel by pixel, the optical flow estimation obtained from the previous inference $\dot{\mathrm{x}}_m$ can only achieve pixel-level accuracy. We refer interested readers to [42] for more details. To prevent erroneous optical flow from being amplified through upsampling and propagated to the next pyramid level, the pixel-level optical flow field $\dot{\mathrm{x}}_m$ is refined to sub-pixel accuracy. Specifically, we project $\mathcal{F}_n$ to $\widetilde{\mathcal{F}}_n$ through the optical flow estimation $\dot{\mathrm{x}}_m$.

The sub-pixel refinement unit S calculates the residual flow $\Delta\dot{\mathrm{x}}_s$ to minimize the feature space distance between $\mathcal{F}_m$ and $\widetilde{\mathcal{F}}_n$, thus obtaining a more accurate optical flow field $\dot{\mathrm{x}}_s$. The formula can be computed as follows:

$$\dot{\mathrm{x}}_s = \underbrace{S(\mathcal{F}_m, \widetilde{\mathcal{F}}_n, \dot{\mathrm{x}}_m)}_{\Delta\dot{\mathrm{x}}_s} + \dot{\mathrm{x}}_m. \tag{6}$$

The photometric error L_α calculates the pixel difference between image $\boldsymbol{I}_m$ and synthesised view $\boldsymbol{I}_n^m$ warped from the source image $\boldsymbol{I}_n$. The formula can be written as:

$$\begin{aligned} L_\alpha(\boldsymbol{I}_m, \boldsymbol{I}_n^m) &= \frac{\tau}{2}\Big(1 - \mathrm{SSIM}(\boldsymbol{I}_m, \boldsymbol{I}_n^m)\Big) + (1-\tau)\,|\boldsymbol{I}_m, \boldsymbol{I}_n^m| \\ \boldsymbol{I}_n^m &= \omega\Big(\boldsymbol{I}_n, p_{re}(\boldsymbol{K}, \boldsymbol{D}_m, \boldsymbol{T}_m^n)\Big), \end{aligned} \tag{7}$$

where structural similarity (SSIM) is an indicator to measure the structure between two images [43], [44]. We set τ = 0.85 to balance the SSIM error and colour intensity error of underwater images. ω represents a projection function. p_{re} is the reprojection of pixel coordinates from the m^{th} image to the n^{th} image. $\boldsymbol{K}$ represents the camera intrinsic. $\boldsymbol{D}_m$ represents the predicted depth map of the m^{th} image. $\boldsymbol{T}_m^n$ represents the relative pose between the two images. The reprojection function p_{re} for a pixel x from the m^{th} image to the n^{th} image can be written as:

$$p_{re}(\boldsymbol{K}, \boldsymbol{D}_m, \boldsymbol{T}_m^n) = \boldsymbol{K}\boldsymbol{T}_m^n\boldsymbol{K}^{-1}\boldsymbol{x}\boldsymbol{D}_m[\boldsymbol{x}]. \tag{8}$$

We define the forward optical flow as $\boldsymbol{L}_m^n$, and the backward optical flow as $\boldsymbol{L}_n^m$. The optical flow network is trained by minimizing the average loss function for each pixel in the whole image. The loss function is written as follows:

$$\begin{aligned} L = \min_n\, & L_\alpha(\boldsymbol{I}_m, \boldsymbol{I}_n^m) + \sigma\, L_\rho(\|\boldsymbol{L}_m^n\|_2, \boldsymbol{I}_m) \\ & + \mu\, L_\theta\Big(\Big| - \boldsymbol{L}_m^n - \boldsymbol{\omega}\Big(\boldsymbol{L}_n^m, p_f(\boldsymbol{L}_m^n)\Big)\Big|\Big), \end{aligned} \tag{9}$$

where $\boldsymbol{I}_n^m = \boldsymbol{\omega}\Big(\boldsymbol{I}_n, p_f(\boldsymbol{L}_m^n)\Big)$, L_ρ represents the depth smoothness error of edge awareness, L_θ represents the bidirectional prediction consistency of forward optical flows and backward optical flows, and p_f represents the correspondence between the m^{th} and n^{th} images established through the optical flow field.

The estimation of trajectory using dense optical flow may be affected by occlusion or dynamic objects, resulting in numerous erroneous matches. The solution is to select high-quality matching relationships from dense images randomly. We leverage forward optical flow and backward optical flow

to predict dense correspondences between 2D-2D images. Optical flow describes the pixel motion variation, which represents the correspondence of all pixels in $\boldsymbol{I}_m$ with $\boldsymbol{I}_n$. The optical flow consistency can be calculated using the following formula:

$$\boldsymbol{C} = -\boldsymbol{L}_m^n - \boldsymbol{\omega}\Big(\boldsymbol{L}_n^m, p_f(\boldsymbol{L}_m^n)\Big). \tag{10}$$

The warping process at pixel x can be represented as:

$$\boldsymbol{\omega}\Big(\boldsymbol{L}_n^m[\boldsymbol{x}], p_f(\boldsymbol{L}_m^n[\boldsymbol{x}]\Big) = \boldsymbol{L}_n^m\Big[\boldsymbol{x} + \boldsymbol{L}_m^n[\boldsymbol{x}]\Big], \tag{11}$$

where $[\boldsymbol{x}]$ is the coordinate index value of the pixel, p_f represents the correspondence established between the m^{th} and n^{th} images through the optical flow field.

The specific meaning of this equation is to predict the corresponding position $\boldsymbol{x} + \boldsymbol{L}_m^n[\boldsymbol{x}]$ in image $\boldsymbol{I}_n$ for a certain pixel point $[\boldsymbol{x}]$ in image $\boldsymbol{I}_m$ based on the forward optical flow. Subsequently, this predicted point is further projected to its pixel position $\boldsymbol{L}_n^m[\boldsymbol{x}+\boldsymbol{L}_m^n[\boldsymbol{x}]]$ in image $\boldsymbol{I}_m$ using the backward optical flow. After calculating the consistency $\boldsymbol{C}$ between the forward and backward optical flow, the optimal N matches are formed by selecting the estimated flow $\hat{\boldsymbol{L}}$ with the least flow inconsistency.

After selecting a set of 2D-2D pixel correspondences ($\boldsymbol{p}_m$, $\boldsymbol{p}_n$) from the image pair, we leverage epipolar geometry to compute the $\boldsymbol{F}$ (Fundamental Matrix) or $\boldsymbol{E}$ (Essential Matrix). Then, the AUV trajectory estimation $[\boldsymbol{R}, \boldsymbol{t}]$ can be obtained through the decomposition of $\boldsymbol{F}$ or $\boldsymbol{E}$. The formula can be written as:

$$\boldsymbol{p}_n^T \boldsymbol{K}^{-T} \boldsymbol{E} \boldsymbol{K}^{-1} \boldsymbol{p}_m = 0, \tag{12}$$

where $\boldsymbol{E} = [\boldsymbol{t}]_\times \boldsymbol{R}$, $\boldsymbol{F} = \boldsymbol{K}^{-\boldsymbol{T}} \boldsymbol{E} \boldsymbol{K}^{-\boldsymbol{1}}$, $[\boldsymbol{t}]_\times$ is the skew-symmetric matrix, $\boldsymbol{K}$ is the camera intrinsic matrix.

C. Self-Supervised Monocular Depth Estimation Network

This section elaborates on how to use monocular colour underwater images as input to predict a depth map in the training of the depth prediction network. The trajectory estimated with optical flow and geometric constraints in Section IV.B may suffer from depth ambiguity, which could result in system drift problems. We adopt the self-supervised training approach to jointly train the depth and pose by minimizing the mean objective function value of each pixel in the whole image. The monocular depth estimation network is based on the general U-Net [40] architecture and is suitable for images with low resolution in underwater scenes.

The self-supervised depth estimation network regards learning problems as one of the novel view synthesis methods. We can extract interpretable depth from the model by constraining the network with intermediate variables for image synthesis. However, due to the relative orientation between these two views, each pixel may have a significant amount of incorrect depth. We formulate the problem as the minimization of a photometric reprojection error during training, with the photometric error calculated using Equ. (7).

When calculating the reprojection error of multi-source images, a strategy employed by some self-supervised depth estimation methods is to evenly distribute the reprojection error among each source image. This may lead to an issue where certain pixels are visible in the target image but not visible in some of the source images. In such cases, when the network predicts the correct depth for these pixels, the corresponding colours in the occluded parts of the source images may mismatch the target, resulting in higher photometric error. The main cause of this phenomenon is the influence of underwater occluding objects or pixels outside the imaging field caused by the motion of AUVs. To address the issues above, instead of averaging the photometric error of all source images, taking the minimum value can significantly reduce artifacts at the image borders, improve the clarity of occlusion boundaries, and enhance accuracy.

The regularization form of the edge-aware depth smoothness term L_β is represented as follows:

$$L_\beta(\boldsymbol{D}_m, \boldsymbol{I}_m) = |\partial_x \boldsymbol{D}_m| e^{-|\partial_x \boldsymbol{I}_m|} + |\partial_y \boldsymbol{D}_m| e^{-|\partial_y \boldsymbol{I}_m|}, \tag{13}$$

where ∂_x and ∂_y represent gradients in horizontal and vertical directions, respectively.

We define the inverse depth consistency error L_χ as follows:

$$L_\chi(\boldsymbol{D}_m, \boldsymbol{D}_n^m) = \Big|\frac{1}{\boldsymbol{D}_m} - \frac{1}{\boldsymbol{D}_n^m}\Big|. \tag{14}$$

Finally, we combine the photometric loss, smoothness loss, and depth consistency loss as L_{sum}, and jointly train the network by minimizing the mean of the objective function. The final training loss function is written as follows:

$$\begin{aligned} L_{\text{sum}} = \min_n L_\alpha\left(\boldsymbol{I}_m, \boldsymbol{I}_n^m\right) + \delta\, L_\beta\left(\boldsymbol{D}_m, \boldsymbol{I}_m\right) \\ + \min_n \gamma\, L_\chi\left(\boldsymbol{D}_m, \boldsymbol{D}_n^m\right), \end{aligned} \tag{15}$$

where L_α represents photometric loss, L_β represents depth smoothness loss, L_χ represents depth consistency loss, and both δ and γ represent loss weightings.

Perspective-n-Point (PnP) is a method for solving the motion of 3D to 2D point pairs. After obtaining n 3D spatial points and their corresponding projection positions, the trajectory of AUVS can be estimated. For example, in two images, if the 3D position of one feature point is known, only three-point pairs are needed to estimate the AUV trajectory. When obtaining the 3D-2D correspondences, let the projection of a 3D point ($\boldsymbol{X}_m$) on the m^{th} image to its corresponding point on the n^{th} image as ($\boldsymbol{X}_m, \boldsymbol{p}_n$), and the PnP method can estimate the AUV trajectory by minimizing the reprojection error. The equation for solving it is as follows:

$$e = \sum_x \|\boldsymbol{K}(\boldsymbol{R}\boldsymbol{X}_m[\boldsymbol{x}] + \boldsymbol{t}) - \boldsymbol{p}_n[\boldsymbol{x}]\|_2, \tag{16}$$

where $[\boldsymbol{x}]$ is the coordinate index value of the pixel.

V. Experiments and Analysis

To validate the performance of the proposed depth-consistent monocular visual trajectory estimation for AUVs in complex underwater environments, experimental tests were conducted in both a fire pool and the Yellow Sea. In this

section, we first introduce the training process of the optical flow and monocular depth estimation network, followed by specific implementation details. Finally, we provide a detailed analysis of the results obtained under different experimental conditions in underwater environments. The code can be found in our project webpage[1].

A. Experimental Setup

In the optical flow network, 3×3 filters are used in each convolutional layer, except for the last layer of the descriptor matching M. In the sub-pixel refinement section, 5×5 filters are used. We trained the optical flow network using PyTorch, leveraging the default parameters of the Adam optimizer. The input and output resolution of the images were set to 640×192. We set the learning rate to 0.0001 and trained the whole network for 30,000 epochs using the KITTI dataset. The monocular depth estimation network was also trained using PyTorch with the default parameters of the Adam optimizer. The input and output resolution of the images was set to 640×192. The learning rate was set to 0.0001, and the whole network was trained for 20,000 epochs using the KITTI dataset. Both network models mentioned above were trained using an NVIDIA GeForce RTX 2080 Ti graphics card.

We also collected underwater data close to the Zhangzi Island in the Yellow Sea. The region maintains a natural marine ecosystem rich in marine life, including sea cucumbers, sea urchins, scallops, etc. After pre-training the model with the KITTI dataset, the underwater dataset was fed into the optical flow network and monocular depth estimation network in the same manner for further learning. In addition, the brightness, contrast, and saturation of the images were randomly varied to simulate changes in underwater environments at different times of the day. We trained the optical flow network 35,000 epochs and the monocular depth estimation network 25,000 epochs, ultimately obtaining a high-performance underwater monocular visual trajectory estimator.

We tested the proposed method in unstable marine scenarios to evaluate its performance. The depth-consistent monocular visual trajectory estimation for the AUV system consists of the following components:

- NVIDIA Jetson TX2,
- Forward-looking monocular global shutter camera,
- LED illumination,
- Autonomous picking systems,
- Rechargeable Lithium-ion battery (24VDC, 10Ah).

The custom-made sensor suite was designed for the target application of achieving marine environmental visual trajectory estimation and completing autonomous object picking tasks. Extensive experiments were carried out in a fire pool and the Yellow Sea. The size of the whole underwater vehicle system is 886 × 622 × 758 mm^3 in dimensions. It consists of a forward-looking global shutter camera, a pair of LEDs, a rechargeable Lithium-ion battery (Charging Voltage: 220VAC), etc. The AUV is driven by a Lithium-ion battery and is equipped with four thrusters, including two plane omnidirectional thrusters and two vertical thrusters, as shown in Fig. 3.

Fig. 3. The picture of the underwater vehicle testbed that we used for the experiment. It includes a forward-looking global shutter camera, a pair of LEDs, etc.

The marine environment experiment is an arduous task, constrained by various conditions. We also conducted laboratory experiments to evaluate the proposed scheme and facilitate comparison with the ground truth. In the indoor experiment, the size of the fire pool is 15×8×4 m^3, We arranged some scallops, sea urchins, sea cucumbers and observation devices at the bottom of the pool. The water in this pool is not static but flowing. Therefore, the influence of ocean currents in the Yellow Sea can be simulated in this dynamic underwater scenario.

B. Experimental Results

We set up various experimental scenarios, and all data were collected by a custom-made AUV. The detailed results of the different experiments are as follows:

In the complex environment of the pool, the 2D-2D matching between adjacent frames of underwater images shows good performance. It can effectively predict the depth map and the consistency of forward and backward optical flow is also satisfactory. Detailed experimental results are shown in Fig. 4. The intermediate deep neural network outputs in Fig. 4 indicate that our proposed method effectively deals with featureless images and predicts the image depth.

Apart from testing the proposed network in complex areas, we also conducted tests in non-complex areas within the pool. Detailed experimental results are shown in Fig. 5. There are no larger targets as references in the environment, only some smaller reference objects such as sea cucumbers and sea urchins. In addition, the image has a phenomenon of featureless images in this area. In this scenario, the 2D-2D matching between adjacent frames of underwater images also shows good performance. From the experimental results, it can be seen that the depth prediction performance and optical flow consistency in non-complex environments at the bottom of the pool are not as good as those in complex environments. Especially in the prediction results of forward and backward optical flow, there is almost no significant difference through the images.

To validate the performance of the proposed monocular visual trajectory estimation system, we analyzed the complexity of the proposed scheme by using NVIDIA Jetson

[1] https://github.com/AlwinWang/AUV-Trajectory-Estimation

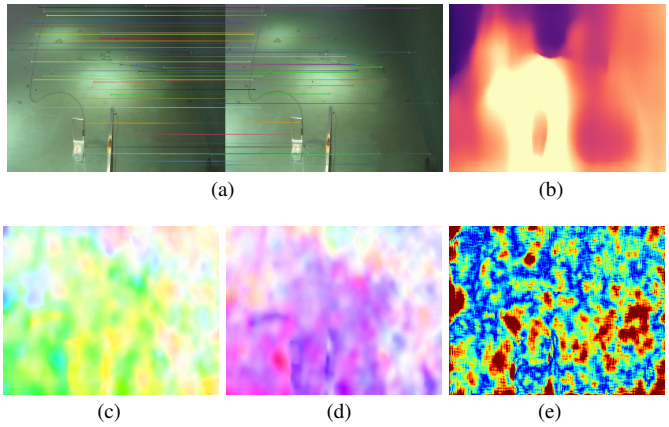


Fig. 4. Intermediate deep neural network outputs of underwater visual trajectory estimation in the complex environment of the pool: (a) Current and previous monocular underwater images with examples of auto-selected 2D-2D matches; (b) Monocular depth prediction; (c) Forward optical flow prediction; (d) Backward optical flow prediction; (e) Forward-backward flow consistency, red means high inconsistency, and blue means low inconsistency.

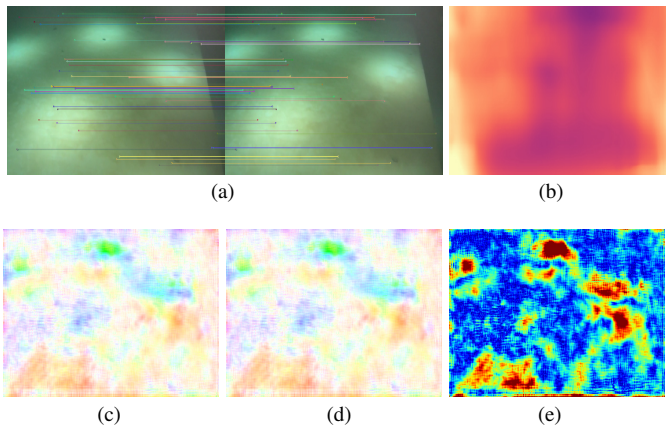


Fig. 5. Intermediate deep neural network outputs of underwater visual trajectory estimation in the uncomplicated environment of the pool: (a) Current and previous monocular underwater images with examples of auto-selected 2D-2D matches; (b) Monocular depth prediction; (c) Forward optical flow prediction; (d) Backward optical flow prediction; (e) Forward-backward flow consistency, red means high inconsistency, and blue means low inconsistency.

TX2 in experimental scenarios. The evaluation considers the average processing time, CPU and memory load, and Table II summarizes the algorithm efficiency. We also analyzed the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The RMSE of our system is 0.44 m and the MAE is 0.51 m. Meanwhile, we compare our method with some related underwater localization methods proposed by Zhao et al. [17], Jung et al. [45], Burguera et al. [46], Lee et al. [47] and Wang et al. [48], all of which are monocular vision-based localization for AUVs. The comparison of localization accuracy and average processing time is shown in Table III. From the results, it is evident that the proposed method exhibits a smaller RMSE compared to other works. Although our average processing time is not the lowest, it remains adequate to fulfil the operational tasks in underwater scenarios. Fig. 6 shows the localization performance of the proposed visual trajectory estimation. Due to our method only using a monocular camera and not including a loop closure detection module, it does not form a complete closed trajectory when the AUV returns to its initial point. Fig. 7 illustrates the cumulative distribution functions (CDFs) of the localization errors.

TABLE II
STATISTICS FOR ALGORITHM EFFICIENCY

Average Processing Time (ms)		CPU Usage (%)	Memory Usage (%)
Feature Detector	Trajectory Estimation		
50	327	36	52

Note: CPU Usage is represented as a percentage of a single CPU core on the given platform. Memory Usage is represented as a percentage of the available RAM on the given platform.

In order to comprehensively evaluate the proposed method, we have also carried out experiments in the Yellow Sea area. In this complex marine environment, it is quite hard for us to get the ground truth. We use the designed AUV to intuitively evaluate the proposed visual trajectory estimation by navigating around the visible rock piles or sea urchins. Fig. 8 shows the results of the proposed network in the rock pile environment of the Yellow Sea. From Fig. 8 (a), it can be seen that the 2D-2D matching effect between adjacent frames of underwater images is still good. Fig. 8 (c) and Fig. 8 (d) demonstrate effective forward and backward optical flow, while the inconsistency between forward and backward optical flow is also relatively low in Fig. 8 (e). Based on the above results, both our proposed optical flow network and depth prediction network can achieve good performance.

Apart from testing the proposed network in the rock pile environment of the Yellow Sea, we also conducted tests in the complex environment of the Yellow Sea. This environment mainly contains rocks plies, sea urchins, and aquatic plants, as shown in Fig. 9. From Fig. 9 (a), it can be seen that the 2D-2D matching effect between adjacent frames of underwater images can also achieve good results. Especially in Fig.9 (b), the colour in the upper right corner is mostly black, indicating that the network predicts deeper depths in this area. This region corresponds to a farther distance in the input image, suggesting that the network can predict a correct depth map. Fig.9 (c) and Fig.9 (d) also demonstrate effective forward and backward optical flow. From Fig.9 (e), it can be seen that the optical flow consistency of the network in the complex environment of the Yellow Sea is not as good as in the rock pile environment, but the inconsistency between forward and backward optical flow is also relatively low. Fig. 10 shows the experimental scenario and the trajectory generated by the proposed visual trajectory estimation in the Yellow Sea area. The above experiments indicate that the proposed depth-consistent monocular visual trajectory estimation method for AUVs can also be applied to complex marine environments.

C. The Picking Application of Visual Trajectory Estimation

In underwater scenarios, accurate trajectory estimation of AUV provides the fundamental functionality for the au-

TABLE III
COMPARISON OF LOCALIZATION ACCURACY AND AVERAGE PROCESSING TIME

Method	Zhao et al. [17]	Jung et al. [45]	Burguera et al. [46]	Lee et al. [47]	Wang et al. [48] (Without IMU)	Ours
RMSE (m)	0.94	0.67	0.83	0.52	0.57	0.44
Processing Time (ms)	372	-	233	-	305	327

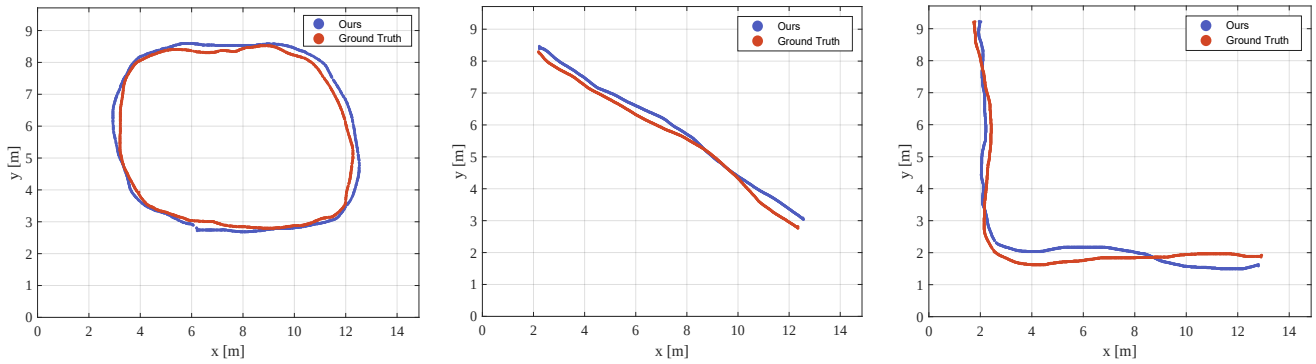


Fig. 6. Comparison of trajectory estimation results (blue) and ground truth (red). Different trajectories are shown in (a), (b) and (c). Best viewed in color.

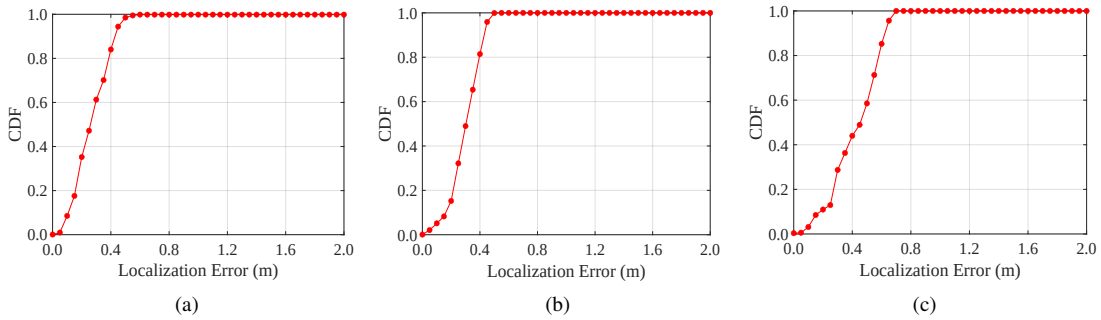


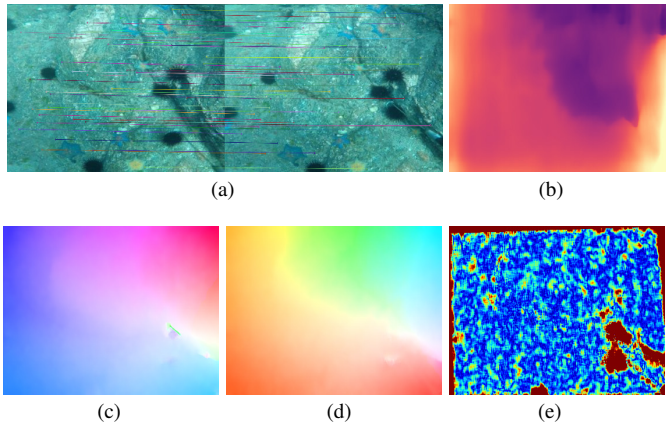
Fig. 7. CDF of localization errors. The localization errors corresponding to different trajectories in Fig. (6) are shown in (a), (b) and (c).

tonomous picking system. To complete autonomous object picking tasks in marine ranching, we have developed a low-cost, agile, and portable underwater autonomous picking system and carried out underwater experiments in the Yellow Sea, where the seafloor mainly contains a large number of rocks, scallops, sea urchins and sea cucumbers. We leverage the YOLOv5 model to perform real-time detection of marine objects, and the detection results are shown in Fig. 11. The blue rectangle at the bottom of the image represents the defined range area. When a detected object is within this set area, an autonomous picking operation is performed. The complete object picking steps are shown in Fig. 12. The experimental results show that the autonomous picking success rate is 80%, and the average time for the system to complete a cycle of picking tasks is about 26 s.

D. Influence of different experimental conditions on results

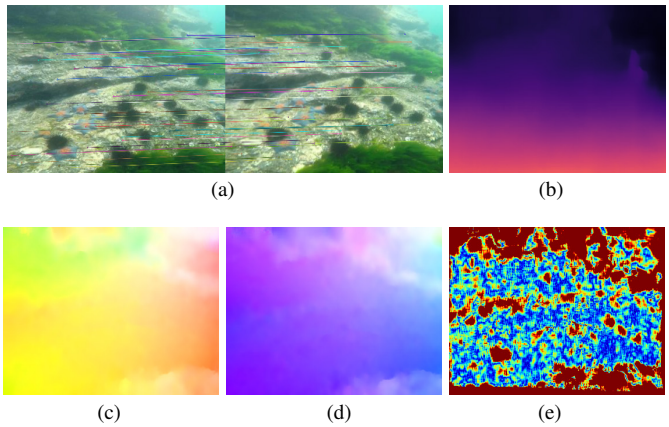
In order to compare the influence of different conditions on the experimental results, we tested the proposed system at different environments. In complex underwater environments, significant variations in lighting conditions can pose challenges for underwater visual trajectory estimation. Specifically, during experiments in the fire pool experiments, different environmental conditions led to variations in the trajectory estimation accuracy. When the natural light is sufficient, the average accuracy of trajectory estimation is 0.44 m. When the natural light is insufficient, it is necessary to use auxiliary LED illumination to supplement the brightness and provide sufficient illumination for target identification and visual trajectory estimation. However, using LED illumination can create some reflective areas in underwater environments, as shown in Fig. 5 (a). These areas show a brighter effect, with some pixels exhibiting particularly bright and shaded areas. The assumption that the grayscale remains unchanged is easily unfounded in such complex environments. This phenomenon may affect the output results of optical flow estimation and depth prediction, leading to lower accuracy of underwater visual trajectory estimation. In low light conditions, the trajectory estimation accuracy is 0.79 m. Sometimes, after a sudden right-angle turn, the camera is prone to fail to track the feature points, resulting in AUV trajectory estimation failure.

In addition, through many experiments, we found that



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

Fig. 8. Intermediate deep neural network outputs of underwater visual trajectory estimation in the rock pile environment of the Yellow Sea: (a) Current and previous monocular underwater images with examples of auto-selected 2D-2D matches; (b) Monocular depth prediction; (c) Forward optical flow prediction; (d) Backward optical flow prediction; (e) Forward-backward flow consistency, red means high inconsistency, and blue means low inconsistency.



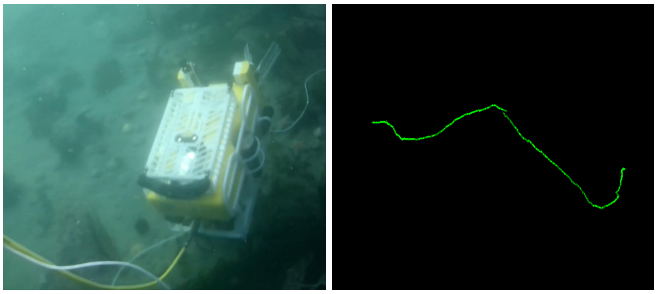
(a) (b) (c) (d) (e)

Fig. 9. Intermediate deep neural network outputs of underwater visual trajectory estimation in the complex environment of the Yellow Sea: (a) Current and previous monocular underwater images with examples of auto-selected 2D-2D matches; (b) Monocular depth prediction; (c) Forward optical flow prediction; (d) Backward optical flow prediction; (e) Forward-backward flow consistency, red means high inconsistency, and blue means low inconsistency.

the optical flow network also has high requirements for the navigation speed of AUVs. Optical flow is the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer and a scene. If the camera carried by the AUV suddenly moves too fast, the optical flow estimation may fail, thereby affecting the accuracy of the visual trajectory estimation. Therefore, it is necessary to pay attention to the navigation speed of AUVs to ensure the accuracy of visual trajectory estimation.

VI. Conclusion

In this paper, to solve the featureless images and depth ambiguity problems caused by underwater multiple mediums environments, we proposed a depth-consistent monocular visual trajectory estimation for AUVs. It is a novel underwater monocular visual trajectory estimation scheme through self-supervised learning. Specifically, we integrate bidirectional



(a) (b)

Fig. 10. Experimental results of AUV monocular visual trajectory estimation in the Yellow Sea area. (a) The experimental scenario. (b) Trajectory estimated by monocular vision.

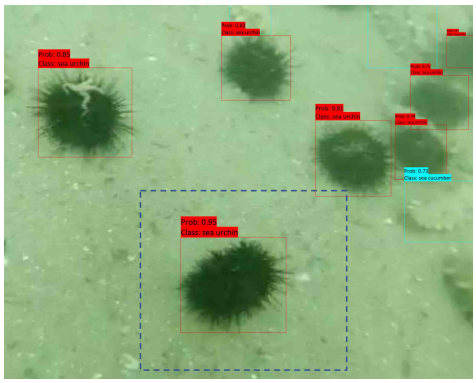


Fig. 11. The detection results of marine object. We set a range area with a blue rectangle at the bottom of the image. When a detected object is within this set area, an automatic picking operation is performed.

optical flow consistency and epipolar constraints strategy to address the trajectory estimation issue of AUVs in featureless scenarios. Then, we solve the depth ambiguity problem in the monocular vision by aligning geometrically triangulated depths to the scale-consistent deep depths. In addition, we develop a low-cost, agile, and portable underwater autonomous picking system with real-time trajectory estimation capabilities, and carry out extensive maritime scenarios experiments to test its performance. Experimental results show that the proposed method can implement the localization task with an accuracy of 0.44 m, and the average processing time of trajectory estimation is 327 ms. In particular, our proposed trajectory estimation method provides an effective localization solution for underwater autonomous picking tasks.

In future work, we will focus on the theoretical analysis of the proposed algorithm. It should be noted that there is still a lot of work to be done for underwater autonomous picking tasks. For instance, how to improve the accuracy of visual trajectory estimation in challenging unstable marine scenarios? How to improve the stability and picking efficiency of AUVs? We will try to address these challenges in our future research.

References

[1] X. Yu, H. Qin, and Z. Zhu, "Underwater localization of AUVs in motion using two-way travel time measurements with unknown sound velocity," *IEEE Trans. Veh. Technol.*, vol. 72, no. 9, pp. 11 358–11 373, Sep. 2023.

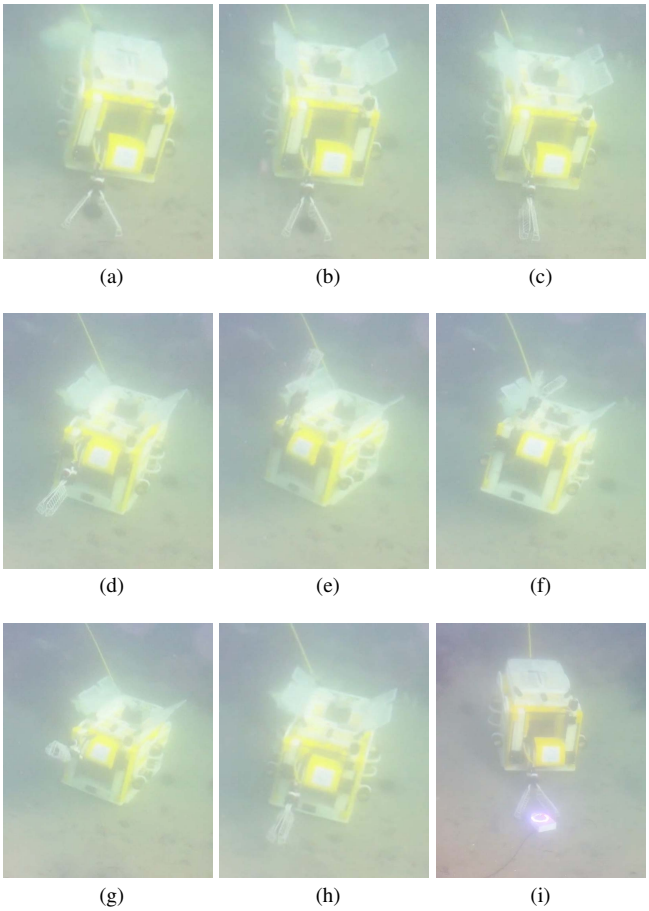


Fig. 12. Snapshot of the AUV picking system. (a) Approaching the sea cucumber. (b) Opening the collection box. (c) Closing the clamp. (d) Moving up the clamp. (e) Turning over the clamp. (f) Opening the clamp and putting down the sea cucumber. (g) Turning over the clamp. (h) Moving down the clamp. (i) Closing the collection box and approaching the next target. The video of AUV picking in marine environments can be found here: https://www.youtube.com/shorts/I8ySycUgVGQ.

[2] Y. Shi, W. Xie, G. Zhang, H. Dong, and W. Zhang, "Event-triggered saturation-tolerant control for autonomous underwater vehicles with quantitative transient behaviors," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 9857–9867, Aug. 2023.

[3] H. Wang, G. Han, A. Gong, A. Li, and Y. Hou, "A backbone network construction-based multi-AUV collaboration source location privacy protection algorithm in UASNs," *IEEE Internet Things J.*, vol. 10, no. 20, pp. 18 198–18 210, Oct. 2023.

[4] M. Cheng, Q. Guan, F. Ji, J. Cheng, and Y. Chen, "Dynamic-detection-based trajectory planning for autonomous underwater vehicle to collect data from underwater sensors," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13 168–13 178, Aug. 2022.

[5] Y. Wang, D. Gu, X. Ma, J. Wang, and H. Wang, "Robust real-time AUV self-localization based on stereo vision-inertia," *IEEE Trans. Veh. Technol.*, vol. 72, no. 6, pp. 7160–7170, Jun. 2023.

[6] X. Mei, D. Han, N. Saeed, H. Wu, B. Han, and K.-C. Li, "Localization in underwater acoustic IoT networks: Dealing with perturbed anchors and stratification," *IEEE Internet Things J.*, vol. 11, no. 10, pp. 17 757–17 769, May. 2024.

[7] C. Zhang, W. Shi, Z. Gong, Q. Zhang, and C. Li, "Trajectory optimization for target localization using time delays and Doppler shifts in bistatic sonar-based Internet-of-Underwater-Things," *IEEE Internet Things J.*, vol. 10, no. 18, pp. 16 427–16 439, Sep. 2023.

[8] Y. Su, Y. Xu, Z. Pang, Y. Kang, and R. Fan, "HCAR: A hybrid coding-aware routing protocol for underwater acoustic sensor networks," *IEEE Internet Things J.*, vol. 10, no. 12, pp. 10 790–10 801, Jun. 2023.

[9] Y. Wang, X. Ma, J. Wang, S. Hou, J. Dai, D. Gu, and H. Wang, "Robust AUV visual loop closure detection based on variational auto-encoder network," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 8829–8838, Dec. 2022.

[10] Z. Hu, W. Qi, K. Ding, G. Liu, and Y. Zhao, "An adaptive lighting indoor vSLAM with limited on-device resources," *IEEE Internet Things J.*, vol. 11, no. 17, pp. 28 863–28 875, Sep. 2024.

[11] D. Shen, G. Liu, T. Li, F. Yu, F. Gu, K. Xiao, and X. Zhu, "ORB-NeuroSLAM: a brain-inspired 3D SLAM system based on ORB features," *IEEE Internet Things J.*, vol. 11, no. 7, pp. 12 408–12 418, Apr. 2023.

[12] J. Chen, Q. Li, W. Zhang, B. Wang, and D. Zhang, "A degradation-robust keyframe selection method based on image quality evaluation for visual localization," *IEEE Internet Things J.*, vol. 11, no. 10, pp. 18 421–18 434, May. 2024.

[13] K. Ebadi, L. Bernreiter, H. Biggie, G. Catt, Y. Chang, A. Chatterjee, C. E. Denniston, S.-P. Deschênes, K. Harlow, S. Khattak *et al.*, "Present and future of SLAM in extreme environments: The DARPA SubT challenge," *IEEE Trans. Robot.*, vol. 40, no. 0, pp. 936–959, Oct. 2023.

[14] P. Huang, L. Zeng, X. Chen, K. Luo, Z. Zhou, and S. Yu, "Edge robotics: Edge-computing-accelerated multirobot simultaneous localization and mapping," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 14 087–14 102, Aug. 2022.

[15] W. He, Z. Lu, X. Liu, Z. Xu, J. Zhang, C. Yang, and L. Geng, "A real-time and high precision hardware implementation of RANSAC algorithm for visual SLAM achieving mismatched feature point pair elimination," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 71, no. 11, pp. 5102–5114, Nov. 2024.

[16] R. Gou, G. Chen, X. Pu, X. Liao, and R. Chen, "A visual SLAM with tightly-coupled integration of multi-object tracking for production workshop," *IEEE Internet Things J.*, vol. 11, no. 11, pp. 19 949–19 962, Jun. 2024.

[17] C. Zhao, H. Dong, J. Wang, T. Qiao, J. Yu, and J. Ren, "Dual-type marker fusion-based underwater visual localization for autonomous docking," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–11, Nov. 2023.

[18] W. Zhao, S. Liu, Y. Shu, and Y.-J. Liu, "Towards better generalization: Joint depth-pose learning without posenet," in *Proc. IEEE Int. Conf. Comp. Vis. Patt. Recogn.*, 2020, pp. 9151–9161.

[19] H. Zhan, C. S. Weerasekera, J.-W. Bian, and I. Reid, "Visual odometry revisited: What should be learnt?" in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 4203–4210.

[20] X. Ma, S. Hou, Y. Wang, J. Wang, and H. Wang, "Multiscale and dense ship detection in SAR images based on key-point estimation and attention mechanism," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–11, Jan. 2022.

[21] B. Xu, J. Hu, and Y. Guo, "An acoustic ranging measurement aided SINS/DVL integrated navigation algorithm based on multivehicle cooperative correction," *IEEE Trans. Instrum. Meas.*, vol. 71, no. 1, pp. 1–15, Aug. 2022.

[22] P. Vial, N. Palomeras, J. Solà, and M. Carreras, "Underwater pose SLAM using GMM scan matching for a mechanical profiling sonar," *J. Fields Robot.*, vol. 41, no. 1, pp. 511–538, Dec. 2023.

[23] Z. Dong, W. Yan, R. Cui, L. Lei, and Y. He, "An adaptive grouping sonar-inertial odometry for underwater navigation," *Ocean Engineering*, vol. 294, p. 116688, Feb. 2024.

[24] X. Zhang, B. He, and S. Gao, "An integrated navigation method for small-sized AUV in shallow-sea applications," *IEEE Trans. Veh. Technol.*, vol. 72, no. 3, pp. 2878–2890, Mar. 2022.

[25] J. Guo, S. Song, J. Liu, H. Chen, B. Lin, and J.-H. Cui, "An efficient Geo-routing aware MAC protocol based on OFDM for underwater acoustic networks," *IEEE Internet Things J.*, vol. 10, no. 11, pp. 9809–9822, Jun. 2023.

[26] G. Han, Z. Zhou, Y. Zhang, M. Martínez-García, Y. Peng, and L. Xie, "Sleep-scheduling-based hierarchical data collection algorithm for gliders in underwater acoustic sensor networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9466–9479, Sep. 2021.

[27] T. Qiu, Y. Li, and X. Feng, "Distributed channel sensing MAC protocol for multi-UUV underwater acoustic network," *IEEE Internet Things J.*, vol. 11, no. 9, pp. 16 119–16 133, May. 2024.

[28] Y. Wang, S. Song, X. Guo, J. Liu, Q. Ye, and J. Cui, "An efficient localization scheme with velocity prediction for large-scale underwater acoustic sensor networks," *IEEE Internet Things J.*, vol. 11, no. 4, pp. 6508–6520, Feb. 2024.

[29] Y. Su, L. Guo, Z. Jin, and X. Fu, "A mobile-beacon-based iterative localization mechanism in large-scale underwater acoustic sensor networks," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3653–3664, Mar. 2020.

[30] J. Yan, D. Guo, X. Luo, and X. Guan, "AUV-aided localization for underwater acoustic sensor networks with current field estimation," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8855–8870, Aug. 2020.
[31] J. He, M. Li, Y. Wang, and H. Wang, "OVD-SLAM: An online visual SLAM for dynamic environments," *IEEE Sensors J.*, vol. 23, no. 12, pp. 13 210–13 219, Jun. 2023.
[32] M. Li, J. He, Y. Wang, and H. Wang, "End-to-end RGB-D SLAM with multi-MLPs dense neural implicit representations," *IEEE Robotics Autom. Lett.*, vol. 8, no. 11, pp. 7138–7145, Nov. 2023.
[33] C. Chen, B. Wang, C. X. Lu, N. Trigoni, and A. Markham, "Deep learning for visual localization and mapping: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 12, pp. 17 000–17 020, Dec. 2023.
[34] H. Cui, X. Gao, and S. Shen, "MCSfM: Multi-camera-based incremental Structure-from-Motion," *IEEE Trans. Image Processing*, vol. 32, pp. 6441–6456, Nov. 2023.
[35] Z. Ye, C. Bao, X. Zhou, H. Liu, H. Bao, and G. Zhang, "EC-SfM: Efficient covisibility-based structure-from-motion for both sequential and unordered images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 1, pp. 110–123, Jan. 2024.
[36] J. Jiang, J. Yuan, X. Zhang, and X. Zhang, "DVIO: an optimization-based tightly coupled direct visual-inertial odometry," *IEEE Trans. Ind. Electron.*, vol. 68, no. 11, pp. 11 212–11 222, Nov. 2021.
[37] H. Luo, G. Li, D. Zou, K. Li, X. Li, and Z. Yang, "UAV navigation with monocular visual inertial odometry under GNSS-denied environment," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–15, Oct. 2023.
[38] W. Wang, B. Joshi, N. Burgdorfer, K. Batsosc, A. Q. Lid, P. Mordohaia, and I. Rekleitisb, "Real-time dense 3D mapping of underwater environments," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2023, pp. 5184–5191.
[39] Z. Fan, Z. Song, J. Xu, Z. Wang, K. Wu, H. Liu, and J. He, "Object level depth reconstruction for category level 6D object pose estimation from monocular RGB image," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 220–236.
[40] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[41] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.
[42] T.-W. Hui, X. Tang, and C. C. Loy, "Liteflownet: A lightweight convolutional neural network for optical flow estimation," in *Proc. IEEE Int. Conf. Comp. Vis. Patt. Recogn.*, 2018, pp. 8981–8989.
[43] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[44] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2019, pp. 3828–3838.
[45] J. Jung, Y. Lee, D. Kim, D. Lee, H. Myung, and H. T. Choi, "AUV SLAM using forward/downward looking cameras and artificial landmarks," in *Proc. IEEE Underwater Technol.*, 2017, pp. 1–3.
[46] A. Burguera Burguera and F. Bonin-Font, "A trajectory-based approach to multi-session underwater visual SLAM using global image signatures," *J. Mar. Sci. Eng.*, vol. 7, no. 8, p. 278, Aug. 2019.
[47] D. Lee, D. Kim, S. Lee, H. Myung, and H. T. Choi, "Experiments on localization of an AUV using graph-based SLAM," in *Proc. 10th Int. Conf. Ubiquitous Robots Ambient Intell.*, 2013, pp. 526–527.
[48] Y. Wang, X. Ma, J. Wang, and H. Wang, "Pseudo-3D vision-inertia based underwater self-localization for AUVs," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 7895–7907, Jul. 2020.

Yangyang Wang (M'22) received the Ph.D. degree in signal and information processing from the Dalian University of Technology, Dalian, China, in 2023. From October 2021 to October 2022, he was a Visiting Fellow with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, U.K. He is currently a Postdoctoral Researcher at Dalian Maritime University, Dalian, China. His research interests include underwater localization, underwater simultaneous localization and mapping, robotics, machine learning and computer vision.

Xiaokai Liu received her B.S. degree from School of Information Engineering, Dalian Maritime University (DMU), PR China, in 2010, and Ph.D degree from Dalian University of Technology, China, in 2017, all in Information and Communication Engineering. She is currently an associate professor in DMU. From September 2013 to March 2015, she was a visiting researcher with University of California at Merced. Her research interests include computer vision, device-free activity recognition and multi-modal learning.

Dongbing Gu (SM'07) received the B.Sc. and M.Sc. degrees in control engineering from the Beijing Institute of Technology, Beijing, China, in 1985 and 1988, respectively, and the Ph.D. degree in robotics from the University of Essex, Colchester, U.K., in 2004.

From October 1996 to October 1997, he was an Academic Visiting Scholar with the Department of Engineering Science, University of Oxford, Oxford, U.K. In 2000, he joined the University of Essex as a Lecturer. He is currently a Professor of Robotics with the School of Computer Science and Electronic Engineering, University of Essex. His research interests include robotics, multiagent systems, cooperative control, model predictive control, visual simultaneous localization and mapping, wireless sensor networks, and machine learning.

Jie Wang (M'12, SM'18) received his B.S. degree from Dalian University of Technology, Dalian, China, in 2003, M.S. degree from Beihang University, Beijing, China, in 2006, and Ph.D degree from Dalian University of Technology, Dalian, China, in 2011, all in Electronic Engineering. He is currently a Professor with the School of Information Science and Technology, Dalian Maritime University. Before that, he worked at Dalian University of Technology from 2011 to 2021. He was a visiting researcher with University of Florida from 2013 to 2014. His research interests include intelligent wireless sensing, machine learning, wireless localization and tracking, ISAC, and cognitive radio networks. He is an Editor for ACM Computing Surveys and IEEE Transactions on Cognitive Communications and Networking, and was an Editor for IEEE Transactions on Vehicular Technology.

Xianping Fu is a Full Professor at Dalian Maritime University, China. He received a Ph.D. degree from Dalian Maritime University in 2005. He previously worked as a Postdoctoral Researcher at Tsinghua University in 2008 and Senior Research Fellow at Harvard University. His major research interests are image processing for content recognition, multimedia technology and underwater robot vision.

He has authored over 100 journal and conference papers in these areas, which have been published in IJCAI, TMM, TITS, TCSVT, TVT, ICME, ICMR, OCEANs, etc. He won the American RPB international scholar research award in 2009. His group was included in the Liaoning Revitalization Talents Program. Now, he is working as the Dean of the College of Information Science and Technology in Dalian Maritime University and director of the Liaoning Underwater Robot Engineering Research Center.