applied sciences

MDPI

Article

Coastline Identification with ASSA-Resnet Based Segmentation for Marine Navigation

Yuhan Wang, Weixian Li, Zhengxun Zhou and Ning Wu *

Key Laboratory of Beibu Gulf Offshore Engineering Equipment and Technolog, Beibu Gulf University, Qinzhou 535011, China; 2302020110@stu.bbgu.edu.cn (Y.W.); liweixian9016@163.com (W.L.)
* Correspondence: n.wu@bbgu.edu.cn

Abstract

Real-time and accurate segmentation of coastlines is of paramount importance for the safe navigation of unmanned surface vessels (USVs). Classical methods such as U-Net and DeepLabV3 have been proven to be effective in coastline segmentation tasks. However, their performance substantially degrades in real-world scenarios due to variations in lighting and environmental conditions, particularly from water surface reflections. This paper proposes an enhanced ResNet-50 model, namely ASSA-ResNet, for coastline segmentation for vision-based marine navigation. ASSA-ResNet integrates Atrous Spatial Pyramid Pooling (ASPP) to expand the model's receptive field and incorporates a Global Channel Spatial Attention (GCSA) module to suppress interference from water reflections. Through feature pyramid fusion, ASSA-ResNet reinforces the semantic representation of features at various scales to ensure precise boundary delineation. The performance of ASSA-ResNet is validated with a dataset encompassing diverse brightness conditions and scenarios. Notably, mean Pixel Accuracy (mPA) and mean Intersection over Union (mIoU) of 98.90% and 98.17%, respectively, have been achieved on the self-constructed dataset, with corresponding values of 99.18% and 98.39% observed on the USVInland unmanned vessel dataset. Comparative analyses reveal that ASSA-ResNet outperforms the U-Net model by 1.78% in mPA and 2.9% in mIOU relative to the DeepLabV3 model. It also demonstrates enhancements of 1.85% in mPA and 3.19% in mIoU. On the USVInland dataset, ASSA-ResNet exhibits superior performance compared to U-Net, with improvements of 0.41% in mPA and 0.12% in mIoU, while surpassing DeepLabV3 by 0.33% in mPA and 0.21% in mIoU.

Keywords: autonomous ship; coastal segmentation; ResNet-50; attention mechanism; multi-scene recognition

check for updates

Academic Editor: Atsushi Mase

Received: 9 July 2025
Revised: 8 August 2025
Accepted: 15 August 2025
Published: 19 August 2025

Citation: Wang, Y.; Li, W.; Zhou, Z.; Wu, N. Coastline Identification with ASSA-Resnet Based Segmentation for Marine Navigation. *Appl. Sci.* **2025**, *15*, 9113. https://doi.org/10.3390/app15169113

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the advancement of autonomous driving technology, unmanned surface vehicles (USVs) have found wide applications such as hydrological survey and mapping [1], water quality monitoring [2], floating waste removal [3], and maritime search and rescue operations [4]. Environmental perception technology is a critical component for USV autonomy, and the ability to accurately segment the boundaries of navigation water areas is an important prerequisite for autonomous tasks [5]. However, images in the water environment are affected by factors such as light changes and reflections, making accurate segmentation challenging. Therefore, a flexible segmentation algorithm with high precision is of great significance for real-world applications [6].

At present, a variety of research methods for water area segmentation have been developed. For traditional water area images, methods such as clustering segmentation and threshold segmentation, as well as segmentation with maximum entropy, are mainly used for image segmentation. The clustering segmentation method [7] divides the image into several clusters with similar features and partitions the image into different regions ac-cording to the similar attributes of pixels. Park et al. [8] proposed a method based on the density based spatial clustering of applications with noise (DBSCAN) to segment the im-age into homogeneous regions. The *i*th wavelet texture analysis is then combined to extract the water body region, thus achieving effective segmentation of the water area in the image. Despite the high segmentation efficiency, this algorithm shows obvious differences when dealing with the bright and shadow areas of the water area in the image. The threshold segmentation method [9] divides the pixels of the image into different regions by setting one or more thresholds according to the threshold range of the pixel gray values. Yu et al. [10] proposed a semiautomatic threshold segmentation method based on local image patches. First, the image is preprocessed by Gaussian filtering and median filtering, and the image is divided into multiple rectangular regions. Subsequently, an adaptive binarization threshold is set within each rectangular region to achieve the segmentation of the water area. Although the threshold segmentation requires a small amount of calculation, it is very sensitive to image noise. Especially when the gray values of the background and the target are close, a suitable segmentation threshold is difficult to determine. In order to find the best segmentation level, a maximum entropy segmentation method was introduced to analyze the distribution of pixels to discover the threshold that maximizes the total entropy of the image region [11]. Han et al. [12] proposed an active contour model based on a modified symmetric cross-entropy to define the external energy constraint term. The segmentation accuracy can be effectively improved by applying the median of the pixel values in the segmented object and the background region. However, this method depends on a large number of parameters, resulting in complicated parameter adjustment.

In recent years, deep learning models have also been applied to semantic segmentation, and Fully Convolutional Networks (FCN) outperform fully connected Neural Networks in pixel-level image segmentation [13]. However, the continuous downsampling operation in FCN leads to the loss of small target features, leading to segmentation errors for smaller objects. Subsequently, U-Net [14] and SegNet [15] were proposed to adopt "U"-shaped encoder–decoder architecture and skip connections to fuse shallow semantic information with high level semantic features and overcome lost details caused by downsampling. How-ever, when processing feature maps, skip connections often assign the same weight to each channel and key regions cannot be emphasized based on the importance of different chan-nels. In this regard, Jonnala et al. proposed a multi scale residual and attention-enhanced U-Net model (AER U-Net), which combined the residual block, attention mechanism, and dropout layer to improve accuracy and generalizability for largescale waterbody segmentation [16]. In a following contribution, a DSIA U-Net model was designed to combine deep and shallow interaction mechanisms and the attention module to improve segmentation accuracy for small objects [17]. Li et al. later proposed a boundary attention (BA) module and combined it with an adaptive-weight multi-task learning (AWML) model to capture useful boundary information [18]. However, the smaller convolutional kernel in the CNN can only focus on a small region in the image at each convolution and is less capable of recognizing objects of different sizes. CNN pooling operations and larger strides significantly reduce feature resolution for small-scale feature extraction [19]. Coastal border identification methods with underwater images, such as the SAM model, are still limited by conditions like lighting and background interference [20]. Hong et al. proposed WaterSAM, which introduces LoRA technology to reduce computational and annotation

costs, while increasing the segmentation efficiency of small objects and fuzzy boundary recognition [21]. To improve the operating efficiency of SAM, Zhang et al. also proposed Efficient ViT-SAM, which uses an efficient encoder and knowledge distillation method to significantly accelerate reasoning while maintaining segmentation accuracy [22]. Fu et al. proposed the Lite-SAM model to achieve real-time full image segmentation with a lightweight encoder LiteViT and an AutoPPN module [23].

An attention mechanism has been proposed with two sub-modules of spatial attention and a channel attention, and the weights of feature maps can be adaptively adjusted to improve attention to key regions [24]. However, the small convolutional kernels in the CNN network can only focus on a small area in the image and are less flexible with object size. ASPP can help to obtain a larger receptive field and acquire the context information of features of different sizes through the combination of multiple atrous convolutional kernel sizes [25,26]. Another method uses the Feature Pyramid Network (FPN), which fuses high-level semantic information with low-level detail information through a top-down path and can obtain information features of different scales. However, during the feature fusion process, some high-frequency components may be lost or weakened [27].

The accurate delineation of coastlines in complex scenarios with highly variable water conditions demands a more flexible and robust modeling approach. Factors such as light reflections on the water surface, reflections from vessels and port structures, as well as variations in illumination intensity at different times of day, all introduce significant challenges to reliable coastal segmentation [28]. To address these issues, this study introduces a novel model, termed ASSA-ResNet, specifically designed to mitigate the effects of such visual complexities and enhance the accuracy of coastline identification.

The proposed ASSA-ResNet model adopts ResNet-50 [29] as its backbone for feature extraction, leveraging its residual connections to effectively mitigate the vanishing gradient problem that arises as network depth increases. This design enables the network to capture rich multi-scale features critical for precise coastal boundary delineation. The architecture incorporates several targeted modules tailored to the unique challenges of coastal imagery. First, to address interference from water surface reflections and glare—often manifested as localized, high-frequency noise—the model integrates a GCSA attention mechanism. This module adaptively enhances features associated with stable land–water boundaries while suppressing transient reflection artefacts by jointly evaluating feature importance across both channel and spatial dimensions, thereby improving the separation of true targets from background noise. Second, to handle substantial illumination variability (e.g., backlighting, shadows), an Atrous Spatial Pyramid Pooling (ASPP) module is employed to capture contextual information across multiple receptive fields, allowing the network to infer the consistent structural patterns of coastlines regardless of global lighting changes. Finally, by fusing features across a pyramid of scales, the model enhances semantic representation at different resolutions, significantly improving segmentation accuracy. This integrated approach offers a robust technical framework to support the automated and intelligent development of unmanned surface vessels.

2. General Model Architecture

Figure 1 presents the overall architecture of the proposed ASSA-ResNet model, specifically designed for high-precision coastline segmentation in water-area imagery. The network follows an end-to-end encoder–decoder paradigm, with the data processing flow conceptually divided into an upper encoder and a lower decoder pathway. In the encoder stage, an input image depicting a real coastal scene is first processed by a standard convolution (CONV) layer and a max-pooling (MAXPOOL) layer, which perform initial feature extraction and spatial down-sampling. The resulting features are then passed

into a backbone constructed from a series of ResNet-50 bottleneck modules, whose deep residual connections facilitate effective multi-scale feature extraction while mitigating the vanishing gradient problem. Following backbone processing, a Global Channel-Spatial Attention (GCSA) module is applied to the encoded feature maps. By jointly assessing feature salience across both channel and spatial dimensions, this attention mechanism guides the network to emphasize critical target regions—such as stable land–water boundaries—while suppressing background interference arising from reflections on the water surface.

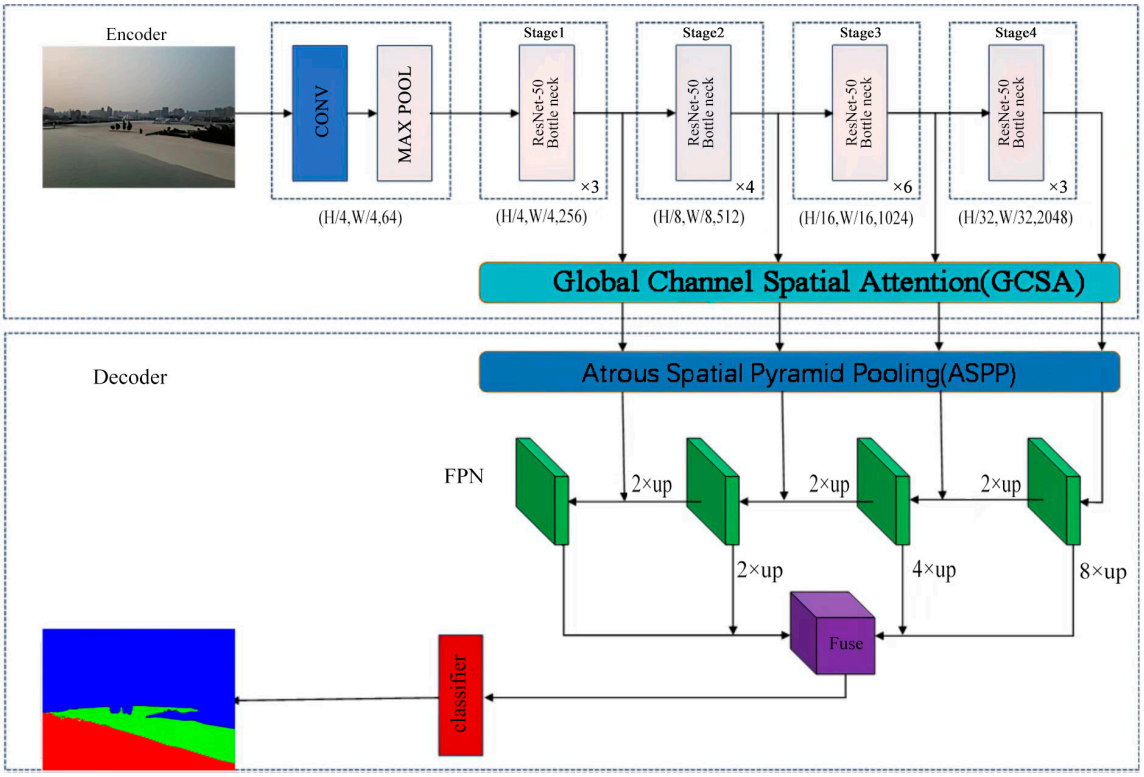


Figure 1. The ASSA-ResNet structure.

The feature maps, enhanced by the encoder and the attention module, are then passed to the ASPP module, which employs parallel atrous convolutions with different dilation rates to significantly expand the model's receptive field without a substantial increase in computational cost. In this way, both local details and global contextual information from the image can be simultaneously captured. As shown in Figure 1, the multi scale features (green blocks) extracted from the different branches of ASPP are fused to form a comprehensive feature map, which will ultimately be fed into a decoder that utilizes an FPN to combine high level semantic information with low level detail information. Finally, a classifier performs pixel-wise classification on the refined feature map from the decoder to generate the final segmentation result. The output image uses blue, green and red colors to identify semantic categories of water, background, and sky, respectively.

2.1. The Encoder Component

Due to the complex backgrounds commonly present in coastline imagery—such as shoreline reflections and varying illumination conditions—coastline identification tasks often suffer from reduced accuracy. In this regard, an attention-enhanced encoder is designed to suppress irrelevant background features and selectively emphasize target regions. As illustrated in Figure 2, the encoder integrates a GCSA module to enhance feature representations across multiple scales extracted by the backbone network. The GCSA module combines channel attention, channel shuffling, and spatial attention mechanisms,

enabling the model to focus more effectively on the salient target regions. By jointly evaluating spatial and channel-wise feature importance, the encoder significantly improves the network's ability to distinguish coastal boundaries from visually similar background interference.

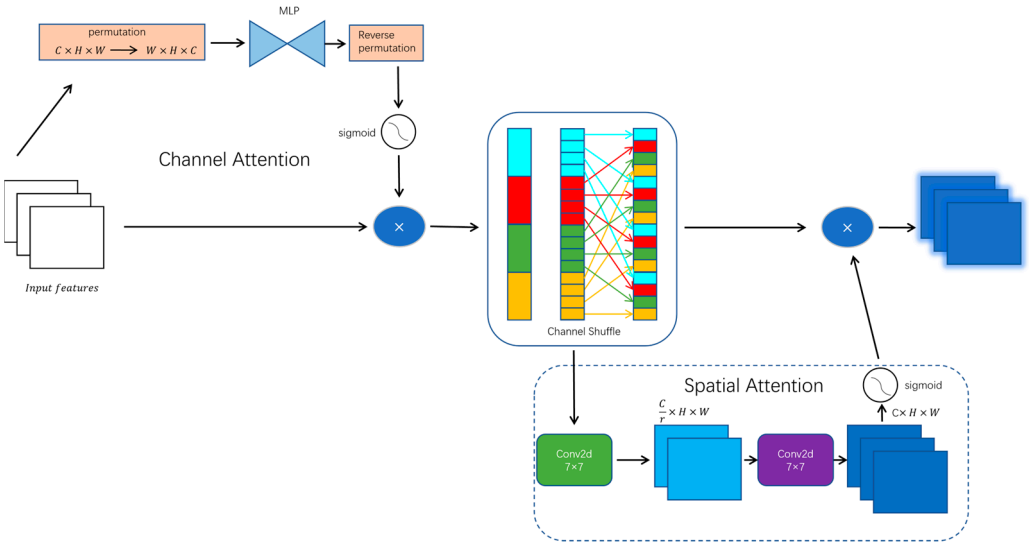


Figure 2. Implementation of the attention mechanism.

First, the input feature map is passed through the channel attention sub-module, which adaptively recalibrates the importance of each channel. This is followed by the spatial attention sub-module, which further refines the feature representation by emphasizing informative spatial locations. The original feature map comprises multiple channels, each with a spatial resolution of H × W, where H and W denote height and width, respectively. This sequential attention mechanism enables the model to capture both inter-channel dependencies and spatial saliency, thereby enhancing its ability to focus on relevant regions in the coastal imagery.

In the channel attention sub-module, the input feature map with dimensions C × H × W is first permuted to a shape of W × H × C to facilitate channel-wise processing. A two-layer Multi-Layer Perceptron (MLP) is then employed to model inter-channel dependencies. Specifically, the first MLP layer reduces the channel dimensionality to one-fourth of its original size, followed by a ReLU activation function to introduce non-linearity. The second MLP layer subsequently restores the channel dimension to its original size. After this transformation, an inverse permutation is applied to revert the feature map to its original shape of C × H × W. A Sigmoid activation function is then applied to produce the final channel attention map. This attention map is element-wise multiplied with the original input feature map, resulting in an enhanced representation that emphasizes informative channels while suppressing irrelevant ones, such that

$$F_{channel} = \sigma(MLP(Permute(F_{input}))) \odot F_{input} \quad (1)$$

where $F_{channel}$ is the augmented feature map, σ denotes the Sigmoid function, ⊙ signifies element-wise multiplication, and F_{input} is the original input feature map.

A channel shuffling operation is then applied to further enhance the mixing and dissemination of information across channels. The enhanced feature map is first divided into four groups, each containing C/4 channels. A transposition operation is then performed on these groups to reorder the channels within each group, effectively disrupting the original channel arrangement. This reorganization facilitates improved interaction among features from different channel subsets. Finally, the shuffled feature map is reshaped back to its

original dimensions of C × H × W. This operation promotes a more comprehensive integration of feature information across the network, thereby improving the model's overall feature representation capability, such that

$$F_{shuffle} = ChannelShuffle(F_{channel}) \quad (2)$$

where $F_{shuffle}$ represents the shuffled feature map, and the number of channels in the input feature map is denoted as $F_{channel}$.

In the spatial attention sub-module, the spatial resolution is maintained by the application of a 7 × 7 convolutional layer to the input feature map with stride of 1 and padding of 3. The number of channels can then be reduced to 1/4 of its initial value. Subsequently, the feature map then undergoes a non-linear transformation through batch normalization followed by a ReLU activation function. In order to restore the channel dimension to its original size *C*, a subsequent 7 × 7 convolutional layer is applied, followed by another batch normalization layer to stabilize the feature distribution. The resulting output is passed through a Sigmoid activation function to generate the spatial attention map. Finally, an element-wise multiplication is performed between the spatial attention map and the previously shuffled feature map, yielding the final output feature map. This operation enables the model to selectively emphasize informative spatial regions while preserving refined channel-wise representations, such that

$$F_{spatial} = \sigma(Conv(BN(ReLU(Conv(\ F_{shuffle}))))) \odot F_{shuffle} \quad (3)$$

where $F_{spatial}$ represents the feature map subsequent to the application of spatial attention.

2.2. The Decoder Component

In aquatic imagery, the boundaries between components such as shoals, man-made structures, and bridges and the surrounding water regions are often indistinct. Relying solely on high-level semantic features extracted by the network may result in segmentation errors, particularly in complex or visually ambiguous areas, due to interference from background elements. A multi-scale feature fusion decoder can be developed to integrate the ASPP with an FPN to capture contextual information across different scales. Specifically, the ASPP module employs atrous convolutions to expand the receptive field, allowing the network to extract rich contextual features ranging from fine local details to broader global structures. The FPN is then utilized to effectively fuse these multi-scale features, resulting in a more informative and spatially aware feature representation. The implementation of the proposed multi-scale feature fusion decoder is presented in Figure 3.

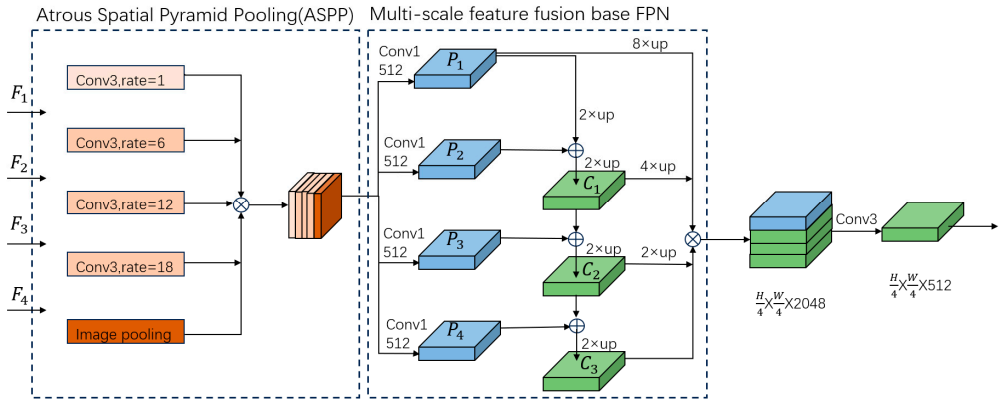


Figure 3. Multi-scale feature fusion decoder.

The image processed through the watershed algorithm experiences four steps within the encoder. First, the feature map F = {F_1, F_2, F_3, F_4} is obtained and serves as the input for the decoder. At every step of the residual block, the input image undergoes downsampling and convolution operations. Consequently, the resolution of the output feature map F_i becomes {1/4, 1/8, 1/16, 1/32} of the original input, and the number of output channels turns out to be {256, 512, 1024, 2048}.

Prior to multi-scale feature fusion, the ASPP module processes the intermediate feature map F_i to extract contextual information across multiple receptive fields. As illustrated in the dashed box on the left side of Figure 3 the ASPP module consists of five parallel branches: one 1 × 1 convolution, three 3 × 3 dilated convolutions with dilation rates of 6, 12, and 18, respectively, and one global average pooling operation. The 1 × 1 convolution adjusts the number of channels to ensure dimensional consistency across branches. The three dilated convolutions capture features at different spatial scales, thereby enabling the model to aggregate both local and global context. For the global average pooling branch, a 1 × 1 convolutional layer is also appended to refine channel dimensions. This design enables the ASPP module to effectively encode multi-scale contextual information prior to the fusion stage.

Furthermore, each feature map F_i is passed through a 1 × 1 convolutional layer to standardize the number of channels to 512, resulting in a set of transformed feature maps denoted as P P = {P_1, P_2, P_3, P_4}. Beginning with the highest-level feature map P_4, a top-down pathway is established in which each higher-level feature is upsampled by a factor of two and fused with the corresponding lower-level feature via pixel-wise addition. This process produces the fused feature set C = {C_1, C_2, C_3}. Subsequently, each level of the FPN is upsampled by factors of {1, 2, 4, 8}, respectively, to match a spatial resolution of 1/4 of the original input image. A 3 × 3 convolution is then applied to each upsampled feature map for further refinement. Finally, all refined feature maps are concatenated along the channel dimension to generate the final fused feature representation, which effectively integrates multi-scale contextual and semantic information.

3. Experimental Data and Processing

Datasets play a crucial role in deep learning tasks, as they directly influence the quality of model training, generalization performance, and overall accuracy. The rapid advancement of autonomous driving technologies has been largely driven by the availability of large-scale, high-quality datasets that support research, training, and evaluation. Similarly, for visual perception tasks in USVs, the availability of relevant datasets is equally essential. Although a few datasets for scene segmentation and recognition in USV applications have been proposed in recent years, their scale and diversity remain limited, constraining the development and evaluation of robust perception algorithms. To address this gap, this study constructs two specialized datasets through independent data collection tailored to the real-world operational scenarios of unmanned ships: a water surface scene segmentation dataset and a water surface scene matching dataset. These datasets are designed to offer greater diversity and be more representative, thereby providing more comprehensive resources for advancing visual perception research in the field of autonomous maritime systems.

Most existing outdoor datasets for visual scene matching are primarily focused on urban and rural street environments and are commonly utilized in research on autonomous driving technologies. Representative examples include the Pitts250 k dataset [30], the Nordland dataset [31], and the Oxford RobotCar dataset [32]. A limited number of datasets have also been collected through aerial or underwater imaging, serving visual perception research for UAVs and unmanned underwater vehicles (UUVs). The USVInland dataset,

developed jointly by OKO Tech, Tsinghua University, and Northwestern Polytechnical University, is a notable contribution specifically tailored for inland waterway USVs. Designed to address the scarcity of publicly available data in this domain, the dataset was collected over a four-month period and consists of 27 segments of raw data, covering more than 26 km of diverse inland waterway environments using multi-sensor configurations.

Nevertheless, research on scene matching datasets for unmanned ships remains limited. This is primarily due to two major challenges. First, compared to unmanned vehicles and drones, unmanned ships have not achieved widespread adoption. The primary demand comes from marine water conservancy and river management sectors, where stringent safety regulations and complex approval processes significantly hinder experimental deployments. Second, even when safety conditions are met, conducting experiments in real water environments requires the transportation and assembly of vessels, which introduces high logistical uncertainty. Moreover, maritime navigation routes are often broader and less structured than terrestrial roads, making it difficult to collect repeated data from the same location.

In this research, we introduce a new dataset, the Beihai Scene Matching dataset, which was constructed through extensive field experiments. This dataset includes maritime images categorized into three semantic classes such as water, background, and sky. The explicit inclusion of the sky category is intended to reduce boundary ambiguity and suppress interference from both the background and water surface, thereby improving the performance of boundary-aware scene matching algorithms.

The Beihai Scene Matching dataset comprises 1000 maritime images captured during a 20 km coastal voyage along the South China Sea. The images encompass a wide range of lighting conditions—including frontlighting, backlighting, and variable ambient brightness—and diverse scene types such as ports, vessels, and coastline structures. The original dataset is divided into 600 training images, 200 validation images, and 200 test images. To enhance dataset diversity and support model generalization, images were captured at multiple resolutions (e.g., 240×240, 480×480, and 1280×1208 pixels), and a variety of data augmentation techniques, including rotation and cropping, were applied to expand the dataset to 3000 images.

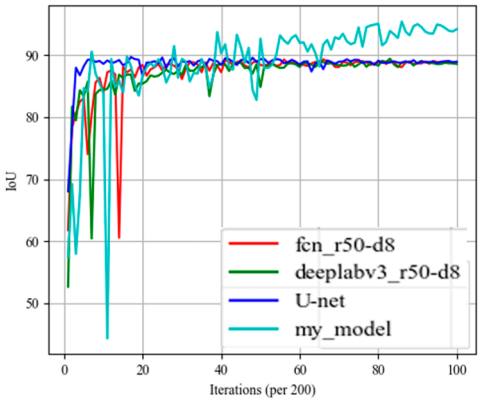
In this experiment, a subset of the dataset was utilized, consisting of 518 images for training, 182 images for testing, and 91 images selected from the test set for validation. This selected dataset configuration ensures balanced coverage of scenes and lighting conditions, enabling robust evaluation of scene matching performance in maritime environments.

4. Experimental Results and Analysis

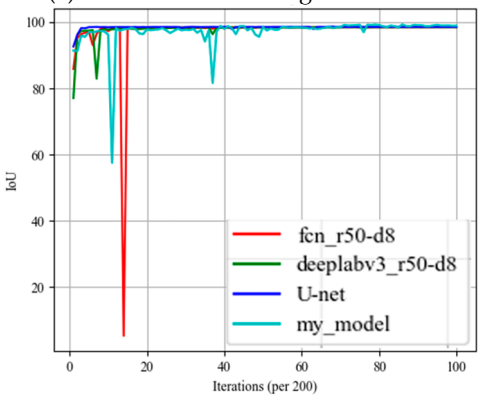
4.1. The Performance of Diverse Models During the Training Stage

Figure 4 presents the evolution of Intersection over Union (IoU) metrics for three target categories—water, sky, and background (land)—across various models, measured at every 200 training iterations. The x-axis denotes the iteration count (Iteration/200), while the y-axis represents the IoU values. As shown, traditional models such as U-Net, DeepLabV3, and FCN demonstrate significant fluctuations in IoU across all categories during the early stages of training. These models experience unstable performance, with noticeable declines in accuracy at certain iteration points, although they gradually achieve convergence in the later stages.

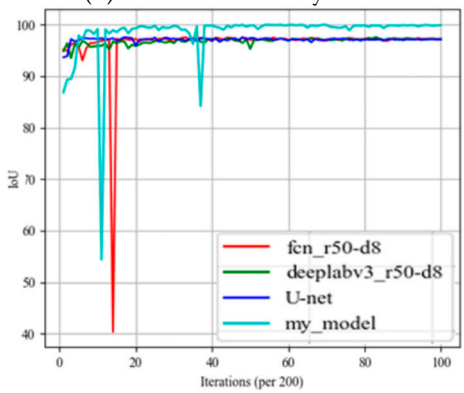
In contrast, the proposed ASSA-ResNet model exhibits superior stability and accuracy throughout the training process. Its IoU values remain consistently high with minimal variation and converge steadily after approximately 10,000 iterations. This reflects the model's robust learning capacity and enhanced precision in segmenting complex maritime scenes.



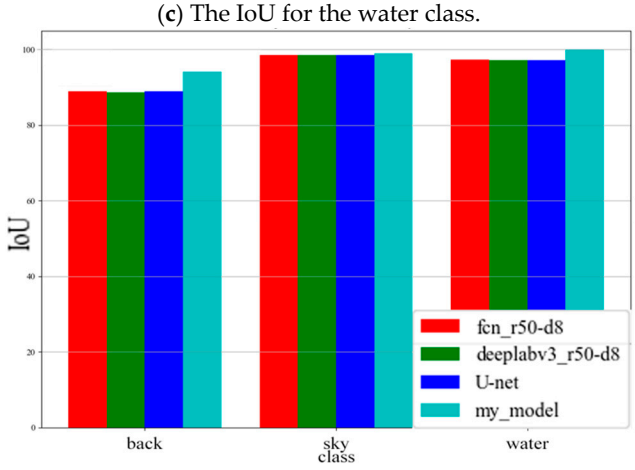
(**a**) The IoU for the background class.



(**b**) The IoU for the sky class.



(**c**) The IoU for the water class.



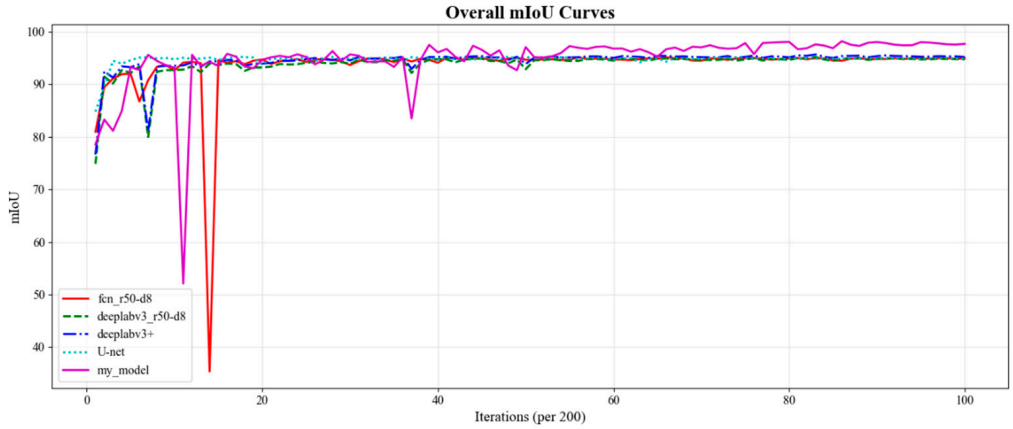
(**d**) The IoU comparison for three categories with the existing models.

Figure 4. Performance comparison with the existing models on the validation set.

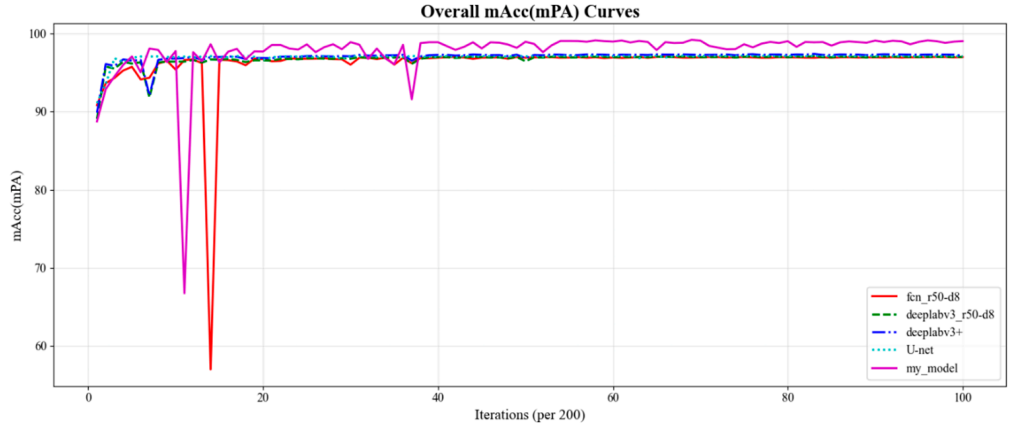
By the end of the training phase, ASSA-ResNet achieved an IoU of 99.79% for the water category (compared to 96.51–97.37% for baseline models), 99.31% for the sky category (vs. 98.41–98.56%), and 95.42% for the background/land category (vs. 89.08–89.71%). These results demonstrate the model's strong generalization ability and high segmentation accuracy, particularly in distinguishing water bodies under complex environmental conditions.

4.2. Comparative Experimental Results Analysis

The proposed ASSA-ResNet model was trained and evaluated on the Beihai Scene Matching dataset, using mIoU and mPA as the primary evaluation metrics. The performance was compared against several representative classical segmentation models, including FCN, U-Net, and DeepLabV3, under identical experimental settings and parameter configurations. All models were trained over multiple epochs, and the optimal weights can be used for final evaluation. The comparative performance results between ASSA-ResNet and the baseline models are presented in Figure 5.



(**a**) The comparison of overall mIoU with the existing models.



(**b**) Comparison of overall mPA with the existing models.

Figure 5. Performance comparison with the existing models.

Figure 5 demonstrates that among the evaluated methods, DeepLabV3 and U-Net achieved relatively high mIoU scores. Both models are based on deep CNNs with extensive layer architectures. DeepLabV3 leverages dilated convolutions and an ASPP module to capture multi-scale contextual information, while U-Net employs feature pyramids and global context fusion to integrate semantic features across different resolutions. However, their high performance comes at the cost of increased model complexity and parameter size.

In contrast, the proposed ASSA-ResNet model enhances segmentation efficiency and accuracy by integrating an attention mechanism and a spatial pyramid pooling structure. Compared to U-Net, ASSA-ResNet achieves an improvement of 2.90% in mIoU and 1.78% in mPA. Relative to DeepLabV3, the gains are 3.19% in mIoU and 1.85% in mPA. When compared to FCN, the improvements are 3.07% in mIoU and 1.90% in mPA. These results confirm that ASSA-ResNet consistently outperforms classical segmentation models in terms of both accuracy and robustness.

Moreover, to further validate the superiority of the proposed architecture, a comparison with the enhanced model DeepLabV3+ was also conducted. ASSA-ResNet still achieves notable performance gains, with an increase of 2.42% in mIoU and 1.73% in mPA, underscoring its effectiveness in complex coastal scene segmentation tasks.

As illustrated in Figures 6 and 7, the proposed ASSA-ResNet model demonstrates superior performance across all environments by effectively fusing multi-scale features and integrating contextual information, particularly in intricate marine environments, where the water surface area can be accurately delineated while mitigating issues of over- and under-segmentation. Specifically, under backlit conditions with significant water surface reflection, the model can effectively capture fine details and clearly outline complex boundaries. This enables it to achieve state-of-the-art results on mIoU and mPA metrics while maintaining a relatively small model parameter count.

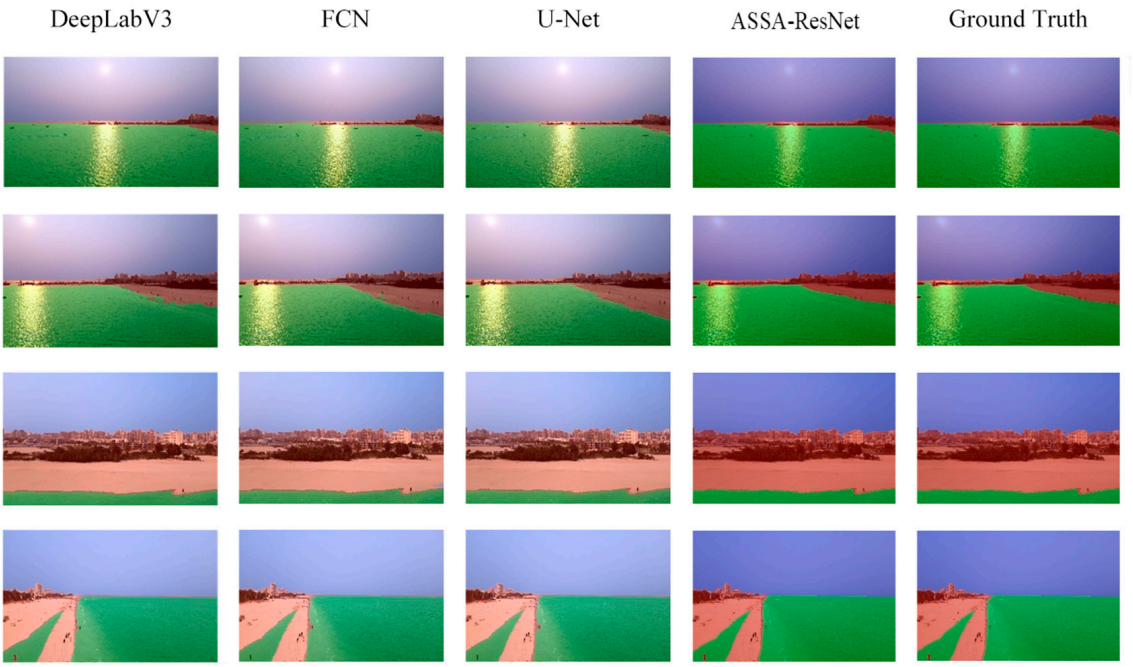


Figure 6. Comparison of the segmentation performance of different models with Ground Truth.



Figure 7. Comparison of segmentation effects of different models with Ground Truth.

In contrast, the ASPP module in DeepLabV3 demonstrates limitations when processing regions affected by intense specular reflections, such as sun glare. These high-intensity

areas, characterized by elevated pixel values, can be misinterpreted by the model as non-marine objects, resulting in voids or 'holes' within the segmentation mask. Similarly, the FCN suffers from substantial spatial information loss due to its core design, which relies heavily on aggressive downsampling via pooling or strided convolutions. Although upsampling techniques are employed to restore the original image resolution, the spatial granularity is not fully recovered, leading to blurred, coarse, and misaligned segmentation outputs, particularly around fine boundaries such as the horizon or coastline.

While U-Net addresses some of these issues through its skip connections, which preserve intermediate spatial features, it is still constrained by a limited effective receptive field. This limitation reduces its ability to capture global context, making it less suitable for accurately segmenting large, homogeneous regions such as open sea. Moreover, in scenarios where the sea–sky boundary is ambiguous or lacks clear contrast, U-Net may struggle to delineate the full extent of the maritime area. Figures 6 and 7 present visual comparisons of the segmentation results, highlighting the shortcomings of these models and the advantages of the proposed approach.

ASSA-ResNet's generalization capability is further evaluated with additional experiments on the USVInland dataset under identical training configurations. This dataset, characterized by diverse inland waterway environments, provides a robust benchmark for validating segmentation performance across varying scenarios. As shown in Table 1, ASSA-ResNet consistently outperforms several classical segmentation models, highlighting its superior adaptability, accuracy, and robustness in real-world applications beyond the marine coastal domain.

Table 1. Performance comparison with the existing models.

Model	PA/%		IOU/%		mPA/%	mIOU/%
	Water	Land	Water	Land		
FCN	99.1	98.9	98.21	97.07	98.92	97.54
U-Net	98.27	99.27	98.0	98.54	98.77	98.27
DeepLabV3	98.79	98.91	98.14	97.21	98.85	98.18
ASSA-ResNet	99.16	99.21	98.70	98.08	99.18	98.39

4.3. Ablation Experiments

In order to verify the effectiveness of ASSA-ResNet in a more comprehensive way, ablation experiments are also conducted based on different modules. A total of five groups of experiments ① to ⑥ are designed to test the importance of the roles of the attention enhancement modules such as GCSA, FPN, and ASPP. The experimental environment configuration and parameters are consistent with all comparative experiments. The ablation experiment results are shown in Table 2, where Experiment ⑥ is the result of a complete ASSA-ResNet structure, with the mIoU and mPA reaching 98.17% and 98.90%, respectively.

Table 2. Ablation experiment results of the ASSA-ResNet model.

Configuration	mIoU (%)	ΔmIoU	mPA (%)	ΔmPA
ASSA-ResNet	98.17	0	98.90	0
No GCSA Module	97.32	−0.85	98.88	−0.02
No ASPP Module	95.66	−2.51	97.68	−1.22
No FPN Module	96.98	−1.19	98.09	−0.81
Backbone (ResNet-50)	93.54	−4.63	96.82	−2.08

In Experiment ①, the GCSA attention module was removed from the architecture. This resulted in a performance decrease, with the mIoU dropping by 0.85% and the mPA

by 0.02% compared to the complete model in Experiment ⑥. This drop in performance, particularly in the mIoU, is more sensitive to boundary accuracy, demonstrating the critical role of the attention mechanism. The GCSA module is designed to dynamically focus on the most relevant features of the target, such as the coastline, while suppressing interference from background noise like water surface reflections. By selectively amplifying target features across both channel and spatial dimensions, it significantly boosts the model's ability to precisely identify and segment the target in complex scenarios, thereby enhancing both the efficiency and precision of information processing. The results confirm that this targeted focus is essential for achieving the highest level of segmentation accuracy.

In Experiment ②, the multi-scale feature fusion is eliminated, and solely the high-level semantic features are segregated subsequent to traversing the ASPP modules. Consequently, the model forfeits its capacity to detect objects of varying scales. In comparison to Experiment ⑤, the mIoU and mPA exhibit reductions of 1.19% and 0.81%, respectively. This indicates that multi scale feature fusion can significantly enhance the target characteristics, which is essential for the model to augment its segmentation performance.

Experiment ③ is for the removal of the ASPP module, with features from the backbone network being fed directly into the multi-scale fusion component. This change led to a substantial drop in performance, with the mIoU decreasing by 2.51% and the mPA by 1.22% relative to the complete model. This significant degradation underscores the importance of the ASPP, which plays a crucial role in expanding the model's receptive field. This allows the model to capture multi-scale contextual information more effectively. The results clearly indicate that this enhanced contextual understanding is vital for robust and accurate segmentation, especially in identifying the coastline under varying environmental conditions.

In Experiment ④, segmentation is directly carried out using the backbone network. Compared with the complete model, the mIoU and mPA decrease by 4.63% and 2.08%, respectively. This shows that the models of ASPP and GCSA achieve complementary advantages. That is, the multi-scale feature fusion ability of ASPP is utilized to ensure the feature integrity of objects of different sizes, and the attention mechanism of GCSA is employed to enhance the sensitivity to the key features of the target, minimizing interference from the background to the greatest extent.

In Experiment ⑤, a binary classification dataset based on USVInland has been used for the recognition task of two-class images such as water area and background. Compared with the complete model in Experiment ⑥, the mIoU and mPA decrease by 2.36% and 0.91%, respectively. This indicates that the three-class classification model significantly enhances the ability to semantically understand and distinguish features of different targets in complex scenes by simultaneously learning the features of water, land, and sky. The model can accurately identify the junction area between the sky and the water area, avoiding misjudgment caused by areas with similar colors such as beaches and wetlands.

4.4. Computational Complexity Analysis

The suitability of the proposed ASSA-ResNet model for real-time applications is also tested with a comparative analysis of computational efficiency against several existing segmentation models. Specifically, we evaluated the number of trainable parameters, floating-point operations (FLOPs) for a 480×480 input resolution, and the average inference time per image on a single NVIDIA RTX 3090 GPU (NVIDIA Corporation, Santa Clara, CA, USA). The comparative results in Table 3 demonstrate the model's computational advantages while maintaining high segmentation accuracy.

Table 3. Comparison of computational complexity.

Model	mIoU (%)	mPA (%)	Parameters (M)	FLOPs (G)	Inference Time (ms)
FCN	97.54	98.92	25.8	55.2	35
U-Net	98.27	98.77	31.1	68.5	42
DeepLabV3	98.18	98.85	41.5	75.3	48
ASSA-ResNet	98.39	99.18	35.4	71.2	45

As presented in Table 3, the FCN model exhibits the lowest computational overhead; however, this comes at the cost of significantly reduced segmentation accuracy. While DeepLabV3 and U-Net demonstrate higher accuracy, they require substantially more parameters and computational resources. In contrast, the proposed ASSA-ResNet model achieves superior segmentation performance (in terms of mIoU and mPA) with fewer parameters and lower FLOPs than DeepLabV3. ASSA-ResNet also delivers competitive inference speed, achieving a strong balance between accuracy and computational efficiency. These characteristics of ASSA-ResNet are particularly well-suited for real-time deployment on resource-constrained platforms such as USVs, where processing efficiency is critical.

5. Discussion and Future Work

While the proposed ASSA-ResNet model demonstrates high segmentation accuracy and robustness across a variety of lighting conditions, such as direct sunlight, shadows, and specular reflections, there are still several important areas where further enhancement is both necessary and promising. One critical avenue for future research lies in optimizing the model for real-time deployment in unmanned ship systems, where computational efficiency is a key constraint. Specifically, although the model achieves excellent segmentation performance, it incorporates attention mechanisms and multi-scale fusion modules (e.g., GCSA, FPN, and SPP) that contribute to an increased computational load. To address this, future work should explore techniques for model compression, such as pruning, quantization, or knowledge distillation, in order to reduce the number of trainable parameters without significantly compromising accuracy. Furthermore, hardware-aware neural architecture search (NAS) and lightweight backbone networks (e.g., MobileNet and ShuffleNet) may be useful to design more compact and efficient variants of ASSA-ResNet, which can meet the strict latency and energy requirements of embedded systems typically used onboard USVs.

In addition to computational concerns, environmental robustness remains a largely unexplored challenge. While this study evaluates model performance across typical daytime conditions, it does not systematically assess segmentation quality under extreme or dynamic maritime environments, such as dense fog, heavy rain, low-light conditions, or rough sea states involving large waves and splashes. These scenarios often introduce visual noise, occlusions, and distortions that can severely degrade the performance of standard computer vision algorithms. Therefore, future research should incorporate techniques such as adversarial data augmentation, domain adaptation techniques, and synthetic-to-real training strategies to improve the model's generalization ability in such adverse settings. Furthermore, assembling or simulating a more diverse and annotated benchmark dataset that reflects real-world maritime operational variability will be crucial for validating model robustness.

In summary, enhancing both the computational efficiency and environmental resilience of the ASSA-ResNet model will significantly advance its practical viability for autonomous maritime navigation, coastal surveillance, and intelligent port operations, making it a critical step toward the broader adoption of AI-driven solutions in smart shipping and oceanographic applications.

6. Conclusions

This study introduces ASSA-ResNet, a multi-scale feature fusion model designed to address the limitations of existing semantic segmentation methods in water surface imagery, particularly in environments affected by background noise from water surface illumination and shoreline reflections. ASSA-ResNet leverages ResNet-50 as the backbone for feature extraction and incorporates a GCSA module to dynamically recalibrate feature representations across both channel and spatial dimensions. This mechanism enables the network to selectively focus on relevant regions, thereby effectively suppressing background interference. To further enhance semantic integration and receptive field capacity, an FPN is employed alongside an SPP module for robust multi-scale feature fusion and precise identification of complex water area boundaries. Experimental validation on a self-constructed Beihai Scene Matching dataset demonstrates the model's superior segmentation performance, achieving 98.17% mIoU and 98.90% mPA, outperforming several established benchmark methods. The proposed model also exhibits strong generalizability when evaluated on the USVInland unmanned vessel dataset, attaining 99.18% mPA and 98.39% mIoU.

Author Contributions: Software: Y.W.; validation: Y.W. and W.L.; writing—original draft: Y.W., Z.Z. and N.W.; formal analysis: Y.W. and W.L.; conceptualization, Y.W. and N.W.; investigation, N.W.; resources, N.W.; writing—review and editing, Y.W. and N.W.; supervision, N.W.; methodology, Y.W., W.L. and Z.Z.; project administration: N.W.; funding acquisition: N.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported partially by the Guangxi Science and Technology Joint Special Program (2025GXNSFHA069258); by the Guangxi Science and Technology Major Program (2024AA29055); and by the 100 Scholar Plan of the Guangxi Zhuang Autonomous Region of China (2018).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in Zenodo at https://zenodo.org/records/16608640. (accessed on 30 July 2025).

Acknowledgments: The authors wish to express their sincere gratitude to Ivan Lee of the University of South Australia for his valuable insights and constructive discussions.

Conflicts of Interest: The authors declare that this research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

1. Peng, Y.; Yang, Y.; Cui, J.; Li, X.; Pu, H.; Gu, J.; Xie, S.; Luo, J. Development of the USV 'JingHai-I' and sea trials in the Southern Yellow Sea. *Ocean. Eng.* **2017**, *131*, 186–196. [CrossRef]
2. Ferri, G.; Manzi, A.; Fornai, F.; Ciuchi, F.; Laschi, C. The HydroNet ASV, a Small-Sized Autonomous Catamaran for Real-Time Monitoring of Water Quality: From Design to Missions at Sea. *IEEE J. Ocean. Eng.* **2015**, *40*, 710–726. [CrossRef]
3. Ruangpayoongsak, N.; Sumroengrit, J.; Leanglum, M. A Floating Waste Scooper Robot On Water Surface. In Proceedings of the 2017 17th International Conference on Control, Automation and Systems (ICCAS), Jeju, Republic of Korea, 18–21 October 2017; pp. 1543–1548.
4. Mendonca, R.; Marques, M.M.; Marques, F.; Lourenco, A.; Pinto, E.; Santana, P.; Coito, F.; Lobo, V.; Barata, J. A cooperative multi-robot team for the surveillance of shipwreck survivors at sea. In Proceedings of the OCEANS 2016 MTS/IEEE Monterey, Monterey, CA, USA, 19–23 September 2016; pp. 1–6.
5. Wang, W.; Gheneti, B.; Mateos, L.A.; Duarte, F.; Ratti, C.; Rus, D. Roboat: An Autonomous Surface Vehicle for Urban Waterways. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 6340–6347.

6. Bae, I.; Hong, J. Survey on the Developments of Unmanned Marine Vehicles: Intelligence and Cooperation. *Sensors* **2023**, *23*, 4643. [CrossRef] [PubMed]
7. Liu, C.; Yang, J.; Yin, J.; An, W. Coastline detection in SAR images using a hierarchical level set segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 4908–4920. [CrossRef]
8. Park, C.; Jeon, J.; Moon, Y.; Eom, I. Single image based algal bloom detection using water areas segmentation and probabilistic algae indices. *IEEE Geosci. Remote Sens. Lett.* **2019**, *7*, 8869–8878.
9. Amitrano, D.; Ciervo, F.; Di Martino, G.; Papa, M.N.; Iodice, A.; Koussoube, Y.; Mitidieri, F.; Riccio, D.; Ruello, G. Modeling watershed response in semiarid regions with high-resolution synthetic aperture radars. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 2732–2745. [CrossRef]
10. Yu, T.; Xu, S.W.; Tao, B.Y.; Shao, W.Z. Coastline detection using optical and synthetic aperture radar images. *Adv. Space Res.* **2022**, *70*, 70–84. [CrossRef]
11. Wang, B.; Chen, L.L.; Cheng, J. New Result on Maximum Entropy Threshold Image Segmentation Based on P System. *Optik* **2018**, *163*, 81–85. [CrossRef]
12. Han, B.; Wu, Y. A novel active contour model based on modified symmetric cross entropy for remote sensing river image segmentation. *Pattern Recognit.* **2017**, *67*, 396–409. [CrossRef]
13. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
14. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the 2015 International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
15. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
16. Jonnala, N.S.; Siraaj, S.; Prastuti, Y.; Chinnababu, P.; Babu, B.P.; Bansal, S.; Upadhyaya, P.; Prakash, K.; Faruque, M.R.I.; AlMugren, K.S. AER U-Net: Attention-enhanced multi-scale residual U-Net structure for water body segmentation using Sentinel satellite images. *Sci. Rep.* **2025**, *15*, 16099. [CrossRef] [PubMed]
17. Jonnala, N.S.; Bheemana, R.C.; Prakash, K.; Bansal, S.; Jain, A.; Pandey, V.; Faruque, M.R.I.; Al-Mugren, K.S. DSIA U-Net: Deep shallow interaction with attention mechanism UNet for remote sensing satellite images. *Sci. Rep.* **2025**, *15*, 549. [CrossRef] [PubMed]
18. Li, A.; Jiao, L.; Zhu, H.; Li, L.; Liu, F. Multitask Semantic Boundary Awareness Network for Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5400314. [CrossRef]
19. Leissaee, A.A.; Kumar, A.; Anwer, R.M.; Khan, S.; Cholakkal, H.; Xia, G.S.; Khan, F.S. Transformers in Remote Sensing: A Survey. *Remote Sens.* **2023**, *15*, 1860. [CrossRef]
20. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2023, Paris, France, 1–6 October 2023; pp. 4015–4026.
21. Hong, Y.; Zhou, X.; Hua, R.; Lv, Q.; Dong, J. WaterSAM: Adapting SAM for Underwater Object Segmentation. *J. Mar. Sci. Eng.* **2024**, *12*, 1616. [CrossRef]
22. Zhang, Z.; Cai, H.; Han, S. Efficientvit-sam: Accelerated segment anything model without performance loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024, Seattle, WA, USA, 16–22 June 2024; pp. 7859–7863.
23. Fu, J.; Yu, Y.; Li, N.; Zhang, Y.; Chen, Q.; Xiong, J.; Yin, J.; Xiang, Z. Lite-sam is actually what you need for segment everything. In Proceedings of the European Conference on Computer Vision, Milano, Italy, 29 September–4 October 2024; Springer Nature: Cham, Switzerland, 2024; pp. 456–471.
24. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; Springer: Munich, Germany, 2018; pp. 3–19.
25. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs [EB/OL]. Available online: https://arxiv.org/pdf/1412.7062.pdf (accessed on 7 May 2015).
26. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation [EB/OL]. Available online: https://arxiv.org/pdf/1706.05587.pdf (accessed on 5 December 2017).
27. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434.
28. Li, X.; Wang, X.; Ye, H.; Qiu, S.; Liao, X. Multinetwork Algorithm for Coastal Line Segmentation in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–12. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]

30. Warburg, F.; Hauberg, S.; Lopez-Antequera, M.; Gargallo, P.; Kuang, Y.; Civera, J. Mapillary street-level sequences: A dataset for lifelong place recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2626–2635.
31. Sünderhauf, N.; Neubert, P.; Protzel, P. Arewethereyet Challenging SeqSLAM on a 3000 km journey across all four seasons. In Proceedings of the 2013 IEEE International Conference on Robotics and Automation (ICRA), Karlsruhe, Germany, 6–10 May 2013.
32. Maddern, W.; Pascoe, G.; Linegar, C.; Newman, P. 1year, 1000km: Theoxford robotcar dataset. *Int. J. Robot. Res.* **2017**, *36*, 3–15. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.