University of Essex

Research Repository

A lightweight dual-branch semantic segmentation network for enhanced obstacle detection in ship navigation

Accepted for publication in Engineering Applications of Artificial Intelligence.

Research Repository link: https://repository.essex.ac.uk/38818/

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

www.essex.ac.uk

Research paper

A lightweight dual-branch semantic segmentation network for enhanced obstacle detection in ship navigation

Hui Feng [a,b], Wensheng Liu [a,b], Haixiang Xu [a,b], Jianhua He [c]

[a] *Key Laboratory of High Performance Ship Technology (Wuhan University of Technology), Ministry of Education, Wuhan, Hubei, China*
[b] *School of Naval Architecture, Ocean and Energy Power Engineering, Wuhan University of Technology, Wuhan, Hubei, China*
[c] *School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, United Kingdom*

ARTICLE INFO

Keywords:
Smart ship
Maritime obstacle detection
Real-time semantic segmentation
Light-weight network

ABSTRACT

Semantic segmentation is essential for ship navigation as it enables the identification and understanding of semantic regions, thereby enhancing the navigational capabilities of smart ships. However, current deep learning techniques encounter challenges in balancing model size and segmentation accuracy due to the complexity of water surface features. In response, we propose a novel lightweight dual-branch semantic segmentation network. The model initially utilizes a specially designed dual-branch backbone to independently extract local details and global semantics from water surface images. The detail branch compresses and reconstructs feature information to mitigate interference from water dynamics, while the semantic branch efficiently expands the receptive field to capture global object relationships. Additionally, we introduce an aggregation module that holistically guides the feature responses to facilitate the sufficient aggregation of dual-branch information. Furthermore, a cascaded fusion approach is proposed to restore diminished localization precision, while also ensuring fusion accuracy by leveraging the segmentation attributes of deep features. Experimental results on visible light datasets from real navigation scenarios demonstrate that our network achieves approximately a 10% improvement in obstacle detection precision compared to existing advanced maritime models. Moreover, within the domain of the latest lightweight and real-time research, our network attains an optimal balance among accuracy, parameter efficiency, and real-time performance. This contributes to enhancing the navigation safety of intelligent vessels and promotes adaptability for onboard deployment.

1. Introduction

The shipping industry plays an indispensable role in facilitating international trade. With advancements in modern technologies such as communication, the Internet of Things, and autopilot systems, there is a growing research momentum on smart ships with intelligent and autonomous navigation (Guo et al., 2023; Liu et al., 2024; Yang et al., 2023). To fulfill maritime market demands for safe and efficient navigation, smart ships require sensor systems to identify and evade obstacles while ensuring navigation within designated areas. Visual sensor-based obstacle segmentation systems are crucial for the safety of smart ships.

Semantic segmentation aims to assign labels to each pixel in an image to distinguish various objects. The advancements in autonomous driving have spurred extensive research into semantic segmentation (Chen et al., 2017b, 2018; Gao, 2023). However, applying these models directly to navigation scenarios proves challenging (Bovcon et al., 2019; Cane and Ferryman, 2018), as illustrated in Fig. 1. In more complex scenarios, there is often an increase in false positive predictions (FPs) and a failure to detect small waterborne targets, which poses a significant risk in real-world navigation.

Compared to traditional algorithms that rely on image grayscale distribution characteristics (Jin et al., 2019; Kristan et al., 2015; Liu et al., 2021; Lv et al., 2017), deep learning-based semantic segmen-tation approaches automatically acquire knowledge of image features. They demonstrate superior generalization performance and represent the current research mainstream (Bovcon and Kristan, 2021; Chen et al., 2021; Teršek et al., 2023). Moreover, the increasing prominence of foundation models has sparked growing interest in established vision foundation models for their ability to perform semantic segmenta-tion across various common scenarios (Kirillov et al., 2023; Zhang et al., 2023b). Nevertheless, applying existing deep learning models that solely rely on visible light images faces challenges when perform-ing semantic segmentation in navigation scenarios. On the one hand, models tailored for navigation scenarios require high parameters to achieve accurate segmentation, which presents deployment challenges

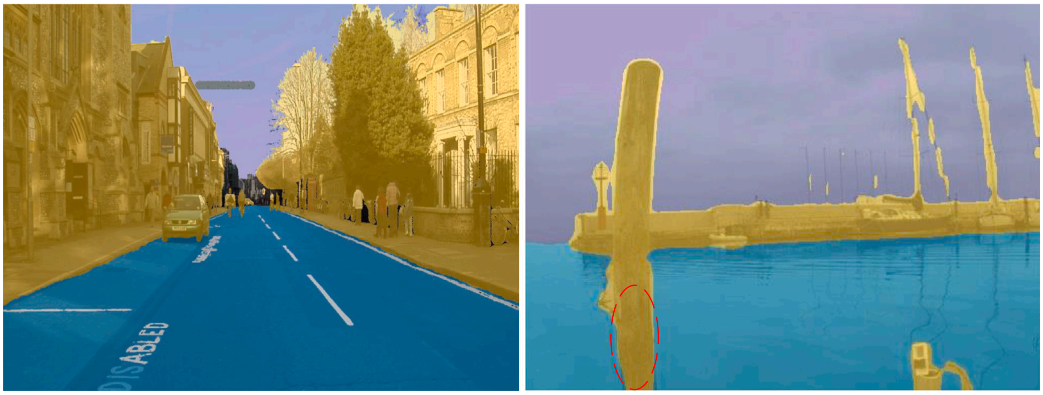

Fig. 1. Through comparison, it is evident that the water surface, serving as the navigational area for ships, is more irregular than fixed road surfaces, characterized by dynamic and uneven features. Additionally, reflections and glare pose significant disturbances to water surface segmentation.

on ship edge platforms, especially for foundation models that demand substantial training and hardware resources (Zhang et al., 2024; Bommasani et al., 2021). On the other hand, in contrast to the autonomous vehicle domain, semantic segmentation datasets collected in naviga-tion scenarios are often limited in size (Bovcon et al., 2019; Lambert et al., 2022; Žust et al., 2023), hindering the generalization perfor-mance of models. Even with vision foundation models, maintaining high segmentation accuracy in complex navigation scenarios remains a challenging task. Hence, developing a segmentation algorithm that maintains lightweight characteristics while achieving high accuracy and demonstrating superior generalization in navigation scenarios has emerged as a pressing issue to address.

To effectively address the challenges and application demands of obstacle segmentation models in navigation scenarios, this study introduces a novel model, AF-BiSeNet, based on a dual-branch architecture (Yu et al., 2021), as shown in Fig. 2. The model extracts semantic and detail features from the input through dual branches, aggregates them using an aggregation module, and finally fuses the multi-scale features for output. Unlike existing research, both the aggregation and fusion processes in AF-BiSeNet are lightweight. In the detail branch, our proposed Embedded Feature Refinement Module (EFRM) compresses feature vectors into *n*-dimensions instead of 1-dimension to better mitigate interference from high-frequency information in water surface images. Moreover, it employs asymmetric convolution processing to facilitate interaction among non-compressed direction vectors. In the semantic branch, the enhanced capability of a larger receptive field to extract feature semantics is utilized by employing dilated convolution for selected channel features in the Dilated Gather and Expansion (DGE) layer. The Bilateral Refinement Aggregation (BRA) module is designed to aggregate dual-branch features. It utilizes dual-branch information to generate aggregation weights for each branch, thereby enhancing the accuracy of aggregating both local details and global semantics. Finally, the proposed Cascading Top-down Enhanced Fusion (C-TEF) method segments the features of each stage to reduce channel counts and utilizes semantic properties during segmentation to achieve enhanced fusion of multi-scale information from top to bottom. The main contributions of this paper can be summarized as follows:

(1) The EFRM used for detail branch is capable of mitigating water wave interference, while improving the branch's capacity to retain and encode important features. In addition, the semantic branch's DGE layer efficiently expands the receptive field to better address issues related to water surface reflections and glare, while also maintaining computational efficiency.
(2) We introduce a novel feature aggregation mechanism, BRA, which comprehensively guides the response of dual-branch features in both channel and spatial dimensions. It is designed to provide more effective features for subsequent segmentation heads.
(3) The proposed C-TEF employs channel reduction and enhanced fusion to restore accurate positioning information of shallow layers in a more lightweight manner, thereby improving segmentation accuracy for boundaries and small targets.
(4) The experimental results demonstrate AF-BiSeNet's superior precision and generalization performance on navigation scenario datasets, with approximately 10% higher precision in obstacle detection compared to existing advanced maritime models. It also exhibits enhanced parameter efficiency, making it more adaptable for practical applications.

The rest of this article is organized as follows. Section 2 presents a survey of existing works on water surface semantic segmentation, feature aggregation, and multi-scale fusion. Section 3 provides a detailed introduction to the structural design and principles of AF-BiSeNet. Section 4 provides experimental comparison and analysis of AF-BiSeNet and existing methods. In Section 5, we conclude the work on AF-BiSeNet and discuss future research directions. Our code, model, and data can be accessed on the following website: https://www.alipan.com/s/Tx4phbHo792.

2. Related work

2.1. Semantic segmentation of navigation scenarios

Semantic segmentation serves not only to differentiate the contours of obstacles on the water surface but also to partition navigable areas, thereby enabling a global understanding of the navigation environment. This can provide strong support for local path planning of smart ships (Hong et al., 2019; Ni et al., 2020), thus laying the necessary foundation for achieving autonomous navigation.

Traditional water surface image segmentation algorithms can be categorized into two main groups: threshold-based and graph theory-based. Li et al. (2020) proposed a threshold segmentation method based on uniformity measurement by combining the one-dimensional Otsu method with a uniformity measurement strategy. Bovcon et al. (2018) extended the graphical model for probabilistic semantic segmentation by converting ship attitude information from the Inertial Measurement Unit (IMU) into a prior horizon position.

In recent years, with the increasing emphasis on smart ships, there has been a notable enhancement in the performance of deep learning segmentation methods. To achieve higher accuracy, existing meth-ods often employ architectures with an encoder followed by a high-parameter decoder, such as Water Segmentation and Refinement (WaSR) (Bovcon and Kristan, 2021), its variant eWaSR (Teršek et al., 2023), and Water Obstacle Detection Network Based on Image Segmentation (WODIS) (Chen et al., 2021). There are also architectures that utilize a significant number of dilated convolutions (Xue et al.,

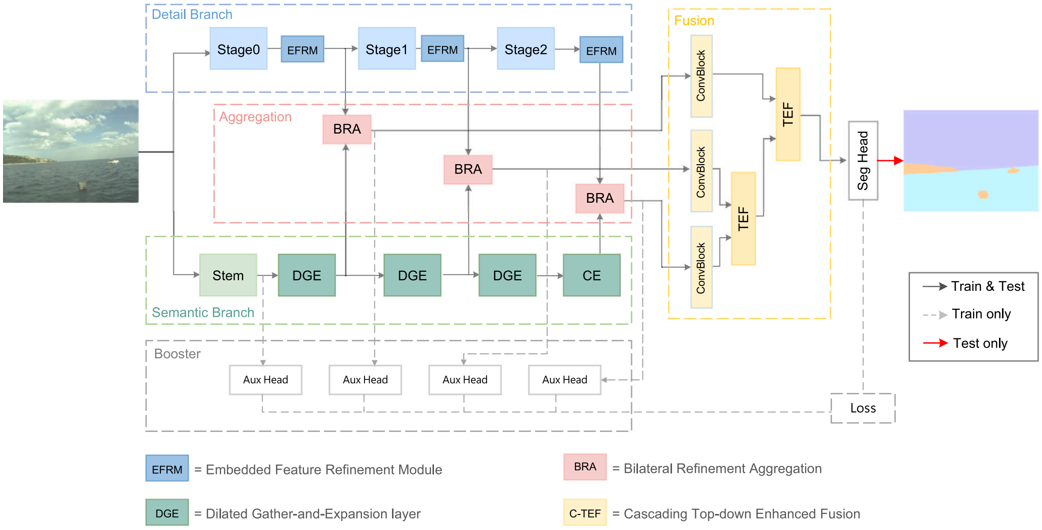

Fig. 2. The overall network structure of AF-BiSeNet. Stages 0 through stage 2 all denote stacks of convolutions. "Seg Head" and "Aux Head" denotes the segmentation head, which consists of multiple layers of convolutions. The structure of these heads, Stem layer and CE layer are consistent with BiSeNetV2 (Yu et al., 2021).

2021). These models typically have significantly large sizes and require extensive resources for training and deployment. More recently, the Vision Transformer (Dosovitskiy et al., 2021), which excels at capturing remote contextual information, has exhibited promising performance (Zhang et al., 2023a). Nevertheless, its extensive computational demands have constrained its practical utility. Furthermore, despite considerable attention on foundational models (Kirillov et al., 2023; Zhang et al., 2023b), their high deployment requirements, inadequacy of navigation scenario datasets, and issues related to semantic granularity levels (Li et al., 2023b) impede their direct applicability to perception tasks in navigation scenarios.

In the realm of lightweight and real-time semantic segmentation research, the Bilateral Segmentation Network (BiSeNet) series of models (Yu et al., 2018, 2021; Tsai and Tseng, 2023) pioneered the dual-branch model by specifically designing semantic and detail branches based on channel and network depth considerations. Over recent years, numerous variant models have been derived from the BiSeNet series. Short-Term Dense Concatenate (STDC) (Fan et al., 2021) has rethought BiSeNet by incorporating a detailed guidance module that leverages edge information to direct shallow layers in learning spatial information through a single-stream manner. However, irregular water surfaces in navigation scenarios may introduce detrimental effects on detail guidance. Proportional-Integral-Derivative Net (PIDNet) (Xu et al., 2023b) incorporates control theory into semantic segmentation by designing a three-branch network, but it also increases the training and inference burdens. Ranjbarzadeh et al. (2023) employs multiple encoding approaches to generate 11 distinct images, enabling the proposed efficient cascade Convolutional Neural Network (CNN) to analyze input textures more effectively, thus eliminating the need for deep CNN models. With the increasing research, there have been recent real-time and lightweight semantic segmentation networks based on Transformer (Wang et al., 2022; Xu et al., 2023a). Nonetheless, achieving a satisfactory balance between parameter efficiency and accuracy remains challenging for these models.

Although lightweight architectures may reduce the training and deployment requirements, their segmentation accuracy still has great potential for improvement due to insufficient consideration of the specific characteristics of water surface images.

2.2. *Feature aggregation*

For maritime navigation, accurate identification of obstacle positions is fundamental for safe sailing, while the effective extraction of water surface boundaries serves as an indicator of a vessel's navigational status. Consequently, the precise aggregation of semantic and detailed feature information from images is crucial for semantic segmentation in navigation scenarios.

Existing research has basically adopted adaptive weighting methods to selectively integrate different information. The Bilateral Guided Aggregation (BGA) of BiSeNetV2 employed the output of the semantic branch as attention weights to amplify the responses of the detail branch. However, this approach somewhat reduced the contribution of high-level semantic information contained in the semantic branch. The Feature Fusion Module (FFM) (Bovcon and Kristan, 2021; Chen et al., 2021; Yu et al., 2018) employed channel attention to automatically learn aggregation strategies in the channel dimension. However, this approach does not consider the differences in spatial information among various feature maps. The Bilateral Fusion module, proposed by Zhang et al. (2021), employs spatial and channel attention mechanisms to enhance features within each branch. The enhancement within a single branch has a limited effect on the aggregation of features across branches. Li et al. (2023) posited that targets of varying scales require different contextual information for accurate discrimination. Therefore, a spatial selection mechanism is employed to enhance only the regions in the feature map that are adapted to its receptive field, and finally complement different features' target information through addition.

Efficient aggregation techniques must take into account the variations in information conveyed by different features. However, current methods, despite adaptively acquiring weights, do not comprehensively account for both spatial and channel dimensions in aggregation.

2.3. *Multi-scale fusion*

Excessive downsampling of the input may lead to diminished accuracy in locating sea-sky lines and shorelines in water surface images, as well as result in the loss of feature information for small obstacles at longer distances. To address these issues, it is essential to fuse multi-scale features to restore shallow details.

The existing multi-scale fusion methods basically maintain the high channel characteristics of features throughout the fusion process. For instance, the Joint Pyramid Upsampling (JPU) module (Wu et al., 2019) processes the last three layers of the Fully Convolutional Network (FCN) (Long et al., 2015) features. It utilizes convolutions with varying dilated ratios to capture feature information at different stages. Similarly, Gao (2023) obtained a decoder structure with relatively optimal accuracy through extensive experiments. However, due to the high-channel nature, the fusion structure often contains a large number of parameters. In contrast, Zha et al. (2021) initially conducted feature segmentation and then fused the segmentation outcomes to reduce the number of feature channels, a process termed segmentation fusion. Nevertheless, the effectiveness of the channel attention employed in this method is somewhat limited when dealing with a small number of channels. Van Quyen and Kim (2023) leveraged the low confidence of ambiguous regions and computed semantic attention by subtracting the confidence from 1, aiming to direct the model's focus towards these regions.

The multi-scale fusion method for semantic segmentation models often leads to an excessive number of parameters because of the high-channel nature of features. However, despite streamlining the fusion process through the segmentation fusion mentioned above, the substantial reduction in feature channels makes it challenging to effectively integrate multi-scale information.

3. Proposed method

The overall structure of AF-BiSeNet is shown in Fig. 2. The segmentation process for water surface images comprises three main steps. Initially, the input water surface image undergoes dual-branch processing to extract semantic and detail information separately. Subsequently, the proposed BRA is applied across multiple stages of the model to thoroughly aggregate the dual-branch information. Finally, a top-down multi-scale enhanced fusion is used to fuse the aggregated features and obtain the final output.

The distinction between the AF-BiSeNet model and existing models lies in the design of its individual modules. In the feature extraction section, the semantic branch extracts high-level semantics through the DGE layer, which efficiently expands the receptive field, while the detail branch incorporates the EFRM designed to acquire spatial attention, aiming to reduce high-frequency interference on the water surface. In the aggregation section, the BRA module comprehensively directs the response of dual-branch features and fully aggregates the feature information from both branches. In the multi-scale fusion section, the C-TEF module compresses the channels of BRA output features and leverages the semantic properties of deep features to achieve enhanced fusion. It is noteworthy that both the aggregation and fusion modules of AF-BiSeNet are highly lightweight. Further details about the model can be found in the annotations in Fig. 2.

3.1. Detail branch and semantic branch

Detailed image information is indispensable for semantic segmentation models. In addition, a global perception of the image is necessary to ensure comprehensive connectivity among regions belonging to the same category. Therefore, the backbone of AF-BiSeNet comprises the detail branch, which is dedicated to extracting local details, and the semantic branch, which focuses on extracting global semantics.

Detail branch. The design of the detail branch incorporates fewer 3×3 convolution downsampling stages and wider channel dimensions to focus more on the underlying details. In navigation situations, the dynamic nature of the water surface often gives rise to noticeable interference, including clutter and flashes. This interference poses a significant challenge for the current detail branch due to its small receptive fields. In response to this challenge, we propose the EFRM, as shown in Fig. 4, which empowers the detail branch to filter out high-frequency interference and concentrate more on the principal objects.

The design of the EFRM module is inspired by word embedding in natural language processing (Bengio et al., 2000). As depicted in Fig. 3, word embedding transforms the sparse vectors of one-hot encoding into dense vectors. This enables the representation of internal relationships between word vectors while reducing dimensionality. Similarly, we posit that the vector of the image along a single direction also exhibits sparsity because of the similarity in pixel distribution within areas such as the water surface and sky. The EFRM module initially compresses these vectors into *n*-dimensions through maximum and average pooling. Subsequently, asymmetric convolution processing is employed to facilitate vector interactions in the uncompressed direction during encoding. The process of handling feature vectors in EFRM can be regarded as the extraction of vector principal components. By compressing and encoding the vectors horizontally or vertically, it tends to extract the invariant components while disregarding irregular noise. Consequently, regardless of the intricate nature of water waves, EFRM tends to overlook high-frequency dynamic disturbances in the image, resulting in consistent outputs and a heightened emphasis on the relatively stable visual characteristics of objects on the water surface. Furthermore, the encoded vectors in both directions are used to reconstruct spatial attention through matrix multiplication. The spatial attention is then used to refine the detailed branch features, thereby directing the branch's focus towards the principal objects on the water surface. In comparison to existing spatial attention mechanisms (Hou et al., 2021; Tsai and Tseng, 2023), the proposed EFRM effectively preserves spatial information while enhancing the module's capacity to represent crucial information. This is achieved by pooling features axially into *n*-dimensional vectors and utilizing asymmetric convolutional encoding.

In addition, we incorporate convolutional reparameterization from ACNet (Ding et al., 2019) into the convolution of the detail branch. By integrating reparameterized convolutional blocks, we can increase the number of parameters during model training without affecting the inference speed, thereby effectively enhancing the performance of the detail branch in extracting local details.

Semantic branch. The semantic branch efficiently captures global semantic information through multiple network layers and narrower channel dimensions. In the context of water surface images, a larger receptive field can capture the spatial and semantic relationships between the water surface area and obstacles, thereby enhancing the network's overall understanding of the scene. The widely employed dilated convolution has been proven effective in increasing the receptive field. However, due to its discontinuous memory access in contrast to regular convolutions, the excessive use of dilated convolutions significantly undermines the operational efficiency of the model (Gao, 2023). To address this issue, we propose the DGE as an efficient approach to expand the receptive field.

The structure of the DGE layer is shown in Fig. 4. Initially, the grouped semantic branch features are processed separately using regular convolution and dilated convolution. This strategy is designed to leverage the high receptive field properties of dilated convolution while mitigating excessive memory access costs. Then, the feature channels are expanded and reduced to extract information in a high-dimensional space (Sandler et al., 2018). Finally, considering the varied receptive fields of the features across different channels, we incorporate the Efficient Channel Attention (ECA) (Wang et al., 2020) to enable the module to autonomously select the parts of features that are effective for segmentation.

The DGE layer effectively enhances the receptive field of branches by using dilated convolution on partial channels. Compared with existing research on expanding the receptive field of images (Chen et al., 2017a; Dosovitskiy et al., 2021; Gao, 2023), the DGE layer demonstrates higher operational efficiency.

3.2. Bilateral refined aggregation

The model in this paper employs a dual-branch architecture as its backbone, leading to significant distinctions in the output features. To ensure that the segmentation head has access to features containing both global semantics and local details, it is necessary to merge the information from dual-branch features.

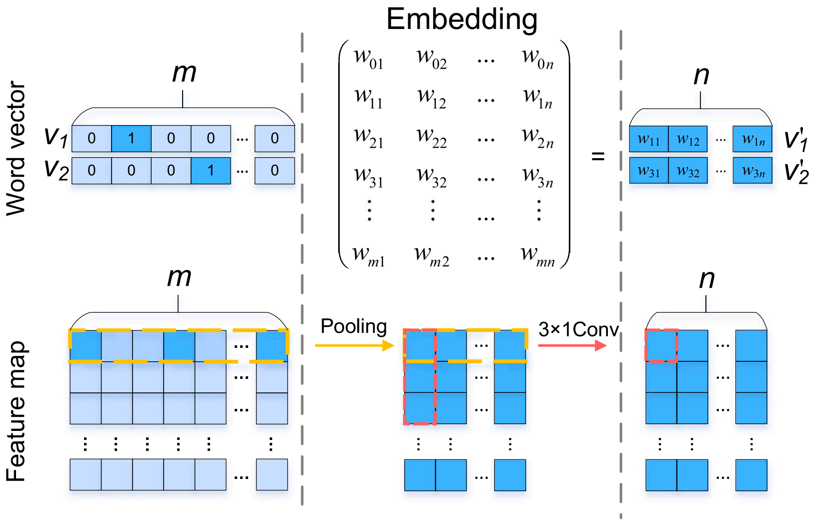
Embedding
Word vector
m
v_1 0 1 0 0 ... 0
v_2 0 0 0 1 ... 0
w_{01} w_{02} ... w_{0n}
w_{11} w_{12} ... w_{1n}
w_{21} w_{22} ... w_{2n}
w_{31} w_{32} ... w_{3n}
w_{m1} w_{m2} ... w_{mn}
=
n
w_{11} w_{12} ... w_{1n} v'_1
w_{31} w_{32} ... w_{3n} v'_2
Feature map
m
Pooling
3×1Conv
n

Fig. 3. Word embedding and EFRM processing (taking horizontal calculation as an example).

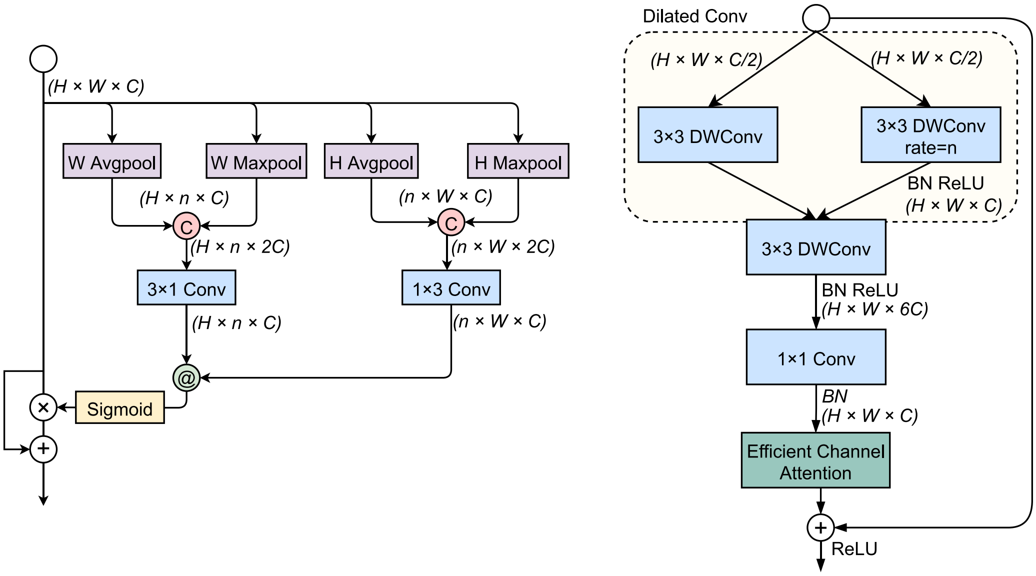

Fig. 4. The Embedded Feature Refinement Module (EFRM) on the left and the Dilated Gather and Expansion (DGE) layer on the right. “C” denotes concatenation; “+” denotes element-wise addition; “×” denotes element-wise multiplication; “@” denotes matrix multiplication, the same below.

Current aggregation methods generate a solitary adaptive factor to weigh multiple features, resulting in the inability to flexibly enhance complementary information and suppress redundant information during aggregation. Furthermore, the aggregation of these methods is limited to a single channel or spatial dimension. Considering these factors, we propose the BRA module, as shown in Fig. 5, which fully utilizes dual-branch information to generate multiple specific weights for guiding feature responses and aggregation. Initially, this module preprocesses the features from dual branches using depthwise separable convolution (Howard et al., 2017), followed by concatenating them to generate attention weights for the channel and spatial dimensions through convolutional and attention mechanisms. These attention weights are then split and assigned to each corresponding branch. Finally, the dual-branch features are weighted to refine the features, facilitating the selective aggregation of semantic and detail information through element-wise addition.

The BRA module utilizes concatenated features to generate adaptive weights instead of employing them directly as subsequent outputs, allowing for the simultaneous integration of dual-branch information to generate attention. Furthermore, dividing the acquired attention weights enables specialized refinement in different branches. The specificity is manifested by the fact that, after the learning process, the attention dedicated to refining the semantic branch can better focus on global features, such as water surface areas and shore regions. Similarly, the attention devoted to refining the detail branch is directed towards local details like water obstacles and segmentation boundaries. The summation of weighted dual-branch features ultimately facilitates the effective aggregation of localized texture details and global semantic information within each respective branch. In comparison to existing research on feature aggregation (Bovcon and Kristan, 2021; Yu et al., 2021; Yu et al., 2018), the BRA module distinguishes itself by maintaining efficiency while comprehensively considering features in both the channel and spatial dimensions.

Additionally, the segmentation loss of the output from the BRA is computed by utilizing the “Aux Head” as mentioned in Fig. 2, to provide supervision for the module’s output during the training process.

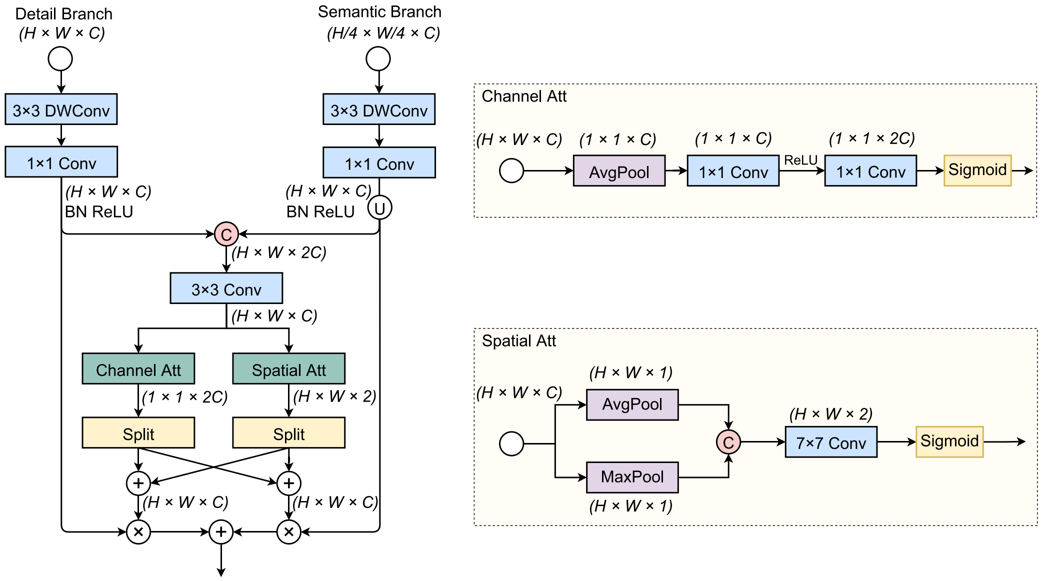

Fig. 5. The composition of Bilateral Refinement Aggregation (BRA) and its attention mechanisms pertaining to both channel and spatial aspects. "U" denotes upsampling operation, the same below.

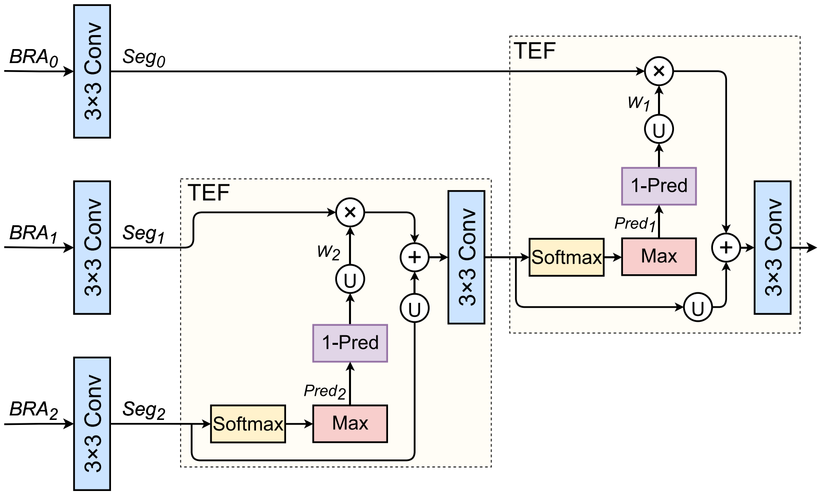

Fig. 6. Cascading Top-down Enhanced Fusion (C-TEF) involves cascading TEF modules. This module efficiently facilitates enhanced fusion of features from deep to shallow layers in the network. Here, BRA_i denotes the output of the ith BRA module in the model; Seg_i represents the result of channel compression; "Softmax" denotes the normalization calculation of inputs on each channel; "Max" represents obtaining the prediction result by taking the maximum value in the channel dimension; "1-Pred" signifies subtracting the predicted result from 1 to obtain semantic weights for enhancing shallower features.

These headers can be removed during inference, thereby improving segmentation accuracy without sacrificing inference speed.

3.3. Cascading top-down enhanced fusion

Excessive downsampling of images may result in reduced accuracy when localizing water surface boundaries and lead to the loss of small obstacles near the sea-sky line. Hence, it is essential to conduct multi-scale fusion at various stages to recover position and feature information embedded in shallow features.

Current multi-scale fusion methods inadvertently incorporate a large number of parameters to accommodate the high-channel nature of features. Conversely, lightweight fusion methods suffer from a significant weakening of the representation capacity of multi-scale features due to the substantial reduction in channels. To maintain the model's lightweight design while fusing features from different scales, we propose the C-TEF module, as shown in Fig. 6. Firstly, the module efficiently condenses the channel dimensions of outputs BRA_i from multiple BRA modules, as shown in Fig. 2, by means of convolution to align with the number of semantic categories. This process obtains Seg_i, with a higher value of 'i' denoting a deeper layer in the network. This strategy significantly reduces computational complexity. Secondly, we employ the computational procedure outlined in Eq. (1) to derive the semantic weight W_i by utilizing the feature Seg_i. W_i is a single-channel feature map with the identical shape as Seg_{i-1}. Following Eq. (2), W_i is used to perform element-wise multiplication with Seg_{i-1}. Afterwards, the weighted Seg_{i-1} is then fused by addition with the upsampled Seg_i, resulting in Seg'_{i-1}, which is utilized to calculate W_{i-1} for the next round of enhanced fusion. Ultimately, when the shallowest Seg_0 is involved in the cascading operation of the C-TEF module, the result is an enhanced fusion of the model's top-down features. This result will serve as input to the "Seg Head" depicted in Fig. 2 to generate the final segmentation output. A detailed explanation of the calculation of semantic weights W_i will be provided later.

$$W_i = Up_{i-1}(1 - Max_{channel}(\text{softmax}(Seg_i))) \quad (1)$$

$$Seg'_{i-1} = \mathrm{Conv}_{3\times 3}[Seg_{i-1} \cdot W_i + Up_{i-1}(Seg_i)] \tag{2}$$

Seg_i represents the output after channel compression of BRA_i; $Up_{i-1}(\cdot)$ represents upsampling the feature size to match Seg_{i-1} through bilinear interpolation; $Conv_{3\times3}$ represent 3 × 3 convolutions; $Max_{channel}$ represents taking the maximum value on the channel dimension. The obtained Seg'_{i-1} will be used for the calculation of new semantic weights W_{i-1} for the next enhancement fusion process.

In water surface images, the classification of regions like sea-sky lines, water shorelines and adjacent small targets is often ambiguous, resulting in low confidence in segmenting these areas. Thus, based on this principle, Eq. (1) calculates semantic weights by subtracting the segmentation's confidence score from 1. Higher semantic weights indicate regions that are challenging to recognize, and incorporating these weights into shallow features enables a more concentrated focus on these areas during fusion. The process of enhanced fusion is systematically applied from the top of the network to the shallow layer, facilitating the comprehensive fusion of multi-scale features.

4. Experimental results and analysis

4.1. Implementation details

In this paper, BiSeNetV2 (Yu et al., 2021) is used as a baseline to carry out the study. To validate the effectiveness of the proposed method, two classic datasets, MaSTr1478 (Bovcon et al., 2019; Zust and Kristan, 2022) and LaRS (Žust et al., 2023), were carefully selected for verification. The MaSTr1478 dataset is a collection of images depicting coastal scenes and complex navigational scenes. 1034 images in the dataset were used for training, and the remaining 444 were used for testing. Due to incomplete IMU information in the dataset and to ensure fairness, the IMU information was not used in the model training process. The LaRS dataset aims to address the lack of diverse datasets in maritime obstacle detection to fully capture the complexity of typical maritime environments. In the dataset, 2102 images are used for training, while the remaining 701 images are used for testing. The semantic categories for segmentation are defined as three types: water, obstacle, and sky.

In this study, AF-BiSeNet is trained using an SGD optimizer with a batch size of 8 and a learning rate of 0.05. The training process was conducted for a total of 200 epochs, and a learning rate decay strategy with a factor of 0.9 was used. In terms of the models selected for comparison, their training parameters are chosen to match those described in the respective papers as much as possible. All experiments were trained on a single NVIDIA A100 GPU and tested on a single NVIDIA RTX 3060 Ti GPU. The image resolution in the experiments is 384 × 512. The loss function uses the commonly employed cross-entropy loss. Meanwhile, the segmentation loss and the auxiliary loss weight coefficients are both set to 1 at each stage. This setting helps the network generate accurate segmentation results to ensure fusion accuracy.

4.2. Evaluation metric

For the multiclass semantic segmentation problem, the evaluation metrics used to assess the performance of the model are Pixel Accuracy (PA) and Mean Intersection over Union (MIoU) (Long et al., 2015). The formulas are as follows:

$$PA = \frac{\sum_i n_{ii}}{\sum_i t_i} \tag{3}$$

$$mIoU = \frac{1}{n_{cls}} \cdot \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}} \tag{4}$$

In the formula, n_{ij} is the number of pixels that predict category i as category j; n_{cls} indicates the number of target categories; $t_i = \sum_j n_{ij}$ represents the total number of pixels in the ground truth for the target category i.

Table 1
Discussion of experiments on the detail branch. The term "Baseline" represents BiSeNetV2. Group 1 represents the segmentation accuracy of EFRM using different pooling dimensions, denoted as *n*. Group 2 represents ablation experiments on the detail branch. (1) denotes the EFRM with pooling dimension 2, and (2) denotes the addition of the convolutional reparameterization method.

Method		IoU (%)	
		Obstacle	Water
	n		
Group 1	1	90.16	97.91
	2	**90.44**	**98.09**
	3	90.14	98.00
	4	90.01	97.95
	8	89.92	97.93
Group 2	Baseline	89.79	97.85
	Baseline +(1)	90.44	98.09
	Baseline +(1)+(2)	**90.55**	**98.10**

PA and MIoU primarily evaluate the accuracy of image segmentation. However, in the context of water surface scenes, achieving precise segmentation of obstacles may not significantly improve these metrics due to their limited proportion within the image. Therefore, this paper also employs the maritime obstacle detection benchmark MODS (Bovcon et al., 2021) to evaluate the model's performance. The precision of water surface boundary lines evaluated through mean square error, while obstacle detection is measured by precision (Pr) and recall (Re), as well as the F1 score, which represents the harmonic mean of the two.

4.3. Quantitative and qualitative results

In this section, a series of experiments is conducted on navigation scenario datasets to evaluate the effectiveness of the proposed methodologies. The intricate appearance and diverse distribution of obstacles, coupled with their relatively small proportion within water surface images, make obstacle segmentation accuracy crucial in reflecting the model's performance. If not otherwise specified, the following experimental results are based on the MaSTr1478 dataset.

4.3.1. Discussion of experiments on the detail branch

In order to impress our EFRM, we demonstrate the effect of pooling dimension *n* on model performance in Group 1 of Table 1. The results indicate that utilizing different pooling dimensions all leads to improved segmentation accuracy compared to the baseline model. Furthermore, it can be noted that as the pooling dimension *n* increases, the IoU accuracy of obstacle and water surface segmentation initially increases and then decreases. The rationale behind this phenomenon lies in the positive correlation between dimensionality and the capacity of feature vectors to effectively represent crucial information, leading to improvements in segmentation accuracy with the integration of EFRM. However, as the dimension continues to increase, the optimization difficulty will also escalate because of the growing volume of data, and the compressed and encoded vectors become more susceptible to noise. Consequently, the overall accuracy tends to exhibit an initial rise followed by a decline. The experimental results indicate that in navigation scenarios, setting the pooling dimension *n* of EFRM to 2 can optimally capture the primary information of features in a single direction. Furthermore, as evidenced by the experimental results of different attention modules in Table 2, our proposed EFRM demonstrates greater competitiveness in enhancing segmentation accuracy.

The incorporation of the convolutional reparameterization method contributes to an improvement in segmentation accuracy, as shown in Group 2 of Table 1. This indicates that adding a higher number of reparameterizable convolutions during training effectively enhances the capability of the detail branch to extract local features.

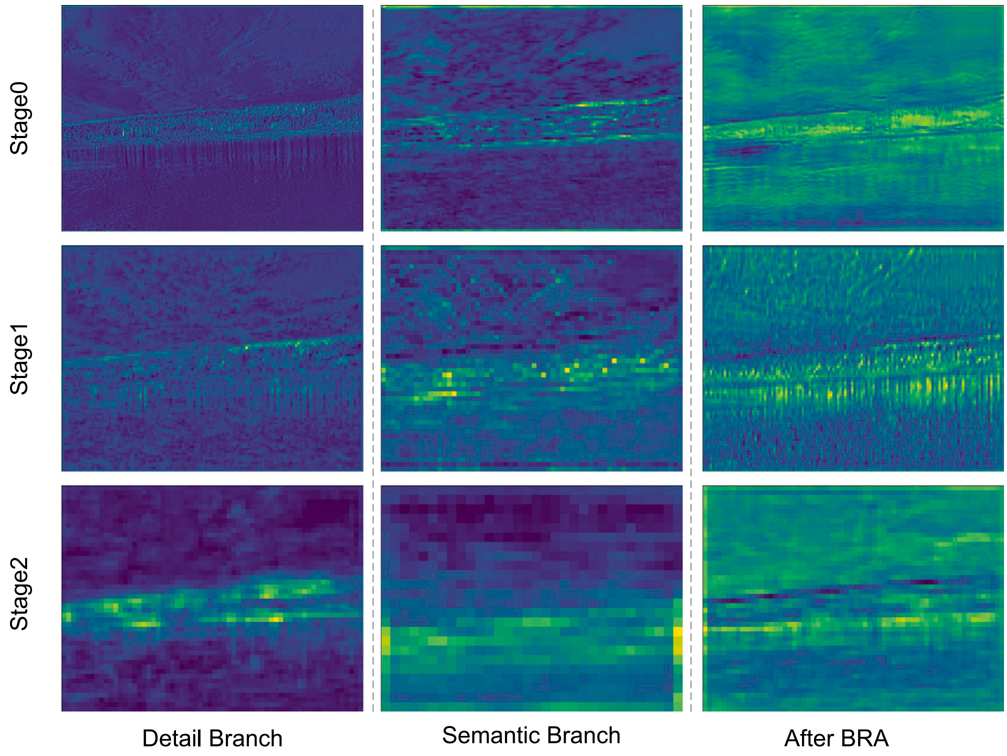

Fig. 7. The features are represented by calculating the average of multiple channels. Observing the features at different stages reveals that the detail branch extracts more high-frequency information, but fewer regions are activated. Conversely, the semantic branch exhibits a heightened feature response, yet its capacity for localization is limited, and the extracted local features are sketchy. After the proposed BRA module aggregates the information, the features not only preserve the detailed information but also demonstrate an enhanced overall feature response.

Table 2
The comparative experimental results of integrating different modules independently into the baseline, which represents BiSeNetV2. The attention module is applied to the detail branch, the feature aggregation method is added to the final stage of the baseline, and the multi-scale fusion method integrates features from three stages.

Method	Module	IoU (%)	
		Obstacle	Water
Attention mechanism	CA (Hou et al., 2021) CFRM	89.89	97.94
	(Tsai and Tseng, 2023)	89.84	97.87
	Ours EFRM	**90.44**	**98.09**
Feature aggregation	BGA (Yu et al., 2021)	89.79	97.85
	FFM (Yu et al., 2018)	89.79	97.87
	Ours BRA	**90.35**	**97.96**
Multi-scale fusion	RegSeg fusion (Chen et al., 2017b)	90.46	97.96
	Ours C-TEF	**90.76**	**98.02**

4.3.2. *Comparative experiments of different aggregation and fusion methods*

To demonstrate the effectiveness of the proposed BRA module and C-TEF method, we conducted comparative experiments with existing methods on the baseline model. The IoU results of different feature aggregation modules are presented in the “Feature aggregation” group in Table 2, with FFM serving as a commonly used in contemporary water surface semantic segmentation models. The findings indicate that the BRA module in this study outperforms other methods by 0.56% in terms of obstacle IoU accuracy. This indicates that the aggregated features of the BRA module contribute more effectively to the accuracy of segmentation. Additionally, Fig. 7 illustrates the disparities before and after the aggregation of dual branch features in the BRA module at different network stages. In summary, the features generated by the BRA module adeptly preserve local information, such as water surface boundaries and obstacle textures, while utilizing the advanced semantics of semantic branches to activate more regions.

The experimental results of the “Multi-scale fusion” group in Table 2 indicate that our proposed cascade-enhanced fusion method is more effective in improving segmentation accuracy when utilizing features with channels only equal to the number of categories, compared to the RegSeg decoding method. Furthermore, the computed semantic weights in Fig. 6 are visualized for further clarification. As shown in Fig. 8, the water-shore boundaries, sky-shore boundaries, and regions of small targets in the input images are all marked. It is evident that the weight maps exhibit brightness at these marked locations, indicating that semantic weights assign higher values to these areas, thereby placing greater emphasis on challenging boundaries and small target regions during the fusion process.

4.3.3. *Ablation study of AF-BiSeNet*

Table 3 presents the ablation study results for the proposed AF-BiSeNet, with BiSeNetV2 (Yu et al., 2021) as the baseline. The study investigates three innovative research aspects: semantic branch design (1), detail branch design (2), and the proposed BRA and C-TEF modules (3). Experimental findings indicate that, compared to the baseline, the obstacle IoU accuracy improves by 0.73% and 0.88% after systematically incorporating the semantic and detail branch design. Through multiple experimental validations, we found that setting the dilation rates of the three DGEs in the semantic branch to 4, 4, and 2, respectively, achieves an optimal balance between model accuracy and real-time performance. The dilation rate for the last DGE stage was set to 2 because increasing the dilation rate for deeper DGEs did not significantly enhance accuracy. This could be attributed to the fact that, at the same dilation rate, convolutional kernels capture more pronounced discrepancies in feature information across different positions for deeper and smaller-scale feature maps. Dilated convolutions encounter challenges in capturing pixel relationships across distant spatial distances in navigation scenarios, resulting in minimal improvements or even potential declines in segmentation precision.

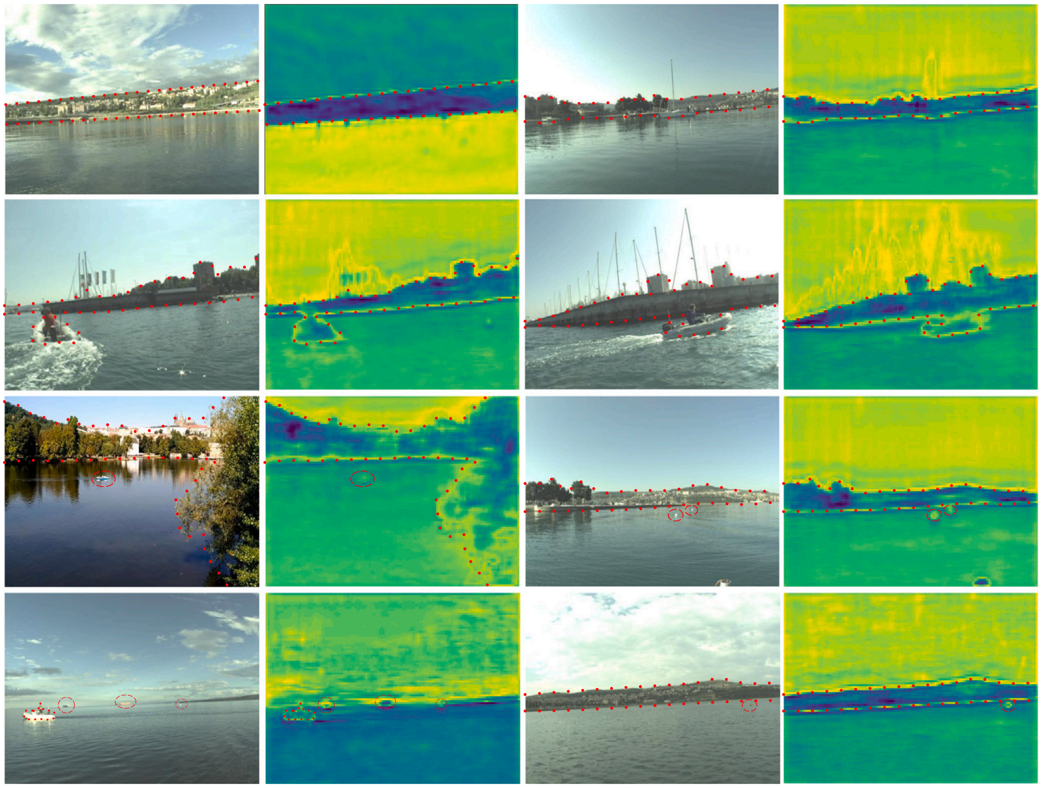

Fig. 8. The visualization of navigation scenario images paired with their corresponding C-TEF semantic weight W_1, as shown in Fig. 6. In this visualization, the red dots denote boundaries with increased weight values, while red dashed lines delineate areas with small water surface targets. It is observed that, in various scenarios, elevated weights are not only assigned to boundaries but also to regions with small water surface targets, demonstrating a heightened attention of the semantic weights on these challenging areas for classification purposes.

Table 3
Baseline refers to BiSeNetV2 (Yu et al., 2021). (1) denotes the semantic branch design in 3.1 of this paper. Through experimental validation, setting the dilation rates of the three DGEs in the model to 4, 4, and 2, respectively, has proven to significantly enhance accuracy; (2) denotes the detail branching design; (3) denotes the experimental results with the addition of the BRA module of 3.2 and the C-TEF method of 3.3.

Method	(1)	(2)	(3)	IoU (%)			MIoU (%)	PA (%)	Params (M)
				Obstacle	Water	Sky			
Baseline				89.79	97.85	98.64	95.42	98.53	2.16
Method 1	√			90.52	97.98	98.75	95.75	98.63	**1.89**
Method 2	√	√		90.67	98.01	98.75	95.81	98.65	2.18
AF-BiSeNet	√	√	√	**91.14**	**98.13**	**98.79**	**96.02**	**98.72**	2.13

Upon implementing the BRA aggregation module and C-TEF method proposed in this study, the obstacle IoU improves by 1.35%. Furthermore, due to the adoption of group convolution in DGE, and the adjustment of the features in the first stage of the detail branch to match those of the corresponding features of the semantic branch in terms of channel dimensions, it can be observed that the improvements made to the detail and semantic branches do not introduce excessive parameters. Importantly, due to the extremely lightweight nature of both the BRA and C-TEF modules, the parameter count of AF-BiSeNet is even reduced compared to the baseline model. In summary, our proposed approach significantly improves obstacle segmentation accuracy while maintaining high accuracy in water surface segmentation, and also reduces the parameter count of the model.

4.3.4. Comparative experiments of different model

In this section, the proposed AF-BiSeNet will be compared on various benchmarks with recent semantic segmentation models tailored for navigation scenarios, as well as models designed for lightweight and real-time applications.

Evaluation on the MaSTr1478 and LaRS datasets. The training details of the dataset are provided in Section 4.1. Table 4 demonstrates the comparison of the models on different datasets, with AF-BiSeNet achieving the highest accuracy in segmentation metrics. Notably, in terms of the crucial obstacle IoU metrics, AF-BiSeNet outperforms WODIS and WaSR by 2.23% and 2.54% respectively on the MaSTr1478 dataset. On the LaRS dataset, it exceeds WODIS and WaSR by margins of 2.57% and 1.69%, respectively. This indicates that the research presented in this paper can achieve superior accuracy compared to existing advanced maritime segmentation models. Furthermore, a comparison with recent lightweight and real-time models demonstrates that AF-BiSeNet exhibits strong competitiveness on navigation scenario datasets. The qualitative analysis of the segmentation results is shown in Fig. 9.

Evaluation on the MODS dataset. The improvement in segmentation metrics solely signifies a more accurate comprehension of the entire image, while the capability to precisely detect obstacles within the navigable range is essential for the safe navigation of smart ships. Thus, to further demonstrate the superior precision in obstacle detection and generalization performance of the proposed AF-BiSeNet, we assessed multiple models using 2169 images from the MODS dataset. It is worth emphasizing that the models were exclusively trained on

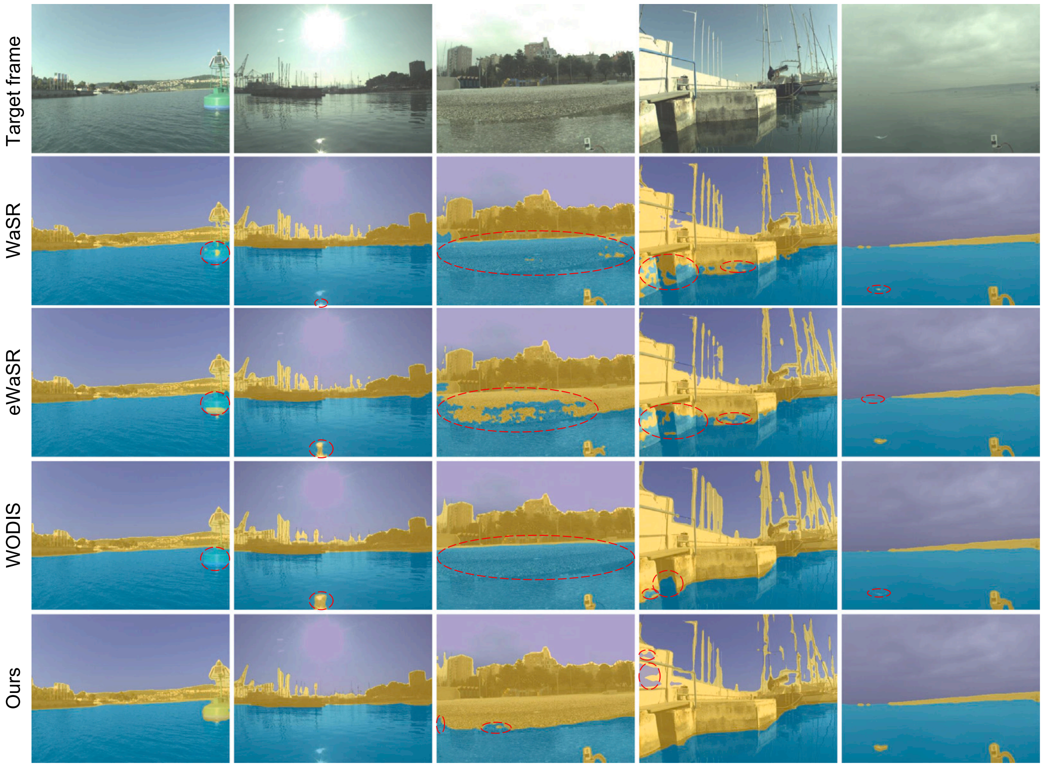

Fig. 9. Qualitative comparison of segmentation results. Mis-segmented regions are marked by red dashed lines. The qualitative results demonstrate that AF-BiSeNet significantly improves the segmentation of flashing and small targets, and the segmentation of water surface boundaries is more accurate compared to other models.

Table 4
The proposed AF-BiSeNet is compared with other models. Among them, WODIS, WaSR, and eWaSR are all advanced water surface semantic segmentation models in recent years. The remaining models represent the latest research in lightweight and real-time semantic segmentation.

Model	MaSTr1478			LaRS		
	IoU (%)		MIoU (%)	IoU (%)		MIoU (%)
	Obstacle	Water		Obstacle	Water	
WODIS	88.91	97.49	95.01	91.65	97.31	95.70
WaSR	88.60	97.25	94.88	92.53	97.48	96.10
eWaSR	88.18	97.16	94.66	91.76	97.10	95.74
RegSeg	86.98	96.78	94.12	92.20	97.33	95.98
STDC1	88.65	97.68	94.92	91.76	97.38	95.76
STDC2	88.11	97.49	94.66	91.37	97.17	95.56
LETNet	88.87	97.52	95.01	91.23	97.04	95.51
RTFormer	89.59	97.72	95.33	93.63	98.08	96.70
PIDNet	89.48	97.64	95.26	93.83	98.16	96.80
Ours	**91.14**	**98.13**	**96.02**	**94.22**	**98.21**	**97.02**

Table 5
The performance of AF-BiSeNet was compared with other models on the MODS dataset. The performance within the hazardous zones (i.e., the water surface areas within 15 meters of the vessels) is presented in parentheses. μ_R denotes the water edge detection robustness, which is calculated based on the water edge labels. Pr and Re denote the precision and recall of obstacle detection, and F1 is the reconciled value of both.

Model	μ_R (%)	Pr (%)	Re (%)	F1 (%)	Params (M)	Times (s)
WODIS	94.1	81.7(74.1)	89.1(93.2)	85.3(82.6)	49.10	0.020
WaSR	95.3	81.4(46.2)	90.2(89.7)	85.6(60.9)	52.47	0.066
eWaSR	95.4	82.4(54.4)	90.0(89.8)	86.0(67.7)	60.24	0.013
RegSeg	92.8	85.8(69.4)	85.3(86.2)	85.5(76.9)	3.33	0.019
STDC1	94.5	88.9(78.1)	82.9(89.2)	85.8(83.3)	5.32	**0.006**
STDC2	94.4	90.9(**80.5**)	81.1(89.2)	85.7(84.6)	9.35	0.012
LETNet	95.6	87.5(71.2)	86.8(92.8)	87.1(80.6)	**0.95**	0.055
RTFormer	94.5	90.8(69.7)	86.6(93.1)	88.6(79.7)	18.69	0.012
PIDNet	95.5	90.0(68.9)	91.7(96.1)	90.8(80.3)	28.76	0.015
Ours	**96.4**	**91.6**(79.1)	**91.8**(**96.8**)	**91.7**(**87.1**)	2.13	0.016

the MaSTr1478 dataset, and none of the images used for evaluation in MODS were included in the training data.

According to the results presented in Table 5, our AF-BiSeNet accomplishes the most accurate detection of water surface boundaries. Compared to maritime semantic segmentation models such as WODIS, WaSR, and eWaSR, our AF-BiSeNet achieves approximately a 10% increase in global precision while maintaining a higher global recall rate. Moreover, within a hazardous navigational range of 15 m, our AF-BiSeNet attains a recall rate of 96.8%, significantly enhancing ship navigation safety. Besides, AF-BiSeNet strikes the optimal balance between accuracy, parameter efficiency, and real-time performance among recent studies on lightweight and real-time semantic segmentation. For example, compared to STDC1 and STDC2, although they are close in precision and offer better real-time performance, our AF-BiSeNet exhibits approximately 10% higher global recall rate. Compared to the lighter LETNet, our approach demonstrates significant advantages in both detection metrics and real-time performance. While RTFormer and PIDNet are comparable to AF-BiSeNet in terms of global F1 score, their accuracy within a 15-meter range is approximately 7% lower. Additionally, their parameter counts are 8.77 and 13.50 times higher than our approach, respectively. Therefore, it is evident that our AF-BiSeNet outperforms other models in obstacle segmentation accuracy and generalization capability across multiple navigation scenarios. Furthermore, it maintains high levels of lightweight design and real-time inference performance.

5. Conclusion

In the realm of autonomous navigation for smart ships, effectively detecting obstacles in the scene and delineating navigable areas through deep learning semantic segmentation is an essential issue. However, the intricate nature of water surfaces complicates feature extraction, and excessive downsampling leads to the loss of essential features and reduces localization accuracy. Additionally, lightweight design is a critical factor for deploying and applying models on ships. In this paper, we propose a lightweight semantic segmentation network to enhance water surface obstacle detection. Unlike previous studies, we utilize a dual-branch architecture that incorporates lightweight feature aggregation and multi-scale fusion methods to enhance obstacle detection accuracy and generalization performance with fewer parameters.

The core factor that enables the superiority of the AF-BiSeNet model over existing models is the design of its novel modules. The EFRM, utilized in the detail branch, has the capacity to preserve and encode crucial information while filtering out high-frequency disturbances. The DGE layer enhances the receptive field to enable the semantic branch to more effectively capture relationships between objects on the water surface. Additionally, the BRA module can comprehensively guide the feature response of dual branches in both channel and spatial dimensions, leading to a more accurate integration of detailed and semantic information. Furthermore, the proposed C-TEF achieves lightweight multi-scale fusion by reducing channels, and it also leverages semantic properties in segmentation to ensure accurate fusion.

Experimental results demonstrate AF-BiSeNet's exceptional precision and generalization capabilities compared to advanced maritime semantic segmentation models, which can enhance the navigation safety of smart ships. Furthermore, in the realm of lightweight and real-time research, it strikes the optimal balance between accuracy, parameter efficiency, and real-time performance, making it more accessible and adaptable within the constraints of limited computing resources on ships. Although experiments have demonstrated the comprehensive competitiveness of our research, there is still potential for further improvement in segmentation accuracy and real-time performance of the model. Furthermore, the challenge of achieving high generalization performance across a wider range of navigation scenarios persists due to constraints in data availability, underscoring the need for continued investigation in this area.

CRediT authorship contribution statement

Hui Feng: Writing – review & editing, Methodology, Conceptualization. **Wensheng Liu:** Writing – original draft, Validation, Methodology, Conceptualization. **Haixiang Xu:** Writing – review & editing, Conceptualization. **Jianhua He:** Writing – review & editing, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors appreciate the constructive suggestions from reviewers and the Associate Editor. This work is supported by the National Natural Science Foundation of China under Grant No. 52371374, 51979210. This work was partly funded by EPSRC with RC Grant reference EP/Y027787/1, UKRI under grant number EP/Y028317/1, the Horizon Europe MSCA programme under grant agreement No. 101086228. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

References

Bengio, Y., Ducharme, R., Vincent, P., 2000. A neural probabilistic language model. Adv. Neural Inf. Process. Syst. 13.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Bovcon, B., Kristan, M., 2021. WaSR—a water segmentation and refinement maritime obstacle detection network. IEEE Trans. Cybern. 52, 12661–12674.

Bovcon, B., Muhovič, J., Perš, J., Kristan, M., 2019. The mastr1325 dataset for training deep usv obstacle detection models. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 3431–3438.

Bovcon, B., Muhovič, J., Vranac, D., Mozetič, D., Perš, J., Kristan, M., 2021. Mods—a usv-oriented object detection and obstacle segmentation benchmark. IEEE Trans. Intell. Transp. Syst. 23, 13403–13418.

Bovcon, B., Perš, J., Kristan, M., 2018. Stereo obstacle detection for unmanned surface vehicles by imu-assisted semantic segmentation. Robot. Auton. Syst. 104, 1–13. Cane, T., Ferryman, J., 2018. Evaluating deep semantic segmentation networks for object detection in maritime surveillance. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance. AVSS, IEEE, pp. 1–6.

Chen, X., Liu, Y., Achuthan, K., 2021. WODIS: Water obstacle detection network based on image segmentation for autonomous surface vehicles in maritime environments. IEEE Trans. Instrum. Meas. 70, 1–13.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017b. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder–decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818.

Ding, X., Guo, Y., Ding, G., Han, J., 2019. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1911–1920.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale. In: ICLR 2021 - 9th International Conference on Learning Representations.

Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J., Wei, X., 2021. Rethinking bisenet for real-time semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9716–9725.

Gao, R., 2023. Rethink dilated convolution for real-time semantic segmentation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 4675–4684.

Guo, J., Feng, H., Xu, H., Yu, W., shuzhi Ge, S., 2023. D3-net: Integrated multi-task convolutional neural network for water surface deblurring, dehazing and object detection. Eng. Appl. Artif. Intell. 117, 105558.

Hong, X., Wei, X., Huang, Y., Liu, Y., Xiao, G., 2019. Local path planning method for unmanned surface vehicle based on image recognition and vfh+. J. South China Univ. Technol. (Nat. Sci. Ed.) 47, 24–33.

Hou, Q., Zhou, D., Feng, J., 2021. Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13713–13722.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Jin, X., Niu, P., Liu, L., 2019. A gmm-based segmentation method for the detection of water surface floats. IEEE Access 7, 119018–119025.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R., 2023. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026.

Kristan, M., Kenk, V.S., c, S.Kovaˇciˇ, Perš, J., 2015. Fast image-based obstacle detection from unmanned surface vehicles. IEEE Trans. Cybern. 46, 641–654.

Lambert, R., Chavez-Galaviz, J., Li, J., Mahmoudian, N., 2022. Rosebud: A deep fluvial segmentation dataset for monocular vision-based river navigation and obstacle avoidance. Sensors 22 (4681).

Li, Y., Hou, Q., Zheng, Z., Cheng, M.M., Yang, J., Li, X., 2023. Large selective kernel network for remote sensing object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16794–16805.

Li, N., Lv, X., Xu, S., Li, B., Gu, Y., 2020. An improved water surface images segmentation algorithm based on the otsu method. J. Circuits Syst. Comput. 29, 2050251.

Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Yang, J., Li, Zhang, L., Gao, J., 2023b. Semantic-sam: Segment and recognize anything at any granularity. arXiv abs/2307.04767.

Liu, J., Li, H., Liu, J., Xie, S., Luo, J., 2021. Real-time monocular obstacle detection based on horizon line and saliency estimation for unmanned surface vehicles. Mob. Netw. Appl. 26, 1372–1385.

Liu, T., Zhang, Z., Lei, Z., Huo, Y., Wang, S., Zhao, J., Zhang, J., Jin, X., Zhang, X., 2024. An approach to ship target detection based on combined optimization model of dehazing and detection. Eng. Appl. Artif. Intell. 127, 107332.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Lv, J., Wu, Y., Chen, X., 2017. Segmentation optimization simulation of water remote congestion image of the ship. Multimedia Tools Appl. 76, 19605–19620.

Ni, H., Guan, W., Wu, C., 2020. Usv obstacle avoidance based on improved watershed and vfh method. In: 2020 11th International Conference on Prognostics and System Health Management. PHM-2020 Jinan, IEEE, pp. 543–546.

Ranjbarzadeh, R., Jafarzadeh Ghoushchi, S., Tataei Sarshar, N., Tirkolaee, E.B., Ali, S.S., Kumar, T., Bendechache, M., 2023. Me-ccnn: Multi-encoded images and a cascade convolutional neural network for breast tumor segmentation and recognition. Artif. Intell. Rev. 56, 10099–10136.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4510–4520.

Teršek, M., Žust, L., Kristan, M., 2023. eWaSR—an embedded compute ready maritime obstacle detection network. Sensors 23 (5386).

Tsai, T.H., Tseng, Y.W., 2023. BiSeNet v3: Bilateral segmentation network with coordinate attention for real-time semantic segmentation. Neurocomputing 532, 33–42.

Van Quyen, T., Kim, M.Y., 2023. Feature pyramid network with multi-scale prediction fusion for real-time semantic segmentation. Neurocomputing 519, 104–113.

Wang, J., Gou, C., Wu, Q., Feng, H., Han, J., Ding, E., Wang, J., 2022. Rtformer: Efficient design for real-time semantic segmentation with transformer. Adv. Neural Inf. Process. Syst. 35, 7423–7436.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2020. Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11534–11542.

Wu, H., Zhang, J., Huang, K., Liang, K., Yu, Y., 2019. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. arXiv preprint arXiv: 1903.11816.

Xu, G., Li, J., Gao, G., Lu, H., Yang, J., Yue, D., 2023a. Lightweight real-time semantic segmentation network with efficient transformer and cnn. IEEE Trans. Intell. Transp. Syst..

Xu, J., Xiong, Z., Bhattacharyya, S.P., 2023b. Pidnet: A real-time semantic segmentation network inspired by pid controllers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19529–19539.

Xue, H., Chen, X., Zhang, R., Wu, P., Li, X., Liu, Y., 2021. Deep learning-based maritime environment segmentation for unmanned surface vehicles using superpixel algorithms. J. Mar. Sci. Eng. 9 (1329).

Yang, D., Solihin, M.I., Zhao, Y., Yao, B., Chen, C., Cai, B., Machmudah, A., 2023. A review of intelligent ship marine object detection based on rgb camera. IET Image Process..

Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N., 2021. BiSeNet v2: Bilateral network with guided aggregation for real-time semantic segmentation. Int. J. Comput. Vis. 129, 3051–3068.

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N., 2018. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 325–341.

Zha, H., Liu, R., Yang, X., Zhou, D., Zhang, Q., Wei, X., 2021. Asfnet: Adaptive multiscale segmentation fusion network for real-time semantic segmentation. Comput. Animat. Virt. Worlds 32, e2022.

Zhang, Y., Doughty, H., Snoek, C.G., 2024. Low-resource vision challenges for foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21956–21966.

Zhang, J., Gao, J., Liang, J., Wu, Y., Li, B., Zhai, Y., Li, X., 2023a. Efficient water segmentation with transformer and knowledge distillation for usvs. J. Mar. Sci. Eng. 11 (901).

Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S., Hong, C.S., 2023b. Faster segment anything: Towards lightweight sam for mobile applications. arXiv preprint arXiv:2306.14289.

Zhang, Y., Liu, H., Hu, Q., 2021. Transfuse: Fusing transformers and cnns for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1 2021, Proceedings, Part I, Vol. 24. Springer, pp. 14–24.

Zust, L., Kristan, M., 2022. Temporal context for robust maritime obstacle detection. In: 2022 IEEE, RJS International Conference on Intelligent Robots and Systems. IROS, pp. 6340–6346.

Žust, L., Pers, J., Kristan, M., 2023. LaRS: A diverse panoptic maritime obstacle detection dataset and benchmark. In: 2023 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 20247–20257.