# Robust AUV Visual Loop Closure Detection Based on Variational Auto-Encoder Network

Yangyang Wang, Xiaorui Ma, *Member*, *IEEE*, Jie Wang, *Senior Member*, *IEEE*, Shilong Hou, Ju Dai, Dongbing Gu, *Senior Member*, *IEEE*, and Hongyu Wang, *Member*, *IEEE*

***Abstract*—The visual loop closure detection for Autonomous Underwater Vehicles (AUVs) is a key component to reduce the drift error accumulated in simultaneous localization and mapping tasks. However, due to viewpoint changes, textureless images, and fast-moving objects, the loop closure detection in dramatically changing underwater environments remains a challenging problem to traditional geometric methods. Inspired by strong feature learning ability of deep neural networks, we propose an underwater loop closure detection method based on a variational auto-encoder network in this paper. Our proposed method can learn effective image representations to deal with the challenges caused by dynamic underwater environments. Specifically, the proposed network is an unsupervised method, which avoids the difficulty and cost of labeling a great quantity of underwater data. Also included is a semantic object segmentation module, which is utilized to segment the underwater environments and assign weights to objects in order to alleviate the impact of fast-moving objects. Furthermore, an underwater image description scheme is used to enable efficient access to geometric and object-level semantic information, which helps to build a robust and real-time system in dramatically changing underwater scenarios. Finally, we test the proposed system under complex underwater environments and get a recall rate of 92.31% in the tested environments.**

***Index Terms*—AUV SLAM, loop closure detection, semantic segmentation, deep neural network.**

## I. Introduction

AUTONOMOUS Underwater Vehicles (AUVs) are now used for a variety of tasks, for instant, pipeline detecting, oceanographic survey, deep-sea exploration, and seafood products monitoring [1], [2]. AUV localization is such a significant task with wide applications that numerous methods have been proposed in recent years. Loop closure detection refers to the assignment of deciding whether or not a vehicle has, after an excursion of arbitrary length, returned to a previously visited location. It plays a key role in eliminating error accumulation over long-term operations. Correct loop detection results contribute to the relocation of AUVs or promote the algorithm to obtain more consistent and accurate results, as illustrated in Fig. 1, and similar demonstration is also given in [3].

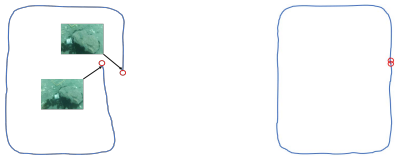


Fig. 1. An underwater trajectory map. (a) Without loop closure detection. (b) With loop closure detection.

Although some vision-based loop closure (or place recognition) algorithms [4], [5] were proposed to eliminate the drifts for terrestrial environments, they cannot directly handle the large-scale environment with real-world underwater data. In addition, due to the nature of underwater environments, changing viewpoints, blur motions, textureless images, and fast-moving objects further increase the difficulty of solving the problem. Therefore, exploring a significant loop closure detection method for AUVs has become a current research hotspot.

In previous works, many underwater localization methods relying on the use of high-accuracy underwater sensors, such as sonar, Doppler Velocity Log (DVL), and 3-axis compass, could provide good localization results, but the high cost limits their widespread deployments [1]. In recent years, underwater acoustic sensor networks (UASNs) have been extensively studied. Yan et al. [6] developed an AUV aided localization solution for UASNs, subject to asynchronous clock, stratification effect and mobility constraints in cyber channels. At the same time, there has been a growing interest in underwater optical wireless networks for localization. In [7], the authors investigated underwater relaying and routing techniques and provided their end-to-end performance analysis under the location uncertainty. Saeed et al. [8] proposed a robust 3D localization method for partially connected underwater optical wireless sensor networks, which can accommodate the outliers and optimize the placement of the anchors to improve the localization accuracy. Nevertheless, it requires a tedious and costly installation process before deployment, which is time consuming. Meanwhile, sensor nodes often have passive motions caused by water current or tide. In this case, it is hard

This work was supported in part by the National Natural Science Foundation of China under Grants 61801078, 62071081, and U1933104, and in part by the Fundamental Research Funds for the Central Universities under Grant DUT21GF204, and in part by the China Postdoctoral Science Foundation under Grant 2020M682827. Yangyang Wang is currently a Visiting Fellow with the University of Essex, supported by the China Scholarship Council program. (*Corresponding author: Hongyu Wang.*)

Yangyang Wang, Xiaorui Ma, Shilong Hou, and Hongyu Wang are with School of Information and Communication Engineering, Dalian University of Technology, Dalian, 116024, P.R. China (e-mail: yyw@mail.dlut.edu.cn, maxr@dlut.edu.cn, hsldlut@163.com, whyu@dlut.edu.cn).

Jie Wang is with the School of Information Science and Technology, Dalian Maritime University, Dalian 116026, P.R. China (e-mail: wang_jie@dlmu.edu.cn).

Ju Dai is with Peng Cheng Laboratory. No.2, Xingke 1st Street, Nanshan District, Shenzhen, Guangdong Province, P.R. China. (email:daij@pcl.ac.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)

to calculate the actual distance between the nodes, which has a great impact on localization accuracy.

Recent developments in vision research areas have witnessed dramatically increasing performance in localization on terrestrial applications. Most early loop closure detection methods for mobile robots are based on the assumption of appearance invariability. However, when they face a long-term self-localization task in outdoor environments, dynamic scene, illumination change, viewpoint varying, and seasonal variation will greatly reduce the precision and recall rate of system detection [9]. These issues pose a severe challenge to the robustness of algorithms. Without loop closure detection, the state estimation from visual simultaneous localization and mapping (SLAM) systems could lead to serious deviation for long-term and large-scale environments. A true positive loop closure detection can significantly reduce the error of the system, but a false positive loop closure detection could make the graph optimization algorithm converge a completely wrong value. Hence, correct loop closure detection is a crucial step to the optimization algorithm in a SLAM system.

Some underwater self-localization solutions have been proposed to address the challenge issues, but they determine the loop closure detection through hand-crafted landmarks. In [10], Jung et al. proposed a method that utilizes the optical camera measurement to implement an underwater SLAM system. Artificial landmarks (3D geometric landmarks, tag-type landmarks, etc) were arranged to appear in view of the forward and downward-looking cameras, which limits the wide application of AUVs in unknown scenarios. To get rid of the dependence on external equipment for the localization of AUVs, our previous work [11] developed a low-cost, portable, and small volume sensor suite for underwater vehicle localization. However, its loop closure detection module only uses the geometric information of images. Exploiting the information provided by images like semantic, geometric and visual appearance is still an open issue to be solved.

Traditional visual localization related methods are not robust under dramatically changing underwater environments. Firstly, the underwater environment has a single structure, repeated texture and poor specificity of descriptors. Secondly, traditional feature points have a requirement on the perspective of observation. When AUVs observe from a different perspective, it may appear to be a completely different feature point. Due to the influence of water flow, the angle change caused by AUVs during navigation is very disadvantageous to the identification of underwater feature points. Therefore, it is not enough to rely on geometric information only. Fortunately, neural networks are proved to be more efficient in feature extraction. And semantic features, such as underwater stones, are salient both spatially and temporally. Currently, most deep learning based methods require users-chosen parameters for loop closure detection. Moreover, the size of datasets with annotated ground-truth labels is limited and they are expensive to collect.

Inspired by successful applications of deep networks in feature extraction, we propose an underwater variational auto-encoder (UVAE) network, which can robustly detect loop closures in dramatically changing underwater environments. We have successfully implemented the proposed network on the NVIDIA TX2 platform. Experimental results demonstrate that our method produces a high recall rate of 0.92 on the fire pool underwater dataset and demonstrates a competitive performance. We summarize our contributions as follows:

- We propose an underwater loop closure detection method based on a variational auto-encoder network, which can learn effective image representations to solve the robustness problem in complex dynamic underwater scenarios.
- We propose an unsupervised semantic object segmentation method to avoid the dependence on large-scale underwater labeled dataset, which can segment the underwater scenario and assign weights to objects, so as to alleviate the impact of fast-moving objects on detection.
- We propose an underwater image description scheme that enables efficient access to geometric and object-level semantic information. This improves the robustness and real-time performance of the system in challenging underwater scenarios.

The paper is structured as follows. Section II gives a brief introduction of some related works. Section III presents an overview of the proposed robust AUV visual loop closure detection based on variational auto-encoder network. In Section IV, we introduce the proposed unsupervised variational auto-encoder network, including network architecture and network training. Section V describes the implementation details and experimental results on the self-collected underwater datasets in the fire pool as well as the Yellow Sea. Finally, the conclusions and directions for future work are summarized in Section VI.

## II. Related Work

For harsh underwater environments, the localization of AUVs is the basis for the completion of oceanographic survey, real-time seabed mapping, and routine seafood product monitoring. At present, the great majority of underwater navigation algorithms are based on acoustic sensors such as sonar, long baseline (LBL), ultrashort baseline (USBL), and DVL [1]. At the same time, the UASNs have received much attention due to a wide range of applications including marine resources exploration and ocean data acquisition [2]. However, it is too expensive to use sonar, USBL and DVL to collect data for fishery work. Sometimes it also takes long time to deploy the underwater acoustic sensor nodes equipment.

As far as vision is concerned, there are relatively few researches on underwater environments. Even in the field of underwater image processing, a large proportion is in the study of underwater image enhancement (UIE). An UIE algorithm mainly includes: histogram equalization and fusion techniques, as well as data-driven frameworks with end-to-end convolutional neural networks (CNNs) to learn essential parameters or transmission maps from the degraded inputs. CNNs do have the ability to deal with special circumstances [12], [13], for instance, Woźniak et al. [13] introduced a soft tree decision architecture which works well for decision processes over cultural heritage. With the increasing attention to the development of marine resources, Liu et al. [14]

established a benchmark dataset with real world sea images. But in general, current underwater datasets are rare, which could be one of the reasons why there are not many vision-based underwater localization studies.

Recent developments in visual localization research have witnessed dramatically increasing performance in service robots [15], autonomous micro aerial vehicles, autonomous driving and augmented reality, which promoted the rapid development of SLAM technology. With the advance of deep learning, a large number of learning based methods have emerged. In [16], the authors used a learning based method to solve the problem of estimating the 6D pose of specific objects from a single RGB-D image. Visual localization is closely related to a wide range of applications, especially re-localization [17], loop closure detection in SLAM and Structure-from-Motion systems [18]. In the field of visual place recognition, the state-of-the-art system is NetVLAD [4]. In this method, local descriptors, a flat convolution feature map, are assigned to different local learning clusters to construct global image descriptors. It is a highly versatile system that represents the state-of-the-art accuracy, but it fails to provide the ability to close the loop precisely.

Semantic information has the potential to make positioning systems more robust. Naseer et al. [19] proposed a learning approach to robust binary segmentation and feature aggregation of deep networks, which exploits the image content to create a dense and salient scene description. In this work, the authors provided a coarsely labeled dataset for semantic saliency in dynamic and perceptually changing urban environments which captures long-term weather and structural changes. Kingma et al. [20] introduced a stochastic variational inference and learning algorithm and proved that a reparameterization of the variational lower bound yields a lower bound estimator that can be straightforwardly optimized using standard stochastic gradient methods.

Long term unmanned equipment operation shows that the appearance change may be an important factor in visual place recognition failures. Cadena et al. [21] proposed a place recognition algorithm for SLAM systems using stereo cameras that considers both appearance and geometric information of points of interest in images. In [22], the authors proposed how to choose the best environment similarity criterion, that is, if each environment is described by its co-occurrence characteristics, then the similarity between environments can be determined by comparing their co-occurrence matrices. Similarly, Chen et al. [23] trained two CNN architectures to complete specific place recognition tasks, and used a multi-scale feature coding method to generate condition invariant and viewpoint invariant features.

To avoid wasting time using pixel-level supervision, Gao et al. [24] presented a network based on the stacked denoising auto-encoder (SDA) that learns a nonlinear representation of raw image in an unsupervised way. But it is mainly used to handle indoor environments. In [25], the authors proposed an unsupervised auto-encoder network for robust and fast loop closure detection. However, it does not take advantage of the semantic information, and a threshold needs to be set to determine a true positive. In the field of self-supervised, Schönberger et al. [26] proposed a generative model of visual localization based on 3D geometric and semantic information. Yet, it takes around one second per frame to relocalise, which is a bit slow. Merrill et al. [27] proposed a new auto-encoder network called CALC2.0, but the network does not apply for underwater applications.

In view of the rapid growth of underwater resource development and utilization, it is necessary to accomplish an accurate and efficient localization solution for AUVs. However, visual loop closure detection of AUVs is known to be a complex and challenging problem. Corke et al. [28] compared acoustic and visual methods for underwater localization showing the viability of using visual methods in some underwater scenarios. In order to solve the robustness problem in complex dynamic underwater scenarios, our previous work [11] proposed an underwater self-localization method based on Pseudo-3D vision-inertia for AUVs, which merges depth information into 2D visual image to achieve continuous and robust localization. In addition, we also optimized the 4 DOF pose graph to enhance the global consistency and designed an online loop closure detection module to realize the relocalization. However, our previous work did not leverage the semantic information in underwater images.

Different from other existing works, our proposed method, as described in the following section, integrates semantic, geometric and visual appearance information from underwater images to improve the performance of underwater self-localization estimation. In particular, the proposed network is an unsupervised method, which avoids the difficulty and cost of labeling a great quantity of underwater data.

## III. System Overview

The architecture of the proposed robust AUV visual loop closure detection based on variational auto-encoder network is shown in Fig. 2. The whole network starts with the input of training data, and then we send a pair of images to the encoder part of the network at a time. After passing through a $3 \times 3$ convolution layer, the convolutional features is fed into five blocks of two convolution, each with a max pooling behind it. Then, two $1 \times 1$ convolution layers are used to computer the latent variables $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\sigma}^{(i)}$. In the second half of the network, the latent variable is divided into $K+1$ local descriptors. One is used to reconstruct the RGB image, while the remaining $K$ are sent to the decoder for concatenation and then used to predict a full resolution semantic segmentation label.

We choose to extract keypoints from the maximally activated regions outputted from the *conv5* layer. After the meaningful number of keypoints is computed, the proposed network extracts the keypoint descriptors. Next, the network extracts the residual of feature vectors around the point of interest in a $3 \times 3$ window. Then, we connect the residuals above to obtain the keypoint descriptors. Here, the Euclidean distance metric is used to compare these descriptors during keypoint matching. Once keypoints are matched, a loop closure is detected. The detection process is as follows: Firstly, we perform a K-Nearest Neighbor search with $k$ = 5 on the global image

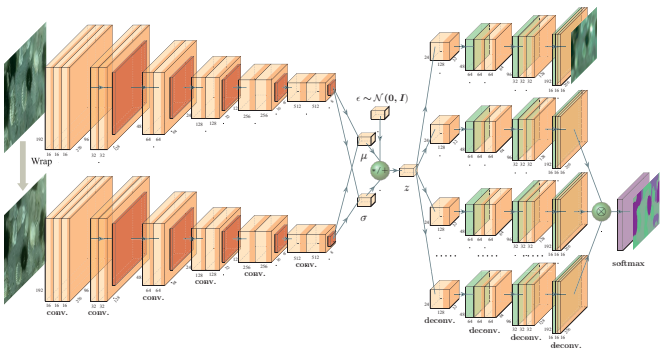



Fig. 2. The architecture of the proposed network.

descriptor database. Next, we filter $k$ candidates by matching keypoints. If a valid fundamental matrix can be computed with the matched keypoints using the Random Sample Consensus (RANSAC) algorithm [29], a loop closure is considered to be detected. Otherwise, if there are not enough valid matches to estimate a fundamental matrix, the candidate will be rejected. The details of network architecture and its training will be introduced in Section IV. Table I shows the main notations used in this paper.

TABLE I
NOTATION DEFINITIONS

| Name | Description |
|---|---|
| $N_i$ | The set of training images |
| $H$ | The height of the image |
| $W$ | The width of the image |
| $\mathcal{D}_{KL}$ | The Kullback-Leibler (KL) divergence |
| $T$ | The number of descriptors corresponding to the decoder |
| $K$ | The number of concatenation sent to the decoder |
| $\odot$ | The element-wise product |
| $D_S$ | The dense pixel-wise semantic segmentation image |
| $L_{\text{sum}}$ | The overall objective function |

## IV. UNSUPERVISED VARIATIONAL AUTO-ENCODER NETWORK

This section describes the proposed unsupervised UVAE network. In general, our model learns the latent variables in an underwater image through the UVAE network, which can extract more abundant features and make the model more accurate. The following two subsections will introduce the network in detail.

### A. Network Architecture

We consider the image dataset $X = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ to be used is composed of $N$ independent and identically distributed sample variable $\boldsymbol{x}$. We assume that the data is generated by a random process, which contains an unobservable continuous random variable $\boldsymbol{z}$. With the variational bound of our model, the marginal likelihood consists of the sum of marginal likelihoods of individual data points $\log p_\theta(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots, \boldsymbol{x}^{(N)}) = \sum_{i=1}^{N} \log p_\theta(\boldsymbol{x}^{(i)})$, which can be written as:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}^{(i)}) = \mathcal{D}_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{x}^{(i)})) \\ + \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}^{(i)})\end{aligned} \tag{1}$$

where the first right-hand side term is the Kullback-Leibler (KL) divergence of the approximate from the true posterior. Due to the KL divergence is non-negative, $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}^{(i)})$ is called the (variational) lower bound on the marginal likelihood of data point $i$, the formula can be written as:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}^{(i)}) \geq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}^{(i)}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[-\log q_\phi(\boldsymbol{z}|\boldsymbol{x}) \\ + \log p_\theta(\boldsymbol{x}, \boldsymbol{z})]\end{aligned} \tag{2}$$

The above equation can also be written as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}^{(i)}) = -\mathcal{D}_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)}) \parallel p_\theta(\boldsymbol{z})) \\ + \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})}[\log p_\theta(\boldsymbol{x}^{(i)}|\boldsymbol{z})]\end{aligned} \tag{3}$$

$\mathcal{D}_{KL}$ represents the KL divergence, $\theta$ represents the generative parameters, and $\phi$ is the variational parameters. $q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})$ and $p_\theta(\boldsymbol{z})$(the prior information) are both Gaussian.

The input to our proposed network is a set of RGB images. Before the training begins, each input training image $N_i$ is wrapped randomly and resized to 192×256, which is used to create an image pair. We send a pair of images to the encoder part of the network at a time, which consists of 3× 3 convolutional layers. Then the convolutional features are fed into five blocks of two convolutional layers, with one max pooling layer after each block. The upper and lower encoders share the weights for the true positive images in training. After that, there are two separate 1×1 convolutional layers to computer the latent variables $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\sigma}^{(i)}$. Let the prior over the latent variables be the centered isotropic multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}^{(i)}, diag(exp(\boldsymbol{\sigma}^{(i)})))$, we sample from the posterior $\boldsymbol{z}^{(i,l)} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})$ using $\boldsymbol{z}^{(i,l)} = \boldsymbol{\mu}^{(i)} + \boldsymbol{\sigma}^{(i)} \odot \boldsymbol{\epsilon}^{(l)}$, where $\boldsymbol{\epsilon}^{(l)}$ is an auxiliary noise variable $\boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. With $\odot$ we signify an element-wise product. Let $J$ denote the dimensionality of $\boldsymbol{z}$, and let $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\sigma}^{(i)}$ be the variational mean and standard deviation evaluated at data point $i$. Finally, let $\sigma_j$ and $\mu_j$ simply denote the $j^{th}$ element of these vectors. The estimation results of this model and data point $\boldsymbol{x}^{(i)}$ can be written as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}^{(i)}) \simeq \frac{1}{2}\sum_{j=1}^{J}(1 + \log((\sigma_j^{(i)})^2) - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2) \\ + \frac{1}{L}\sum_{l=1}^{L}\log p_\theta(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i,l)})\end{aligned} \tag{4}$$

where

$$\boldsymbol{z}^{(i,l)} = \boldsymbol{\mu}^{(i)} + \boldsymbol{\sigma}^{(i)} \odot \boldsymbol{\epsilon}^{(l)}, \quad \boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{5}$$

In our proposed network, semantic information and visual appearance are encoded into separate feature maps of the latent space $\boldsymbol{z}$. After passing through the encoder, the latent variable

$\boldsymbol{z}$ is divided into $K+1$ local descriptors with the shape of $T \times \frac{H}{16} \times \frac{W}{16}$ on the channel. This can also be interpreted as $T$ local descriptors. One of them is used to reconstruct the full resolution RGB image, while the remaining $K$ descriptors are sent to the decoder for concatenation and then used to predict a full resolution semantic segmentation label. In this way, all the information contained in the image is encoded by $K+1$ local descriptors, so that the network can automatically put the relevant features into the corresponding feature map of $\mu$ and $\sigma$. The latent variables here are both vector quantities, which are obtained by flattening the $T \times \frac{H}{16} \times \frac{W}{16} \times (K+1)$ three-dimensional arrays.

We use the KL divergence loss to optimize the latent variables and construct the standard normal distribution, which can be constructed as:

$$\begin{aligned} L_\alpha &\simeq \mathcal{D}_{KL}[\mathcal{N}(\boldsymbol{\mu}^{(i)}, diag(exp(\boldsymbol{\sigma}^{(i)}))) \parallel \mathcal{N}(\mathbf{0}, \boldsymbol{I})] \\ &= \frac{1}{2}((\sum_i \exp(\boldsymbol{\sigma}^{(i)}) - \boldsymbol{\sigma}^{(i)} - \dim(\boldsymbol{\mu}^{(i)}) + (\boldsymbol{\mu}^{(i)})^\top \boldsymbol{\mu}^{(i)}) \end{aligned} \tag{6}$$

After sampling with a standard normal distribution, the latent variable $\boldsymbol{z}$ is sliced into $K+1$ groups of feature maps corresponding to $K$ object classes and one visual appearance. The slice part of latent variable $\boldsymbol{z}$ is sent to $K+1$ independent decoders to decode the features and object classes corresponding to appearance respectively. Then, the output information of the decoder for appearance is sent to the generated image loss function:

$$L_\beta = -\sum_i (I_i \log(G_i) + (1 - I_i) \log(1 - G_i)) \tag{7}$$

where $I_i$ and $G_i$ represent the input image and generated image respectively, and $i$ is the index.

At the same time, the remaining $K$ outputs of the decoders for semantic segmentation are concatenated channel-wise. After that, we send the concatenated result to a standard pixel-wise softmax, and define the cross entropy loss function as follows:

$$L_\gamma = -\sum_{i=1}^{M} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \tag{8}$$

where $M$ represents the number of pixels for an image. $\hat{y}^{(i)}$ is the predicted result, and $y^{(i)}$ represents the truth-value. We only use the limited number of labeled underwater images for training. The unsupervised auto-encoder network is able to effectively train the segmentation decoder with the limited labelled data and avoid the difficulty to label large amount of the underwater images.

In the convolution layer of encoder, we use the exponential linear unit activation layer. At the end of each layer, we use 3×3 max pooling with a stride of 3 to downscale the features. However, the layers for computing the latent variables and the last layer of the semantic segmentation in the decoder have no activation. In addition, the generated image decoder has a sigmoid activation. In the convolution layer of decoder, we use the sub-pixel convolutional neural network to upscale the decoder features.

We define $D_\chi$ as the normalized overall image descriptors for the database images, $D_\Upsilon$ as the positive sample, and $D_\Phi$ as the negative sample. Therefore, the objective function is defined as follows:

$$L_\xi = \max(0, D_\chi^\top (D_\Phi - D_\Upsilon) + \Theta) \tag{9}$$

where $\Theta$ represents the margin hyperparameter. Using this loss function, the UVAE network learns to separate the similarity between positive and negative samples through the margin $\Theta$. By optimizing this loss function, the cosine similarity between the positive samples representation and database is maximized, while the cosine similarity between the negative image representation and database is minimized.

### *B. Network Training*

We use an NVIDIA GeForce RTX 2080 Ti GPU to train our model. Since we have limited self-collected underwater dataset (only about 5,000 different underwater images), we first pre-trained our model with the COCO dataset [30]. In this dataset, the COCO-Stuff dataset includes total 164,000 images from COCO 2017. Furthermore, in order to alleviate the dependence on large-scale underwater labeled dataset, we randomly wrap the self-collected underwater images and send them to the network for learning. At the same time, due to our dataset does not contain the positive samples, we also use random rotation, translation and stretching transformation to generate corresponding images. By using these methods, we are able to simulate the influence of changing pitch, roll and yaw angles on the vision images during AUV navigation.

Light illumination at different times of the day could result in significant changes in the underwater appearance. Texture changes could affect the scene geometry. Intense viewpoint changes in underwater environment could cause serious perspective distortions, and often lead to structural overlap between query and database. Due to the low number of underwater training examples, we augmented our data with a series of transformations. We applied rotation, warp images, color distortion, and skew for the augmentation. We calculate the dense pixel-wise semantic segmentation $D_S = \{D_{S_i}\}$ for all input images, where each pixel of $D_{S_i}$ is assigned a semantic class label $l_s \in \{1, 2, \cdots, K\}$. Then, we fuse the images into semantic 3D voxel maps for the database $M_d$ and query images $M_q$. Each voxel in the semantic 3D map has one of $K+2$ labels, that is, a voxel is occupied by one of the $K$ semantic classes, or marked as unobserved space $K_u$ or free space $K_f$.

Our model leverages the latent space as the descriptor and captures the high-level geometric and semantic information. After the semantic segmentation of underwater images, we assign weights to the dynamic and static segmentation results according to the speed of moving objects. The details of weight allocation will be introduced in Section V. The overall objective function is defined as:

$$L_{\text{sum}} = L_\alpha 10^{-4} + L_\beta 10^{-4} + L_\gamma + L_\xi \tag{10}$$

After training 200,000 epochs with the COCO dataset, we get a satisfactory result of semantic segmentation. Then, we

continue to put the underwater dataset into the network for further learning. On this basis, we randomly darken images with an average intensity above the threshold $\Omega = 0.3$ to compensate for the lack of night-time images in our underwater dataset. In addition, we also randomly add blue and green light to simulate the scene changes at different times of the day. We retrained the network 10,000 epochs in the same way mentioned above, and finally trained the whole network 210,000 epochs.

## V. Experiments and Analysis

### A. Experiment Setup

We trained the proposed network with TensorFlow using the default parameters of the Adam optimizer and set the learning rate to 0.0001, each holding a batch size of 6. We tested the proposed method in the real-world underwater environment to evaluate its performance.

The custom-made sensor suite for AUV loop closure detection is $1470\times650\times800\mathrm{mm}^3$ in dimensions. It includes a NVIDIA Jetson TX2, a forward-looking global shutter camera, two light sources, etc. Our AUV is driven by four thrusters, including two vertical thrusters and two plane omnidirectional thrusters. We carried out underwater experiments in the Yellow Sea, where the seafloor mainly contains a large number of rocks, sea cucumbers, sea urchins, scallops. In addition, we also set up a number of observation devices under the water, as shown in Fig. 3.

(a) (b) (c)

Fig. 3. The underwater vehicle that we used for the Yellow Sea experiment. (a) The appearance of the AUV testbed. (b) The custom-made sensor suite collecting dada in the Yellow Sea. (c) The real scene of the experiment in the Yellow Sea, where the data in this paper was collected. There are a large amount of rocks, sea cucumbers, sea urchins, scallops and other objects under the white iron rectangular area observed in zoomed-in image.

### B. Experiment Results

After training, we used 12 representative underwater images to test the network. The generated images are very similar to the input original images, and the generated image contains the main objects captured underwater, as shown in Fig. 4. Although some unimportant scenes may be lost, it will not have a significant impact on the detection of loop closure. On the other hand, noise interference can also be filtered out, for example, it can make up for the absorption and scattering of particles suspended in the medium before reaching the camera in the underwater scene, resulting in low contrast and fog effect. In addition, it can compensate for the color difference caused by the optical wavelength, dissolved organic compounds and salinity of water.

We assign the weights to the dynamic and static segmentation results according to the speed of the moving objects. Furthermore, we use different colors to represent different semantic information, as shown in Fig. 5. As you can see, our method can achieve good semantic segmentation results. In the complex underwater environment, we roughly classified the substances in the water into five categories. Stones, aquatic plants and artificial structural scenario are marked as background, whose weights are set as 1.0, and represented by darkseagreen. The weights for objects like sea urchines, which barely move, are set to 0.8, and we use purple to represent them. Orchid represents sea cucumbers moving in similar pace, whose weights are set to 0.7. Due to the fact that scallops could suddenly move several meters, we set their weights to a relatively low level, 0.5. Lightgreen represents scallops. Finally, the weights for fishes and AUVs moving in the water are set to 0.2, and we use coral to represent them. In the task of loop closure detection, the objects with high weights have high reliability. Even when there are only fishes and AUVs in the view field, the rough relative position recognition can be completed to ensure the robustness of the system.

In order to eliminate the effect of false positive, we use the Euclidean distance metric during keypoint matching to compare the descriptors. By combining the fast and discriminative global image descriptors with the geometric features of keypoints, we are able to precisely complete the loop closure detection without the need for the use of thesholding techniques. If there are not enough valid matches to estimate the fundamental matrix using the RANSAC algorithm, the candidate will be rejected. The algorithm needs at least five matches. On this basis, the final candidate is regarded as the candidate with the highest global descriptor similarity score and effectively matched. In Fig. 6, we tested it in different scenarios, which can be used as a good geometric check for loop closure detection. Outlier rejection is handled by the uncertainty derived from false positives of both semantic and geometric consistencies. It can eliminate the influence of the uncertainty derived from false positives. In the underwater environment with a single color background, the matching can be completed correctly, with a matching score of 0.98, as shown in Fig. 6 (a). It can be done successfully even in the presence of aquatic plants disturbances, as shown in Fig. 6 (b). In Fig. 6 (c), the matching effect is very good near the artificial structural scenario. Moreover, we also tested in an indoor pool, and the matching score is 0.95, as shown in Fig. 6 (d). The average processing time of matching point detection is 11 ms. In the above four representative scenarios, these matching keypoints can provide a good geometric check for underwater loop closure detection.

In order to verify the effectiveness of our method in the loop closure detection, we also evaluated our algorithm by driving the AUV around the fire pool. In the indoor experiment, the fire pool size is $20\times8\times5$ $\mathrm{m}^3$. There is only one wellhead in the pool that can be put into the underwater vehicle. At the bottom of the pool, we arranged sea cucumbers, sea urchins and scallops along the way to provide underwater features. We control the AUV to move along the edge of the pool in a rectangular trajectory. The water in this pool is not static, but flowing. Therefore, in this dynamic environment,

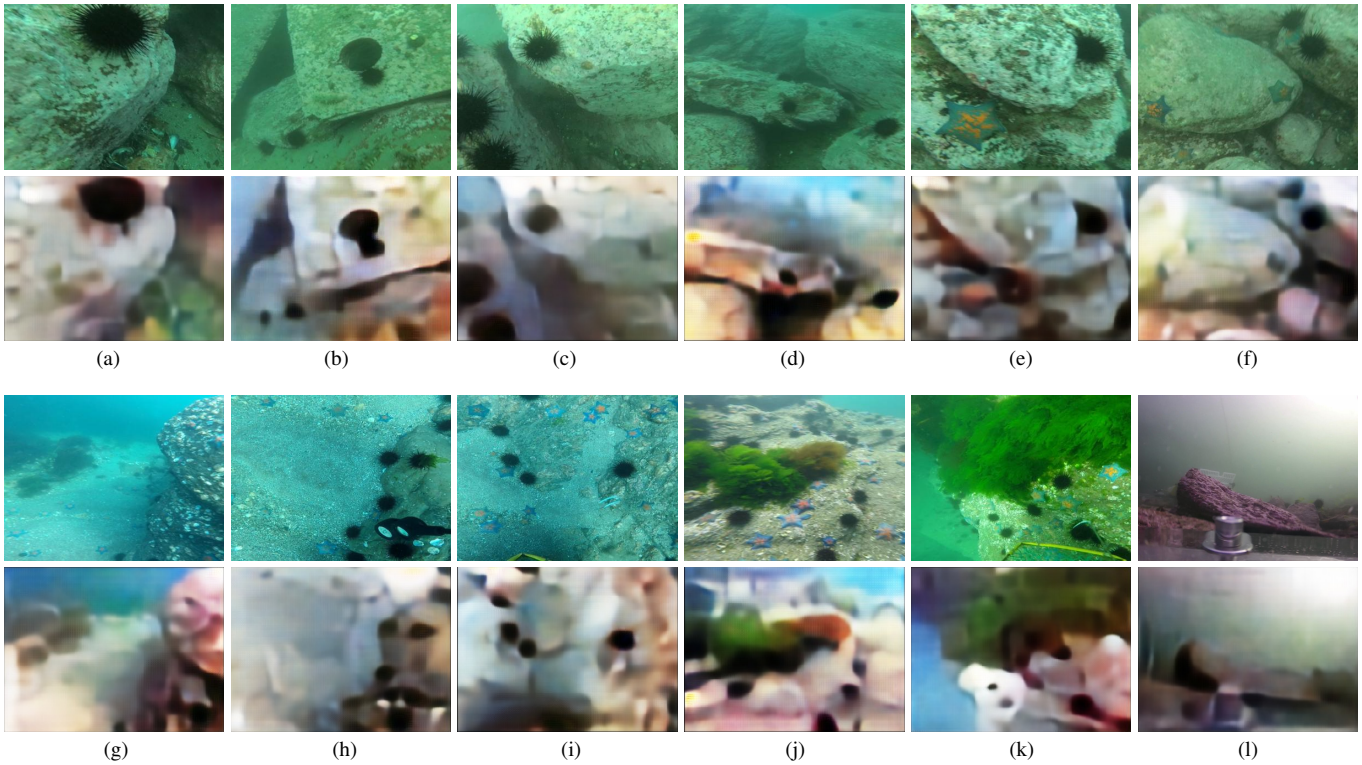



Fig. 4. Comparison of the input and generated images. Among them, the first line is the input images, and the second line is the generated images successively through the UVAE network. Sea urchins, stones, starfish, and artificial targets were generated in the underwater environment, while some floating particles, insignificant sand or small shells in the water may be ignored.

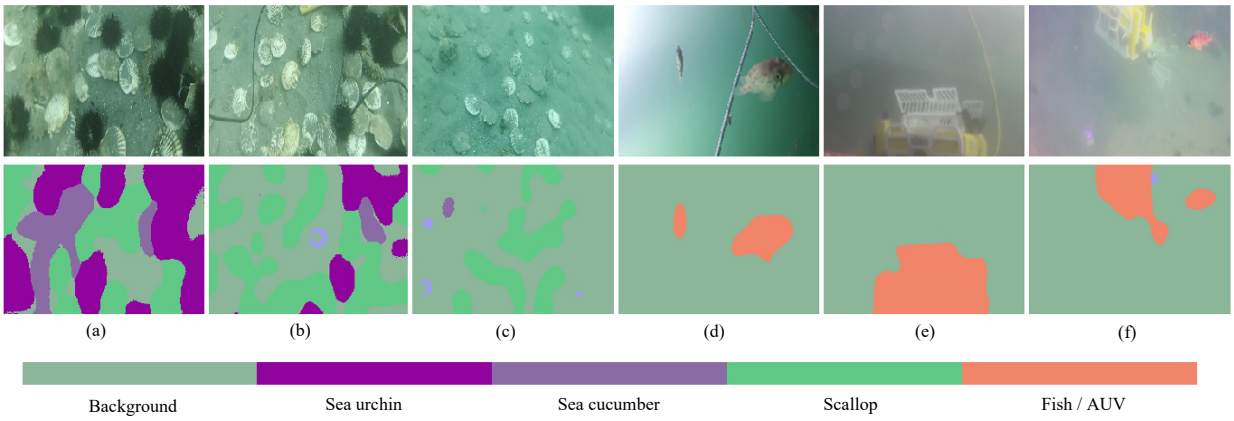



Fig. 5. The results of semantic segmentation. Among them, the first line is the input images, and the second line is the semantic segmentation images successively through our network, where darkseagreen represents the background, purple represents sea urchins, orchid represents sea cucumbers, lightgreen represents the scallops, and coral represents fast-moving objects (AUV or fish). We can segment the underwater scenes and assign the weights to the target, eliminating the influence of fast-moving objects.

the influence of ocean currents in the Yellow Sea can be simulated, as show in Fig. 7 (a). Fig. 7 (b) shows the detailed loop closure detection results. We used X-axis and Y-axis to represent the position information, and Z-axis to represent the keyframe index. Our system successfully detects that the AUV returns to the starting position at the frame 788, and can perform the matching correctly. It achieves over 91 frames per second (FPS) on a single-core desktop for underwater loop closure detection. From these results, it shows that the proposed network is capable of accurate loop closure detection in practice.

The scenarios in the Yellow Sea underwater dataset we collected are challenging and have a lot of viewpoints changes. We proved the discriminate capability of our local image descriptors. We extracted two principal components of descriptors for the scallop and sea cucumber classes, where the principal component analysis (PCA) whitening matrices are trained for each class of the whole underwater dataset respectively. As shown in Fig. 8, (a), (b) and (c) represent an image in the database, true positive image and true negative image respectively. In Fig. 8 (d), blue represents the database descriptor, green represents the true positive, and

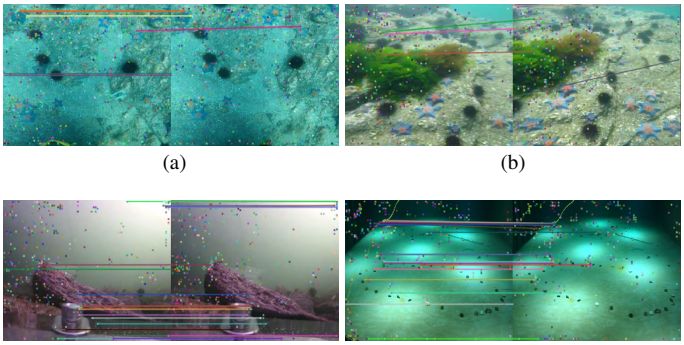


Fig. 6. Comparison results of matching points after performing geometric methods in different underwater scenarios. (a) A seabed with a single color background (matching score: 0.98). (b) A colorful seabed (matching score: 0.73). (c) An artificial structural scenario (matching score: 0.97). (d) A scenario in an indoor fire pool (matching score: 0.95).

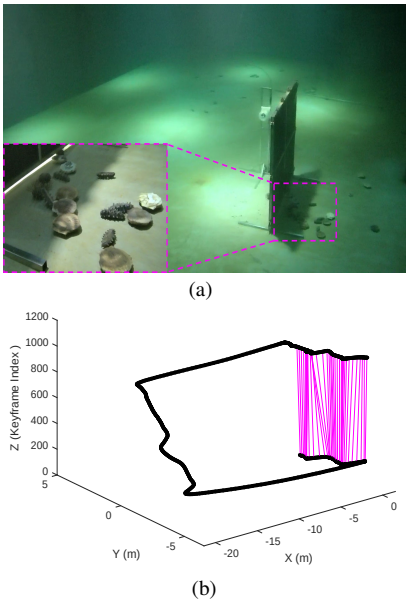



Fig. 7. Experimental results of our system in underwater loop closure detection. (a) The picture of the fire pool experiment. We arranged sea cucumbers, sea urchins and scallops along the way. The zoomed-in part of the picture is the starting position of the AUV. (b) The results of our system performance on the fire pool dataset. Clearly, the loop closure is correctly detected in the complex dynamic underwater scenario.

red represents the true negative. This effect shows the benefit from using semantic information in the image descriptors. For example, the visual appearance method can not clearly distinguish between two different descriptors. The blue line in the figure is basically in the middle of green and red. The reason for this could be that the true negative images are similar to the images in the database in terms of texture, which would cause the system to misjudge them. However, by using the semantic information, we could avoid the disadvantage that the visual appearance might be indistinguishable, and our descriptors could maintain an obvious distinction. We can see that after adding the semantic descriptors of scallops and sea cucumbers, the blue line is closer to the green line on the unit circle. The aforementioned results confirm that the potential information extracted from our network is very effective.

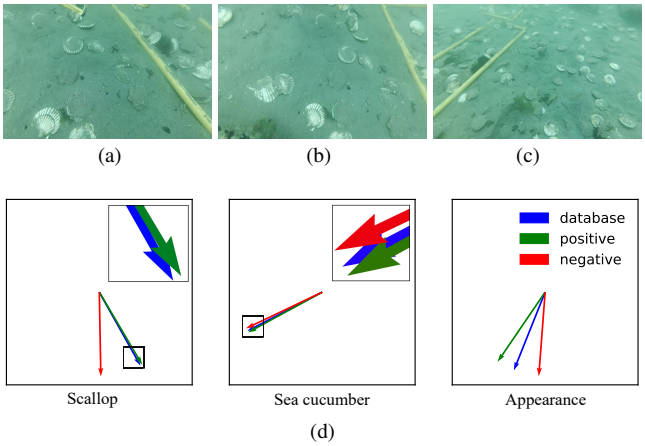



Fig. 8. The result of computing the descriptors for underwater seabed images in the Yellow Sea area. (a) An image in the database. (b) A true positive query image. (c) A true negative query image. (d) The computed results of two principal components of the corresponding local normalized residual descriptors. In the upper right corner of the picture are details and illustrations. Best viewed in color and with zoom in.

### C. Discussion

Robust camera-only AUV localization in difficult underwater environments is a challenging problem. Water flow or tide, plankton and water particles all would cause different imaging backgrounds, which could bring noise in the system. In order to deal with the influence of dynamic water current, colour shift and low illuminations on the experimental results, the Yellow Sea underwater dataset specifically contains the images taken at different times of the day. Especially in the pool environment, we not only carried out experiments at different times, but also collected underwater images in very calm and flowing scenarios respectively.

Due to the difference between underwater and ground terrestrial environments, some terrestrial algorithms SeqLPD [5], ORB-SLAM2 [18] were prone to collapse in underwater datasets, and so far we cannot directly compare the proposed method with related exiting loop closure detection algorithms. So in this paper, we only compare it with the state-of-the-art method NetVLAD [4] and CALC [25]. In order to achieve a fair comparison, we implemented the NetVLAD and CALC algorithms in our system. On the above two models that were trained separately, we also added our underwater dataset to train them again, to avoid the possibility that they have not seen any underwater scenarios during training. We choose three underwater datasets to validate the robustness of the three algorithms, and the results of different scenarios are shown in Table II and Fig. 9. In the table, the performance indicator used is the highest recall rate at 1.0 precision. The mixed scenario refers to 100 pairs of underwater images of the Yellow Sea and fire pool.

In the pool, the recall rates of three methods are 55.93%, 36.36% and 92.31% respectively, since the pool is small and there are fewer underwater dynamic objects. When facing the complex Yellow Sea environment, the recall rate of NetVLAD decreased considerably. However, the recall rate of our method is still high, which is 80.00%. Because the proposed method uses the object-level semantic information, it could alleviate

TABLE II
STATISTICS FOR RECALL RATE

| Underwater Dataset | NetVLAD | CALC | Ours |
|---|---|---|---|
| Fire Pool | 55.93% | 36.36% | 92.31% |
| Yellow Sea | 11.36% | 26.67% | 80.00% |
| Mixed Scenario | 18.52% | 52.63% | 81.82% |

the impact of fast-moving objects. Even if there are moving objects in the water, such as fishes and AUVs, our method can improve the robustness of the system in challenging underwater scenarios through semantic features.

The precision-recall curves are shown in Fig. 9, which includes the recall rate, area under curve (AUC), average precision (AP) and F1-score. We can see that our method outperforms the NetVLAD and CALC algorithms in all three scenarios. The AUC values of our method are 0.99, 0.98 and 0.98 respectively, and the F1-scores are 0.96, 0.89 and 0.90 respectively, indicating that our method has a better performance in underwater place recognition.

## VI. CONCLUSION

In this paper, to solve the robustness problem of loop closure detection in dynamically changing underwater environments, we proposed an UVAE network, which mainly extracts semantic, geometric and visual appearance information from underwater images. It is a novel underwater visual loop closure detection scheme based on an unsupervised deep network. The proposed method is verified by our custom-made underwater sensor suite on real-world maritime space. Specifically, our system successfully detects that the AUV returns to the starting position at the frame 788, and can perform the matching correctly in the complex dynamic underwater scenario. The average processing time of matching point detection is 11 ms. Moreover, when comparing the proposed method against NetVLAD and CALC in the pool, our method has a higher recall rate, which is 92.31%. Hence, the results show that the network has better performance.

It should be noted that there is still a lot of work to do to further improve the performance of localization and mapping tasks in underwater environments through autonomous navigation, e.g., how to improve accuracy of loop closure detection in complex underwater environments? How to apply decentralized visual SLAM algorithms to multiple AUVs in large scale environments? We will try to address these issues in future studies.

## REFERENCES


[1] T. Qiu, Z. Zhao, T. Zhang, C. Chen, and C. P. Chen, "Underwater internet of things in smart ocean: System architecture and open issues," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4297–4307, Jul. 2020.

[2] I. Jawhar, N. Mohamed, J. Al-Jaroodi, and S. Zhang, "An architecture for using autonomous underwater vehicles in wireless sensor networks for underwater pipeline monitoring," *IEEE Trans. Ind. Informat.*, vol. 15, no. 3, pp. 1329–1340, Mar. 2019.

[3] K. L. Ho and P. Newman, "Detecting loop closure with scene sequences," *Int. J. Comput. Vis.*, vol. 74, no. 3, pp. 261–286, Sep. 2007.

[4] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1437–1451, Jun. 2018.

[5] Z. Liu, C. Suo, S. Zhou, H. Wei, Y. Liu, H. Wang, and Y.-H. Liu, "SeqLPD: Sequence matching enhanced loop-closure detection based on large-scale point cloud description for self-driving vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2019, pp. 1218–1223.

[6] 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.

[7] A. Celik, N. Saeed, B. Shihada, T. Y. Al-Naffouri, and M. Alouini, "End-to-end performance analysis of underwater optical wireless relaying and routing techniques under location uncertainty," *IEEE Trans. on Wirel. Commun.*, vol. 19, no. 2, pp. 1167–1181, Feb. 2020.

[8] N. Saeed, T. Y. Al-Naffouri, and M. Alouini, "Outlier detection and optimal anchor placement for 3-D underwater optical wireless sensor network localization," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 611–622, Jan. 2019.

[9] H. Chen, G. Zhang, and Y. Ye, "Semantic loop closure detection with instance-level inconsistency removal in dynamic industrial scenes," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2030–2040, Mar. 2021.

[10] 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 *2017 IEEE Underwater Technology (UT)*, 2017, pp. 1–3.

[11] 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.

[12] D. Połap, M. Wozniak, M. Korytkowski, and R. Scherer, "Encoder-Decoder based CNN structure for microscopic image identification," in *Int. Conf. Neural Inf. Process.*, 2020, pp. 301–312.

[13] M. Woźniak and D. Połap, "Soft trees with neural components as image-processing technique for archeological excavations," *Pers. Ubiquit. Comput.*, vol. 24, no. 3, pp. 363–375, 2020.

[14] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo, "Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light," *IEEE Trans. Circuits Syst. Video Technol.*, DOI:10.1109/TCSVT.2019.2963772.

[15] R. C. Luo and C. C. Lai, "Multisensor fusion-based concurrent environment mapping and moving object detection for intelligent service robotics," *IEEE Trans. on Ind. Electron.*, vol. 61, no. 8, pp. 4043–4051, Aug. 2014.

[16] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6D object pose estimation using 3D object coordinates," in *Proc. Euro. Conf. on Comp. Vision*, 2014, pp. 536–551.

[17] J. Xiao, D. Xiong, Q. Yu, K. Huang, H. Lu, and Z. Zeng, "A real-time sliding-window-based visual-inertial odometry for MAVs," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 4049–4058, Jun. 2020.

[18] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[19] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, "Semantics-aware visual localization under challenging perceptual conditions," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 2614–2620.

[20] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[21] C. Cadena, D. Gálvez-López, J. D. Tardós, and J. Neira, "Robust place recognition with stereo sequences," *IEEE Trans. Robot.*, vol. 28, no. 4, pp. 871–885, Aug. 2012.

[22] M. Mohan, D. Gálvez-López, C. Monteleoni, and G. Sibley, "Environment selection and hierarchical place recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 5487–5494.

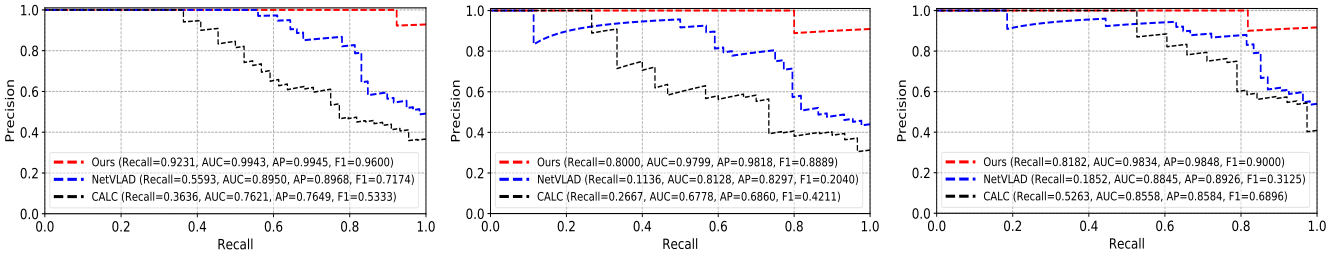
[23] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 3223–3230.

[24] X. Gao and T. Zhang, "Unsupervised learning to detect loops using deep neural networks for visual SLAM system," *Auton. Robots*, vol. 41, no. 1, pp. 1–18, Jan. 2017.

[25] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," in *Proc. Robot., Sci. Syst.*, 2018, pp. 1–10.

[26] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic visual localization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6896–6906.

[27] N. Merrill and G. Huang, "CALC2.0: Combining appearance, semantic and geometric information for robust and efficient visual loop closure," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2019, pp. 4554–4561.




(a) (b) (c)

Fig. 9. Precision-recall curves for the underwater dataset. (a) In the fire pool. (b) In the Yellow Sea. (c) In the mixed scenario. Best viewed in color.

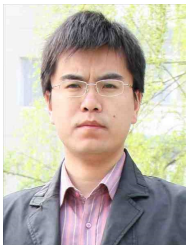

[28] P. Corke, C. Detweiler, M. Dunbabin, M. Hamilton, D. Rus, and I. Vasilescu, "Experiments with underwater robot localization and tracking," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2007, pp. 4556–4561.

[29] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[30] H. Caesar, J. Uijlings, and V. Ferrari, "COCO-Stuff: Thing and stuff classes in context," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 1209–1218.


**Yangyang Wang** received the B.S. and M.S. degrees in electronic information science and technology from the Xinjiang Agricultural University, Urumqi, China, in 2014, and 2017, respectively. He is currently working toward the doctoral degree from the School of Information and Communication Engineering of Dalian University of Technology (DUT), Dalian, China. From November 2021 till now, he is a Visiting Fellow with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, U.K. His research interests include underwater localization, underwater simultaneous localization and mapping (SLAM), machine learning and computer vision.

**Xiaorui Ma** (M'17) received the B.S. degree in applied mathematics from Lanzhou University, Lanzhou, China, in 2008, and Ph.D. degree in communication and information system from Dalian University of Technology, Dalian, China, in 2017. She is currently an Associate Professor at Dalian University of Technology. Her research interests include processing and analysis of remote sensing images, specially hyperspectral image classification and synthetic aperture radar image classification.



**Jie Wang** (M'12, SM'18) received his B.S. degree from the Dalian University of Technology, Dalian, China, in 2003, M.S. degree from Beihang University, Beijing, China, in 2006, and Ph.D degree from the Dalian University of Technology, Dalian, China, in 2011, all in electronic engineering. He is currently a Full Professor at Dalian Maritime University. He used to be an Associate Professor at Dalian University of Technology from 2014 to 2017. He was a visiting researcher with the University of Florida from 2013 to 2014. His research interests include wireless localization and tracking, radio tomography, wireless sensing, machine learning, wireless sensor networks, and cognitive radio networks. He serves as an Associate Editor for IEEE Transactions on Vehicular Technology.

**Shilong Hou** received the B.S. degree in electronic information science and technology from the Nanjing Agricultural University, Nanjing, China, in 2018. He is currently working toward the M.E. degree in the School of Information and Communication Engineering, Dalian University of Technology. His research interests include synthetic aperture radar image target detection and scene interpretation, and machine learning.

**Ju Dai** is currently a post doc in Peng Cheng Laboratory, Shenzhen, China. She received her B.S. degree and M.Sc. degree in Electronic Engineering, China University of Geosciences (CUG), Wuhan, China, in 2011 and 2014 respectively, and the Ph.D. degree in Signal Processing in Dalian University of Technology (DUT), Dalian, China, in 2020. Her research interests include person re-identification, saliency detection, human 3D pose estimation and movement behavior analysis.

**Dongbing Gu** (Senior Member, IEEE) 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.

**Hongyu Wang** (M'98) received the B.S. degree in electronic engineering from the Jilin University of Technology, Changchun, China, in 1990, the M.S. degree in electronic engineering from the Graduate School, Chinese Academy of Sciences, Beijing, China, in 1993, and the Ph.D. degree in precision instrument and optoelectronics engineering from Tianjin University, Tianjin, China, in 1997. He is currently a Professor with the Dalian University of Technology, Dalian, China. His research interests include image processing, and underwater localization.