Review

Marine Vision-Based Situational Awareness Using Discriminative Deep Learning: A Survey

Dalei Qiao ^{1,2} , Guangzhong Liu ^{1,*} , Taizhi Lv ² , Wei Li ³ and Juan Zhang ²¹ College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China; qdly0406@163.com² College of Information Engineering, Jiangsu Maritime Institute, Nanjing 211170, China; lvtai zhi@163.com (T.L.); 20060529@jmi.edu.cn (J.Z.)³ Nanjing Marine Radar Institute, Nanjing 211153, China; lesli e28@163.com

* Correspondence: gzhliu@shmtu.edu.cn; Tel.: +86-021-38282803

Abstract: The primary task of marine surveillance is to construct a perfect marine situational awareness (MSA) system that serves to safeguard national maritime rights and interests and to maintain blue homeland security. Progress in maritime wireless communication, developments in artificial intelligence, and automation of marine turbines together imply that intelligent shipping is inevitable in future global shipping. Computer vision-based situational awareness provides visual semantic information to human beings that approximates eyesight, which makes it likely to be widely used in the field of intelligent marine transportation. We describe how we combined the visual perception tasks required for marine surveillance with those required for intelligent ship navigation to form a marine computer vision-based situational awareness complex and investigated the key technologies they have in common. Deep learning was a prerequisite activity. We summarize the progress made in four aspects of current research: full scene parsing of an image, target vessel re-identification, target vessel tracking, and multimodal data fusion with data from visual sensors. The paper gives a summary of research to date to provide background for this work and presents brief analyses of existing problems, outlines some state-of-the-art approaches, reviews available mainstream datasets, and indicates the likely direction of future research and development. As far as we know, this paper is the first review of research into the use of deep learning in situational awareness of the ocean surface. It provides a firm foundation for further investigation by researchers in related fields.

Keywords: marine situational awareness; deep learning; autonomous surface vessel; ship detection; ship tracking; vessel re-identification; multimodality fusion

Citation: Qiao, D.; Liu, G.; Lv, T.; Li, W.; Zhang, J. Marine Vision-Based Situational Awareness Using Discriminative Deep Learning: A Survey. *J. Mar. Sci. Eng.* **2021**, *9*, 397. <https://doi.org/10.3390/jmse9040397>

Academic Editor: Fablet Ronan

Received: 11 March 2021

Accepted: 6 April 2021

Published: 8 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 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

Nearly 70% of the earth's surface is covered by oceans, inland rivers, and lakes. Increasingly fierce and complicated maritime rights disputes between countries have made integrated maritime surveillance, rights protection, and law enforcement more difficult and arduous and have necessitated improvements in omnidirectional observation and situational awareness in specific areas of the ocean. Massive marine surveillance videos require real-time monitoring by crew members or traffic management staff, which consumes a great deal of manpower and material resources, and there are also missed inspections due to factors such as human fatigue. Automatic detection, re-identification, and tracking of ships in surveillance videos will transform maritime safety.

Society is slowly accommodating the shift to fully automated vehicles, such as airplanes, road vehicles, and ships at sea. The rapid development of information and communication technologies such as the internet of things, big data, and artificial intelligence (AI) has led to a demand for increased automation in ships of different types and tonnages. Marine navigation is becoming more complex. Intelligent computer vision will improve automated support for navigation. It is necessary to combine intelligent control of shipping

with intelligent navigation systems and to strengthen ship–shore coordination to make worthwhile progress in developing autonomous ship navigation technology.

A moving ship must detect and track the ships around it. This task can be regarded as the problem of tracking a moving target using a motion detecting camera. Collision avoidance on the water may require maneuvers such as acceleration and turning, possibly urgently. Onboard visual sensors may be subject to vibrations, swaying, changes in light, and occlusion. The detection and tracking of dim or small targets on the water is very challenging. According to the latest report of Allianz global corporate & specialty (AGCS) [1], 75–96% of collisions at sea are related to crew errors. There is therefore an urgent need to use advanced artificial intelligence technology for ocean scene parsing and auxiliary decision-making on both crewed and autonomous vessels, with a longer term goal of gradually replacing the crew in plotting course and controlling the vessel [2]. If human errors can be reduced, the probability of marine collisions will be greatly reduced.

In general, the rapid development of intelligent shipping requires improved computerized visual perception of the ocean surface. However, current visual perception, which uses deep learning, is inadequate for control of an autonomous surface vessel (ASV) and associated marine surveillance. There are many key issues to be resolved [3], such as:

- (1) The visual sensor on a vessel monitors a wide area. The target vessel is usually far away from the monitor and thus accounts for only a small proportion of the entire image. It is considered to be a dim and small target at sea. On a relative scale, the product of target width and height is less than one-tenth of the entire image; on an absolute scale, the target size is less than 32×32 pixels [4].
- (2) Poor light, undulating waves, interference due to sea surface refraction, and wake caused by ship motion, among other factors, cause large changes in the image background.
- (3) Given the irregular jitter as well as the sway and the heave of the hull, the onboard surveillance video inevitably shows high frequency jitter and low frequency field-of-view (FoV) shifts.
- (4) Image definition and contrast in poor visibility due to events such as sea-surface moisture and dense fog are inadequate for effective target detection and feature extraction.

Deep learning (DL) is a branch of artificial intelligence that was pioneered by Hinton [5]. It creates a deep topology to learn discriminative representation and nonlinear mapping from a large amount of data. Deep learning models fall into two categories: supervised discriminative models and unsupervised generative models. We used the discriminative model in this study. It is widely used in visual recognition in maritime situations and learns the posterior probability distribution of visible data. Research into visual perception in the marine environment provides data that can be mined for purposes of ship course planning and intelligent collision avoidance decision-making. It has significant practical value both for improving monitoring of the marine environment from onshore and for improving automated ASV navigation.

2. Research Progress of Vision-Based Situational Awareness

Research into automated visual interpretation of the marine environment lags research into aerial visualization (e.g., in drones) and terrestrial visualization (e.g., autonomous vehicles). Shore-based surveillance radar, onboard navigation radar, and automated identification systems (AIS) display target vessels only as spots of light on an electronic screen and, therefore, there are no visual cues about the vessel. Ships are generally large, thus treating them as point objects is inadequate for determining courses or avoiding collision, as has been confirmed by the occurrence of many marine accidents in recent years. The experimental psychologist Treichler showed through his famous learning and perception experiment that the perception information obtained by humans through vision accounts for 83% of all sensing modalities (vision, hearing, touch, etc.) [6]. Monocular vision sensor has a number of advantages (easy installation, small size, low energy consumption, and strong real-time performance) and provides reliable data for use in integrated surveillance systems at sea or ASV navigation systems.

Typical discriminative deep learning models include convolutional neural networks, time series neural networks, and attention mechanisms. They transform the search for an objective function into the problem of optimization of the loss function. In this review, we define the objects of marine situational awareness (MSA) as targets and obstacles in the local environment of the ASV or in the waterways monitored by the vessel traffic service (VTS). To further facilitate the future research in MSA field, we used discriminative deep learning as the main approach to vision-based situational awareness in maritime surveillance and autonomous navigation. On the basis of 85 published papers, we reviewed current research in four critical areas: full scene parsing of the sea surface, vessel re-identification, ship tracking, and multimodal fusion of perception data from visual sensors. The overall structure of marine situational awareness technology used in this paper is shown in Figure 1. For simplicity but without loss of generality, here, we do not consider the input of multimodal information.

Figure 1. The overall structure of marine situational awareness (MSA).

The relationship between full scene parsing, ship recognition and re-identification, and target tracking is as follows. The detection and the classification module detect the target vessel and strip it from the full scene. This module is the core of the three situational awareness tasks; it is linked to the two tasks of re-identification and target tracking. Semantic segmentation separates foreground objects (ships, buoys, and other obstacles) and background (sea, sky, mountains, islands, land) at the pixel level to eliminate distractions from target detection. Instance segmentation complements target detection and classification. Target detection provides prior knowledge for re-identification (i.e., direct re-identification of the detected vessel) in the recognition and re-identification module and provides features of the vessel's appearance to the tracking-by-detection pipeline. The combination of the three processes results in the upstream tracking-by-detection task continuously outputting the states and the positions of surrounding vessels and other dynamic or static obstacles to navigation. This information is the basis for the final prediction of target ship behavior and ASV collision avoidance and navigational decision-making.

2.1. Ship Target Detection

2.1.1. Target Detection

It is likely that the automatic identification system (AIS) is not installed or deactivated on small targets, such as marine buoys, sand carriers, wooden fishing boats, and pirate vessels. Because these objects are small, their radar reflectivity is slight, and their outline may be difficult to detect by a single sensor due to surrounding noise. Low signal strength, low signal-to-noise ratios, irregular and complex backgrounds, and undulating waves increase the difficulty of detecting and tracking dim and small targets. All these factors can lead to situations requiring urgent action to avoid collision. The early detection of dim or small targets on the sea surface at a distance provides a margin of time for course correction and collision avoidance calculations in the ASV.

Conventional image detection commonly uses three-stage visual saliency detection technology consisting of preprocessing, feature engineering, and classification. Although these techniques detect objects and track them accurately, they are slow in operation and do not perform well in changing environments. Their use in a complex and changeable marine environment produces information lag, thus endangering the ASV and the surrounding vessels. Rapid detection and object tracking technology based on computer vision can increase the accuracy of perception of the surrounding environment and improve understanding of the navigational situation of an ASV and thus improve autonomous navigation. The required technology has been advanced by the incorporation of deep learning.

Deep learning in target detection uses a convolutional neural network (CNN) to replace the conventional sliding window and hand-crafted features. The detection is based on deep learning from the traditional dual-stage methods (faster R-CNN [7–9], R-FCN [7], cascade R-CNN [8], mask R-CNN [9]), to single-stage methods (SSD [7,10], RetinaNet [8], YOLO [11], EfficientDet [12]), and then to the state-of-the-art (SOTA) anchor-free methods (CenterNet [12], FCOS [13]), which are widely used as marine ship detectors. In general, the speed of the single-stage method is obviously faster than that of the dual-stage method, but the accuracy is close to or even better than that of the latter. Therefore, the single-stage method, especially the anchor-free model, is becoming more and more popular. For dim and small targets on the sea, STDNet built a regional context network to pay attention to its region of interest and the corresponding context and realized the detection of ships less than 16×16 pixel area on 720p videos [14]. DSMV [8] introduced a bi-directional gaussian mixture model on the input multi-frame images with temporal relationship and combined it with the deep detector for ship detection. In addition, combining the robust feature extraction capabilities of CNN with other methods, such as saliency detection [15], background subtraction [16], and evidence theory [17] has been proven to be effective in improving the accuracy of detection and classification. Table 1 shows the comparison results of the SOTA ship target detection algorithms.

Table 1. Comparison of the state-of-the-art (SOTA) ship target detection algorithms.

Method	Anchor-Free or Not	Backbone	Dataset	Performance
Mask R-CNN [9]	no	ResNet-101	SMD	F-score: 0.875
Mask R-CNN [12]	no	ResNet-101	MarDCT	AP(IoU = 0.5): 0.964
Mask R-CNN [12]	no	ResNet-101	IPATCH	AP(IoU = 0.5): 0.925
RetinaNet [13]	no	ResNet-50	MSCOCO + SMD	AP(IoU = 0.3): 0.880
Faster R-CNN [10]	no	ResNet-101	Seaships	AP(IoU = 0.5): 0.924
SSD 512 [10]	no	VGG16	Seaships	AP(IoU = 0.5): 0.867
CenterNet [12]	yes	Hourglass	SMD + Seaships	AP(IoU = 0.5): 0.893
EfficientDet [12]	no	EfficientDet-D3	SMD + Seaships	AP(IoU = 0.5): 0.981
FCOS [13]	yes	ResNeXt-101	Seaships + private	AP: 0.845
Cascade R-CNN [13]	no	ResNet-101	Seaships + private	AP: 0.846
YOLOv3 [11]	no	Darknet-53	private	AP(IoU = 0.5): 0.960 (specific ship)

SMD: Singapore maritime dataset; CNN: convolutional neural network; AP: average precision; IoU: intersection over union.

When a vessel is under way, distant targets (other vessels or obstacles) first appear near the horizon or the shoreline [18], thus pre-extracting the horizon or the shoreline significantly reduces the potential area for target detection. In recent years, a method of ship detection and tracking for vessels appearing on the horizon was proposed that uses images

collected by onboard or buoy-mounted cameras [19], as shown in Figure 2. Jeong et al. identified the horizon by defining it as the region of interest (ROI) [20]. Zhang et al. used a discrete cosine transform for horizon detection to dynamically separate out the background and performed target segmentation on the foreground [21]. Sun et al. developed an onboard method of coarse–fine stitching to detect the horizon [22]. Steccanella et al. used CNN to divide the surface viewed pixel by pixel into water area and non-water area [23] in order to generate the horizon and the shoreline by curve fitting. Shan et al. investigated the orientation of the visual sensor as means to extract the horizon and detect surrounding ships [24]. Su et al. developed a gradient-based iterative method of detecting the horizon using an onboard panoramic catadioptric camera [25]. It is impractical to install X-band navigation radars, binocular cameras, and other ranging equipment on a small ASV due to constraints of size and cost. Gladstone et al. use planetary geometry and the optical characteristics of a monocular camera to estimate the distance of the target vessel on the sea [26], using the detected horizon as the reference; the average error they obtained after many experiments was 7.1%.

Figure 2. Block diagram for ship detection and classification.

2.1.2. Image Segmentation

The process of image segmentation is to divide the image into several specific regions with unique properties and extract regions of interest. Image segmentation of sea scenes is mostly performed as semantic segmentation in a process which does not distinguish specific targets and performs pixel-level classification. Bovcon et al. performed semantic segmentation and used an inertial measurement unit (IMU) to detect three-dimensional obstacles on the water surface [27] and created a semantic segmentation dataset containing 1325 manually labeled images for use by ASVs [28]. They conducted experiments to compare the SOTA algorithms such as U-Net [29], pyramid scene parsing network (PSPNet) [30], and Deeplabv2 [31]. Bovcon et al. designed a horizon/shoreline segmentation and obstacle detection process using an encoder–decoder model to process onboard IMU data [32]; this approach significantly improved the detection of dim or small targets. Cane et al. [33] compared the performance of conventional semantic segmentation models such as SegNet, Enet, and ESPNet using four public maritime datasets (MODD [27], Singapore maritime dataset (SMD) [34], IPATCH [35], SEAGULL [36]). Zhang et al. designed a two-stage semantic segmentation method for use in a marine environment [37] that initially identified interference factors (e.g., sea fog, ship wakes, waves on the sea) and then extracted the semantic mask of the target ship. Jeong et al. used the PSPNet to parse the sea-surface [38] and then extract the horizon using a straight line fitting algorithm. Experiments showed that this method was more accurate and more robust than the conventional method. Qiu et al. designed a real-time semantic segmentation model for ASVs using the U-Net framework [39] in which the sea scene was divided into five categories: sky, sea, human, target ship, and other obstacles. Kim et al. [40] developed a lightweight skipZ_ENet model on the Jetson TX2, an embedded AI platform, to semantically segment sea obstacles in real-time. The sea scene was divided into five categories: sky, sea surface, target ships, island or dock, and small obstacles. For the possible interrelation and interaction between the three tasks of detection, classification, and segmentation, it can be further assumed that the features required by the above three tasks have a shared feature subspace, and a new trend will unite them for multi-task learning [41].

Instance segmentation and panoptic segmentation [42], building on recent progress in target detection and semantic segmentation, have been used to make pixel level distinctions

between objects with different tracks and objects that behave differently from each other. Panoptic segmentation is a synthesis of semantic segmentation and instance segmentation. Through the mutual promotion of the two tasks, it can provide a unified understanding of the sea scene, which is worthy of further exploration. However, to our knowledge, there is no relevant literature in this field. Current research into instance segmentation of ships on the sea is concentrated mainly on remote sensing, for example [43,44], and only a few studies that use visible light can be found [45,46]. Examples of target detection, semantic segmentation, and instance segmentation are shown in Figure 3.

Figure 3. Target detection, semantic segmentation, and instance segmentation from the same image.

2.2. Ship Recognition and Re-Identification

Both onshore video surveillance systems and ASVs use more than one camera (aboard ship, on the bow, stern, port, and starboard of the vessel) to monitor the waters and the surface around the vessel. Re-identification (ReID) is the need to recognize and retrieve the same target object across different cameras and different scenes. A ship can be considered to be a rigid body, and a change in attitude from different viewpoints is much larger than that of pedestrians and vehicles. Changes in the luminosity of the sea surface and of the background as well as changes in perceived vessel size due to changes in distance from the vision sensor can make clusters of images of a single vessel appear to be of different vessels and clusters of images of different vessels appear to be images of the same vessel. Thus, there are significant difficulties in ensuring accurate identification of vessels and other objects on the sea surface.

Target re-identification is an area of image retrieval that mostly focuses on pedestrians and vehicles. Google Scholar searches using keywords such as “Ship + Re-identification”, “Ship + Re identification”, “Vessel + Re-identification”, and “Vessel + Re identification” show that there are few reports on related achievements in the ship field [47–50], two of which were published by authors of this paper. We define vessel re-identification (also referred to as ship-face recognition) in a manner analogous to the re-identification of pedestrians and vehicles and of the target ship images in a given probe set as the process of using computer vision to retrieve the ship in the gallery set of cross frame or cross camera images. Current ReID research focuses on resolving issues of intra-class differences and

inter-class similarity. However, ships are homogeneous, rigid bodies that are highly similar, and these properties increase the difficulty of re-identification.

Re-identification of cross-modal target ships for onshore closed circuit television (CCTV) or ASV monitoring can be performed on detected targets to compensate for the single perspective of a vessel captured by a single camera. Such re-identification can be introduced into the multitarget tracking pipeline for long term tracking. Figure 4 shows the four key steps of vessel re-identification [49]: ship detection, feature extraction, feature transformation, and feature similarity metric learning.

Figure 4. Flow chart of typical vessel re-identification.

Visual information of ships is different from the visual information of pedestrians, vehicles, and even rare animals in two significant areas when considering re-identification. Ships are much greater in size and differ in perspective due to changes in viewpoint. Thus, we cannot simply adopt or modify a target re-identification algorithm from another application to migrate it to identification of a target ship. Wang et al. focused on feature extraction in their review of deep network models for maritime target recognition, which provides a valuable reference for research in this field [51]. For feature extraction, introducing orthogonal moment methods, such as Zemike moment [52] and Hu moment [53], for feature extraction and fusing them with features extracted by CNN can significantly improve the accuracy of ship recognition. Resnet-50, VGG-16, and DenseNet-121 were used as feature extractors on the Boat Re-ID dataset built by Spagnolof et al. [54], and the results show that ResNet-50 achieved the best results among the three. By estimating the position and the orientation of the target-ship in the image, Ghahremani et al. re-identified it using a TriNet model [48]. On the common object dataset (MSCOCO) and two maritime-domain datasets (i.e., IPATCH and MarDCT), Heyse et al. explored a domain-adaption method for the fine-grained identification of ship categories [55]. Groot et al. re-identified target ships imaged from multiple cameras by matching vessel tracks and time filtering, among other methods [56]. They divided the known distance between fixed point cameras by the navigation time to estimate the target speed. The authors in [49] proposed a multi-view feature learning framework that combined global and discriminative local features. The framework combined the cross-entropy loss with the proposed orientation quintuple (O-Quin) loss. The inherent features of ship appearance were fully exploited, and the algorithm for learning multiview representation was refined and optimized for vessel re-identification. The discriminative features detection model and the viewpoint estimation model were embedded to create an integrated framework.

Introduction of ship-face recognition that is based on re-identification into mobile VTS can confirm the identity and the legal status of nonconforming vessels, as shown in Figure 5. The process typically followed is that the AIS is initially registered with the camera coordinate system; the pre-trained re-identification model is then used for feature extraction, and the re-identification result is finally compared for similarity with the AIS message. If the similarity is greater than some threshold value, the target vessel is judged to be illegal for tampering with AIS information.

Figure 5. Ship face recognition process using re-identification (ReID).

2.3. Ship Target Tracking

It is not sufficient in tracking a vessel to give discrete vessel position information in a complex marine environment, especially when the target is under way. The course and the speed of the target vessel are required in order to make real-time collision avoidance decisions. The environment may be cluttered, and visibility may be poor, thus it is necessary to correlate and extrapolate data to provide a stable track of the vessel's course. As shown in Figure 6, the tracking-by-detection framework usually consists of two stages. In the first stage, the detection hypothesis is given by the pre-trained target detector, and in the second stage, the detection hypothesis is correlated on the timeline to form a stable track.

Figure 6. Block diagram of typical tracking-by-detection pipeline.

A generative model tracking pipeline that consists of a particle filter [57], a Kalman filter [58], and a meanshift algorithm [59] ignores correlation with features having a sea background and other nontargets while ship tracking, which leads to low accuracy. The real-time performance of the model is also less than desirable due to the complexity of calculation. Researchers have introduced advanced correlation filtering and deep learning methods for ship tracking in sea surface surveillance videos in recent years. Correlation filtering [60] is renowned for its fast tracking speed. However, most of its success is in single target tracking due to defects in multitarget tracking and the fact that extracted features are greatly affected by illumination and attitude. Deep learning is more pervasive in tracking, and its tracking accuracy is significantly better due to the depth of feature extraction and more detailed representation. The advantages of accuracy in hand-crafted features make it unwise to discard correlation filtering or deep learning. Zhang et al. combined features extracted by deep learning with a histogram of oriented gradients (HOG), local binary patterns (LBP), scale-invariant feature transform (SIFT), and other hand-crafted features in tracking target ships [61].

In recent work on tracking marine targets, researchers combined appearance features with motion information in data fusion, a move that significantly improved tracking stability. Yang et al. introduced deep learning into the detection of vessels on the sea surface and embedded deep appearance features in the ensuing tracking process in combination with the standard Kalman filtering algorithm [62]. Leclerc et al. used transfer learning

to classify ships, which significantly improved the predictive ability of the tracker [63]. Qiao et al. proposed a multimodal, multicue, tracking-by-detection framework in which vessel motion information and appearance features were used concurrently for data association [50]. Re-identification was used for ship tracking, and the appearance features were used as one of the multicues for long term target vessel tracking. Shan et al. [24] adapted the Siamese region proposal network (SiamRPN) and introduced the feature pyramid network (FPN) to accommodate the characteristics of the marine environment in detecting and tracking dim or small targets on the sea surface (pixels accounted for less than 4%) and created a real-time tracking framework for ships, sea-SiamFPN. Although the accuracy has been significantly improved compared with correlation filtering methods, this framework also belongs to the category of single target tracking, and its application in actual marine scenes has certain limitations. It is worth noting that the features extracted by CNN in the detection stage can be directly reused in subsequent instance segmentation and tracking tasks to save computing resources. For example, Scholler directly used the features in the detection phase for appearance matching during tracking [64].

3. Multimodal Awareness with the Participation of Visual Sensors

3.1. Multimodal Sensors in the Marine Environment

Modality refers to different perspectives or different forms of perceiving specific objects or scenes. There is high correlation between multimodal data because they represent aspects of the same item, thus it is advantageous to implement multimodal information fusion (MMIF). This process uses multimodal perception data acquired at different times and in different spaces to create a single representation through alignment, conversion, and fusion and initiates multimodal sensor collaborative learning to maximize the accuracy and the reliability of the environmental awareness system.

The improved perception capability is constrained by the equipment used for perception and the techniques used to interpret the data recorded. Mobile VTS, which is a complex information management system, needs a variety of shipborne sensors that obtain data for the surrounding environment in real-time, learn from each other, and improve the robustness and the fault tolerance of the sensing system through redundancy. The performance of various sensors that operate in a marine environment is shown for comparison in Table 2.

Table 2. Comparison of visible light and other multimodal sensors for marine use.

Sensor	Frequency Range	Work Mode	Advantage	Disadvantage
RGB camera	visible light (380–780 nm)	passive	low cost rich appearance high spatial resolution	small coverage sensitive to light and weather lack of depth
infrared camera	far infrared (9–14 μ m)	passive	strong penetrating power not affected by light and weather	low imaging resolution weak contrast
navigation radar	X-band (8–12 GHz), S-band (2–4 GHz)	active	independent perception robust for bad weather long distance	low precision low data rate high electromagnetic radiation have blind area
millimeter wave radar	millimeter wave (30–300 GHz)	active	wide frequency band high angular resolution high data rate	short detection range have blind area
automated identification systems (AIS)	very high frequency (161.975 or 162.025 MHz)	passive	robust for bad weather long distance wide coverage	low data rate only identify ships need to report actively
Lidar	Laser (905 nm or 1550 nm)	active	3D perception high spatial resolution good electromagnetic resistance	sensitive to weather high cost

Individual sensors can be combined depending on the pros and the cons of the modal sensors shown in Table 1, and multiple sensors are combined to form an integrated sensing system to create a more reliable sea surface traffic monitoring system. For example, AIS and X-band radar can be used together in medium and long range surveillance (1–40 km), while millimeter-wave radar and RGB or infrared cameras provide accurate perception of an environment within 1 km of the sensor. The detection range of each sensor is shown in Figure 7.

Figure 7. Areal coverage of multimodal sensors.

3.2. Multimodal Information Fusion

Multisensor fusion is inevitable in marine environmental perception, similar to its development in drone operation and autonomous road vehicles. Video or images can be captured at low cost in all weather and in real-time and thus provide a large amount of reliable data. However, the use of a single sensor produces a high rate of false or missing detection due to poor light, mist or precipitation, and ship sway. This can be corrected if data from more than one source for the same target is fused to produce a single, more accurate, and more reliable data record. There is an urgent need to address this issue, that is, to gain the benefits of using multimodal sensors to reduce the probability of errors in detection and tracking of surrounding target vessels and thus improve perception of the surrounding environment in an autonomous navigation or marine surveillance system. Chen et al. merged onshore surveillance video and onboard AIS data for a target vessel. They adjusted the video camera attitude and focal length according to the Kalman filtered AIS position by linking AIS and the camera to co-track target vessels [65]. Thompson linked lidar and camera to detect, track, and classify marine targets [66]. Helgesen et al. linked visible light camera, infrared camera, radar, and lidar for target detection and tracking on the sea surface [67] and used joint integrated data association (JIPDA) to fuse data in the association stage. Haghighbayan et al. used probabilistic data association filtering (PDAF) to fuse multisensor data from the same sources as Hegelsen et al. [68]. A visible light camera provides the most accurate data for bounding box extraction, and Haghighbayan et al. used a CNN for subsequent target classification. Farahnakian et al. compared visible light

images with thermal infrared images (with coast, sky, and other vessels in the background) and found that, in day and night sea scenes, two deep learning fusion methods (DLF and DenseFuse) were clearly better than six conventional methods (VSM-WLS, PCA, CBF, JSR, JSRDS, and ConvSR) [69]. In the 2016 Maritime robotX challenge, Stanislas et al. fused data from camera, lidar, and millimeter-wave radar [70] and developed a method to concurrently create an obstacle map and a feature map for target recognition. Farahnakian et al. fused visible and infrared images at pixel level, feature level, and decision level to overcome the problems associated with operating in a harsh marine environment [71]. Comparative experiments have shown that the feature level fusion produces the best results. Farahnakian et al. used a selective search to create a large number of candidate regions on the RGB image [72] and then used other modal sensor data to refine the selection to detect targets on the sea surface. In addition, for underwater vehicle and its interaction with surface vehicle [73], the fusion of acoustic sonar or even geomagnetic sensor and visible light camera effectively expands the space for three-dimensional awareness of the marine environment [74,75].

Visible light images or videos are the basic data for navigation in congested ports and short range reactive collision avoidance. Data from other modal sensors, such as X-band radar or thermal infrared imaging, and AIS information can be fused with visible light images. Bloisi et al. fused visual images and VTS data [76]. They used the remote surveillance cameras as the primary sensor in locations where radar and AIS do not operate (e.g., radar will set a specific sector as silent if it includes a residential area). Figure 8 shows the process for detection and tracking with fusion of visible images and millimeter-wave radar [77]. Low level and high level bimodal fusion occur concurrently. A deep CNN can be used for feature compression to ensure that features that meet the decision boundary are maximally extracted. Please note that, if the visible light camera in Figure 8 is replaced with a thermal one, the millimeter wave radar is replaced with lidar, AIS, or navigation radar, or even if the above two sets of sensors are combined, the tracking pipeline will still work well.

Figure 8. Tracking process including multimodal data fusion.

4. Visual Perception Dataset on Water Surface

4.1. Visual Dataset on the Sea

Although there is a rich available selection of video or image datasets of pedestrians and vehicles, there are few available ship image datasets, especially full scene annotated datasets. Lack of such data severely restricts in-depth research.

Table 3 summarizes available datasets of ship targets that have been published to date (2 March 2021). The most representative datasets are SMD, SeaShips, and VesselID-539. The Singapore maritime dataset (SMD) was assembled and annotated by Prasad et al. [34]. SMD contains 51 annotated high definition video clips, consisting of 40 onshore videos and 11 onboard videos. Gundogdu et al. built a dataset containing 1,607,190 annotated ship images obtained from a ship spotting website that were binned into 197 categories. Annotations include the location and the type of ship, such as general cargo, container, bulk carrier, or passenger ship [78]. Shao et al. created the large ship detection dataset SeaShips [10]. They collected data from 156 surveillance cameras in the coastal waters of Zhuhai, China. They took account of factors such as background discrimination, lighting, visible scale, and occlusion to extract 31,455 images from 10,080 videos. They identified and annotated in VOC format six ship types: ore carriers, bulk carriers, general cargo ships, container ships, fishing boats, and passenger ships. Unfortunately, only a very small number of video clips have been published. Bovcon et al. created a semantic segmentation dataset consisting of 1325 manually annotated images for ASV research [28]. VesselID-539, the ship re-identification dataset created by Qiao et al., contains images that represent the four most common challenging scenarios; examples are given in Figure 9. It contains 539 vessel IDs, 149,363 images, and marks 447,926 boundary boxes of vessels or partial vessels from different viewpoints.

Table 3. Available image datasets of ship targets.

Datasets	Shooting Angle	Usage Scenarios	Resolution (Pixels)	Scale	Open Access
MARVEL [78]	onshore, onboard	classification	512 × 512, etc.	>140k images	yes
MODD [27]	onboard	detection and segmentation	640 × 480	12 videos, 4454 frames	yes
SMD [34]	onshore, onboard	detection and tracking	1920 × 1080	36 videos, >17k frames	yes
IPATCH [35]	onboard	detection and tracking	1920 × 1080	113 videos	no
SEAGULL [36]	drone-borne	detection, tracking, and pollution detection	1920 × 1080	19 videos, >150k frames	yes
MarDCT [79]	onshore, overlook	detection, classification, and tracking	704 × 576, etc.	28 videos	yes
SeaShips [10]	onshore	detection	1920 × 1080	>31k frames	partial
MaSTr1325 [28]	onboard	detection and segmentation	512 × 384, 1278 × 958	1325 frames	yes
VesselReID [47]	onboard	re-identification	1920 × 1080	4616 frames, 733 ships	no
VesselID-539 [49]	onshore, onboard	re-identification	1920 × 1080	>149k frames, 539 ships	no
Boat ReID [54]	overlook	detection and re-identification	1920 × 1080	5523 frames, 107 ships	yes
VAIS [80]	onshore	multimodal fusion	vary in size	1623 visible light, 1242 infrared images	yes

Figure 9. Four challenging scenarios from the VesselID-539 dataset.

4.2. Visual Datasets for Inland Rivers

An inland river image typically contains target ships and other obstacles in the foreground against a background of sky, water surface, mountains, bridges, trees, and onshore buildings. This is a more complex image than would be produced on open water. An inland river dataset and research based on it can provide valuable reference for a marine scenario. Teng et al. collected ship videos from maritime CCTV systems in Zhejiang, Guangdong, Hubei, Anhui, and other places in China to create a dataset for visual tracking of inland ships [81]. The researchers considered many factors such as time, region, climate, and weather and carefully selected 400 representative videos. Liu et al. adapted this dataset by removing video frames with no ship or no obvious changes in target ship scale and background to create a visual dataset for inland river vessel detection [82]. The videos in this dataset were cropped to 320×240 pixels a total of 100 segments; each video clip has a duration of 20–90 s and contains 400–2000 frames.

The training of a deep learning model requires a large amount of data. The requirements for diversity in weather conditions and a variety of target scales, locations, and viewpoints makes the acquisition of real world data both extremely difficult and expensive. The introduction of synthetic data has become popular as a method of augmenting a training dataset. Shin et al. used copy-and-paste in Mask R-CNN to superimpose a foreground target ship on different backgrounds [83] to increase the size of the dataset used for target detection. Approximately 12,000 additional synthetic images (only one-tenth of the total) were injected into the training set, which increased the mean average precision (mAP, IoU = 0.3) index by 2.4%. Chen et al. used a generative adversarial network for data augmentation to compensate for the limited availability of images of small target vessels (such as fishing boats or bamboo rafts) on the sea and used YOLOv2 for small boat detection [84]. Milicevic et al. obtained good results when they used conventional methods, such as horizontal flipping, cut/copy-and-paste, and changing RGB channel values, to increase the data for use in fine-grained ship classification [85]. The classification accuracy

of the additional 25,000 synthetic images has been increased by 6.5%. Although the gain of this index is lower than the 10% increase in the direct use of 25,000 original images, it is of practical significance to alleviate the shortage of datasets for MSA.

5. Conclusions and Future Work

Visual sensors are commonly used in addition to navigational radar and AIS to directly detect or identify obstacles or target vessels on the water surface. We summarized some innovations in vision-based MSA using deep learning and especially presented a comprehensive summary and analysis of the current status of research into full scene image parsing, vessel re-identification, vessel tracking, and multimodal fusion for perception of sea scenes. There have been theoretical and technological breakthroughs in these areas of research. Visual perception of the marine environment is a new and therefore immature area of cross-disciplinary research; future work in the field needs to be more directed and more sophisticated, focusing on the solution of two key issues—the stability of perception and the ability to perceive dim and small targets summarized in this paper. The following are areas that will reward future study.

- (1) Multi-task fusion. Semantic segmentation and instance segmentation must be combined architecturally and at the task level. Prediction of the semantic label and instance ID of each pixel in an image is required for full coverage of all objects in a marine scene to correctly divide the scene into uncountable stuff (sea surface, sky, island) and countable things (target vessels, buoys, other obstacles). The combination of object segmentation in fine-grained video with object tracking requires the introduction of semantic mask information into the tracking process. This will require detailed investigation of scene segmentation, combining target detection and tracking using existing tracking-by-detection and joint detection and tracking paradigms, but that is necessary to improve perception of target motion at a fine-grained level.
- (2) Multi-modal fusion. Data provided by the monocular vision technology summarized in this paper can be fused with other modal data, such as the point cloud data provided by a binocular camera or a lidar sensor, to provide more comprehensive data for the detection and the tracking of vessels and objects in three-dimensional space. This will significantly increase the perception capability of ASV navigational systems. The pre-fusion of multimodal data in full-scene marine parsing, vessel re-identification, and tracking (i.e., fusing data before other activities occur) will allow us to make full use of the richer original data from each sensor, such as edge and texture features of images from monocular vision sensors or echo amplitude, angular direction, target size, and shape of X-band navigation radar. We will thus be able to incorporate data from radar sensors into image features and then fuse them with visual data at the lowest level to increase the accuracy of MSA.
- (3) The fusion of DL-based awareness algorithms and traditional ones. Traditional awareness algorithms have the advantages of strong interpretability and high reliability. Combining them with the deep learning paradigm summarized in this paper, a unified perception framework with the advantages of both paradigms will be obtained, which can further enhance the ability of marine situation awareness.
- (4) Domain adaptation learning. Draw lessons from the development experience of unmanned vehicles and UAVs to build a ship–shore collaborative situational awareness system. In particular, it is necessary to start with the improvement of self-learning ability and introduce reinforcement learning to design a model that can acquire new knowledge from sea scenes.

Author Contributions: Investigation, D.Q.; resources, W.L.; writing—original draft preparation, D.Q.; writing—review and editing, D.Q.; supervision, G.L.; project administration, T.L.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded in part by the China postdoctoral science foundation, grant number 2019M651844; in part by the natural science foundation of the Jiangsu higher education institutions of China, grant number 20KJA520009; in part by the Qinglan project of Jiangsu Province; in part by the project of the Qianfan team, innovation fund, and the collaborative innovation center of shipping big data application of Jiangsu Maritime Institute, grant number KJCX1809; and in part by the computer basic education teaching research project of association of fundamental computing education in Chinese universities, grant number 2018-AFCEC-266.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: SMD: <https://sites.google.com/site/dilipprasad/home/singapore-maritime-dataset> (accessed on 1 February 2020); MarDCT: <http://www.diag.uniroma1.it/~labrococo/MAR/> (accessed on 1 March 2020); MODD & MaSTr1325: <https://www.vicos.si/Downloads/MODD> (accessed on 1 March 2020); SEAGULL: <http://vislab.isr.ist.utl.pt/seagull-dataset/> (accessed on 1 March 2020); SeaShips: [http://www.lmars.whu.edu.cn/prof_web/shaozhenfeng/datasets/SeaShips\(7000\).zip](http://www.lmars.whu.edu.cn/prof_web/shaozhenfeng/datasets/SeaShips(7000).zip) (accessed on 1 June 2020); VAIS: <http://vcip-okstate.org/pbvs/bench/Data/12/VAIS.zip> (accessed on 1 July 2020); MARVEL: <http://www.shipspotting.com/gallery/> (accessed on 1 July 2020).

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

References

- Fields, C. *Safety and Shipping 1912–2012: From Titanic to Costa Concordia*; Allianz Global Corporate and Speciality AG: Munich, Germany, 2012; Available online: <http://www.agcs.allianz.com/content/dam/onemarketing/agcs/agcs/reports/AGCS-Safety-Shipping-Review-2012.pdf> (accessed on 20 March 2021).
- Brcko, T.; Androjna, A.; Srše, J.; Boć, R. Vessel multi-parametric collision avoidance decision model: Fuzzy approach. *J. Mar. Sci. Eng.* **2021**, *9*, 49. [\[CrossRef\]](#)
- Prasad, D.K.; Prasath, C.K.; Rajan, D.; Rachmawati, L.; Rajabaly, E.; Quek, C. Challenges in video based object detection in maritime scenario using computer vision. *arXiv* **2016**, arXiv:1608.01079.
- Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296.
- Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [\[CrossRef\]](#) [\[PubMed\]](#)
- Trichler, D.G. Are you missing the boat in training aids. *Film AV Comm.* **1967**, *1*, 14–16.
- Soloviev, V.; Farahnakian, F.; Zelioli, L.; Iancu, B.; Lilius, J.; Heikkonen, J. Comparing CNN-Based Object Detectors on Two Novel Maritime Datasets. In Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (ICMEW), London, UK, 6–10 July 2020; pp. 1–6.
- Marques, T.P.; Albu, A.B.; O'Hara, P.D.; Sogas, N.S.; Ben, M.; Mcwhinnie, L.H.; Canessa, R. Size-invariant Detection of Marine Vessels From Visual Time Series. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 5–9 January 2021; pp. 443–453.
- Moosbauer, S.; Konig, D.; Jakel, J.; Teutsch, M. A benchmark for deep learning based object detection in maritime environments. In Proceedings of the IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 916–925.
- Shao, Z.; Wu, W.; Wang, Z.; Du, W.; Li, C. SeaShips: A Large-Scale Precisely Annotated Dataset for Ship Detection. *IEEE Trans. Multimed.* **2018**, *20*, 2593–2604. [\[CrossRef\]](#)
- Betti, A.; Michelozzi, B.; Bracci, A.; Masini, A. Real-Time target detection in maritime scenarios based on YOLOv3 model. *arXiv* **2020**, arXiv:2003.00800.
- Nalamati, M.; Sharma, N.; Saqib, M.; Blumenstein, M. Automated Monitoring in Maritime Video Surveillance System. In Proceedings of the 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), Wellington, New Zealand, 25–27 November 2020; pp. 1–6.
- Spraul, R.; Sommer, L.; Arne, S. A comprehensive analysis of modern object detection methods for maritime vessel detection. Artificial Intelligence and Machine Learning in Defense Applications II. *Int. Soc. Opt. Photonics* **2020**, 1154305. [\[CrossRef\]](#)
- Bosquet, B.; Mucientes, M.; Brea, V.M. STDnet: Exploiting high resolution feature maps for small object detection. *Eng. Appl. Artif. Intell.* **2020**, *91*, 1–16. [\[CrossRef\]](#)
- Shao, Z.; Wang, L.; Wang, Z.; Du, W.; Wu, W. Saliency-Aware Convolution Neural Network for Ship Detection in Surveillance Video. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 781–794. [\[CrossRef\]](#)
- Kim, Y.M.; Lee, J.; Yoon, I.; Han, T.; Kim, C. CCTV Object Detection with Background Subtraction and Convolutional Neural Network. *KIIE Trans. Comput. Pract.* **2018**, *24*, 151–156. [\[CrossRef\]](#)

17. Debaque, B.; Florea, M.C.; Duclos-Hindie, N.; Boury-Brisset, A.C. Evidential Reasoning for Ship Classification: Fusion of Deep Learning Classifiers. In Proceedings of the 22th International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2–5 July 2019; pp. 1–8.
18. Li, R.; Liu, W.; Yang, L.; Sun, S.; Hu, W.; Zhang, F.; Li, W. DeepUNet: A Deep Fully Convolutional Network for Pixel-Level Sea-Land Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3954–3962. [\[CrossRef\]](#)
19. Fefilat'yev, S.; Goldgof, D.; Shreve, M.; Lembke, C. Detection and tracking of ships in open sea with rapidly moving buoy-mounted camera system. *Ocean Eng.* **2012**, *54*, 1–12. [\[CrossRef\]](#)
20. Jeong, C.Y.; Yang, H.S.; Moon, K.D. Fast horizon detection in maritime images using region-of-interest. *Int. J. Distrib. Sens. Netw.* **2018**, *14*, 14. [\[CrossRef\]](#)
21. Zhang, Y.; Li, Q.Z.; Zang, F.N. Ship detection for visual maritime surveillance from non-stationary platforms. *Ocean Eng.* **2017**, *141*, 53–63. [\[CrossRef\]](#)
22. Sun, Y.; Fu, L. Coarse-fine-stitched: A robust maritime horizon line detection method for unmanned surface vehicle applications. *Sensors* **2018**, *18*, 2825. [\[CrossRef\]](#)
23. Steccanella, L.; Bloisi, D.D.; Castellini, A.; Farinelli, A. Waterline and obstacle detection in images from low-cost autonomous boats for environmental monitoring. *Rob. Auton. Syst.* **2020**, *124*, 103346. [\[CrossRef\]](#)
24. Shan, Y.; Zhou, X.; Liu, S.; Zhang, Y.; Huang, K. SiamFPN: A Deep Learning Method for Accurate and Real-Time Maritime Ship Tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 315–325. [\[CrossRef\]](#)
25. Su, L.; Sun, Y.-X.; Liu, Z.-L.; Meng, H. Research on Panoramic Sea-sky line Extraction Algorithm. *DEStech Trans. Eng. Technol. Res.* **2020**, 104–109. [\[CrossRef\]](#)
26. Gladstone, R.; Moshe, Y.; Barel, A.; Shenhav, E. Distance estimation for marine vehicles using a monocular video camera. In Proceedings of the 24th European Signal Processing Conference (EUSIPCO), Budapest, Hungary, 29 August–2 September 2016; pp. 2405–2409.
27. Bovcon, B.; Perš, J.; Kristan, M. Stereo obstacle detection for unmanned surface vehicles by IMU-assisted semantic segmentation. *Rob. Auton. Syst.* **2018**, *104*, 1–13. [\[CrossRef\]](#)
28. Bovcon, B.; Muhovic, J.; Pers, J.; Kristan, M. The MaSTr1325 dataset for training deep USV obstacle detection models. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 4–8 November 2019; pp. 3431–3438.
29. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical image computing and computer-assisted intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
30. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer vision and Pattern Recognition (CVPR), Honolulu, Hawaii, 21–26 July 2017; pp. 2881–2890.
31. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [\[CrossRef\]](#)
32. Bovcon, B.; Kristan, M. A water-obstacle separation and refinement network for unmanned surface vehicles. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–30 June 2020; pp. 9470–9476.
33. Cane, T.; Ferryman, J. Evaluating deep semantic segmentation networks for object detection in maritime surveillance. In Proceedings of the AVSS 2018—15th IEEE International Conference Advanced Video Signal-Based Surveillance, Auckland, New Zealand, 27–30 November 2018; pp. 1–6.
34. 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. [\[CrossRef\]](#)
35. Patino, L.; Nawaz, T.; Cane, T.; Ferryman, J. Pets 2016: Dataset and Challenge. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR), Las Vegas, NV, USA, 1–26 June 2016; pp. 1–8.
36. Ribeiro, R.; Cruz, G.; Matos, J.; Bernardino, A. A Data Set for Airborne Maritime Surveillance Environments. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 2720–2732. [\[CrossRef\]](#)
37. Zhang, W.; He, X.; Li, W.; Zhang, Z.; Luo, Y.; Su, L.; Wang, P. An integrated ship segmentation method based on discriminator and extractor. *Image Vis. Comput.* **2020**, *93*, 103824. [\[CrossRef\]](#)
38. Jeong, C.Y.; Yang, H.S.; Moon, K.D. Horizon detection in maritime images using scene parsing network. *Electron. Lett.* **2018**, *54*, 760–762. [\[CrossRef\]](#)
39. Qiu, Y.; Yang, Y.; Lin, Z.; Chen, P.; Luo, Y.; Huang, W. Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV. *China Commun.* **2020**, *17*, 46–57. [\[CrossRef\]](#)
40. Kim, H.; Koo, J.; Kim, D.; Park, B.; Jo, Y.; Myung, H.; Lee, D. Vision-Based Real-Time Obstacle Segmentation Algorithm for Autonomous Surface Vehicle. *IEEE Access* **2019**, *7*, 179420–179428. [\[CrossRef\]](#)
41. Liu, Z.; Waqas, M.; Jie, Y.; Rashid, A.; Han, Z. A Multi-task CNN for Maritime Target Detection. *IEEE Signal Process. Lett.* **2021**, *28*, 434–438. [\[CrossRef\]](#)
42. Kirillov, A.; He, K.; Girshick, R.; Rother, C.; Dollár, P. Panoptic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 9404–9413.
43. Huang, Z.; Sun, S.; Li, R. Fast Single-shot Ship Instance Segmentation Based on Polar Template Mask in Remote Sensing Images. *arXiv* **2020**, arXiv:2008.12447.

44. Nie, X.; Duan, M.; Ding, H.; Hu, B.; Wong, E.K. Attention Mask R-CNN for ship detection and segmentation from remote sensing images. *IEEE Access* **2020**, *8*, 9325–9334. [\[CrossRef\]](#)
45. van Ramshorst, A. Automatic Segmentation of Ships in Digital Images: A Deep Learning Approach. 2018. Available online: <https://repository.tudelft.nl/islandora/object/uuid:55de4322-8552-4a2c-84d0-427b2891015b> (accessed on 20 March 2021).
46. Nita, C.; Vandewal, M. CNN-based object detection and segmentation for maritime domain awareness[C]// Artificial Intelligence and Machine Learning in Defense Applications II. *Int. Soc. Opt. Photonics* **2020**, *11543*, 1154306.
47. Ghahremani, A.; Kong, Y.; Bondarev, E.; de With, P.H.N. Towards parameter-optimized vessel re-identification based on IORnet. In Proceedings of the International Conference on Computational Science, Faro, Portugal, 12–14 June 2019; pp. 125–136.
48. Ghahremani, A.; Kong, Y.; Bondarev, E.; De With, P.H.N. Re-identification of vessels with convolutional neural networks. In Proceedings of the 2019 5th International Conference on Computer and Technology Applications (ICCTA), Turkey, Istanbul, 16–17 April 2019; pp. 93–97.
49. Qiao, D.; Liu, G.; Dong, F.; Jiang, S.X.; Dai, L. Marine Vessel Re-Identification: A Large-Scale Dataset and Global-and-Local Fusion-Based Discriminative Feature Learning. *IEEE Access* **2020**, *8*, 27744–27756. [\[CrossRef\]](#)
50. Qiao, D.; Liu, G.; Zhang, J.; Zhang, Q.; Wu, G.; Dong, F. M³C: Multimodel-and-Multicue-Based Tracking by Detection of Surrounding Vessels in Maritime Environment for USV. *Electronics* **2019**, *8*, 723. [\[CrossRef\]](#)
51. Wang, N.; Wang, Y.; Er, M.J. Review on deep learning techniques for marine object recognition: Architectures and algorithms. *Control Eng. Pract.* **2020**, 104458. [\[CrossRef\]](#)
52. Cao, X.; Gao, S.; Chen, L.; Wang, Y. Ship recognition method combined with image segmentation and deep learning feature extraction in video surveillance. *Multimed. Tools Appl.* **2020**, *79*, 9177–9192. [\[CrossRef\]](#)
53. Ren, Y.; Yang, J.; Zhang, Q.; Guo, Z. Ship recognition based on Hu invariant moments and convolutional neural network for video surveillance. *Multimed. Tools Appl.* **2021**, *80*, 1343–1373. [\[CrossRef\]](#)
54. Spagnolo, P.; Filieri, F.; Distanto, C.; Mazzeo, P.L.; D'Ambrosio, P. A new annotated dataset for boat detection and re-identification. In Proceedings of the 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan, 18–21 September 2019; pp. 1–7.
55. Heyse, D.; Warren, N.; Tesic, J. Identifying maritime vessels at multiple levels of descriptions using deep features. *Proc. SPIE 11006 Artif. Intell. Mach. Learn. Multi-Domain Oper. Appl.* **2019**, 1100616. [\[CrossRef\]](#)
56. Groot, H.G.J.; Zwemer, M.H.; Wijnhoven, R.G.J.; Bondarev, Y.; de With, P.H.N. Vessel-speed enforcement system by multi-camera detection and re-identification. In Proceedings of the 15th International Conference on Computer Vision Theory and Applications (VISAPP), Valletta, Malta, 30 November–4 December 2020; pp. 268–277.
57. Kaido, N.; Yamamoto, S.; Hashimoto, T. Examination of automatic detection and tracking of ships on camera image in marine environment. In Proceedings of the 2016 Techno-Ocean (Techno-Ocean), Kobe, Japan, 6–8 October 2016; pp. 58–63.
58. Zhang, Z.; Wong, K.H. A computer vision based sea search method using Kalman filter and CAMSHIFT. In Proceedings of the 2013 Int. Conf. Technol. Adv. Electr. Electron. Comput. Eng (TAECE), Konya, Turkey, 9–11 May 2013; pp. 188–193.
59. Chen, Z.; Li, B.; Tian, L.F.; Chao, D. Automatic detection and tracking of ship based on mean shift in corrected video sequence. In Proceedings of the 2nd International Conference on Image, Vision and Computing (ICIVC), Chengdu, China, 2–4 June 2017; pp. 449–453.
60. Zhang, Y.; Li, S.; Li, D.; Zhou, W.; Yang, Y.; Lin, X.; Jiang, S. Parallel three-branch correlation filters for complex marine environmental object tracking based on a confidence mechanism. *Sensors* **2020**, *20*, 5210. [\[CrossRef\]](#)
61. Zhang, S.Y.; Shu, S.J.; Hu, S.L.; Zhou, S.Q.; Du, S.Z. A ship target tracking algorithm based on deep learning and multiple features. Twelfth International Conference on Machine Vision (ICMV 2019). *Int. Soc. Opt. Photonics* **2020**, *11433*, 1143304.
62. Yang, J.; Li, Y.; Zhang, Q.; Ren, Y. Surface Vehicle Detection and Tracking with Deep Learning and Appearance Feature. In Proceedings of the 5th International Conference on Control, Automation and Robotics (ICCAR), Beijing, China, 19–22 April 2019; pp. 276–280.
63. Leclerc, M.; Tharmarasa, R.; Florea, M.C.; Boury-Brisset, A.C.; Kirubarajan, T.; Duclos-Hindie, N. Ship Classification Using Deep Learning Techniques for Maritime Target Tracking. In Proceedings of the 2018 21st International Conference on Information Fusion (FUSION), Cambridge, UK, 10–13 July 2018; pp. 737–744.
64. Schöller, F.E.T.; Blanke, M.; Plenge-Feidenhans'l, M.K.; Nalpantidis, L. Vision-based Object Tracking in Marine Environments using Features from Neural Network Detections, IFAC-PapersOnLine. 2021. Available online: https://backend.orbit.dtu.dk/ws/portalfiles/portal/219717632/IFAC20_home_papercept_ifac.papercept.net_www_conferences_conferences_IFAC20_submissions_3234_FL.pdf (accessed on 20 March 2021).
65. Chen, J.; Hu, Q.; Zhao, R.; Guojun, P.; Yang, C. Tracking a vessel by combining video and AIS reports. In Proceedings of the 2008 2nd International Conference Future Generation Communication Networking (FGCN), Hainan Island, China, 13–15 December 2008; pp. 374–378.
66. Thompson, D. Maritime Object Detection, Tracking, and Classification Using Lidar and Vision-Based Sensor Fusion. Master's Thesis, Embry-Riddle Aeronautical University, Daytona Beach, FL, USA, 2017. Available online: <https://commons.erau.edu/edt/377> (accessed on 20 March 2021).
67. Helgesen, Ø.K.; Brekke, E.F.; Helgesen, H.H.; Engelhardt, O. Sensor Combinations in Heterogeneous Multi-sensor Fusion for Maritime Target Tracking. In Proceedings of the 2019 22th International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2–5 July 2019; pp. 1–9.

68. Haghbayan, M.H.; Farahnakian, F.; Poikonen, J.; Laurinen, M.; Nevalainen, P.; Plosila, J.; Heikkonen, J. An Efficient Multi-sensor Fusion Approach for Object Detection in Maritime Environments. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 2163–2170.
69. Farahnakian, F.; Movahedi, P.; Poikonen, J.; Lehtonen, E.; Makris, D.; Heikkonen, J. Comparative analysis of image fusion methods in marine environment. In Proceedings of the IEEE International Symposium on Robotic and Sensors Environments (ROSE), Ottawa, ON, Canada, 17–18 June 2019; pp. 535–541.
70. Stanislas, L.; Dunbabin, M. Multimodal Sensor Fusion for Robust Obstacle Detection and Classification in the Maritime RobotX Challenge. *IEEE J. Ocean. Eng.* **2018**, *44*, 343–351. [\[CrossRef\]](#)
71. Farahnakian, F.; Heikkonen, J. Deep learning based multi-modal fusion architectures for maritime vessel detection. *Remote Sens.* **2020**, *12*, 2509. [\[CrossRef\]](#)
72. Farahnakian, F.; Haghbayan, M.H.; Poikonen, J.; Laurinen, M.; Nevalainen, P.; Heikkonen, J. Object Detection Based on Multi-sensor Proposal Fusion in Maritime Environment. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 971–976.
73. Cho, H.J.; Jeong, S.K.; Ji, D.H.; Tran, N.H.; Vu, M.T.; Choi, H.S. Study on control system of integrated unmanned surface vehicle and underwater vehicle. *Sensors* **2020**, *20*, 2633. [\[CrossRef\]](#) [\[PubMed\]](#)
74. Stateczny, A.; Błaszczak-Bak, W.; Sobieraj-Złobińska, A.; Motyl, W.; Wisniewska, M. Methodology for processing of 3D multibeam sonar big data for comparative navigation. *Remote Sens.* **2019**, *11*, 2245. [\[CrossRef\]](#)
75. Remmas, W.; Chemori, A.; Kruusmaa, M. Diver tracking in open waters: A low-cost approach based on visual and acoustic sensor fusion. *J. F. Robot.* **2020**. [\[CrossRef\]](#)
76. Bloisi, D.D.; Previtali, F.; Pennisi, A.; Nardi, D.; Fiorini, M.; Member, S. Enhancing automatic maritime surveillance systems with visual information. *IEEE Trans. Intell. Transp. Syst.* **2016**, *18*, 824–833. [\[CrossRef\]](#)
77. Wang, X.; Xu, L.; Sun, H.; Xin, J.; Zheng, N. On-Road Vehicle Detection and Tracking Using MMW Radar and Monovision Fusion. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2075–2084. [\[CrossRef\]](#)
78. Gundogdu, E.; Solmaz, B.; Yücesoy, V.; Koc, A. Marvel: A large-scale image dataset for maritime vessels. In Proceedings of the Asian Conference on Computer Vision (ACCV), Taipei, Taiwan, 20–24 November 2016; pp. 165–180.
79. Bloisi, D.D.; Iocchi, L.; Pennisi, A.; Tombolini, L. ARGOS-Venice boat classification. In Proceedings of the 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Karlsruhe, Germany, 25–28 August 2015; pp. 1–6.
80. Zhang, M.M.; Choi, J.; Daniilidis, K.; Wolf, M.T.; Kanan, C. VAIS: A dataset for recognizing maritime imagery in the visible and infrared spectrums. In Proceedings of the IEEE Computer Society Conference Computer Vision Pattern Recognition Work (CVPRW), Boston, MA, USA, 7–12 June 2015; pp. 10–16.
81. Teng, F.; Liu, Q. *Visual Tracking Algorithm for Inland Waterway Ships*; Wuhan University of Technology Press: Wuhan, China, 2017; pp. 18–24.
82. Liu, Q.; Mei, L.Q.; Lu, P.P. *Visual Detection Algorithm for Inland Waterway Ships*; Wuhan University of Technology Press: Wuhan, China, 2019; pp. 98–102.
83. Shin, H.C.; Lee, K.I.L.; Lee, C.E. Data augmentation method of object detection for deep learning in maritime image. In Proceedings of the 2020 IEEE International Conference Big Data Smart Computers (BigComp), Pusan, Korea, 19–22 February 2020; pp. 463–466.
84. Chen, Z.; Chen, D.; Zhang, Y.; Cheng, X.; Zhang, M.; Wu, C. Deep learning for autonomous ship-oriented small ship detection. *Saf. Sci.* **2020**, *130*, 104812. [\[CrossRef\]](#)
85. Milicevic, M.; Zubrinic, K.; Obradovic, I.; Sjekavica, T. Data augmentation and transfer learning for limited dataset ship classification. *WSEAS Trans. Syst. Control* **2018**, *13*, 460–465.