Journal of *Marine Science and Engineering*

MDPI

Article

Deep Learning-Based Maritime Environment Segmentation for Unmanned Surface Vehicles Using Superpixel Algorithms

Haolin Xue [1,†], Xiang Chen [2,†], Ruo Zhang [3,4], Peng Wu [1], Xudong Li [5,6] and Yuanchang Liu [1,*]

1 Department of Mechanical Engineering, University College London, Torrington Place, London WC1E 7JE, UK; haolin.xue.20@ucl.ac.uk (H.X.); peng.wu.14@ucl.ac.uk (P.W.)
2 Department of Civil, Environmental and Geomatic Engineering, University College London, Chadwick Building, London WC1E 6BT, UK; xiang.chen.17@ucl.ac.uk
3 Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen 518057, China; ruozhang0608@gmail.com
4 Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen 518055, China
5 School of Mechanical Engineering, Dalian University of Technology, Dalian 116024, China; lixudong2015@mail.dlut.edu.cn
6 Key Laboratory for Micro/Nano Technology and System of Liaoning Province, School of Mechanical Engineering, Dalian University of Technology, Dalian 116024, China
* Correspondence: yuanchang.liu@ucl.ac.uk
† These authors contributed equally to this work.

Abstract: Unmanned surface vehicles (USVs) are receiving increasing attention in recent years from both academia and industry. To make a high-level autonomy for USVs, the environment situational awareness is a key capability. However, due to the richness of the features in marine environments, as well as the complexity of the environment influenced by sun glare and sea fog, the development of a reliable situational awareness system remains a challenging problem that requires further studies. This paper, therefore, proposes a new deep semantic segmentation model together with a Simple Linear Iterative Clustering (SLIC) algorithm, for an accurate perception for various maritime environments. More specifically, powered by the SLIC algorithm, the new segmentation model can achieve refined results around obstacle edges and improved accuracy for water surface obstacle segmentation. The overall structure of the new model employs an encoder–decoder layout, and a superpixel refinement is embedded before final outputs. Three publicly available maritime image datasets are used in this paper to train and validate the segmentation model. The final output demonstrates that the proposed model can provide accurate results for obstacle segmentation.

Keywords: unmanned surface vehicles; image segmentation; deep convolutional neural network; superpixel algorithm; maritime image data

Citation: Xue, H.; Chen, X.; Zhang, R.; Wu, P.; Li, X.; Liu, Y. Deep Learning-Based Maritime Environment Segmentation for Unmanned Surface Vehicles Using Superpixel Algorithms. *J. Mar. Sci. Eng.* **2021**, *9*, 1329. https://doi.org/10.3390/jmse9121329

Academic Editor: Philippe Blondel

Received: 29 September 2021
Accepted: 19 November 2021
Published: 25 November 2021

1. Introduction

1.1. Background and Motivation

Unmanned vehicles, especially drones and autonomous cars, have been widely applied in our daily lives due to the recent advances in robotics and artificial intelligence. Such a trend has attracted increasing attention from the maritime industry with unmanned surface vehicles (USVs) being rapidly developed. USVs can perform various tasks, such as scientific exploration, environmental information collection, search and rescue, and communication, in various scenarios [1]. More importantly, USVs have an inherent advantage in exploring hazardous areas because of their miniaturised design, which can greatly reduce the reliance on human operators and allows autonomous long-duration operations in harsh environments.

In general, a marine environment in which a USV operates is typically dynamic and unpredictable influenced by sudden high-speed incursions on the route, weakness in

environmental awareness due to waves and sea fog, and signal disruptions, etc. Various sensors such as radar, LiDAR and inertial measurement units can be used for environment perception and localisation. However, these sensors may have several disadvantages, such as limited detection accuracy, impaired capability in detecting submerged obstacles, and high prices, especially for LiDAR sensing. Recently, due to the benefits of being able to provide rich texture information and relatively low price, visual cameras have been intensively used for object detection in maritime environments. Among several vision processing algorithms, the classical background subtraction method presents a high false negative rate under the harsh marine environment, which makes it unsuitable to navigate USVs [2]. Target tracking and object detection techniques based on stereo-vision [3] can be generally implemented to cope with dynamic environment perception. However, the accuracy of 3D point cloud map-based algorithms is sensitive to the texture features in the environment. In particular, when the water surface environment has very few features, low-cost cameras can be compromised in providing a reliable detection and consequently generate challenges for environment sensing [4]. For example, a fast semantic segmentation method was proposed in 2006 based on model adaptation using low-cost cameras. The method can achieve an outstanding fast target detection, but it shows a limited accuracy in detecting sunlight reflections and floating objects [5].

1.2. Contributions

Deep learning has raised a lot of attention in recent years due to its powerful feature extraction ability. Semantic segmentation, as one of the key areas using deep learning, is a computer vision task that entails taking raw data (such as images) as input and turning them into masks with highlighted regions of interest. Semantic segmentation-based obstacle detection can assist USVs with identifying potential collision risks when they are conducting tasks. However, existing studies on obstacle detection are typically associated with high false-positive rates and some semantic segmentation algorithms cannot obtain a good obstacle contour from a complex backgrounds. Therefore, to improve the obstacle detection accuracy, especially in environments with complex background, this paper aims to develop a deep learning-based marine environments segmentation algorithm integrated with a superpixel segmentation for a refined object detection result. The contributions of this paper are summarised as follows:

- A superpixel segmentation model called Simple Linear Iterative Clustering (SLIC), has been innovatively integrated with a deep neural network in this paper to improve the segmentation accuracy, especially for obstacle edge detection in maritime environments.
- Enriched cross validations based on three different maritime datasets are conducted with the results proving that the proposed SLIC enabled model has a strong capability in understanding the semantics of the environment.
- Obstacle detection performances are validated using a series of practical maritime datasets with the results showing that a high obstacle detection accuracy can be achieved when using the segmentation generated by the proposed network.

The rest of paper is organised as follows: Section 2 presents a list of related works based on deep learning networks. Section 3 is a description of the Simple Linear Iterative Clustering (SLIC) segmentation algorithm and the construction of the SLIC module is also presented. Section 4 presents all the datasets used in this paper, as well as the results. Section 5 concludes the paper, acknowledges the shortcomings and highlights possible future research directions.

2. Related Work

This section provides a literature review of using computer visions together with machine learning techniques to support autonomous navigation missions of USVs. Given that most USVs are still remotely or semi-autonomously controlled, sensors such as LiDAR [6], radar [7], and visual systems [3], have been intensively used for scene perception. As stated previously, visual systems, as a relatively cheap solution that can provide enriched sensing

information, have been recently adopted. Conventional machine learning approaches, such as Non-parametric Bayesian methods [8] have been applied to process visual signals for semantic labelling. In this work, we limit the review scope to deep convolutional neural network-based methods which are typically implemented to process visual information to assist with USV navigation, for their strong learning capabilities of rich features.

2.1. Image Semantic Segmentation Networks

As a common computer vision issue, semantic segmentation or full pixel semantic segmentation [9] includes image classification, object recognition detection and semantic segmentation. Image semantic segmentation network can provide a pixel-level image interpretation while deferring to human perception, and hence has a broad range of applications. In this section, typical deep networks proposed for semantic segmentation are reviewed with their features highlighted.

Convolutional Neural Networks (CNN) [10] have several fully connected layers after the convolutional layers, which project the feature map generated by the convolutional layers into a fixed-length feature vector. The classical CNN structure represented by AlexNet [11] is suitable for image-level classification and regression tasks since they export a probabilistic description of the whole input image as an output.

Different to CNNs, fully convolutional networks (FCN) can accept input images of any sizes, and use a deconvolutional layer to upsample the feature map. The semantic segmentation network based on FCN using a U-Net structure was proposed for segmentation of medical images in 2015 [12]. U-Net, similar to FCN, has a downsampling and an up-sampling phase and can solve: (1) a pixel localisation problem with shallow information and (2) a pixel classification problem with deep information to accomplish a semantic segmentation. In contrast to the FCN network, the downsampling and up-sampling phases within the U-Net structure employ the same number of convolutional layers. Furthermore, U-Net uses a skip connection structure to send the results of distinct downsampling layers to the appropriate up-sampling layers, allowing the network to obtain more precise pixel space information and increase segmentation accuracy.

SegNet [13] was proposed to improve the scene perception capability within autonomous navigation. The methods within the SegNet and the FCN are rather similar, with the exception that the encoder and decoder within the SegNet (downsampling and up-sampling) use different technologies. As shown in the SegNet architecture in Figure 1, the left side shows a convolutional extraction of features by pooling the perceptual field and shrinking the image (a process known as encoding). Convolution is used here to extract features from the encoder. The deconvolution and up-sampling layers are located on the architecture's right side. After image classification, the features are repeated by deconvolution, then the up-sampling process restores the feature map to the original size of the image (referred to as decoding).

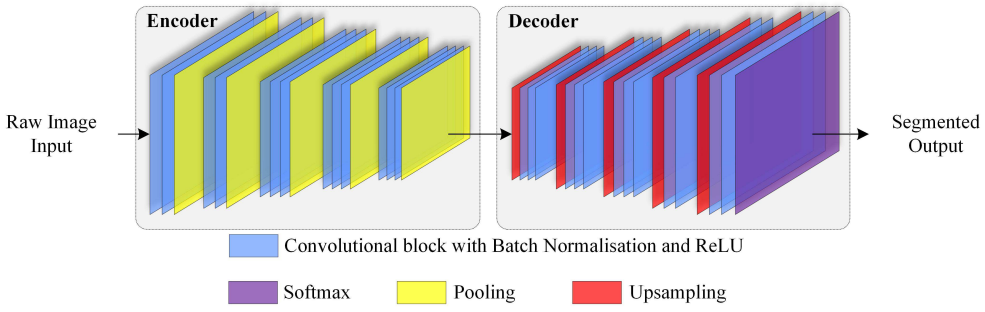

Figure 1. SegNet received inspiration from U-Net and improved using a symmetrical encoder–decoder structure as a pixel-wise end-to-end network where the input image would go through two stages of downsampling and up-sampling.

The DeepLab family [14–17] is a series of semantic segmentation networks proposed by Google. DeepLab v1 was launched in 2014 and achieved the second place in the segmentation task on the PASCAL VOC2012 dataset, followed by DeepLab v2, DeepLab v3 and DeepLab v3+ from 2017 to 2018. Two innovations of DeepLab v1 are the Atrous Convolution and the Fully Connected conditional random field (CRF). DeepLab v2 was thereafter improved by proposing an Atrous Spatial Pyramid Pooling (ASPP). DeepLab v3 further optimises the ASPP by adding a 1×1 convolution, batch normalization operation, etc.

DeepLab v3+ adds an up-sampling decoder module to optimise the accuracy of edges. Considering that the network features of DeepLab v3 do not encompass excessive high-level features, DeepLab v3+ [17] adapts the encoder–decoder structure of Feature Pyramid Networks (FPN) to achieve the fusion of feature maps across blocks. Another contribution of DeepLab V3+ is the adaptation of Aligned Xception network as backbone. As shown in Figure 2, the DCNN part is considered to be encoder, and the part of DCNN output that is upsampled is considered to be decoder, forming an encoder–decoder structure. DeepLab v3+ concatenates the result of the up-sampling of the DeepLab v3 model output with the result of the downsampling of the DCNN output by a factor of 0.25. More refined results are then obtained using a 3×3 convolutional layers and bilinear interpolation of 4-fold up-sampling layers.

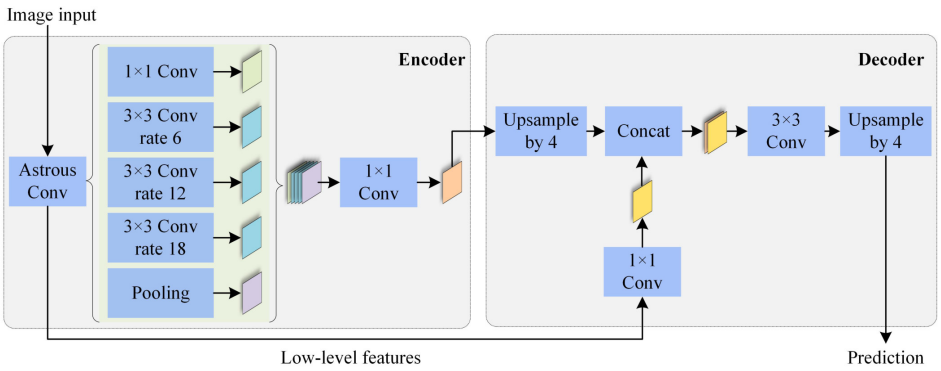

Figure 2. DeepLab v3+ uses an encoder–decoder architecture with two bilinear interpolation up-sampling in the decoder stage to recover the edge information of the image.

2.2. Superpixel Algorithms

Current image processing is largely pixel-based, using a two-dimensional matrix to describe an image without taking into account the spatial organisation relationship between pixels, making the algorithm less efficient. Ren et al. [18] initially presented the notion of superpixels in 2003, which are image blocks made up of surrounding pixels with identical texture, colour, brightness, and other characteristics. It groups pixels based on their similarity in features, allowing it to gather redundant information from the image and greatly minimise the complexity of subsequent image processing operations.

Graph theory-based approaches and gradient descent-based methods are the two most common types of superpixel creation algorithms. The graph-based approach proposed by Felzenswalb et al. [19], the N cut method proposed by Shi et al. [20], and the superpixel lattice method proposed by Moore et al. [21] are the main graph theory-based superpixel segmentation methods. The N cut method uses contour and texture properties to globally minimise the cost function, resulting in regular superpixels, while picture convenience is not fully preserved and is computationally expensive. The graph-based approach, which uses the concept of the least spanning tree to segment the image, can better keep the image border and is faster. However, the size and collision of the created superpixels are uneven. The superpixel lattice technique preserves the image's topological information, but its performance is heavily reliant on the image's pre-extracted bounds.

For gradient descent-based algorithms, important methods include the watershed method [22], the meanshift method [23], the quick-shift method [24] and the Simple Linear

Iterative Clustering (SLIC) method [25]. They are all based on the core concept of clustering. In particular, the SLIC approach is based on colour and distance similarity, and it can produce superpixels of uniform size and shape. More specifically, Achanta et al. [25] proposed a single-minded and easy-to-implement algorithm to convert a colour image into a 5-dimensional feature vector in CIELAB colour space and XY coordinates, and then construct a metric for this feature vector to perform the process of local clustering of image pixels. By initialising the seed points and using a similarity measure, compact, approximately uniform superpixels can be generated relatively quickly. Ren et al. [26] implemented the SLIC algorithm using a GPU and the NVIDIA CUDA framework was able to increase the speed by a factor of 10. Lucchi et al. [27] used the SLIC segmentation algorithm for preprocessing and then built a graph model with superpixels as nodes and spatially adjacent nodes as edges, gave random field definitions for the corresponding conditions and proposed a structural image segmentation using coregulated features.

2.3. Deep Learning-Based Segmentation for USVs

In the past, processing for the detection task of marine semantics data was usually implemented through hardware methods such as LIDAR [28,29]. However, due to the application of machine learning to the navigation of USVs, CNN [30–33], Haar [34], and HOG [35] classifiers are gradually implemented for detection-based tasks. The WaSR [36] network proposed by Borja Bovcon and Matej Kristan achieves target recognition by building fusion blocks in the decoder structure using ASPP modules and FFM modules. Chen et al. [37] achieve accurate target recognition by embedding attention mechanism module to achieve accurate marine semantic segmentation results. In general, most of the existing semantic segmentation networks have been developed and redeveloped based on encoder–decoder structures.

When dealing with datasets containing a large number of samples with diverse features, current deep convolutional neural network approaches are still ineffective, and they typically require additional algorithmic modules, such as an attention mechanism module. In general, larger datasets and more training are required to allow the network model to learn the features of various environments. Nonetheless, realistic circumstances in which the network model has not learned will always exist, and the inference results are often not ideal at this level.

Although superpixel segmentation [38–41] techniques are frequently used in image preprocessing to discover edge characteristics in images, they are rarely used in marine semantic segmentation tasks due to their inability to segment semantically. However, the capability of storing target edge information of superpixel techniques can be used to help with addressing the problem of object edge information loss that occurs in deep convolutional neural networks.

In summary, most of the existing semantic segmentation networks have been developed based on encoder–decoder structures, but the following research gaps can be generated:

- When dealing with datasets containing a large number of samples with diverse features, current deep convolutional neural network approaches are still ineffective.
- Although superpixel segmentation techniques are frequently used in image preprocessing to discover edge characteristics in images, they are rarely used in marine semantic segmentation tasks due to their inability to segment semantically.
- Despite the fact that many studies have used deep learning approaches to solve the segmentation problem of marine semantic datasets, there has been little discussion of the differences between deep convolutional neural networks with different depths in learning marine semantic features.

Therefore, in contrast to literature that investigates new effective deep convolutional network structures for marine semantic segmentation, this paper has proposed a new SLIC enabled segmentation model for refined obstacle detection in maritime environments. It can be demonstrated that by integrating the superpixel algorithm with existing deep neural convolutional networks, the detection accuracy especially for obstacle edges can be well

improved. This can potentially benefit USVs' navigation in a confined area with complex obstacle structures.

3. Method

The proposed semantic segmentation architecture (Figure 3) is assisted by a Simple Linear Iterative Clustering (SLIC) [25] module that refines the segmentation outputs from the deep neural network, with the goal of improving the inference accuracy. Two different types of backbones in DeepLab v3+ the network are used to assess the impacts on semantic segmentation performance.

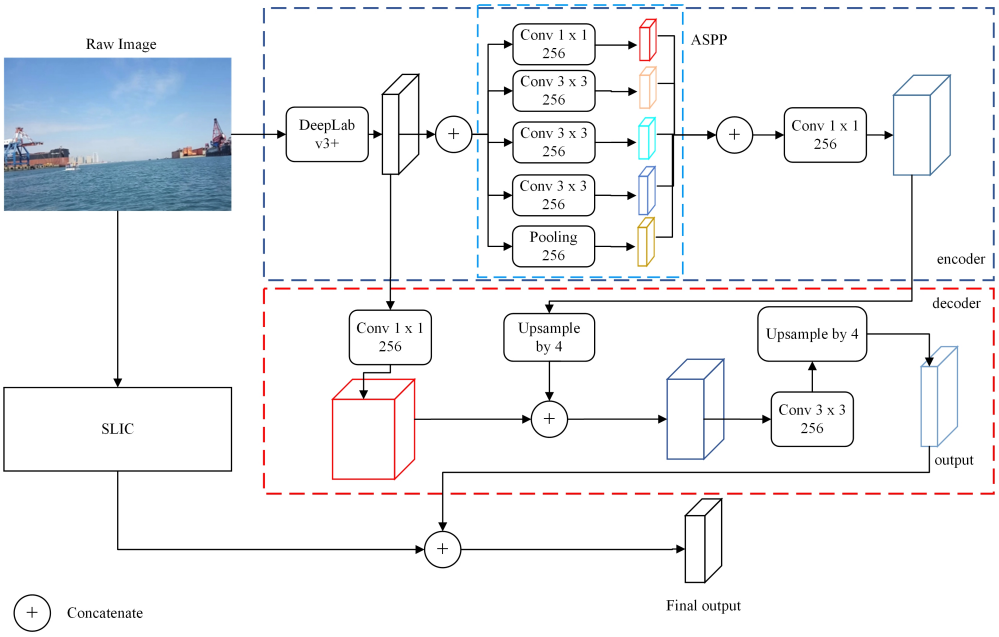

Figure 3. The proposed semantic segmentation architecture: the SLIC algorithm module is mounted on the outside of the deep neural network and refines its results.

3.1. Simple Linear Iterative Clustering Algorithm

The SLIC algorithm [25] performs an iterative categorisation operation on each pixel by converting the colour image into a LAB colour space [42] and combining it with X-Y coordinates to construct a distance function on a five-dimensional vector $C_i = [l_i, a_i, b_i, x_i, y_i]^T$. For pixel i, l_i is the luminance of the pixel; a_i is the colour vector for green to red; b_i is the colour vector for blue to yellow; x_i and y_i indicate the coordinates of the pixel. Please note that differing from RGB colour space, LAB colour space is implemented here as it contains spatial information. Because of the simplicity of the constructed 5D vector C_i, the SLIC algorithm is faster and more efficient than other superpixel algorithms, and can produce a set of superpixels that are compact and have uniform shapes, which facilitate the handling of colour images.

The SLIC algorithm starts from initialising the starting positions of superpixels. When there are N pixels in an image and K superpixels to be assigned, the size of each superpixel is N/K and the distance between each superpixel is $S = \sqrt{N/K}$. For each pixel in an image, the degree of similarity between the pixel and its nearest superpixel is calculated in an iterative way and pixels are grouped to have the same label of the most similar superpixel. To calculate the similarity between each pixel and the superpixel, the distance measure variable D is introduced and calculated as shown in Equation (1).

$$D = \sqrt{{d_c}^2 + (\frac{d_s}{S})^2 m^2} \quad (1)$$

where S is the distance between each pair of superpixels, m indicates the relative importance of space and pixel colour, d_c is the distance of colour information between pixels, d_s is the distance of spatial information between pixels. Figure 4 shows that as the value of m increases, the weight of spatial relations decreases and the weight of colour relations increases, resulting in that the generated superpixels are more irregular in shape but fit closely to the edges in the original image, while at the same time there are more unclassified pixels.

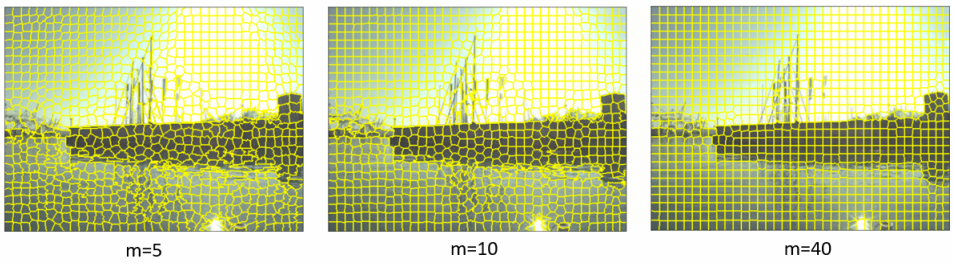

Figure 4. Different values of m lead to different segmentation results for the SLIC algorithm. $m = 5, 10, 40$ were presented.

To improve the algorithm's computational speed, when clustering each pixel, rather than in a whole image, similar pixels are searched only in a $2S \times 2S$ region (as shown in Figure 5) and all pixels within this area will be grouped belonging to the same superpixel.

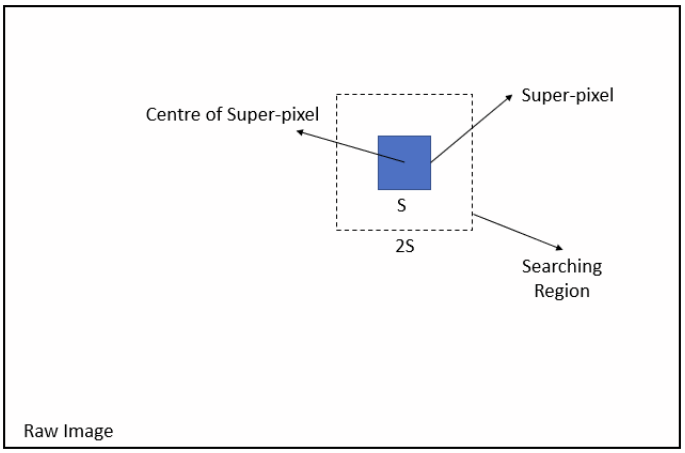

Figure 5. By performing clustering iterations in a $2S \times 2S$ area, the computational burden caused by the increased number of superpixels can be greatly alleviated.

3.2. Structure of the SLIC Enabled Segmentation Model

As explained in the previous section, DeepLab v3+ adds a decoding module to recover some of the detailed information. More specifically, the DeepLab v3+ network employs two bilinear interpolation upsamplings to increase the feature resolution but without the capability of fully restoring the lost information during the up-sampling process. The lost image information is especially centred around edges areas, which is not ideal for an accurate object edge segmentation.

Hence, to address this issue, a new segmentation model with superpixel optimisation is proposed in this paper with the network structure shown in Figure 5. The model adopts a DeepLab v3+ network for an initial segmentation with the encoder module of DeepLab v3+ shown in the blue bounding box and the decoder module shown in the red bounding box. Using such an encoding-decoding structure, semantic segmentation features of images can be extracted to generate a relatively coarse segmentation result. In the meantime, the superpixel optimisation module is used as a subsequent processing part to optimise the semantic segmentation results, especially around edge areas. In this paper, we assume that

a maritime environment only contains three semantic categories including sky, ocean and obstacles, the SLIC algorithm sets various criteria for each category to preserve the sky and ocean while trimming the edges of the obstacle. Last, by fusing the high-level semantic feature information from DeepLab v3+ and the refined object edges information from the SLIC, the final output can be generated. Please note that DeepLab v3+ network is adopted because it is reliable and relatively accurate. Other semantic segmentation networks, such as SegNet [13], BiSeNet [43] and DFANet [44], can also be selected according to specific requirements, which will be discussed in this paper.

The details of the fusion process within the proposed model (Figure 2) is explained here. The superpixel segmentation information from the SLIC algorithm will be projected onto a new image first. Although the SLIC superpixel segmentation information is being mapped, the superpixels are reclassified based on the semantic segmentation results from the DeepLab v3+. Such a process is equivalent to "recolouring" the superpixel segmentation results, i.e., reassigning semantic information to each superpixel. By following such a process, the whole image can be rescanned pixel by pixel until no empty pixels exist.

3.3. Network Implementation

The network implementation includes the selection of backbone and the configuration of the loss function to achieve optimised learning results. To investigate the impact of the SLIC algorithm on the inference time and accuracy, this paper will discuss the adaptability of the proposed model using different types of backbones in the DeepLab v3+ network.

More specifically, Xception [45] and ResNet101 [46] are selected as they are the most commonly used ones in image segmentation. Between them, Xception, as a network structure using a depthwise separable convolution operation, has the feature of reduced network parameters making it computationally efficient. On the contrary, ResNet101 is constructed by combining a forward neural network with shortcut links, to address the problem of degradation caused by deep layers of networks and subsequently increase the learning capacity.

With regards to the loss function construction, as stated in [37], we consider a maritime environment consists of three categories including sky, ocean and obstacles. Training of the model is undertaken using public available datasets such as the MaSTr1325 dataset [47], where there is an issue of label unbalance between obstacles and the other two categories. To accommodate such an issue, L_{final}, a mixture of cross-entropy (Equation (3)) and focal-entropy (Equation (2)) loss functions is adopted in this paper as shown in Equation (4):

$$L_{fl} = -(1 - p_{t_i})^{\gamma} log(p_{t_i}) \tag{2}$$

$$L_{ce} = L(y_{t_i}, \hat{y_{t_i}}) = -ylog(\hat{p}) - (1 - y)log(1 - \hat{p}) \tag{3}$$

$$L_{final} = (1 - \lambda) * L_{fl} + \lambda * L_{ce} \quad \lambda \in (0, 1) \tag{4}$$

where L_{fl} is the focal-entropy loss function; p_{t_i} is the estimated probability of all three types of targets; γ is the hyperparameter; L_{ce} is the cross-entropy loss function; y_{t_i} is the label of the three different categories of targets in the image; $\hat{y_{t_i}}$ is the estimated output of the three categories of targets after passing through the network. Cross-entropy loss is a widely implemented classical loss function for machine learning, and focal-entropy loss can solve unbalanced classes issue by assigning weights. Such a loss function is appropriate for our semantic segmentation task with large-size networks as backbones. It is not useful in practice to find the global optimum for large-size networks when many local optimums exist with good performance; excessive training may lead to overfitting [48].

3.4. Trade-Off: Compactness and Accuracy of Superpixel Segmentation

It should be noted that there is a trade-off in optimising the output of the DeepLab v3+ network using the SLIC algorithm, where a good superpixel segmentation result is

usually deemed to have good compactness and connectivity. In practice, however, when the value of m becomes small, some "holes" might be generated in places that are not classified as any cluster. To eliminate these "holes", the output of DeepLab v3+ must be marked and coloured in the end, which may affect the effectiveness and efficiency of the SLIC optimisation. As the value of m becomes larger, the problem of generating "holes" pixels can be significantly improved, but under-segmentation issue might occur and the superpixels do not adhere well to the edges, especially in those irregularly shaped targets such as masts, mountains, birds etc., resulting in a reduced optimisation.

To resolve the conflicting relationship between the compactness of superpixels and the segmentation accuracy generated by the SLIC algorithm, the number of superpixels usually must be increased when performing a segmentation task. This means that densely clustered regions are generated on the image using a larger number of small-sized superpixels in order to increase the sensitivity of the segmentation algorithm, i.e., using a high value of m to generate high compactness, while still having a good sensitivity of the superpixels to the target edges. Such a procedure will be specifically discussed in the Experiments and Results section, but it is without a doubt that as the number of superpixels increases, the computational time of the SLIC algorithm will increase accordingly, which may limit the on-line computing performance of the SLIC algorithm.

4. Experiments and Results

This section discusses the implementation results of the proposed model. Evaluation metrics will be first discussed followed by data augmentation methods and training setups. The performances of the SLIC algorithm on the maritime datasets will be presented together with the comparative analysis of the proposed model with the original DeepLab v3+ network.

4.1. Dataset and Evaluation Metrics

In this paper, the proposed model is trained using the Marine Semantic Segmentation Training Dataset (MaSTr1325) [47], a large scale marine semantic segmentation training dataset designed for small USVs for target detection. The MaSTr1325 dataset contains 1325 fully annotated high-resolution images of the real marine environments. Ground-truth masks are labelled with different values depending on the categories, for example, obstacles are labelled with 0, sea or water with 1, sky with 2 and unknown areas with 4. A total number of 1221 images in the MaSTr1325 dataset were used for the training, and the remaining images were included in the validation set. Another used dataset is the Multi-modal Marine Obstacle Detection Dataset 2 (MODD2) [49], which is the most comprehensive and voluminous marine obstacle detection dataset available. It contains images of targets at different scales in different weather and extreme conditions, such as sunny days, hazy weather, reflections on the sea surface, and varying sea states. Marine Image Dataset (MID) [50] was also selected to be included in the validation set to test the performance of the model because the MID dataset contains enriched open sea area images. Sample images from different datasets are shown in Figure 6.

Accuracy, recall, F1-measure, and $mIoU$, are used to quantitatively evaluate the performance of the proposed deep neural network model [51]. Precision is calculated by dividing the number of genuine positive samples by the total number of samples. Recall is calculated by dividing the number of true positive samples by the total number of positive samples that should be identified. The F1 score is a mixture of the two indications, where a model with a higher F1 score is more resilient. $mIoU$ is used to compute the ratio of the true and predicted sets' intersection. For semantic segmentation tasks, the $mIoU$ value provides an intuitive representation of how closely the inference results match the ground truth. The greater the value of $mIoU$ (the closer it is to 1), the better the model performs in theory. The definition of these four categories of metrics is shown in Equations (5)–(8).

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

$$mIoU = \frac{1}{k} \sum_{i=1}^{k} \frac{P \cap G}{P \cup G} \tag{8}$$

where TP indicates the number of samples that are true positive. FP indicates the number of samples that are false positive. FN represents false negative. Thus, $TP + FP$ represents the number of samples that are predicted to be positive and $TP + FN$ is the number of samples that are actually positive. As shown in Equation (8), k is the number of semantic types in the dataset, P is the predicted value for each semantic type, and G is the ground-truth value for each semantic type. Again, in this paper, the value of k is 3, representing the three semantic types, i.e., obstacle, ocean and sky.

Figure 6. Some sample images from MaSTr1325, MODD2 and MID datasets.

4.2. Data Augmentation and Training Setups

Data augmentation is a useful technique for boosting data variety by systematically creating extra training samples. The 1221 high resolution (512×384) images used for training in MaSTr1325 are exposed to random data augmentation such as vertical symmetry, brightness change, contrast change, saturation change, etc. The proposed network model was trained using Adam optimiser and the learning rate was initialised to be 10^{-4}. The Resnet101 network used in DeepLab v3+ was pre-trained on a subset of COCO train2017 [52], and the Xception network used was pre-trained on the ImageNet dataset [11]. Each type of network was trained for 50 epochs with a batch size of 2. All training and validation work was done on a node equipped with a Nvidia Tesla V100 at UCL's High Performance Computing platform. All inference work was done on an Intel Core i7-10875H 2.3 GHz workstation with 16 GB of RAM. All parameters are shown in Table 1.

Table 1. Model parameter settings and hardware information.

Drives	Parameters
CPU-inference	i7-10875H 2.3 GHz
GPU	Nvidia Tesla V100
Deep Learning Network API	Pytorch 1.9
Compile Language	Python
Image size	512×384
Training epochs	50
Optimiser	Adam
Learning rate	10×10^{-4}
Batch size	2
Training images	1221
Validation images	324

4.3. SLIC Algorithm Results

Figure 7 shows the output images of the SLIC algorithm under different iterations. The black dots in the figure are the clustering centres of each superpixel and the output is Gaussian blurred for each superpixel for better visual discrimination. When configuring the number of superpixels to be 500 and $m = 10$, the image is segmented in an iterative way. It can be observed in Figure 7 that when the number of iterations reaches 10, the segmentation result is close to the original image showing that the superpixels can provide edges detection. However, it can also be seen that some of the superpixels in the marine area are irregularly shaped and jagged, suggesting that the superpixels are more sensitive to colours than spatial relationships, which can lead to "holes" in the segmentation results.

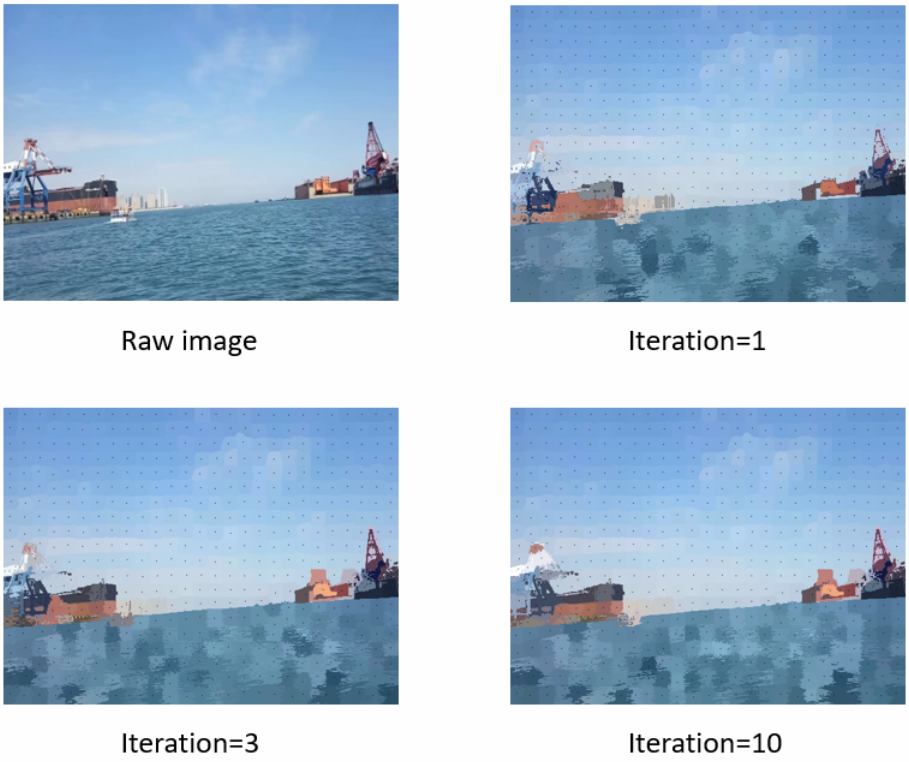

Figure 7. Output image at different iterations for $K = 500$, $m = 10$.

Figure 8 shows the results of segmenting the same image using 10 iterations using 2000 superpixels and 4000 superpixels with $m = 10$. Combining this with the output image shown in Figure 7, it can be seen that as the number of superpixels increases, the size becomes smaller, and the output segmented image is more detailed. For example, in the

left in Figure 4, the white stripe on the hull of the ship can be well preserved, which is not the case when $K = 2000$. In addition, as the number of superpixels increases, each superpixel is more compact with fewer "holes" in the image.

Figure 8. For $m = 10$, iteration=10, the output of the SLIC algorithm module for $K = 2000$ and $K = 4000$.

As can be seen from the above results, the application of the SLIC algorithm achieves a good accuracy when processing images with marine semantic types. Most of the superpixels can detect the edges of objects. One of the foreseeable drawbacks is that the iterative segmentation results using small m-values are too sensitive to some features leading to "holes" due to unworked superpixel edges. Overall, however, the SLIC algorithm achieves fast segmentation of images through a simple logical structure, which has potential and feasibility for refining inferred images from the DeepLab V3+ network. The results from the output of the SLIC optimisation module and the impact of carrying the SLIC module on the overall convolutional neural network model will be shown and discussed in the next section.

4.4. Comparisons with Original DeepLab v3+ Network

Figure 9 shows the output of the DeepLab v3+ network model using Xception as the backbone (labelled as initial results) and the output optimised using the SLIC algorithm (labelled as refined results). It can be seen that the SLIC algorithm can finely trim features with detailed textures. For example, features such as masts and streetlights in the first and third rows of figures have sharper edges after being processed by the SLIC module. The SLIC algorithm module can also rectify several portions in the sky that are mistakenly recognised as barriers in the third row of figures. Furthermore, the railings of the boats on the water in the second row are trimmed finer to get closer to the ground truth, and the trimming is better than the other two sample figures, implying that the SLIC algorithm module has a more robust trimming strategy for complex obstacles such as hole-filled or thinly shaped targets. However, the improvement of the SLIC algorithm is relatively small due to the limited learning capability of Xception.

Figure 10 depicts the DeepLab v3+ network model output using ResNet101 as the backbone (labelled as initial results), as well as the output optimised using the SLIC method (labelled as refined results). The SLIC algorithm module also has a stronger refining impact on the ResNet output mask's edges, such as the railings of the ship in the second row in Figure 10, while the other two rows also show results that are closer to the ground truth. Compared with Xception, ResNet101 has a more powerful learning capability, so the results obtained are more refined. In addition, with the SLIC algorithm, the edges of objects are clearer. For example, some errors (islands on the left of the horizon) in the second row of figures that could affect USVs navigation can be well corrected.

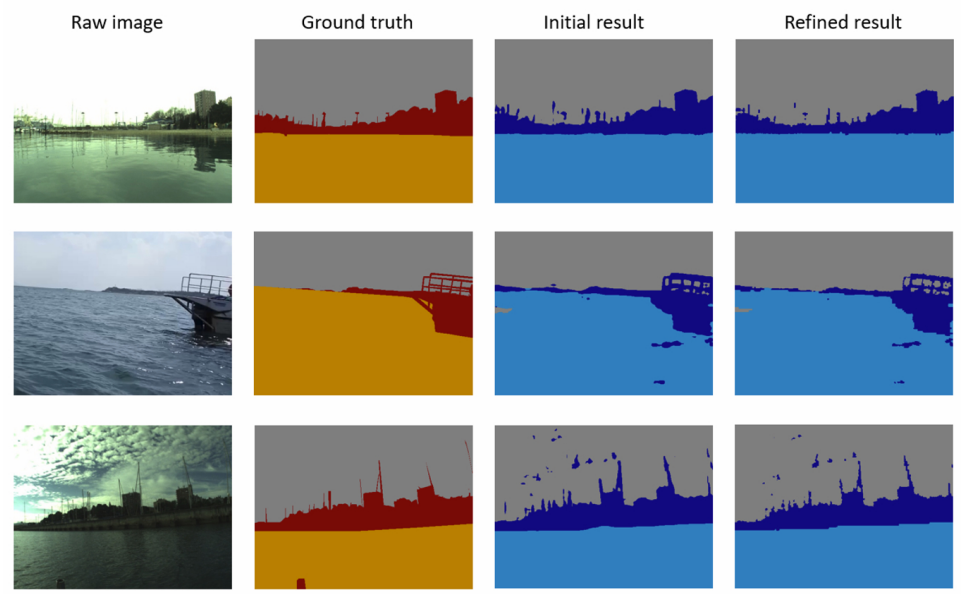

Figure 9. Sample image of the SLIC algorithm module's optimisation of a DeepLab v3+ network with Xception as the backbone. Initial result is the output of the DeepLab network, refined result is the result after processing by SLIC module.

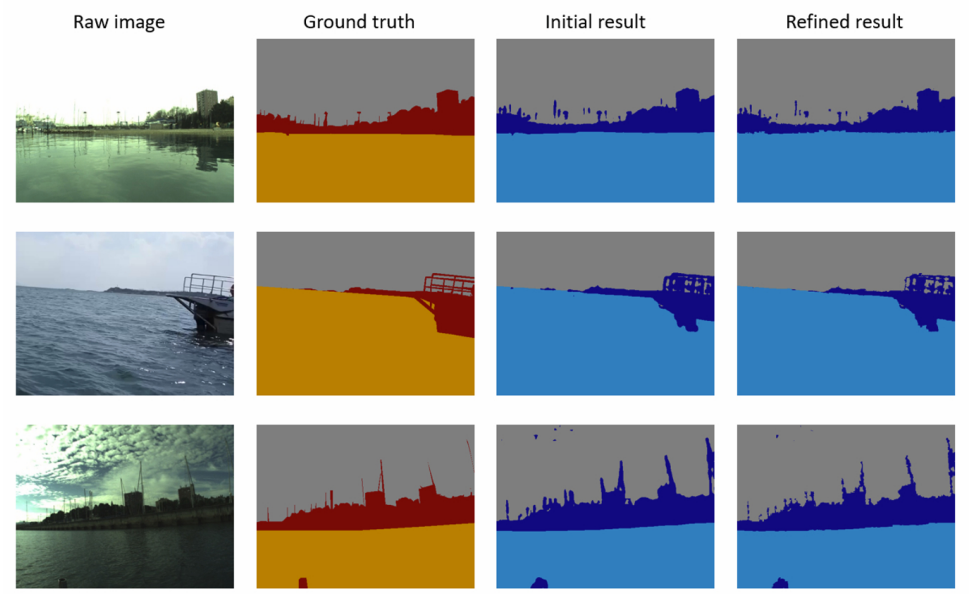

Figure 10. Sample image of the SLIC algorithm module's optimisation of a DeepLab v3+ network with ResNet101 as the backbone. Initial result is the output of the DeepLab network, refined result is the result after processing by SLIC module.

4.5. Quantitative Results

The quantitative results of the SLIC algorithm module's effect on the DeepLab v3+ network model is shown in Table 2. The *mIoU* metric is used to visualise the output of the proposed model. The evaluation metrics used to segment the images using the DeepLab v3+ network model with different backbones, are all derived based on the 324 images in the validation dataset.

In Table 2, according to the inference time of four network models, the DeepLab v3+ network model with ResNet as the backbone outperformed the one with Xception, regardless of whether the SLIC algorithm module was employed. When the SLIC algorithm is adopted, the *mIoU* of the DeepLab v3+_ResNet101 model increased from 89.1% to

90.1% and the $mIoU$ of the DeepLab v3+_Xception model increased from 85.5% to 85.9%. Although such an increase is relatively small, it is mainly generated around edge areas to further improve the high accuracy produced by DeepLab v3+. Additionally, all currently available datasets are taken from open sea areas with limited edge features, to better reveal the superiority of the proposed model, new dataset, especially containing complex features, need to be recorded.

Table 3 compares the $mIoU$ of our DeepLabv3+_ResNet101+SLIC with that of the WODIS [37] on the Singapore Maritime dataset (SMD) [53], MID and MODD2 validation datasets. The WODIS network is a state-of-the-art deep network that has been specifically designed for the maritime environment, which has better performances than other standard networks such as SegNet [13]. Table 3 demonstrates that our proposed DeepLabv3+_ResNet101+SLIC framework is on par with the WODIS, i.e., achieving 89.2% $mIoU$ on the MID dataset (88.1% for WODIS). On the SMD and MODD2 datasets, similar $mIoU$ values (0.2% and 0.1% differences on SMD and MODD2 respectively) can be observed for the two models. Please note that high $mIoU$ values for WODIS are mainly the results of the inclusion of attention modules that are focusing on the objects on water; whereas our proposed network achieves the high $mIoU$ values by having refined results of obstacle edges, which can be better used to ensure navigational safety.

In terms of inference time, the DeepLab v3+_ResNet101 model was about 0.5 seconds quicker than the DeepLab v3+_Xception model. The difference in $mIoU$ between these two network models in inferring the complete validation set was around 3.6%. This is because ResNet101 has a greater depth and learning power, as well as stronger generalisation capability. However, because of its more complicated structure, ResNet takes longer to train and is more prone to overfitting than the Xception network model, thus it is critical to monitor the training process while stated parameters are used.

From a runtime perspective, although the DeepLab v3+_ResNet101 model is faster and more accurate than the DeepLab v3+_Xception network model, the capacity of processing images in real time is slightly limited due to the inherent network architecture limitations. However, we argue that the main focus of this paper is on refining edges detection and the inference time can be improved by implementing light weight structures such as MobileNets [54] into the proposed model.

Table 2. DeepLab v3+ network and the average inference time and $mIoU$ with the SLIC algorithm module on board.

Model	Time(s)	$mIoU$(%)
DeepLab v3+_Xception	1.2301	85.5
DeepLab v3+_ResNet101	0.6725	89.1
DeepLab v3+_Xception + SLIC	1.7865	85.9
DeepLab v3+_ResNet101 + SLIC	1.1256	90.1

Table 3. Comparison of $mIoU$ between the WODIS [37] and our DeepLabv3+_ResNet101+SLIC on the SMD, MID and MODD2 validation datasets.

$mIoU$(%)	SMD	MID	MODD2
WODIS [37]	93.7	88.1	88.2
DeepLabv3+_ResNet101+SLIC	93.5	89.2	88.1

5. Conclusions

This paper proposes a novel deep semantic segmentation model for unmanned surface vehicles operating in the complex maritime environment. The proposed model employs a deep neural network with an auto-encoder layout and a superpixel refinement module which refines the reconstruction output from the deep neural network. Three open-source

maritime image datasets are applied to train and validate the developed semantic segmentation model. The results imply that the deep neural network assisted by the superpixel method has achieved improved semantic segmentation performance with slightly increased inference time. The improved segmentation performance is attributed to the SLIC superpixel method, which is good at distinguishing edges in images. Such a unique feature compensates for the imperfect edge information provided by the deep neural network, consequently leading to an improved mean IoU of 90.1% when the combination of DeepLab v3+_ResNet101 + SLIC is applied. Additionally, we have also observed that the model is less effective in images with a low fraction of obstacles, which is possibly due to the insufficient connectivity between the SLIC and the deep neural network. The SLIC processing also increases the time complexity. In future work, the SLIC will be better integrated with the deep learning processes.

Author Contributions: Conceptualization, H.X., Y.L.; methodology, H.X., X.C., R.Z.; software, H.X., X.C.; validation, H.X., X.L.; formal analysis, H.X., X.C., X.L., Y.L.; investigation, H.X., X.C., Y.L.; resources, R.Z., P.W., Y.L.; data curation, Y.L.; writing—original draft preparation, H.X., X.C., X.L., Y.L.; writing—review and editing, H.X., P.W., Y.L.; visualisation, R.Z.; supervision, Y.L.; project administration, Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by Royal Society (Grant no. IEC-NSFC-191633).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Manley, J.E. *Unmanned Surface Vehicles, 15 Years of Development*; OCEANS 2008; IEEE: Quebec City, QC, Canada, 2008; pp. 1–4.
2. Prasad, D.K. Object Detection in a Maritime Environment: Performance Evaluation of Background Subtraction Methods. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 1787–1802. [CrossRef]
3. Sinisterra, A.J.; Dhanak, M.R.; von Ellenrieder, K. *Stereo Vision-Based Target Tracking System for an USV*; 2014 Oceans-St. John's; IEEE: Quebec City, QC, Canada, 2014; pp. 1–7.
4. Nguyen, A.; Le, B. 3D point cloud segmentation: A survey. In Proceedings of the 2013 6th IEEE Conference on Robotics, Automation and Mechatronics (RAM), Manila, Philippines, 12–15 November 2013; pp. 225–230.
5. Kristan, M.; Kenk, V.S.; Kovačič, S.; Perš, J. Fast image-based obstacle detection from unmanned surface vehicles. *IEEE Trans. Cybern.* **2015**, *46*, 641–654. [CrossRef] [PubMed]
6. Halterman, R.; Bruch, M. Velodyne HDL-64E lidar for unmanned surface vehicle obstacle detection. Unmanned Systems Technology XII. *Int. Soc. Opt. Photonics*, **2010**, *7692*, 76920D.
7. Han, J.; Cho, Y.; Kim, J. Coastal SLAM with marine radar for USV operation in GPS-restricted situations. *IEEE J. Ocean. Eng.* **2019**, *44*, 300–309. [CrossRef]
8. Niemeyer, M.; Arandjelović, O. Automatic semantic labelling of images by their content using non-parametric Bayesian machine learning and image search using synthetically generated image collages. In Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 1–3 October 2018; pp. 160–168.
9. Haralick, R.M.; Shapiro, L.G. Image segmentation techniques. *Comput. Vis. Graph. Image Process.* **1985**, *29*, 100–132. [CrossRef]
10. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6.
11. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
12. Ronneberger, O.; Fischer, P.; Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*; Navab, N., Hornegger, J., Wells, W., Frangi, A., Eds.; Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015; Springer: Cham, Swizterland, 2015; pp. 234–241. Available online: https://arxiv.org/pdf/1505.04597.pdf (accessed on 1 October 2021).
13. 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]
14. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
15. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]

16. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
17. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
18. Malik, R. Learning a classification model for segmentation. In Proceedings of the Proceedings Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; pp. 10–17. [CrossRef]
19. Felzenszwalb, P.F.; Huttenlocher, D.P. Efficient graph-based image segmentation. *Int. J. Comput. Vis.* **2004**, *59*, 167–181. [CrossRef]
20. Shi, J.; Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 888–905.
21. Moore, A.P.; Prince, S.J.; Warrell, J.; Mohammed, U.; Jones, G. Superpixel lattices. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
22. Vincent, L.; Soille, P. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.* **1991**, *13*, 583–598. [CrossRef]
23. Comaniciu, D.; Meer, P. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 603–619. [CrossRef]
24. Vedaldi, A.; Soatto, S. Quick shift and kernel methods for mode seeking. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 705–718.
25. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef] [PubMed]
26. Ren, C.Y.; Reid, I. *gSLIC: A Real-Time Implementation of SLIC Superpixel Segmentation*; Technical Report; University of Oxford, Department of Engineering: Oxford, UK, 2011; pp. 1–6.
27. Lucchi, A.; Li, Y.; Smith, K.; Fua, P. Structured image segmentation using kernelized features. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 400–413.
28. Thompson, D.; Coyle, E.; Brown, J. Efficient LiDAR-based object segmentation and mapping for maritime environments. *IEEE J. Ocean. Eng.* **2019**, *44*, 352–362. [CrossRef]
29. Papadopoulos, G.; Kurniawati, H.; Shariff, A.S.B.M.; Wong, L.J.; Patrikalakis, N.M. 3D-surface reconstruction for partially submerged marine structures using an autonomous surface vehicle. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and SystemsE, San Francisco, CA, USA, 25–30 September 2011; pp. 3551–3557.
30. Nita, C.; Vandewal, M. CNN-based object detection and segmentation for maritime domain awareness. Artificial Intelligence and Machine Learning in Defense Applications II. *Int. Soc. Opt. Photonics* **2020**, *11543*, 1154306.
31. Liu, Y.; Zhang, M.h.; Xu, P.; Guo, Z.W. SAR ship detection using sea-land segmentation-based convolutional neural network. In Proceedings of the 2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP), Shanghai, China, 18–21 May 2017; pp. 1–4.
32. Bousetouane, F.; Morris, B. Fast CNN surveillance pipeline for fine-grained vessel classification and detection in maritime scenarios. In Proceedings of the 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Colorado Springs, CO, USA, 23–26 August 2016; pp. 242–248.
33. Cruz, G.; Bernardino, A. Aerial detection in maritime scenarios using convolutional neural networks. In Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems, Lecce, Italy, 24–27 October 2016; pp. 373–384.
34. Bloisi, D.D.; Previtali, F.; Pennisi, A.; Nardi, D.; Fiorini, M. Enhancing automatic maritime surveillance systems with visual information. *IEEE Trans. Intell. Transp. Syst.* **2016**, *18*, 824–833. [CrossRef]
35. Loomans, M.J.; de With, P.H.; Wijnhoven, R.G. Robust automatic ship tracking in harbours using active cameras. In Proceedings of the 2013 IEEE International Conference on Image Processing, Melbourne, VIC, Australia, 15–18 September 2013; pp. 4117–4121.
36. Bovcon, B.; Kristan, M. WaSR—A Water Segmentation and Refinement Maritime Obstacle Detection Network. *IEEE Trans. Cybern.* **2021**, 1–14. [CrossRef]
37. Chen, X.; Liu, Y.; Achuthan, K. WODIS: Water Obstacle Detection Network Based on Image Segmentation for Autonomous Surface Vehicles in Maritime Environments. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–13. [CrossRef]
38. Elkhateeb, E.; Soliman, H.; Atwan, A.; Elmogy, M.; Kwak, K.S.; Mekky, N. A Novel Coarse-to-Fine Sea-Land Segmentation Technique Based on Superpixel Fuzzy C-Means Clustering and Modified Chan-Vese Model. *IEEE Access* **2021**, *9*, 53902–53919. [CrossRef]
39. Pappas, O.A.; Achim, A.M.; Bull, D.R. Superpixel-guided CFAR detection of ships at sea in SAR imagery. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 1647–1651.
40. Pappas, O.; Achim, A.; Bull, D. Superpixel-level CFAR detectors for ship detection in SAR imagery. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1397–1401. [CrossRef]
41. Chen, Y.; Ming, D.; Lv, X. Superpixel based land cover classification of VHR satellite image combining multi-scale CNN and scale parameter estimation. *Earth Sci. Inform.* **2019**, *12*, 341–363. [CrossRef]
42. Schanda, J. *Colorimetry: Understanding the CIE System*; John Wiley & Sons: Hoboken, NJ, USA, 2007.
43. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.

44. Li, H.; Xiong, P.; Fan, H.; Sun, J. Dfanet: Deep feature aggregation for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9522–9531.
45. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
47. Bovcon, B.; Muhovič, J.; Perš, J.; Kristan, M. The mastr1325 dataset for training deep usv obstacle detection models. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 3431–3438.
48. Choromanska, A.; Henaff, M.; Mathieu, M.; Arous, G.B.; LeCun, Y. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*; 2015; pp. 192–204. Available online: https://proceedings.mlr.press/v38/choromanska15.html (accessed on 2 June 2021).
49. Bovcon, B.; Perš, J.; Kristan, M. Stereo obstacle detection for unmanned surface vehicles by IMU-assisted semantic segmentation. *Robot. Auton. Syst.* **2018**, *104*, 1–13. [CrossRef]
50. Liu, J.; Li, H.; Luo, J.; Xie, S.; Sun, Y. Efficient obstacle detection based on prior estimation network and spatially constrained mixture model for unmanned surface vehicles. *J. Field Robot.* **2021**, *38*, 212–228. [CrossRef]
51. Thoma, M. A survey of semantic segmentation. *arXiv* **2016**, arXiv:1602.06541.
52. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
53. Prasad, D.K.; Rajan, D.; Rachmawati, L.; Rajabally, E.; Quek, C. Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 1993–2016. [CrossRef]
54. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.