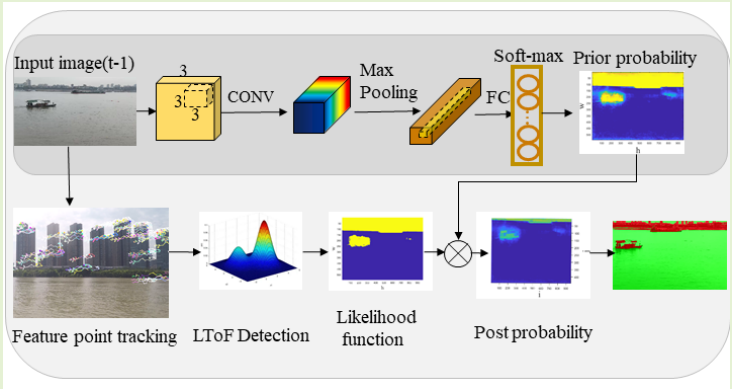
# Knowledge-Driven Semantic Segmentation for Waterway Scene Perception

Qianqian Chen, Changshi Xiao*, Yuanqiao Wen, Haiwen Yuan, Yamin Huang, Yuanchang Liu, Member, IEEE, Wenqiang Zhan, and Qiliang Li, Senior Member, IEEE

***Abstract*—Semantic segmentation as one of the most popular scene perception techniques has been studied for autonomous vehicles. However, deep learning-based solutions rely on the volume and quality of data and knowledge from specific scene might not be incorporated. A novel knowledge-driven semantic segmentation method is proposed for waterway scene perception. Based on the knowledge that water is irregular and dynamically changing, a Life Time of Feature (LToF) detector is designed to distinguish water region from surrounding scene. Using a Bayesian framework, the detector as the likelihood function is combined with U-Net based semantic segmentation to achieve an optimized solution. Finally, two public datasets and typical semantic segmentation networks, FlowNet, DeepLab and DVSNet are selected to evaluate the proposed method. Also, the sensitivity of these methods and ours to dataset is discussed.**




***Index Terms*—semantic segmentation, U-Net, probabilistic fusion, visual perception, waterway scene**

## I. Introduction

IMPROVING the ability of intelligent agents to understand their working environment is one of the current research topics in the field of autonomous vehicles. Advanced sensor technologies for object detection, localization, tracking, and recognition in structured environments [1] have been relatively well developed. Waterway is a typical unstructured scene that lacks of adequate and stable reference information. Thus, there seems to be a challenge for the technologies applied in complex waterway scene. Due to the dynamic, diversity and randomness characteristic of the water surface, the performance of perception algorithm is degraded seriously [2,3].

Different from traditional detection algorithms based on artificial feature or box regression, semantic segmentation that utilizes deep learning to classify each pixel of images shows great promise in such scene perception tasks [4,5]. FCN [6], SegNet [7], U-Net [8,9], and Deeplab [10], as typical deep learning models have been studied for semantic segment. However, these models rely heavily on preliminary collection of volume and quality data, which is hardly possible in actual applications. Furthermore, sample labeling is extremely tedious and time consuming for most semantic segmentation algorithms.

For this purpose, a knowledge-driven semantic segmentation method is proposed for waterway scene perception. Based on the knowledge that water surface with the characteristic of dynamical and unstable texture is almost impossible to track, a Life Time of Feature (LToF) detector is designed to distinguish water from the scenario. Using a Bayesian framework, the detector as the likelihood function is used to calculate a probabilistic map for sematic segmentation, and the poster probabilistic map can be acquired by fusing the output of U-Net as the prior probabilistic map with the likelihood probabilistic

This work was supported in part by the National Natural Science Foundation of China under Grant 52001235 and 52001241; by the 111 Project(B21008); by the Natural Science Foundation of Hubei Province under Grant 2022CFB313; by the Zhejiang Key Research Program under Grant 2021C01010. (Corresponding author: Changshi Xiao).

Qianqian Chen and Wenqiang Zhan are with the School of Navigation, Wuhan University of Technology, Wuhan 430063, China. (e-mail: chenqq@whut.edu.cn; zwq626197298@whut.edu.cn)

Changshi Xiao is with the School of Navigation, Wuhan University of Technology, Wuhan 430063 China, and with the Institute of Ocean Information Technology, Shandong Jiaotong University, Weihai 264200 China. (e-mail: cs_xiao@hotmail.com)

Yuanqiao Wen and Yamin Huang are with the Intelligent Transportation Systems Research Center, Wuhan University of Technology, Wuhan 430063, China and the National Engineering Research Center for Water Transport Safety, Wuhan 430063, China. (e-mail: yqwen@whut.edu.cn; y.huang-3@tudelft.nl)

Haiwen Yuan is with the School of Electrical and Information Engineering, Wuhan Institute of Technology, Wuhan, 430205, China (e-mail: hw_yuan@whut.edu.cn)

Yuanchang Liu is with the Department of Mechanical Engineering, University College London, London, UK(yuanchang.liu.10@ucl.ac.uk)

Qiliang Li is with the Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA 22030 USA. (e-mail: qli6@gmu.edu).

map. To illustrate the performance of the method, comparative experiments with typical semantic segmentation networks FlowNet, DeepLab, and DVSNet are carried out on public maritime datasets. Also, the sensitivity of these methods to datasets will be discussed in this work.

The rest of this article is organized as follows. Section II describes the related works on semantic partitioning and their application in autonomous vehicles. Section III presents the proposed knowledge-based semantic partitioning framework. Section IV conducts extensive experiments to analyze and confirm the effectiveness of the proposed framework. Finally, Section V provides an overview of future work and the conclusions.

## II. RELATED WORK

Advanced sensors and techniques have been considered in various applications for scene perception, e.g., Synthetic aperture radar (SAR) [11], sonar [12], computer vision [13], etc.

In recent years, there has been a growing interest in leveraging deep learning techniques for semantic segmentation. Shi et al. [14] proposed an effective occlusion mask method using clustering and boundary determination to mask occlusions into convex shapes. Various deep neural networks have been developed for semantic segmentation in different scene partitions [15,16]. For example, Choi et al. [17] simulated raw ocean images in real-time using CNN architectures, such as VGGNet, Inception.v3, ResNet, and DenseNet. Additionally, Song et al. [18] introduced a system based on local HSRRSI water features in Mask R-CNN [19] for object recognition and automatic water quality extraction. Moreover, Cheng et al. [20] and Sun et al. [21] have shown that CNNs have achieved good results in image semantic segmentation in certain cases. U-Net as a semantic segmentation framework could be combined with another edge detection framework for detecting coastline from marine scene [22]. Dense skip connections and attention mechanism are performed in U-Net to improve the precision of the segmentation on water areas [23]. Combining U-Net with the level set method is proven outperform U-Net based segmentation [24]. Adaptive dual path learning framework [25] that combines self-supervised learning with image-to-image translation was proven to have great effectiveness in semantic segmentation.

Alternatively, there has been an increasing interest in leveraging multimodal data for semantic segmentation. Huang et al. [26] improved the performance and generalization capability of end-to-end autonomous driving with scene understanding leveraging deep learning and multimodal sensor fusion techniques. Qiu et al. [27] studied semantic segmentation for outdoor scene based on multi-sensor fused data acquired by unmanned ground vehicle (UGV), and a projection algorithm to generate a 2D RGB-DI image from the 3D RGB-DI point cloud was proposed so that the semantic segmentation in RGB-DI cloud points is transformed to the semantic segmentation in RGB-DI images. Ćesić et al. [28] compared images of obstacles detected by visual detection using a stereo camera with those detected by a radar detection technique. Osborne et al. [29] used visible and thermal imaging with time-domain stability characteristics to separate and classify targets. Wang et al. [30] integrated stereo photography with visual obstacle separation technology to detect obstacles below the coastline. Finally, Sinisterra et al. [31] estimated the motion parameters of moving sea vehicles to help USVs separate moving target vehicles.

Overall, these studies demonstrated the potential of traditional, deep learning, and multimodal data-based approaches for semantic segmentation in various marine-related applications. Inspired by the above work, we proposed a knowledge-driven semantic segmentation method for waterway navigation scene perception of autonomous ships. Our method takes the deep learning segmentation result as a priori probability and uses water unstructured feature knowledge as a driver to modify the segmentation result. This hybrid method improves efficiency while ensuring accuracy in segmenting obstacles on the water surface at different locations.

## III. METHODOLOGY

This section presents a novel knowledge-driven semantic segmentation method specifically designed for visual perception of waterway scene. Each pixel can be accurately classified into two categories: water or obstacle based on the prior knowledge that water surface presents dynamic and transient feature while the feature extracted from structured objects (obstacles) mostly is stable and slow-changing. To achieve it, optical flow is calculated from sequential images, and a novel semantic feature is defined by the lifetime of these feature points from optical flow. Thus, a LToF detector is designed to distinguish water from visual scene according to the lifetime statistics of optical flow.

In detail, the proposed LToF detector based on optical flow as the observation model of each pixel is taken as likelihood function. The confidence of semantic segmentation is taken as the prior probability in Bayesian framework. By applying Bayes' rule, the post probabilities of all pixels in the image could be calculated. The details of the proposed method are illustrated in the abstract figure. Either the LToF detection or the semantic segmentation is further introduced in the following sections.

### A. LToF detection based on optical flow

Life Time of Feature (LToF) is defined as the duration during which one feature can be stably tracked between image sequences. By introducing LToF, the temporal characteristic of the features extracted from various objects can be calculated. Classical Kanade-Lucas-Tomasi (KLT) optical flow algorithm [32] is applied to extract and track feature points from scene. The LToF distributions of water surface and non-water are displayed in Fig. 1. It can be seen that the LToFs of water surface and non-water are distinctly different. Experientially two Gaussian functions are utilized to approximate the LToF distributions of water surface and non-water, respectively, as follows:

$$t \sim \begin{cases} f_1(\mu_1, \sigma_1^2) \\ f_2(\mu_2, \sigma_2^2) \end{cases} \tag{1}$$

where $t$ denotes the LToF distribution and composes of two Gaussian functions $f_1$ and $f_2$. The LToF distribution in Fig. 1 can be used to classify the feature into water surface or non-water. Thus, the means and variances $\mu_1$, $\sigma_1^2$, $\mu_2, \sigma_2^2$ of the Gaussian functions should be determined firstly.

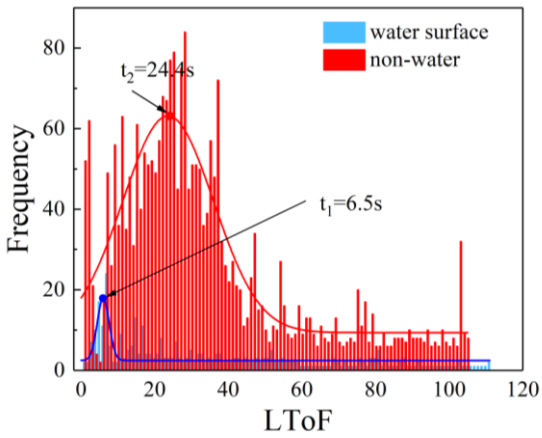



Fig. 1. The graphical model for LToF with two Gaussian maps.

Based on empirical knowledge, it is implied that the water surface in a water navigation scene is irregular and dynamic. Consequently, the feature point tracking time period of the water area, which is measured by the LToF value, is shorter than that of the non-surface area, which may contain various obstacles. To address the issue, we designed a model using Gaussian components with a smaller mean $\mu_1$ to represent the LToF map of the water surface, and Gaussian components with a larger mean $\mu_2$ to represent the LToF map of obstacle areas. $I_{i,j}$ is used to denote the region of the feature point, where $I_{i,j} = 1$ represents the feature point of the water surface area, and $I_{i,j} = 2$ represents the feature point of the obstacle area.

Eq. (2) illustrates the two Gaussian LToF models employed in this study. The observed LToF data comprises time observations $t_{i,j}$, while the hidden data encompasses $f_{i,j}$ (the value of the latent function at input $t_{i,j}$), and the mode identity of the LToF at different samples. The probabilities of LToF are represented by $\partial_{I_{i,j}}$, where $I_{i,j}$=1 or 2.

$$f(t_{i,j},\theta) = \sum_{j=1}^{2} \partial_{I_{i,j}=k} \frac{1}{\sqrt{2\pi\sigma^2_{I_{i,j}=k}}} exp[-\frac{(t_i-\mu_{I_{i,j}=k})^2}{2\sigma^2_{I_{i,j}=k}}] \quad (2)$$

where $\boldsymbol{\theta} = [\mu_1, \mu_2, \sigma_1^2, \sigma_2^2]$ denotes the involved two Gaussian LToF models. The LToF map consists of two regions: water surface and obstacles, as shown below.

$$\mathrm{f}(t_{i,j}, I_{i,j} = \mathrm{k}, \theta) = \frac{1}{\sqrt{2\pi\sigma^2_{I_{i,j}=k}}} \exp[-\frac{(t_{i,j}-\mu_{I_{i,j}=k})^2}{2\sigma^2_{I_{i,j}=k}}] \quad (3)$$

As the mean value of the LToF map on the water surface is generally smaller than that of obstacles, this difference can be utilized to design a suitable model for detecting feature points of obstacles by calculating the LToF threshold $T$. The LToF threshold $T$, as presented in Eq. (4), is defined as the intersection of the two Gaussian maps.

$$f(T, I_{i,j} = 1, \boldsymbol{\theta}) = f(\mathrm{T}, I_{i,j} = 2, \boldsymbol{\theta}) \quad (4)$$

The LToF threshold $T$ is fully determined can be calculated by the Gaussian distributions of water and non-water regions. To better calculate $T$, hyper-parameters $\boldsymbol{l} = [l_1, l_2, l_3]$ is introduced and obtained from the parameters $\boldsymbol{\theta}$ as follows:

$$\begin{cases} l_1 = \sigma_1^2 + \sigma_2^2 \\ l_2 = \sigma_2^2\mu_1 + \sigma_1^2\mu_2 \\ l_3 = \sigma_2^2\mu_1^2 + \sigma_1^2\mu_2^2 + 2\sigma_1^2\sigma_2^2 \ln(\sigma_1/\sigma_2) \end{cases} \quad (5)$$

After the expansion of Eq. (5), the LToF threshold $T$ can be expressed as a polynomial function of the hyper-parameters $l_1$, $l_2$, and $l_3$. Specifically, we have:

$$T = (l_2 + (l_2^2 - l_1 \cdot l_3)^{\frac{1}{2}})/l_1 \quad (6)$$

Here, the hyper-parameters $l_1, l_2, l_3$ are polynomial coefficients that are used to determine the shape of the polynomial function.

To determine the distribution parameters $\mu_i, \sigma_i$ $(i = 1,2)$, Curve Fitting Toolbox (https://ww2.mathworks.cn/ products /curvefitting.html) is used to fit the data acquired from the LToF detector. This toolbox provides the calculation of Gaussian fitting and can obtain the mean and variance of the Gaussian function.

After calculating the threshold $T$ using Eq. (6), feature points with a lifetime exceeding T are defined as longevous feature points, which are generally extracted from obstacles. The two Gaussian models of LToF maps used for detecting longevous feature points can be defined as the likelihood function of the observation model for these feature points. The likelihood function of the longevous feature points can be expressed as:

$$P(Y_{i,j} = t_{i,j} | X_{i,j} \epsilon \mathrm{o}) = \frac{1}{\sqrt{2\pi\sigma^2_{I_{i,j}=k}}} \exp[-\frac{(t_{i,j}-\mu_{I_{i,j}=2})^2}{2\sigma^2_{I_{i,j}=2}}] \quad (7)$$

where $Y_{i,j}$ denotes the observation state of the longevous feature point $(x_i, y_i)$, the LToF $t_{i,j}$ is the measurement, $X_{i,j} \epsilon \{\mathrm{w}, \mathrm{o}\}$ denotes the pixel attributes, and $X_{i,j} \epsilon \mathrm{o}$ represents $(x_i, y_i)$ as in the obstacle area, $X_{i,j} \epsilon \mathrm{w}$ represents $(x_i, y_i)$ as in the water surface area.

Except the longevous feature points, the remaining pixel points in the video have not yet been tracked in the KLT. Clustering [33] is applied to determine the categories of the pixels that were not tracked by KLT. Using the intensity and coordinate of each pixel as feature, the closer to a non-water pixel and the similar with the color of a non-water pixel, the greater the probability that the pixel belongs to non-water.

The model starts from the longevous feature points as seeds and iteratively calculates the distance and gray difference between the untracked pixels and the seed points until the observation value $d_{i,j}$ is obtained for each untracked pixel point.

$$d_{i,j} = \left(1 - \frac{\Delta I_{i,j}}{1080}\right) \cdot \left(1 - \frac{\Delta s_{i,j}}{255}\right) \quad (8)$$

In the equation, $d_{i,j}$ represents the measurement, and the algorithm utilizes $\Delta I_{i,j}$ and $\Delta s_{i,j}$ to describe the gray difference between the new pixels and seed pixels. Fig. 2 illustrates the Likelihood probability based on intensity and distance clustering. Intensity and Distance donates the differences between the intensities and coordinates of untracked pixels and seed pixels, respectively.

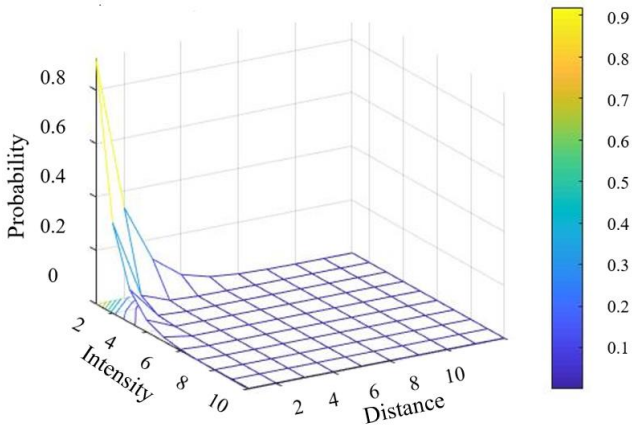



Fig.2. Likelihood probability based on intensity and distance clustering.

By analyzing the standard segmentation labels of each water navigation scene, it is found that the ratio of water area to obstacle area in the water navigation scene image is 4:1. Therefore, the observation value threshold of the untracked pixel points is set to 80%. When the observation value exceeds 80%, the untracked pixel point is judged to belong to the obstacle area.

After performing the clustering algorithm with each longevous feature point as the seed, the probability of the remaining pixels being obstacles gradually accumulates. Based on this, assuming there are $n$ feature points detected, the likelihood function of these pixels is defined as follows:

$$P\left(Y_{i,j} = d_{i,j} \middle| X_{i,j} \epsilon o\right) = \frac{1}{n}\sum_{i}^{n} k_i \cdot \exp(d_{i,j}) \quad (9)$$

where $i$ denotes a specific feature point, $n$ denotes the number of detected feature points, $k_i$ represents the scaling factor, which is positively correlated with the probability of this longevous feature point, as shown below:

$$k_i \propto P\left(Y_{i,j} = t_{i,j} \middle| X_{i,j} \epsilon o\right) \quad (10)$$

The likelihood function is based on two empirical knowledge. Firstly, the water surface in a water navigation scene is irregular and dynamic, resulting in shorter feature point tracking time periods for the water area, as measured by the LToF value, compared to the non-surface area, which may contain various obstacles. Secondly, a closer distance and smaller color difference between the untracked pixels and longevous feature points suggest a higher probability of the untracked pixel belonging to an obstacle area. Therefore, the likelihood function of each pixel in the waterway navigation scene image can be expressed as follows:

$$P\left(Y \middle| X_{i,j} \epsilon o\right) \sim \begin{cases} P\left(Y_{i,j} = t_{i,j} \middle| X_{i,j} \epsilon o\right) \\ P\left(Y_{i,j} = d_{i,j} \middle| X_{i,j} \epsilon o\right) \end{cases} \quad (11)$$

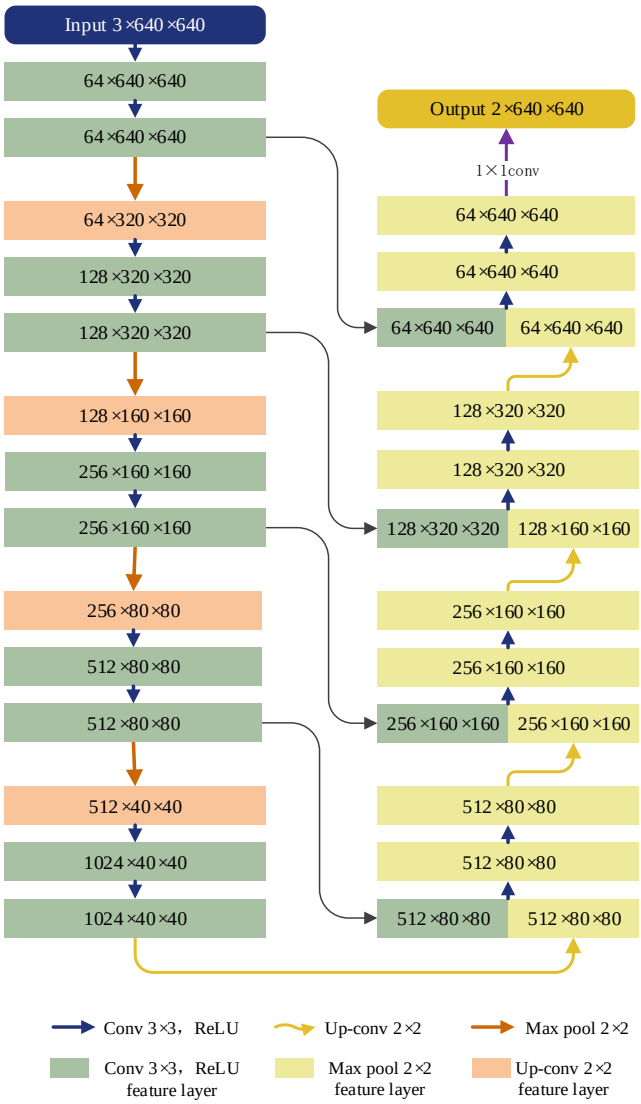



Fig. 3. The U-Net architecture based on the confidence level of segmentation

### B. Semantic Segmentation based on U-Net

A measurable model was developed using U-Net, based on the confidence level of segmentation for ship trajectory problems. Various U-Net models are available in graphical segmentation. U-Net was shown to be a very useful segmentation algorithm in [34], and the network architecture is illustrated by Fig. 3 in details.

The architecture mainly comprises of two iterations of 3×3 convolutions with activation ReLU, and 2×2 maximum pooling for down-sampling after each convolution. As shown in Fig. 3, Conv 3×3 refers to a convolution layer with a convolution kernel 3 × 3, ReLU is an activation function for the convolutional layer, and Max pool 2×2 refers to a down-sampling operation on the acquired feature map and can resize the map to 1/2. Up-conv 2×2 is defined to a up sampling operation by a convolution layer with a convolution kernel 2×2. During the up-sampling stage, the feature channels are increased, and each extended pathway contains a diagram of the upper sampling characteristics, a 2 × 2 up-convolution,

concatenating them with the corresponding feature maps in the compressed route, and performing a 2×3 convolution with each ReLU process. Cropping is required because edge pixels are lost for each convolution.

In the fourth stage, the 64-component feature vector corresponds to classifications using a 1×1 convolution, and the entire network consists of 23 convolutions. To achieve seamless completion of the output profile stitching, it is crucial to select the appropriate input interpolation size, and the 2x2 maximum interpolation operation can be suitable for all X and Y size of the same layer.

Both the input video and the corresponding segmentation graph can be learned by a neural network. As no convolution is performed, the edge width of the output video is smaller than that of the input. To minimize GPU memory, larger blocks are used instead of a large number of batches, reducing the number of batches to a single video. Therefore, 0.99 is currently determined to be in the optimal state, as seen in many training cases before.

The energy of the method was obtained by using the relation between the soft-max and the cross-entropy loss at the pixel level in the final characteristic curve. The maximum software max is given by Eq. (12),

$$\mathrm{P}(X_{i,j}\epsilon\mathrm{o}) = \exp\big(a_k(x_i, y_i)\big) / \big(\sum_{k=1}^{K} \exp(a_k(x_i, y_i))\big) \quad (12)$$

where $a_k(x_i, y_i)$ $X_{i,j}\epsilon\{\mathrm{w}, \mathrm{o}\}$ K is the number of classes and $\mathrm{P}(X_{i,j}\epsilon\mathrm{o})$ is the prior probability of each pixel classification in the waterway navigation scene video.

### C. Bayesian Probability Optimization

In this study, a Bayesian framework is pr oposed to classify pixels in a waterway navigation scene as either water surface, obstacle and sky. The framework is based on the pipeline shown in the abstract figure, where at each video frame, the probability map $\mathrm{P}\big(X_{i,j}\epsilon\mathrm{o}\big|Y\big)$ is first updated for pixel type, and then used to update the segmentation of the waterway navigation scene video.

To obtain $P(X_{i,j}\epsilon\mathrm{o})$ for each pixel type$(x_i, y_i)$, U-Net based segmentation and observation models are used. $P(X_{i,j}\epsilon\mathrm{o})$ represents the probability map of observation in the Bayesian framework. The posterior map indicating whether each pixel belongs to the obstacle area is calculated using Eq. (13):

$$P\big(X_{i,j}\epsilon o\big|Y\big) \propto P\big(Y\big|X_{i,j}\epsilon o\big) * P(X_{i,j}\epsilon\mathrm{o}) \quad (13)$$

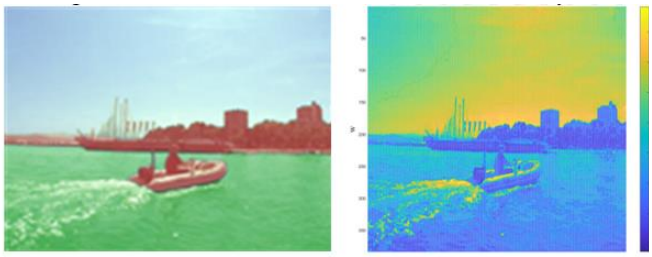
The first factor in Eq. (13) is the likelihood of observation, which can be directly computed using Eq. (11). The second factor is the prior probability density function, which is recursively updated as the previous posterior.

U-Net is utilized for segmenting maritime scene in our work. Firstly, manual labeling training samples are produced and input into the U-Net based semantic segmentation model which is published in our previous work [35]. Then, the initial waterway navigation scene video segmentation based on U-Net and soft-max losses is obtained. Finally, the prior probability $P(X_{i,j}\epsilon\mathrm{o})$ is expressed using the soft-max losses of the segmentation result.

After obtaining the likelihood function and prior probability of each pixel in the waterway navigation scene video, the posterior probability of each pixel can be obtained using Eq. (13). Finally, waterway navigation scene segmentation is achieved by classifying the category of each pixel based on the obtained posterior probability.

Fig. 4 (a) and (b) show priori probability maps generated by knowledge-based semantic segmentation in the two scenarios. The left side is the segmentation based on U-Net, the right side is the prior probability based on U-Net.

## IV. EXPERIMENTS AND DISCUSSION



(a) (b)

Fig. 4. U-Net based sematic segmentation. (a) segmentation image; (b) priori probability map

### A. Waterway Scene Dataset

Two public datasets, Marine Semantic Segmentation Training Dataset (MaSTr1325) [36] and Singapore Maritime Dataset (SMD) [37], are available to perform cross-validation on our knowledge-driven semantic segmentation method. MaSTr1325 is a large-scale marine semantic segmentation training dataset designed specifically for developing maritime obstacle detection algorithms. The dataset composes of 1325 diverse videos and is collected by small coastal unmanned surface vehicles, covering a range of realistic conditions encountered in coastal surveillance tasks. All the images are semantically labeled on a per-pixel basis. The samples of SMD are acquired by cameras fixed on shore and moving vessels, respectively.

The proposed method is trained by SMD and tested by MaSTr1325. To evaluate our method, four typical scenarios have been selected from MaSTr1325, as shown in Table I. Dynamic obstacles, e.g., fast-moving speedboats and waves, are involved in Scene 1. Scene 2 involves large-sized obstacles, e.g., close-up buoy. Floating debris are presented in Scene 3. obstacles of different scales and lighting conditions with different dynamic ranges in waterborne navigation scenes. Strong light and reflected light appear in Scene4, which could reduce seriously the performance of vision algorithms.

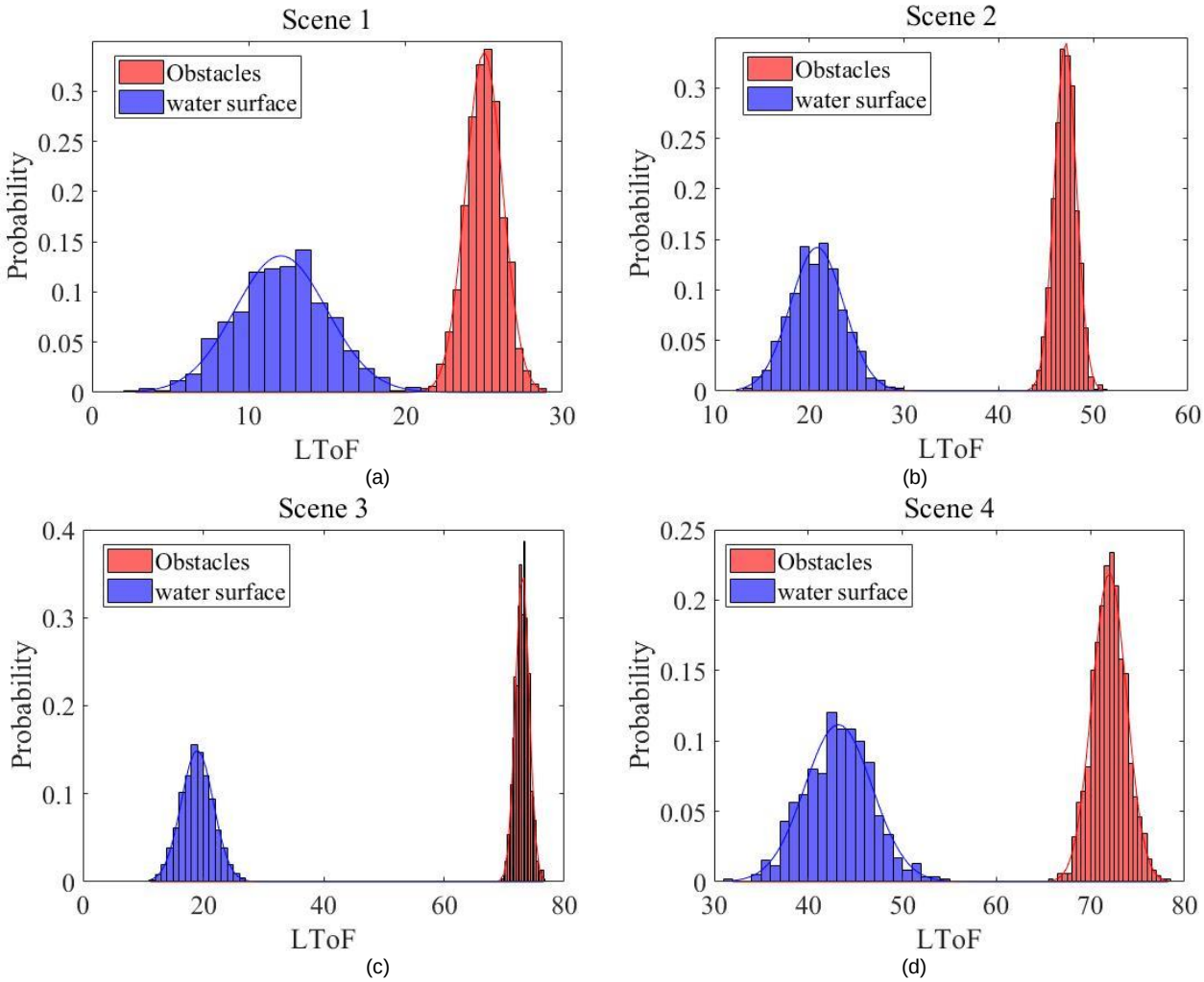



Fig. 5. LToF detection in Scene 1~4. (a) Scene 1; (b) Scene 2; (c) Scene 3; (d) Scene 4

TABLE I Waterway Scene Datasets

| Scene | Description |
|---|---|
| 1 | dynamic obstacles, e.g., fast-moving speedboats and waves |
| 2 | large-sized obstacles, e.g., close-up buoy |
| 3 | small-sized obstacles, e.g., floating debris |
| 4 | strong light and reflected light |

### *B. Segmentation Result*

In this section, LToF statistics are performed on Scene 1-4, and the results are shown in Fig. 5. The red represents the water surface while the blue represents the obstacle. At first, the threshold of the LToF detector can be calculated based on the parameters acquired by Gaussian fitting. It can be observed that the LToF thresholds are apparent in different waterways scenarios.

According to the method presented in Section III, the likelihood probability, prior probability and posterior probability of the category of each pixel in the scenarios can be calculated. One image from Scene 1 and the corresponding probability maps are shown in Fig. 6. Finally, waterway scene sematic segmentation is achieved by the posterior probability map.

The image sequence of the dataset is input to our framework. The LToF detector applies KLT to calculate the optical flow between the images. Subsequently, the threshold of the LToF is calculated as the likelihood probability of each pixel.

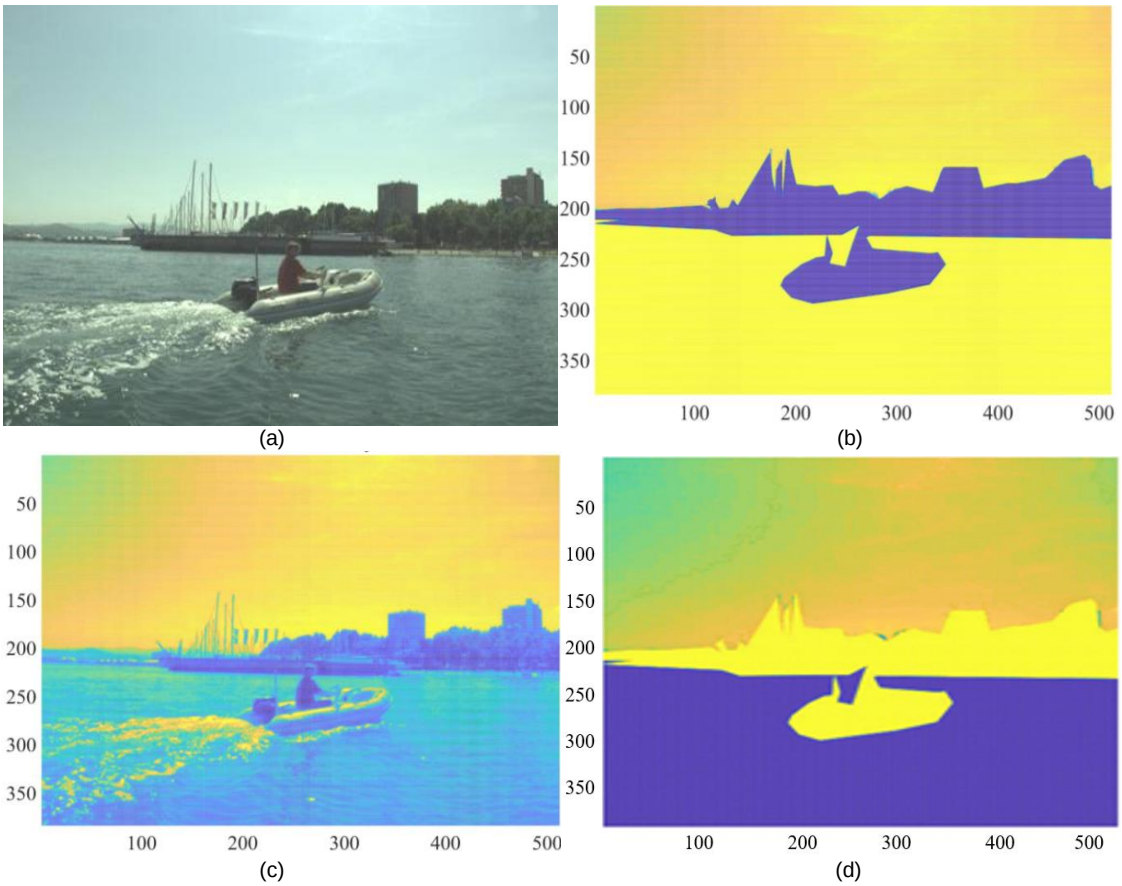



Fig.6. The original image (a) of waterway scene and the likelihood probability map (b), prior probability map (c) and the posterior probability map (d).

Meanwhile, the priori probability map can be obtained by U-Net. The posteriori probability map is calculated in a bayes framework, as the semantic segmentation results for the waterway navigation scene. U-Net and optical flow calculation are the main components of the proposed method in this work, and have been proven to be real-time. In the experiments, one PC with NVIDIA GeForce RTX 2080 Ti GPU is employed to train and evaluate the proposed method, which could process 1920*1080 images at 17 frames per second.

Among popular video semantic segmentation algorithms, DeepLab [10], DVSNet [38], FlowNet [39] have been selected for comparative analysis with our knowledge-driven semantic segmentation method. DeepLab is one of the typical semantic segmentation algorithms based on CNN and selected as a comparison. Atrous convolution is used to increase the receptive field, enabling better capture of contextual information in images without losing resolution. DVSNet is a semantic segmentation algorithm specialized for event based vision. Data is generated by using optical flow to detect motion events in dynamic scenes. FlowNet is a deep learning model that learns the differences between input images for optical flow estimation. Such a model enables better understanding on object structures and movements from video.

The segmentation results of DVSNet, DeepLab, FlowNet and ours are shown in Fig. 7-9. Fig. 7-9 are the segmentation results of scenes 1, 3, 4, respectively. There are no typical obstacles in Scene 2, With sufficient training data, all four methods can achieve good semantic segmentation results. Therefore, it is not used for typical scene analysis. The results show that both the quantity and quality of the training set have a significant impact on the performance of DVSNet, DeepLab, and FlowNet2. In contrast, our semantic segmentation method based on knowledge region is more robust.

Seven metrics are used to evaluate these segmentation methods, including accuracy, precision, recall, F1 score, Intersection over Union (IOU), mean IOU (mIOU), and frequency weighted IOU (fwIOU). IOU is used to measure the intersection of each predicted category with the ground truth, mIOU is calculated by the average IOUs of each category, and FWIOU is calculated as a weighted average according to the IOU of the pixel category.

On this basis, the calculations of these evaluation metrics is as follows:

TP: Correctly classified as water

FP: Misclassified as water

FN: Misclassified as obstacles

TN: Correctly classified as an obstacle

Accuracy=(TP+TN)/(TP+TN+FP+FN)

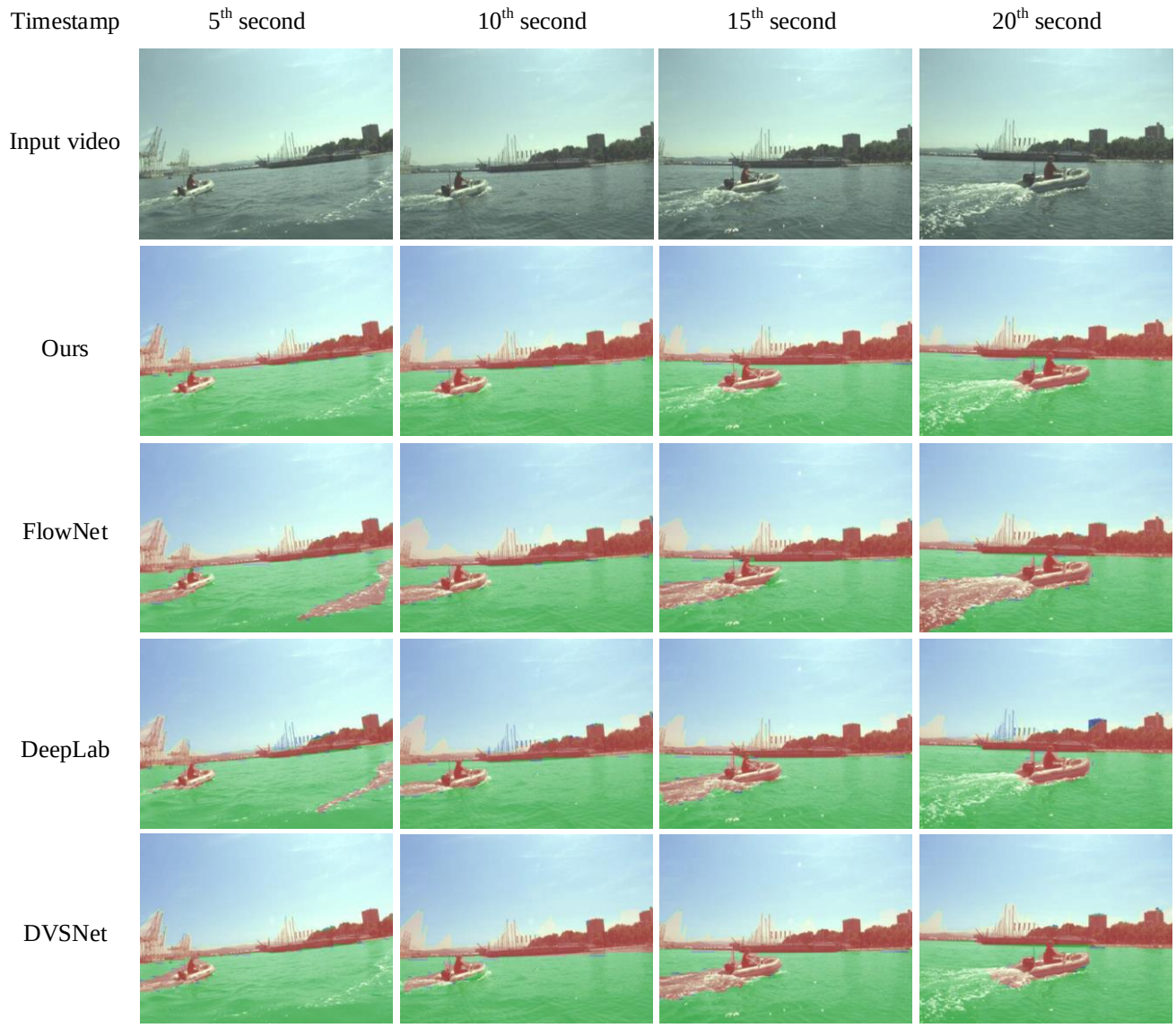



Fig.7. Segmentation Results of Scene 1

Precision=TP/(TP+FP)
Recall=TP/(TP+FN)
F1Score =(2*Precision*Recall)/(Precision+Recall)
IOU= TP/(TP+FP+FN)
mIOU=[TP/(TP+FP+FN)+TN/(TN+FN+FP)]/2
fwIOU=(TP+FN)/(TP+FP+TN+FN)*(TP)/(TP+FP+FN)

TABLE II
SEGMENTATION RESULTS OF SCENE 1

| method | FlowNet | DeepLab | DVSNet | Ours |
|---|---|---|---|---|
| Accuracy(%) | 78.42 | 82.87 | 83.92 | **94.11** |
| Precision(%) | 69.70 | 74.63 | 72.42 | **87.10** |
| Recall(%) | 71.97 | 69.37 | 78.27 | **81.12** |
| F1Score(%) | 70.56 | 43.75 | 73.25 | **83.06** |
| IOU(%) | 61.92 | 63.00 | 77.10 | **78.55** |
| mIOU(%) | 60.52 | 54.82 | 62.31 | **75.58** |
| fwIOU(%) | 72.40 | 75.63 | 81.83 | **89.56** |

In Table II, we present the evaluation of all four methods using different performance metrics. After comparing the segmentation results of the four methods in Scene 1, it was found that the knowledge-driven semantic segmentation method for waterway navigation scenes achieved a higher accuracy in segmenting rapidly moving obstacles. This is because the other three methods are more sensitive to perceiving dynamic targets in the video, which can lead to mistaking water waves for dynamic obstacles. On the other hand, our method can accurately identify the short-lived feature points on the water surface, resulting in more precise semantic segmentation of water area pixels. Therefore, the knowledge-driven semantic segmentation method is more suitable for waterway navigation scenes.

In Table III, we present the evaluation of all four methods using different performance metrics. After comparing the segmentation results of four methods in Scene 3, we found that the knowledge-driven semantic segmentation method for

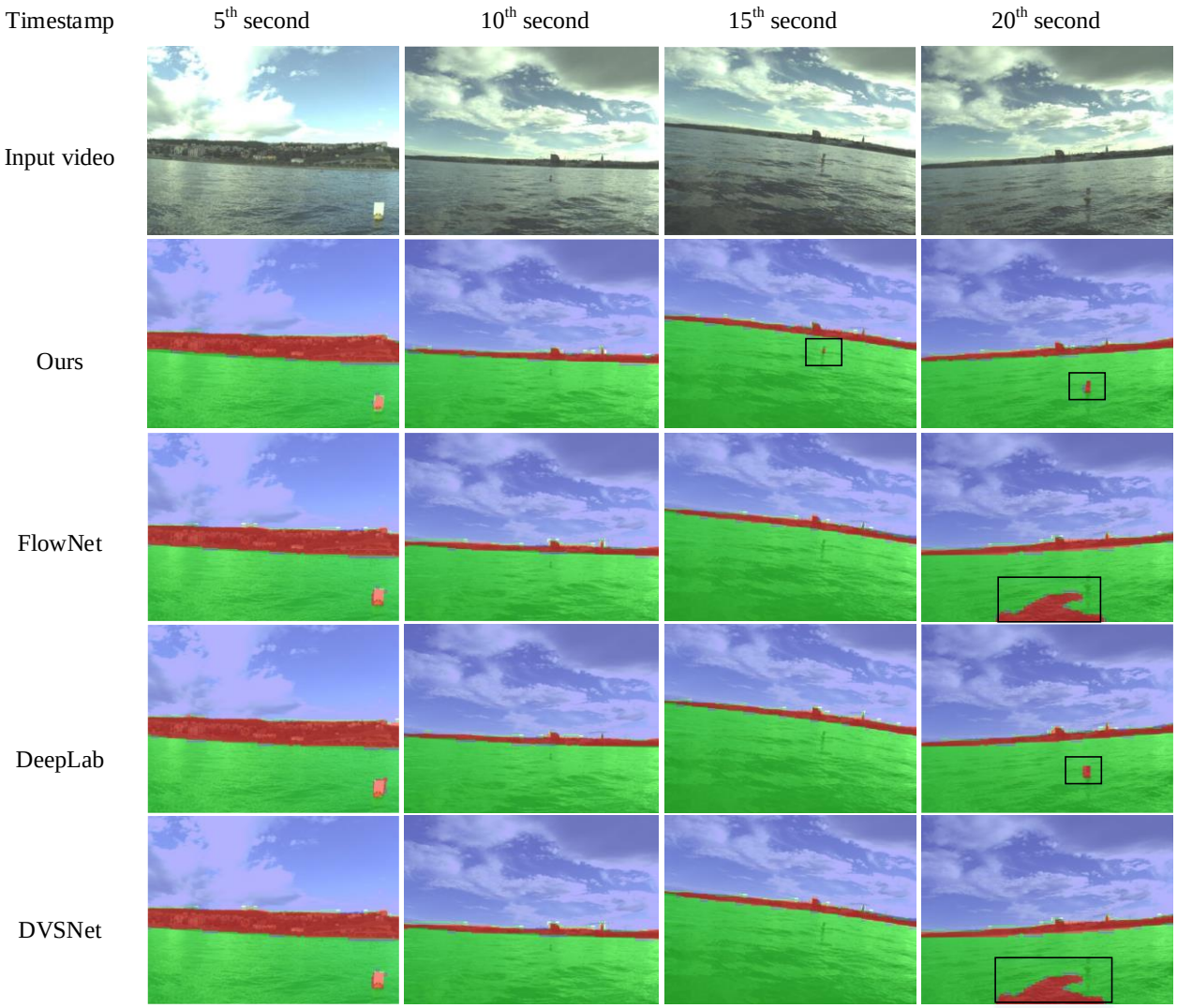



Fig.8. Segmentation Results of Scene 3

TABLE III
SEGMENTATION RESULTS OF SCENE 3

| method | FlowNet | DeepLab | DVSNet | Ours |
|---|---|---|---|---|
| Accuracy(%) | 81.42 | 91.87 | 88.92 | **97.11** |
| Precision(%) | 70.70 | 84.63 | 80.42 | **85.10** |
| Recall(%) | 68.97 | 89.37 | 78.27 | **81.12** |
| F1Score(%) | 57.56 | 63.75 | 63.25 | **73.06** |
| IOU(%) | 69.63 | 78.00 | 64.10 | **87.55** |
| mIOU(%) | 67.52 | 74.82 | 62.31 | **75.58** |
| fwIOU(%) | 51.40 | 75.63 | 61.83 | **89.56** |

waterway navigation scenes has a higher accuracy in segmenting small targets on the water surface, such as floating debris and buoys, while the other three methods have a higher rate of missed detection for small targets. This is because small targets are sparsely distributed in the training dataset, and these targets are static on the water surface, making it difficult for dynamic detection-based semantic segmentation algorithms to segment them. However, our method, which is based on LToF corner detection, can detect the corner points of small obstacles on the water surface, thus improving the accuracy of semantic segmentation for small targets.

In Table IV, we present the evaluation of all four methods using different performance metrics. After comparing the segmentation results of four methods in Scene 4, we found that the knowledge-driven semantic segmentation method for waterborne scenes has a higher accuracy in segmenting the strong glare reflection. The other three methods are affected by the strong glare and tend to segment some areas with large differences in the sky as water areas. In contrast, our method can easily distinguish water surface feature points, resulting in higher accuracy in segmenting water and non-water areas.

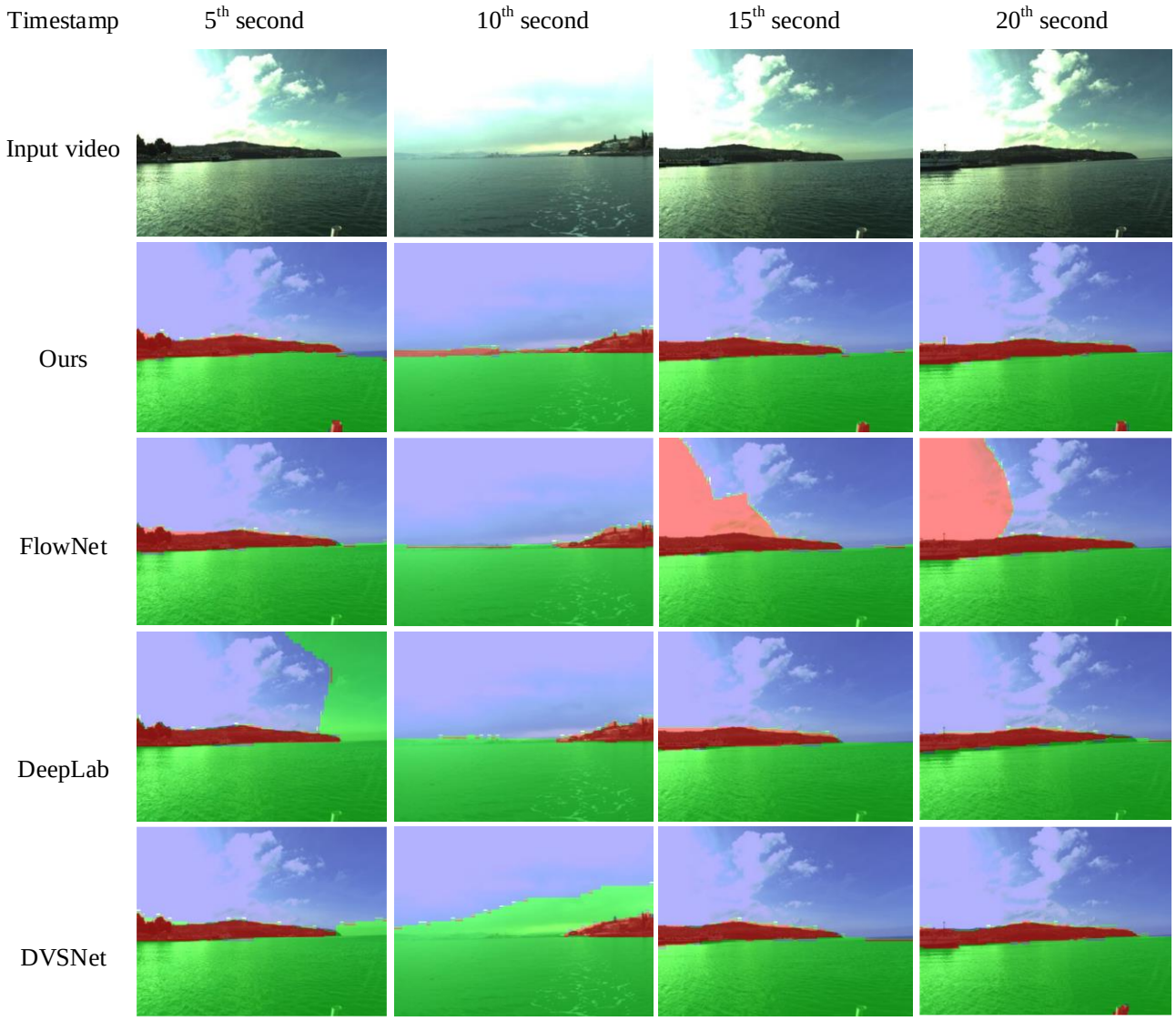



Fig.9. Segmentation Results of Scene 4

TABLE IV
SEGMENTATION RESULTS OF SCENE 4

| method | FlowNet | DeepLab | DVSNet | Ours |
|---|---|---|---|---|
| Accuracy (%) | 48.42 | 82.87 | 83.92 | **90.11** |
| Precision (%) | 10.70 | 34.63 | 50.42 | **75.10** |
| Recall (%) | 48.97 | 59.37 | 68.27 | **71.12** |
| F1Score (%) | 17.56 | 43.75 | 63.25 | **73.06** |
| IOU (%) | 9.63 | 28.00 | 34.10 | **57.55** |
| mIOU (%) | 27.52 | 54.82 | 62.31 | **75.58** |
| fwIOU (%) | 41.40 | 75.63 | 81.83 | **89.56** |

### *C. Stability Analysis*

The success of the knowledge-driven semantic segmentation method mainly relies on experiential knowledge and deep learning. It is discussed whether such knowledge-driven method is affected the quantity and quality of the employed dataset. Thus, two experiments are displayed in the following.

In the first experiment, all samples of SMD and partial samples of MaSTr1325 are used as training set while the remaining samples of MaSTr1325 as testing set. U-Net 1~5 are achieved by training U-Net with 100%, 80%, 60%, 40%, 20% of the training samples. When U-Net 1~5 and our method are evaluated on the remaining samples of MaSTr1325, the percentages of misclassified pixels are shown in Fig. 10 (a). It can be seen that the performance of the U-Net based sematic segmentation algorithm gradually decreased as the number of the training samples from MaSTr1325 decreases. Due to the introduction of prior knowledge, our method could achieve 2.7% percent of misclassified pixels even though only 20% MaSTr1325 used in training set.

In the next, only the samples of SMD are used as training set while the samples of MaSTr1325 as testing set. The definitions of U-Net 1~5 remain consistent with the experiment above. The percentages of misclassified pixels u are shown in Fig. 10 (b), when U-Net 1~5 and our method are evaluated on MaSTr1325. It is found that the misclassified pixels of U-Net semantic segmentation are above 20% once that MaSTr1325 was completely absent from the training set. However, the misclassified pixels of our knowledge-driven semantic segmentation are stabilized below 10%. As a result, it can be proven that such a knowledge-driven semantic segmentation method is more robust in the absence of reliable data.

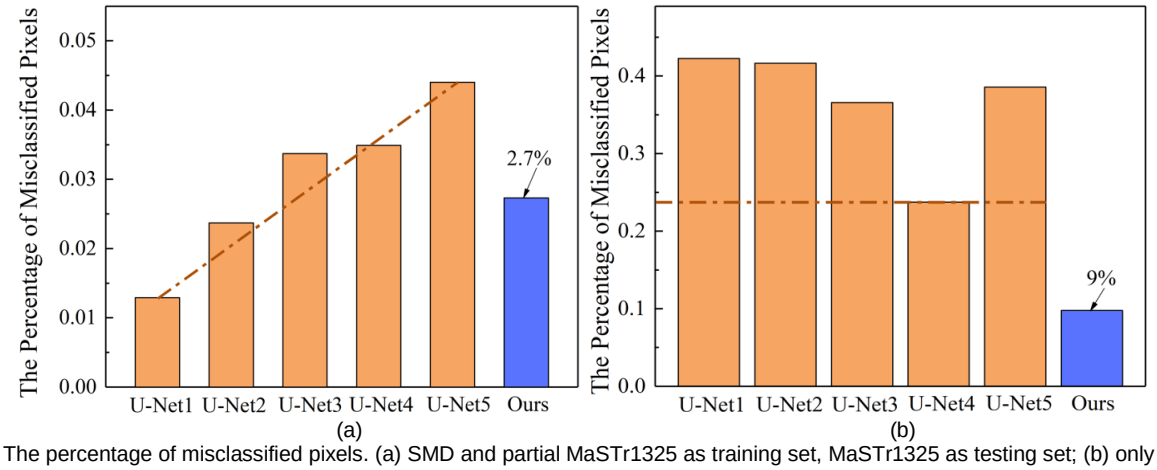



Fig.10. The percentage of misclassified pixels. (a) SMD and partial MaSTr1325 as training set, MaSTr1325 as testing set; (b) only SMD as training set, MaSTr1325 as testing set.

## V. CONCLUSION

A novel knowledge-driven semantic segmentation method is presented for waterway scene perception. Using a Bayesian framework, the prior knowledge acquired by the LToF detector is fused with U-Net based semantic segmentation. The proposed method is evaluated on two public datasets SMD and MaSTr1325. 90% accuracy is achieved by the proposed method. Overall, such a knowledge-driven semantic segmentation method has a better performance in segmenting fast-moving and small objects on the water surface, and is not affected by lighting conditions. It is beneficial for waterway scene perception. In future work, it is of significance to explore the fusion of more prior knowledge and deep learning based semantic segmentation algorithms.

## ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the anonymous reviewers for their contributions and valuable suggestions. Q. Chen and C. Xiao designed the experiments and analyzed them; Y. Wen and H. Yuan performed the experiments and studied the results; W. Zhan and Y Liu were responsible for collecting and organizing the data; Y. Huang and Q Li analyzed the results of the study and summarized them; and Q. Chen wrote the article. (Each of the above contributors agreed on the details of the above information).

## CONSENT FOR PUBLICATION

All authors have read the manuscript and approve for publication.

## COMPETING INTERESTS

The authors declare they have no competing interests.

## REFERENCES

[1] H. Yuan, C. Xiao, W. Zhan, Y. Wang, C. Shi, H. Ye, K. Jiang, Z. Ye, C. Zhou, Y. Wen and Q. Li, “Target detection, positioning and tracking using new UAV gas sensor systems: Simulation and analysis”, *Journal of Intelligent & Robotic Systems*, vol.94, no.3, pp.871-882, Jul.2018.

[2] J. Ni, K. Shen, Y. Chen, W. Cao and S.X. Yang, “An Improved Deep Network-Based Scene Classification Method for Self-Driving Cars”, *IEEE Transactions on Instrumentation and Measurement*, vol.71, pp.1-14, Jan.2022.

[3] H. Yuan, C. Xiao, Y. Wang, X. Peng, Y. Wen and Q. Li, “Maritime vessel emission monitoring by an UAV gas sensor system”, *Ocean Engineering*, vol.218, pp.108206-108215, Dec.2020.

[4] L. V. Tran and H. -Y. Lin, “BiLuNetICP: A deep neural network for object semantic segmentation and 6D pose recognition”, *IEEE Sensors Journal*, vol.21, no.10, pp.11748-11755, May.2021.

[5] N. Panchi, E. Kim and A. Bhattacharyya, “Supplementing remote sensing of ice: Deep learning-based image segmentation system for automatic detection and localization of sea-ice formations from close-range optical images”, *IEEE Sensors Journal*, vol.21, no.16, pp.18004-18018, Aug.2021.

[6] E. Shelhamer, J. Long and T. Darrell, "Fully convolutional networks for semantic segmentation", *IEEE transactions on pattern analysis and machine intelligence*, vol.39, no.4, pp.640-651, Apr.2017.

[7] V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.39, no.12, pp.2481-2495, Dec.2017.

[8] X. Chen, X. Wu, D. K. Prasad, B. Wu, O. Postolache and Y. Yang, "Pixel-Wise ship identification from maritime images via a semantic segmentation Model", *IEEE Sensors Journal*, vol.22, no.18, pp.18180-18191, Sept.2022.

[9] X. Chen, S. Liu, R. W. Liu, H. Wu, B. Han and J. Zhao, "Quantifying arctic oil spilling event risk by integrating an analytic network process and a fuzzy comprehensive evaluation model", *Ocean & Coastal Management*, vol.228, pp.106326-106340, Sept.2022.

[10] L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.40, no.4, pp.834-848, Apr. 2018.

[11] R. Cao, Y. Wang, B. Zhao and X. Lu, "Ship Target Imaging in Airborne SAR System Based on Automatic Image Segmentation and ISAR Technique", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol.14, pp.1985-2000, Jan.2021.

[12] J. Petrich, M.F. Brown, J.L. Pentzer and J.P. Sustersic, "Side scan sonar based self-localization for small Autonomous Underwater Vehicles", *Ocean Engineering*, vol.161, pp.221-226, Aug.2018.

[13] Y. Yang, R. Wang and H. Ren, "Active contour model based on local intensity fitting and atlas correcting information for medical image segmentation", *Multimedia Tools and Applications*, vol.80, pp. 26493-26508, May.2021.

[14] B. Shi, Y. Su, H. Zhang, J. Liu and L. Wan, "Obstacles modeling method in cluttered environments using satellite images and its application to path planning for USV", *International Journal of Naval Architecture and Ocean Engineering*, vol.11, no.1, pp.202-210, Jan.2019.

[15] X. Zhang, L. Wang and Y. Su, "Visual place recognition: A survey from deep learning perspective", *Pattern Recognition*, vol.113, pp.107760-107780, May.2021.

[16] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning", *Nature*, vol.521, pp.436-444, May.2015.

[17] H. Choi, M. Park, G. Son, J. Jeong, J. Park, K. Mo and P. Kang, "Real-time significant wave height estimation from raw ocean images based on 2D and 3D deep neural networks", *Ocean Engineering*, vol.201, pp.107129-107139, Apr.2020.

[18] S. Song, J. Liu, Y. Liu, G. Feng, H. Han, Y. Yao and M. Du, "Intelligent object recognition of urban water bodies based on deep learning for multi-source and multi-temporal high spatial resolution remote sensing imagery", *Sensors*, vol.20, no.2, pp.397-421, Jan.2020.

[19] Y. Fu, J. Fan, S. Xing, Z. Wang, F. Jing and M. Tan, "Image segmentation of cabin assembly scene based on improved RGB-D mask R-CNN", *IEEE Transactions on Instrumentation and Measurement*, vol.71, pp.1-12, Jan.2022.

[20] G. Cheng, P. Zhou and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images", *IEEE Transactions on Geoscience and Remote Sensing*, vol.54, no.12, pp.7405-7415, Dec.2016.

[21] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM", *IEEE Geoscience and Remote Sensing Letters*, vol.15, no.3, pp.474-478, Mar.2018.

[22] K. Heidler, L. Mou, C. Baumhoer, A. Dietz and X. X. Zhu, "HED-UNet: Combined Segmentation and Edge Detection for Monitoring the Antarctic Coastline", *IEEE Transactions on Geoscience and Remote Sensing*, vol.60, pp.1-14, Mar.2022.

[23] M. Xia, Y. Cui, Y. Zhang, Y. Xu, J. Liu and Y. Xu, "DAU-Net: A novel water areas segmentation structure for remote sensing image", *International Journal of Remote Sensing*, vol.42, no.7, pp. 2594-2621, Jan.2021.

[24] Y. Yang, C. Feng and R. Wang, "Automatic segmentation model combining U-Net and level set method for medical images", *Expert Systems with Applications*, vol.153, pp.113419-113427, Sept.2020.

[25] Y. Cheng, F. Wei, J. Bao, D. Chen and W. Zhang, "ADPL: Adaptive dual path learning for domain adaptation of semantic segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.45, no.8, pp.9339-9356, Aug.2023.

[26] Z. Huang, C. Lv, Y. Xing and J. Wu, "Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding", *IEEE Sensors Journal*, vol.21, no.10, pp.11781-11790, May.2021.

[27] Z. Qiu, F. Yan, Y. Zhuang and H. Leung, "Outdoor Semantic Segmentation for UGVs Based on CNN and Fully Connected CRFs", *IEEE Sensors Journal*, vol. 9, no.11, pp.4290-4298, Jun.2019.

[28] J. Ćesić, I. Marković, I. Cvišić and I. Petrović, "Radar and stereo vision fusion for multitarget tracking on the special Euclidean group", *Robotics and Autonomous Systems*, vol.83, pp.338-348, Sept.2016.

[29] C. Osborne, T. Cane, T. Nawaz and J. Ferryman, "Temporally stable feature clusters for maritime object tracking in visible and thermal imagery", *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Karlsruhe, Germany, 2015, pp.1-6.

[30] H. Wang, Z. Wei, S. Wang, C. S. Ow, K. T. Ho and B. Feng, "A vision-based obstacle detection system for Unmanned Surface Vehicle", *2011 IEEE 5th International Conference on Robotics, Automation and Mechatronics (RAM)*, Qingdao, China, 2011, pp. 364-369.

[31] A. J. Sinisterra, M. R. Dhanak and K. Von Ellenrieder, "Stereovision-based target tracking system for USV

operations", *Ocean Engineering*, vol.133, pp.197-214, Mar.2017.

[32] J. Shi and Tomasi, "Good features to track", *IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 1994, pp.593-600.

[33] N. Sundaram and K. Keutzer, "Long term video segmentation through pixel level spectral clustering on GPUs", *IEEE International Conference on Computer Vision Workshops*, Barcelona, Spain, 2011, pp. 475-482.

[34] S. Feng, Q. Zeng, B. Fan, J. Luo, H. Xiao and J. Mao, "Wear Debris Segmentation of Reflection Ferrograms Using Lightweight Residual U-Net", *IEEE Transactions on Instrumentation and Measurement*, vol.70, pp.1-11, Jul.2021.

[35] W. Zhan, C. Xiao, Y. Wen, C. Zhou, H. Yuan, S. Xiu, Y. Zhang, X. Zou, X. Liu and Q. Li, "Autonomous visual perception for unmanned surface vehicle navigation in an unknown environment", *Sensors*, vol.19, no.10, pp.2216-2227, May.2019.

[36] B. Bovcon, J. Muhovič, J. Perš and M. Kristan, "The MaSTr1325 dataset for training deep USV obstacle detection models", *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Macau, China, 2019, pp. 3431-3438.

[37] D. K. Prasad, D. Rajan, L. Rachmawati, E. Rajabaly and C. Quek, "Video Processing from Electro-optical Sensors for Object Detection and Tracking in Maritime Environment: A Survey", *IEEE Transactions on Intelligent Transportation Systems*, vol.18, no.8, pp.1993-2016, Jan.2017.

[38] Y. -S. Xu, T. -J. Fu, H. -K. Yang and C. -Y. Lee, "Dynamic Video Segmentation Network", *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6556-6565.

[39] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks", *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp.1647-1655.

**Qianqian Chen** was born in Wuhan, Hubei Province, China in 1994. She is currently pursuing the Ph.D. degree candidate in School of Navigation, Wuhan University of Technology.
Her research interests include scene segmentation and object recognition from the visual sensor using geometric and machine learning approaches.

**Changshi Xiao** received the M.S. degree in physics from Nanjing University, Nanjing, China, in 1999, and the Ph.D. degree in applied physics from Carnegie Mellon University, USA. He joined the Faculty of Wuhan University of Technology in 2012, where he is currently a Professor with the School of Navigation.

**Yuanqiao Wen** received the M.S. degree in traffic information engineering and control from Wuhan University of Technology, Wuhan, China, in 2002, and the Ph.D. degree in computer sciences from Huazhong University of Science and Technology, Wuhan, China, in 2006, He is currently a Professor with the Intelligent Transportation Systems Research Center at Wuhan University of Technology, Wuhan, China.

**Haiwen Yuan** received the M.S. degree in electrical engineering from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree in traffic information engineering and control from Wuhan University of Technology, Wuhan, China, in 2018. He is currently an Associate Professor in Wuhan Institute of Technology, Wuhan, China.

**Yamin Huang** received the M.S. degree in traffic information engineering and control from Wuhan University of Technology, Wuhan, China, in 2014, and the Ph.D. degree in computer sciences from Technische Universiteit Delft, Delft, Nederland, in 2019. He is currently a Professor with the Intelligent Transportation Systems Research Center at Wuhan University of Technology, Wuhan, China.

**Yuanchang Liu** received the M.Sc. degree in power systems engineering and the Ph.D. degree in marine control engineering from the University College London, London, U.K., in 2011 and 2016, respectively.
He was a Research Fellow in robotic vision and autonomous vehicles with Surrey Space Centre, University of Surrey, funded by UK Space Agency. He is currently a Lecturer with the Department of Mechanical Engineering, University College London. His research interests include automation and autonomy, mainly focusing on the exploration of technologies for sensing and perception, and guidance and control of intelligent and autonomous vehicles.

**Wenqiang Zhan** received the M.S. degree in mechatronic engineering from Wuhan University of Technology, Wuhan, China, in 2016, and the Ph.D. degree in traffic information engineering and control from Wuhan University of Technology, Wuhan, China, in 2021. He is currently a lecturer at Shandong Jiaotong University.

**Qiliang Li** (M'04–SM'14) received the B.S. degree in physics from Wuhan University, Wuhan, China, in 1996, the M.S. degree in physics from Nanjing University, Nanjing, China, in 1999, and the Ph.D. degree in electrical and computer engineering from North Carolina State University, Raleigh, NC, USA, in 2004. From 2004 to 2007, he was with the Semiconductor Electronics Division, National Institute of Standards and Technology, Gaithersburg, MD, USA, as a Research Scientist. He joined the Faculty of George Mason University in 2007, where he is currently a Professor with the Department of Electrical and Computer Engineering. His research interests include semiconductor devices, nanoelectronics, sensor technology, and machine learning.