

Article

A Monocular Vision-Based Safety Monitoring Framework for Offshore Infrastructures Utilizing Grounded SAM

Sijie Xia ¹, Rufu Qin ^{1,*}, Yang Lu ², Lianjiang Ma ² and Zhenghu Liu ²¹ State Key Laboratory of Marine Geology, Tongji University, Shanghai 200092, China; 2231665@tongji.edu.cn² Institute of Deep-Sea Science and Engineering, Chinese Academy of Sciences, Sanya 572000, China; luyang@idsse.ac.cn (Y.L.); malj@idsse.ac.cn (L.M.); liuzh@idsse.ac.cn (Z.L.)

* Correspondence: qinrufu@tongji.edu.cn

Abstract: As maritime transportation and human activities at sea continue to grow, ensuring the safety of offshore infrastructure has become an increasingly pressing research focus. However, traditional high-precision sensor systems often involve prohibitive costs, and the Automatic Identification System (AIS) faces signal loss or data manipulation problems, highlighting the need for a complementary, affordable, and reliable supplemental solution. This study introduces a monocular vision-based safety monitoring framework for offshore infrastructures. By combining advanced computer vision techniques such as Grounded SAM and horizon-based self-calibration, the proposed framework achieves accurate vessel detection, instance segmentation, and distance estimation. The model integrates open-vocabulary object detection and zero-shot segmentation, achieving high performance without additional training. To demonstrate the feasibility of the framework in practical applications, we conduct several experiments on public datasets and couple the proposed algorithms with the Leaflet.js and WebRTC libraries to develop a web-based prototype for real-time safety monitoring, providing visualized information and alerts for offshore infrastructure operators in our case study. The experimental results and case study suggest that the framework has notable advantages, including low cost, convenient deployment with minimal maintenance, high detection accuracy, and strong adaptability to diverse application conditions, which brings a supplemental solution to research on offshore infrastructure safety.

Keywords: object detection; Grounded SAM; monocular distance estimation; maritime safety

Academic Editor: Weicheng Cui
and Shanshan Fu

Received: 4 January 2025

Revised: 10 February 2025

Accepted: 11 February 2025

Published: 13 February 2025

Citation: Xia, S.; Qin, R.; Lu, Y.; Ma, L.; Liu, Z. A Monocular Vision-Based Safety Monitoring Framework for Offshore Infrastructures Utilizing Grounded SAM. *J. Mar. Sci. Eng.* **2025**, *13*, 340. <https://doi.org/10.3390/jmse13020340>

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

1. Introduction

The ocean covers more than 70% of the earth's surface and is a critical part of the global ecosystem. It provides sufficient resources such as fisheries, minerals, and energy [1,2]. It not only plays an indispensable role in climate regulation and biodiversity conservation but also underpins global economic development. With the development of marine science and the expansion of environmental monitoring and resource exploitation activities, an increasing variety of offshore infrastructures has been deployed, including oil platforms [3], wind farms [4], marine ranches [5], nearshore aquaculture systems [6], and ocean observation platforms [7] such as towers, buoys, and moored instruments. In the field of ocean engineering, scholars have proposed various safety and reliability analysis methods to ensure the stable operation of these infrastructures. Specifically, structural reliability analysis is applied during the design phase to ensure offshore structures can withstand wind-induced wave load impact and other extreme climatic conditions, determining fatigue limit states [8,9]. During operation, real-time monitoring and safety analysis are

conducted through sensors to assess the impact of environmental factors such as scouring, corrosion, the growth of marine organisms, and temperature variations, providing timely evaluations and early warnings for potential risks [10,11]. However, human activities such as vessel intrusions and maritime traffic congestion also pose significant risks to offshore infrastructures. These infrastructures particularly suffered from complex vessel activities in coastal waters, where there is high vessel density and complex marine environment conditions [12]. Thus, securing the safety of offshore infrastructures through enhanced real-time vessel monitoring, traffic management, and advanced safety analysis and collision warning systems has become an urgent necessity [13].

To ensure the safety of offshore infrastructures, stakeholders have applied various technologies to study situational monitoring and safety, such as using the Automatic Identification System (AIS) and Geographic Information System (GIS) for vessel tracking and analysis. Ma et al. [14] utilized AIS data, Monte Carlo simulations, and Bayesian networks to calculate the collision probability between bridges and ships, which effectively reduced the risk of collision. Kao et al. [15] proposed a comprehensive GIS-based collision avoidance system for offshore infrastructures, which processes AIS data in the cloud, identifies and filters hazardous vessels, broadcasts alerts to target ships through aid-to-navigation, and displays visual warnings for infrastructure personnel in the Marine Geographic Information System (MGIS). Zhang et al. [16] introduced a method that combines deep learning and AIS data for vessel trajectory prediction, which can be used to provide early warnings of collision risks between vessels and offshore infrastructures. While AIS is a widely used vessel tracking technology that provides real-time position data, its reliability can be compromised by potential signal loss [17] and data manipulation [18].

To address the limitations of AIS-based safety monitoring, many researchers proposed multi-sensor fusion methods for enhancing performance. For instance, Cope et al. [19] introduced a multi-sensor monitoring system that uses X-band radar and optical cameras to monitor vessel activities within protected areas. Such systems maintain reliable vessel positioning even when AIS signals are lost or manipulated, enhancing the redundancy and robustness of the system [20]. However, due to the need to use a variety of expensive sensors, such methods still face the problem of high cost and need to be further optimized.

In recent years, visible-light camera technologies have advanced rapidly, and video surveillance systems, benefiting from their low cost, mature technology, and ease of installation, have become increasingly popular in various fields of monitoring applications [21], including maritime monitoring. These systems are typically incorporated with trending computer vision algorithms. In the context of monitoring offshore targets, there is an urgent demand for efficient and accurate vessel target recognition, localization, and distance estimation. It is known that the computer vision technique has the advantages of object detection [22], instance segmentation [23], and distance estimation [24], which provide practical solutions to meet these demands [25].

Distance estimation techniques based on visible-light cameras can be categorized into monocular and binocular methods, depending on the number of cameras involved [24]. The binocular method, leveraging the disparity principle, can be used to obtain more precise location and distance estimations of targets. Zeng et al. [26] proposed a binocular vision system with both long and short focal lengths designed to measure the height of vessels passing under bridges for bridge security surveillance. However, binocular methods often encounter drawbacks, such as expensive hardware, complex installation, a pre-calibration requirement, and difficulty adjusting stereo calibration parameters after deployment. To address the limitations of binocular methods, scholars have attempted to utilize monocular methods as an alternative, striving to solve the challenge of depth extraction from a single-camera view. In recent years, the accuracy and reliability of monocular distance estimation

methods have gradually improved and have been applied in various fields. For example, Liang et al. [27] proposed a self-supervised depth estimation model that utilizes multi-scale resolution to enhance the expression of depth information, thus improving the accuracy of distance estimation. In maritime applications, Gladstone et al. [28] proposed a method based on monocular vision and a geometric model for estimating the distance to target vessels. This method is highly practical as it detects the horizon and applies the geometric properties of the earth. In addition, Kang et al. [29] combined monocular vision with traditional sensors such as radar to enhance the reliability of the system.

More importantly, images captured by visible-light cameras can be easily analyzed using various computer vision techniques [30]. Many scholars have employed various neural network models for vessel classification, semantic segmentation, and target tracking in coastal regions. For example, Shao et al. [31] proposed a convolutional neural network based on saliency perception for the recognition of nearshore vessels. Ting et al. [32] combined the YOLOv5 model with GhostNet to detect vessels, addressing the issue of the uneven distribution of vertical and horizontal features. These models generally require large amounts of labeled data for training to achieve quick and accurate vessel identification. Technically, these models still face challenges in data dependency and limited generalization ability [33]. To address these limitations, visual foundation models such as the Segment Anything Model (SAM) [34] and Grounding DINO [35] offer a new low-cost solution for instance segmentation and target detection, which require lower data processing and labor costs. Through self-supervised learning and large-scale pretraining, the SAM-based models can perform precise segmentation of various objects without extensive labeled data. Given the advantage of strong zero-shot segmentation capability, SAM has already been applied in fields of medical image segmentation [36,37], remote sensing image segmentation [38], and autonomous driving [39]. Meanwhile, Grounding DINO demonstrates powerful open-vocabulary capabilities, enabling the detection of any target based on human-input text prompts, and it has also been widely applied across various fields [40,41]. However, there are few studies and applications that integrate this novel SAM and Grounding DINO technologies to detect and segment vessels in the maritime domain, thus requiring further investigation.

This paper proposes a monocular vision-based safety monitoring framework for offshore infrastructures, integrating object detection and instance segmentation techniques based on Grounded SAM [42], along with the algorithms of sea horizon detection, monocular distance estimation, and self-extrinsic parameter calibration. By combining Grounded SAM with boundary simplification and convex hull algorithms, the framework generates a more refined detection region for the sea horizon, significantly improving detection accuracy. This approach is particularly effective in scenarios where the sea horizon is partially occluded and is adaptable to other detection methods, offering valuable insights for integrating advanced vision models with traditional algorithms. Additionally, through the application of Grounded SAM, the framework enables zero-shot vessel detection and segmentation without the need for additional training. Moreover, we conducted several experiments to evaluate the capability and performance of the proposed algorithms. Furthermore, to demonstrate the technical feasibility and practicality of the proposed framework, we developed a web-based prototype for the real-time monitoring of offshore infrastructure safety. The experimental results and case study show that the proposed framework enables real-time surveillance and performs accurate vessel detection and segmentation as well as accurate distance estimation, providing a satisfactory complementary solution for improving maritime safety.

The remainder of this paper is structured as follows: Section 2 details the specific methods for monocular vision-based horizon detection, extrinsic parameter estimation,

and distance estimation; Section 3 presents the experimental evaluation and result analysis of each method; and Section 4 concludes the advantages, limitations, and potential improvements of the proposed monocular vision-based safety monitoring framework for offshore infrastructures and the methodology.

2. Methods

2.1. Components of the Proposed Framework

The proposed offshore infrastructure safety monitoring framework is divided into three primary components (Figure 1): a hardware module for video data acquisition on the edge side; the core algorithms run on the server side for real-time video stream processing, computation, and analysis; and a web application platform on the client side designed to visualize the video stream and the results generated by the algorithm modules.

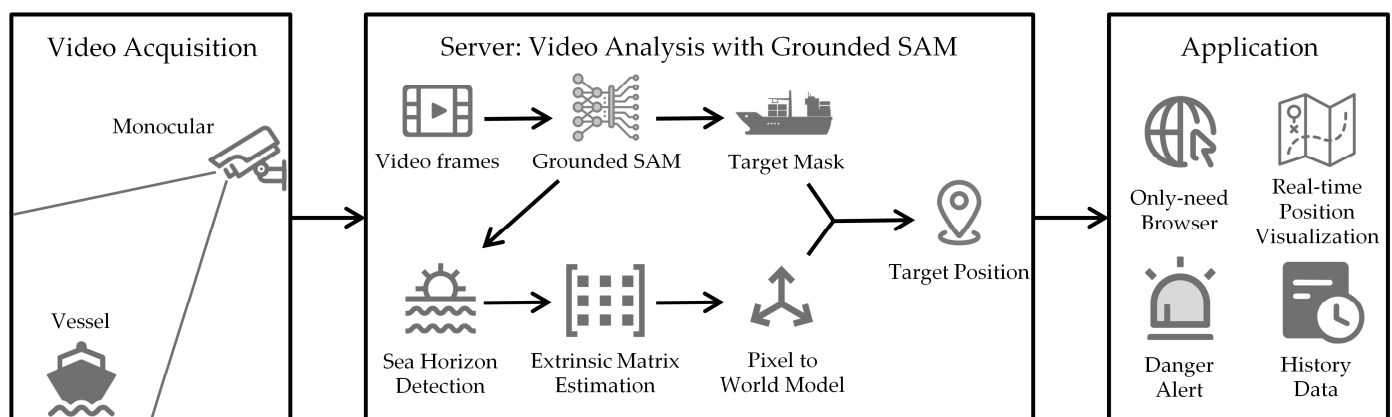


Figure 1. Structural components of monocular vision-based safety monitoring framework for offshore infrastructures.

The video acquisition module utilizes a fixed-focus monocular camera mounted on a fixed platform of the offshore facility. The videos cover the designated monitoring region, ensuring a clear sea horizon line within the view. It should be noted that we need to calibrate the intrinsic camera parameters before installation, while extrinsic parameters do not require pre-calibration.

On the server side, the framework integrates several algorithms to ensure accurate processing and analysis of video data. First, Grounded SAM assists in detecting the sea horizon line from images, and thus, extrinsic parameters can be estimated based on the geometric features of the sea horizon line. Then, these parameters and the intrinsic parameters are used to build a coordinate transformation model to map pixel coordinates on the sea surface to world coordinates. In addition, the framework uses Grounded SAM for marine vessel target detection and instance segmentation to generate pixel masks for vessel targets in the surveillance area. Finally, the pixel positions are converted to world coordinates and distances by combining the coordinate transformation model with vessel masks. This workflow ensures the continuous identification of vessels that may pose a threat in the monitoring area and updates their coordinates and distances in real time.

The client platform provides intuitive interfaces for end users by displaying real-time video streams, video analysis results, and relative information with different visualization modes. In this study, we adopted a B/S architecture [43] to develop the web-based prototype, significantly improving the accessibility and deployment efficiency of the entire system. This comprehensive framework achieves effective situational monitoring, addressing the safety needs of offshore infrastructures. The following sections focus on the detailed introduction of the proposed algorithms and the development of the prototype system,

highlighting their integration and functionality within the safety monitoring framework for offshore infrastructures.

2.2. Object Detection and Instance Segmentation

In this study, we utilize a model called Grounded SAM, which integrates Grounding DINO and SAM to detect and segment objects such as vessels and the sea surface in images. In comparison to other popular models, such as Mask R-CNN, which typically requires extensive retraining on task-specific datasets to achieve optimal performance, and the YOLO, which excels at detecting objects quickly but lacks segmentation abilities, Grounded SAM stands out due to its zero-shot capability and enhanced ability to both detect and segment any targets. This model operates without additional training and generates segmentation masks by simply inputting the targets' names in text form as prompts. Grounding DINO identifies bounding boxes for maritime targets, which SAM uses as prompts for mask segmentation. By combining these models, Grounded SAM improves the precision and robustness of vessel detection in complex oceanic scenarios without any additional training and fine-tuning.

SAM is a generalized promptable segmentation base model trained on a vast dataset consisting of over 1.1 billion segmentation masks and about 11 million images [34]. Extensive pretraining gives SAM a strong zero-shot capability, allowing it to generalize to unseen object types effectively. The model supports the input of multiple types of prompts (e.g., points and bounding boxes) to segment the corresponding targets in an image. Architecturally, SAM consists of an image encoder, a prompt encoder, and a mask decoder. The image encoder extracts detailed visual features, the prompt encoder processes user inputs, and the mask decoder combines these to generate pixel-level masks, enabling flexible and accurate segmentation. However, SAM currently does not support text-based prompts, which means it cannot directly segment human-specified object types in open-set scenes.

In Grounded SAM, the Grounding DINO model is responsible for detecting maritime targets based on prompt text and generating bounding boxes, which are fed into SAM for mask extraction. Grounding DINO is an open-vocabulary object detection model based on the Transformer architecture that can detect any object class in an image and provide accurate bounding boxes based on the user's natural language prompts [35]. The core of the model consists of an image encoder, a text encoder, a feature enhancer, a language-guided query selection module, and a cross-modal decoder. The image encoder extracts visual features from the input image, while the text encoder processes natural language prompts. These features are enhanced and aligned by the feature enhancer and the language-guided query selection module to focus on relevant objects. Finally, the cross-modal decoder integrates these features to generate precise bounding boxes for the targets specified by the text prompt. In this study, Grounded SAM effectively combines the open-vocabulary object detection capability of Grounding DINO with the promptable zero-shot instance segmentation capability of SAM, enabling precise segmentation of maritime vessels by simply inputting images and a text prompt. This not only enhances the accuracy of target detection and distance estimation but also improves the usability of the framework for offshore infrastructure safety monitoring.

2.3. Geometric Method of Monocular Distance Estimation

We employ a geometric coordinate transformation model to estimate the distances of vessel instances detected and segmented by the Grounded SAM. The model's basic principle relies on the pinhole camera model, which establishes the mapping relationship between 2D image coordinates and 3D real-world coordinates.

2.3.1. Pinhole Model and Camera Calibration

The pinhole camera model plays a crucial role in computer vision, describing how the 3D world is projected onto a 2D image plane (Figure 2). However, as the model neglects factors like lens distortion, accurate distance estimation requires camera calibration, which involves estimating both intrinsic and extrinsic parameters.

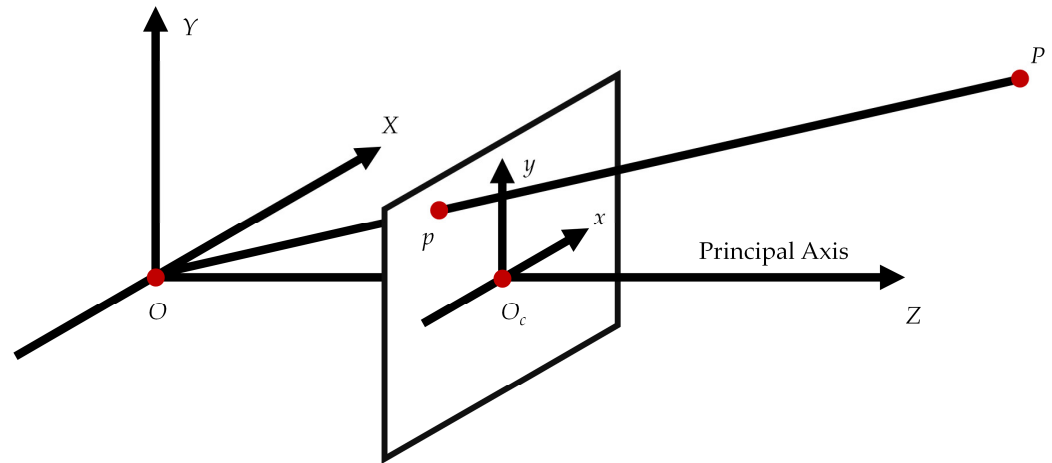


Figure 2. Pinhole camera model.

The intrinsic parameters include the camera's focal length (f_x, f_y) and principal point (c_x, c_y), while the extrinsic parameters involve the rotation vector and translation vector of the camera's coordinate system relative to the world coordinate system. The intrinsic matrix K is calculated as:

$$K = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \quad (1)$$

In this study, Zhang's calibration method [44] is used to calibrate the camera.

2.3.2. Coordinate Transformation

To obtain the position and distance of targets in an image, it is necessary to convert the 2D image coordinate system into the 3D world coordinate system. In this study, the intersection of the camera's plumb line and the sea surface is set as the origin O_w . Based on the camera's perspective, the X_w -axis points horizontally to the right, the Y_w -axis points vertically upward, and the Z_w -axis points horizontally forward. The origin of the camera coordinate system is set at the camera's position, corresponding to $(0, h_c, 0)$ in the world coordinate system, where h_c is the camera's installation height. The axes of the camera coordinate system align with those of the world coordinate system. In the pixel coordinate system, the origin is at the top-left corner of the image, with the u -axis extending to the right and the v -axis extending downward. Figure 3 illustrates the monitoring scene and the basic setup of each coordinate axis while also explaining the geometric relationship between the pixel locations of targets in the image and their actual positions in the 3D world. Deriving the specific transformation method for this geometric relationship is key to calculating the position and distance of the targets.

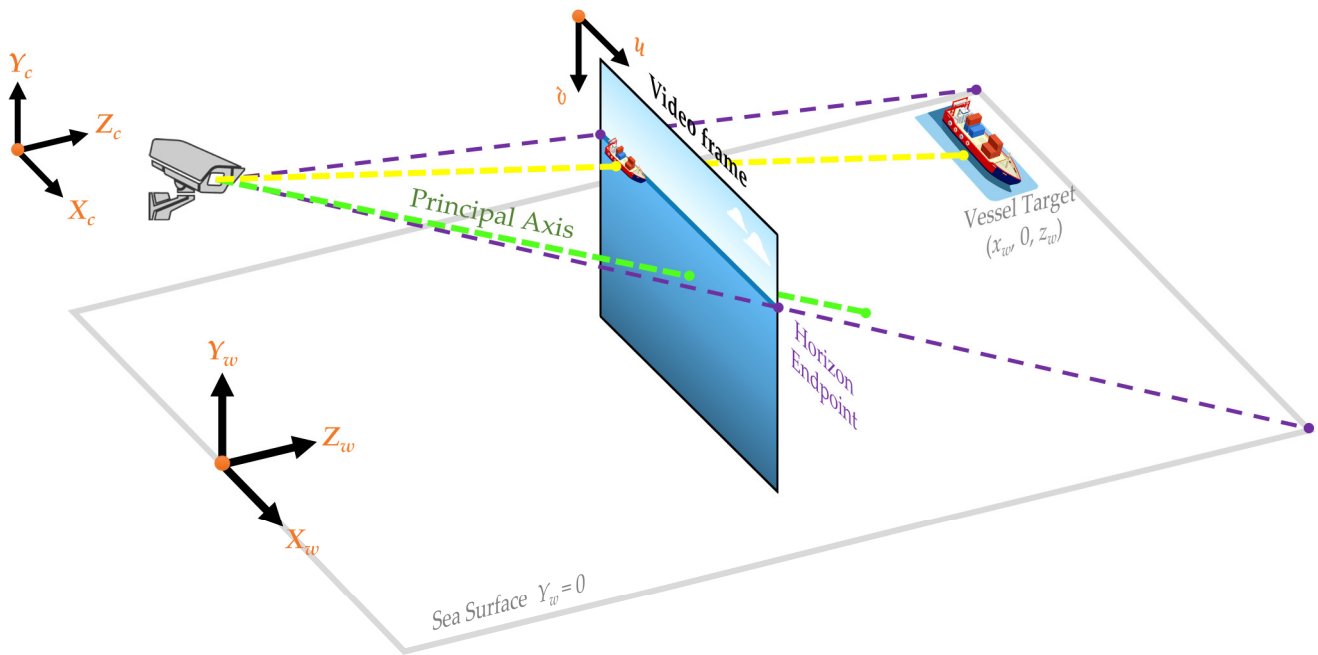


Figure 3. Schematic diagram of camera setup and coordinate system definitions. The dashed lines represent the light path through different points in the frame.

The transformation from 2D image coordinates to 3D world coordinates involves several essential stages. First, pixel coordinates (u, v) are converted to image coordinates (x_i, y_i) . Then, the image coordinates are converted to camera coordinates (X_c, Y_c, Z_c) . Finally, the camera coordinates are converted to world coordinates (X_w, Y_w, Z_w) . These sequential transformations are mathematically defined in Equations (2)–(4).

$$\begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} = K^{-1} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \quad (2)$$

$$\begin{cases} X_c = Z_c \cdot x_i \\ Y_c = Z_c \cdot y_i \end{cases} \quad (3)$$

$$\begin{pmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{pmatrix} = \begin{pmatrix} R & T \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{pmatrix} \quad (4)$$

Since the transformation from a 2D image to 3D coordinates lacks depth information Z_c , this issue is classified as an ill-posed problem. In this study, the observed sea surface is idealized as a plane ($Y_w = 0$), thereby providing the additional parameters needed for the transformation. This allows the computation of the world coordinates of the target point as $(X_w, 0, Z_w)$.

Initially, pixel coordinates (u, v) are converted into normalized camera coordinates (x_n, y_n) . In the camera coordinate system, the direction of the light ray is represented by the vector $(x_n, y_n, 1)$. Utilizing the rotation matrix R , which is estimated by our self-calibration method in Section 2.4.2, this light ray is transformed into the world coordinate system, yielding the direction (X_d, Y_d, Z_d) . When the light ray originates from the camera center at $(0, h, 0)$, the parametric equation of the ray can be established, where the parameter t can be represented by h and Y_d . Substituting t into the ray equation provides the coordinates of the intersection point on the sea surface plane $(X_w, 0, Z_w)$. Finally, the distance d between

the target and the offshore platform is computed based on these coordinates. The complete mathematical model for this transformation process is presented in Equations (5)–(10):

$$\begin{cases} x_n = \frac{u - c_x}{f_x} \\ y_n = \frac{v - c_y}{f_y} \end{cases} \quad (5)$$

$$\begin{pmatrix} X_d \\ Y_d \\ Z_d \end{pmatrix} = R \begin{pmatrix} x_n \\ y_n \\ 1 \end{pmatrix} \quad (6)$$

$$\begin{pmatrix} X_w \\ Y_w \\ Z_w \end{pmatrix} = t \begin{pmatrix} X_d \\ Y_d \\ Z_d \end{pmatrix} + \begin{pmatrix} 0 \\ h \\ 0 \end{pmatrix} \quad (7)$$

$$t = -\frac{h}{Y_d} \quad (8)$$

$$\begin{cases} X_w = -\frac{hX_d}{Y_d} \\ Z_w = -\frac{hZ_d}{Y_d} \end{cases} \quad (9)$$

$$d = \sqrt{X_w^2 + Z_w^2} \quad (10)$$

It is worth noting that the pixels in an image obtained by a visible-light camera are discrete; therefore, it is hard to precisely represent the continuous position of an object, as shown in Figure 4. This limitation is particularly significant at long distances, where each pixel corresponds to a more extensive physical region, resulting in increased computational errors. Consequently, the world coordinate error of distant pixel points is much greater than that of nearby pixel points. Higher image resolution reduces the physical region represented by a single pixel, thereby lowering the error.

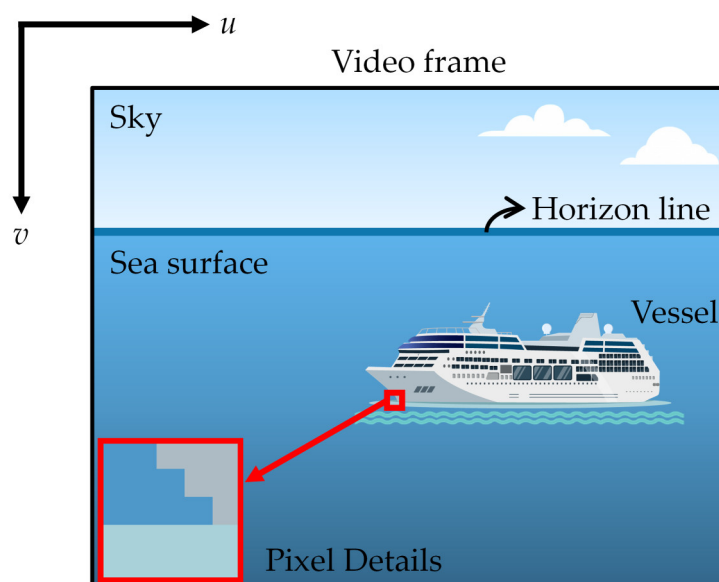


Figure 4. Illustration of a video frame and its pixel discreteness.

2.4. A Self-Calibration Method Based on Sea Horizon Line

In computer vision, traditional camera calibration methods typically require calibration objects, such as chessboards, with known dimensions. However, due to the unique challenges of marine application scenarios, it is difficult to use traditional calibration boards

for extrinsic calibration. Images captured by cameras on offshore infrastructures typically include targets such as the sky, sea surface, maritime vessels, and the boundary between the sea and the sky, known as the sea horizon line, as shown in Figure 4. Therefore, we can analyze the natural features within the scene to solve this issue.

The horizon line serves as a critical reference in such images, acting as the only reliable geometric cue in the vast expanse of the ocean surface. Even without additional calibration objects, it is still possible to use the horizon line to estimate the camera's extrinsic parameters—namely, the rotation and translation matrices—thus enabling the self-calibration of the camera. In this study, the horizon line is used as an alternative solution for extrinsic parameter self-calibration, particularly for cases where camera angles may shift unexpectedly after deployment or when chessboards cannot be used.

2.4.1. Horizon Detection

To perform extrinsic parameter self-calibration, the horizon line in the image must first be detected. In this study, a method based on edge detection and line extraction is used for horizon line detection, combined with Grounded SAM to segment the ocean region, narrowing the detection area and improving the accuracy of horizon line detection. The detailed process is as follows:

To better distinguish the sea surface from the sky and identify the ideal distribution region of the horizon line, the Grounded SAM is used for ocean surface instance segmentation. However, when maritime vessels or other objects occlude the horizon line, the boundaries of ocean surface instances become more complex. This may lead to concave boundaries or the detection of multiple ocean surface instances, significantly obstructing the determination of the final horizon line region of interest (ROI).

To address this issue, we propose an ocean surface instance boundary completion algorithm that combines boundary tracking and simplification to enhance the integrity of ocean surface instances and define an ideal region for horizon line localization. The algorithm first employs the chain code method [45] to connect all boundary pixels, forming closed polygonal boundaries. Next, boundary simplification is performed in two steps: the Douglas–Peucker algorithm [46] recursively removes unnecessary points from the boundary, and Andrew's monotone convex hull algorithm generates the minimal convex set of the boundary point set. This approach ensures a more cohesive and reliable representation of ocean surface instances, facilitating accurate horizon line detection.

Figure 5c,d show the results of completing and simplifying the ocean surface instance boundary. Then, an ROI for horizon line detection is generated within a certain buffer zone of the adjusted boundary (Figure 5e,f).

Finally, the Canny edge detection algorithm is applied to the image within the horizon line detection ROI to extract all edge features. These edge points are then fitted to a single straight line using a RANSAC-based procedure, which identifies the horizon line among the extracted edges. Specifically, RANSAC repeatedly samples pairs of edge points to propose candidate line models, evaluating each by the number of inliers within a distance threshold. The line model with the highest inlier count is ultimately selected as the horizon, ensuring robustness against spurious edges and partial occlusions.

To further enhance detection accuracy, the variation in the detected horizon line across consecutive time frames must remain within a specified threshold. If the variation exceeds the threshold, the algorithm re-executes the detection process for a satisfied result, and the parameters are updated. A scheduled execution mechanism for the algorithm module is implemented to mitigate the impact of potential inaccuracies caused by unexpected shifts in the camera position after installation.

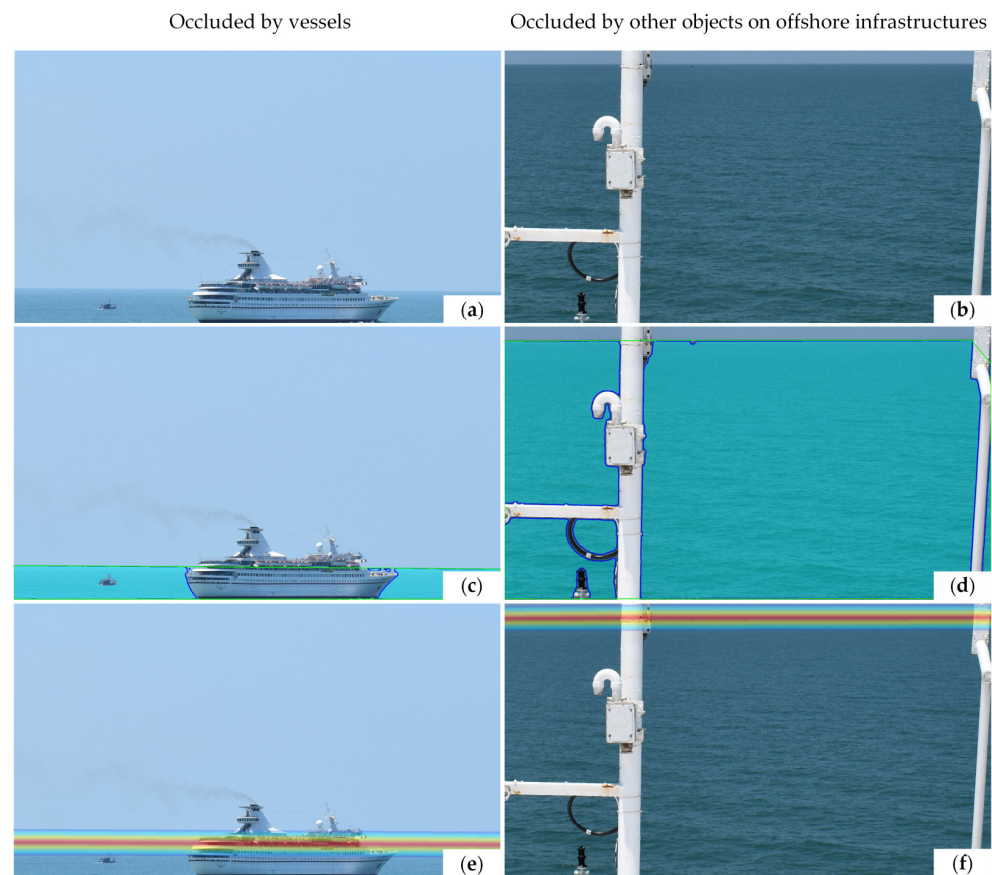


Figure 5. Examples of the sea horizon line ROI extraction steps: (a,b) original images and (c,d) ocean surface instance boundaries. The green line represents the completed boundary of the ocean surface instance after correction, while the blue line indicates the boundary before completion; (e,f) show the ROI (represented as a heatmap overlaid on the image) of the sea horizon line.

2.4.2. Extrinsic Parameter Estimation

The position and angle of the sea horizon line in the image (Figure 6) can be used to estimate the camera's extrinsic parameters, namely its pose. Extrinsic parameters include the rotation matrix and the translation vector relative to the world coordinate system. The translation vector can be directly obtained from the coordinate system settings described in Section 2.3.2.

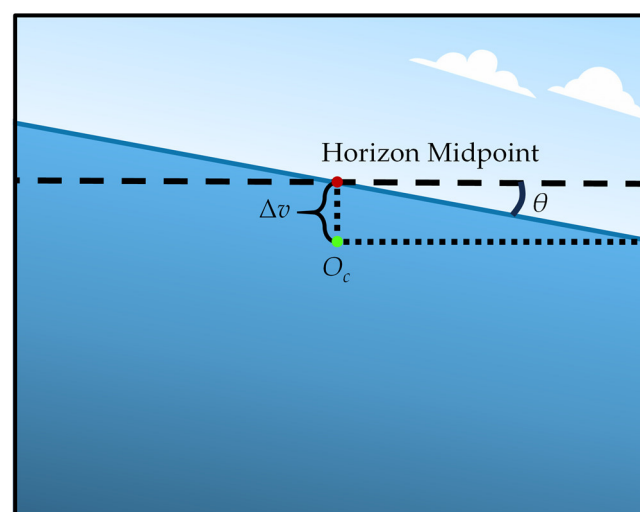


Figure 6. Illustration of angular and positional features of the sea horizon line.

The camera's rotation includes three angles: pitch, roll, and yaw. Since the yaw angle does not affect the characteristics of the horizon line in the image, it is set to 0, corresponding to the camera's default orientation after installation.

Pitch: If the camera's pitch angle is non-zero, the position of the horizon line in the image will shift upward or downward accordingly. By analyzing the vertical position of the horizon line, the pitch angle θ can be calculated, along with the corresponding rotation vector about the x -axis, $R_x(\theta)$, as follows:

$$\theta = \tan^{-1} \frac{v_h - c_y}{f_y} \quad (11)$$

$$R_x(\theta) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{pmatrix} \quad (12)$$

where v_h and c_y represent the vertical coordinates of the midpoint of the horizon line and the optical center, respectively, and f_y is the focal length of the camera along the y -axis.

Roll: If the camera exhibits a roll, the horizon line will appear tilted. By analyzing the angle between the horizon line and the horizontal axis of the image, the roll angle ϕ can be calculated, along with the corresponding rotation vector about the z -axis, $R_z(\phi)$, as follows:

$$\phi = \tan^{-1} \frac{\Delta v}{\Delta u} \quad (13)$$

$$R_z(\phi) = \begin{pmatrix} \cos \phi & -\sin \phi & 0 \\ \sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (14)$$

where Δv and Δu are the vertical and horizontal pixel coordinate differences between the two endpoints of the sea horizon line. Since the rotation matrix about the y -axis is an identity matrix, the final rotation matrix R is expressed as:

$$R = R_z(\phi)R_x(\theta) = \begin{pmatrix} \cos \phi & -\sin \phi \cos \theta & \sin \phi \sin \theta \\ \sin \phi & \cos \phi \cos \theta & -\cos \phi \sin \theta \\ 0 & \sin \theta & \cos \theta \end{pmatrix} \quad (15)$$

The translation vector, which corresponds to the camera's installation height, has minimal impact on the position of the horizon line in the captured image. Given the camera height h and the Earth's radius R , the angle η between the horizon line and the horizontal axis in the camera's field of view can be expressed as:

$$\eta = \cos^{-1} \frac{R}{R+h} \quad (16)$$

Under typical installation conditions, where $h \ll R$, the small error in the translation vector has a negligible effect on the horizon line's position.

The self-calibration method for the camera's extrinsic parameters ensures that even when the camera shifts due to external forces over prolonged operation, the pitch and roll angles can still be accurately extracted. This enhances the robustness and maintainability of the overall monocular distance estimation method for maritime targets.

2.5. Safety Analysis and Visualization

In this section, we describe how the results obtained from monocular distance estimation, vessel segmentation, and horizon detection are analyzed and transformed into

actionable safety information. Furthermore, we explain the visualization techniques used to represent the spatial relationships between maritime objects and infrastructure, enabling operators to assess potential threats better and take timely action.

As shown in Figure 7, we transform all elements in the scene into a two-dimensional plane to simulate a digital maritime map. This visualization method provides an intuitive representation of the following elements:

- Geographic position of the offshore infrastructure: the location of the offshore critical infrastructure or other critical equipment.
- The monitoring area of a monocular camera: the area covered by the camera, which helps operators understand the camera's field of view and its limitations.
- Graded alert zones: circular alert zones centered around the protected offshore infrastructure, representing different levels of risk. These zones can be adjusted in shape or size according to specific operational needs, with each level indicating the proximity of objects to the infrastructure and the corresponding safety response required.
- Detected vessel targets: the positions (calculated from Equations (5)–(9)) of vessels detected within the camera's range, updated in real time to show their movements and proximity to the infrastructure. In addition, the distance information can be displayed near each target.

In our web application system, we utilized Leaflet.js (a lightweight open-source JavaScript library for creating interactive and customizable web-based maps, widely used for dynamic geospatial visualizations) to facilitate the dynamic drawing and real-time updating of all these elements on the map. This allows operators to visualize the current status of the maritime environment and receive up-to-date information about vessel positions and the safety status of critical infrastructure. The system continuously updates in real time, ensuring that any newly detected targets or changes in the monitoring range are immediately reflected, thus enabling rapid response to potential safety threats.

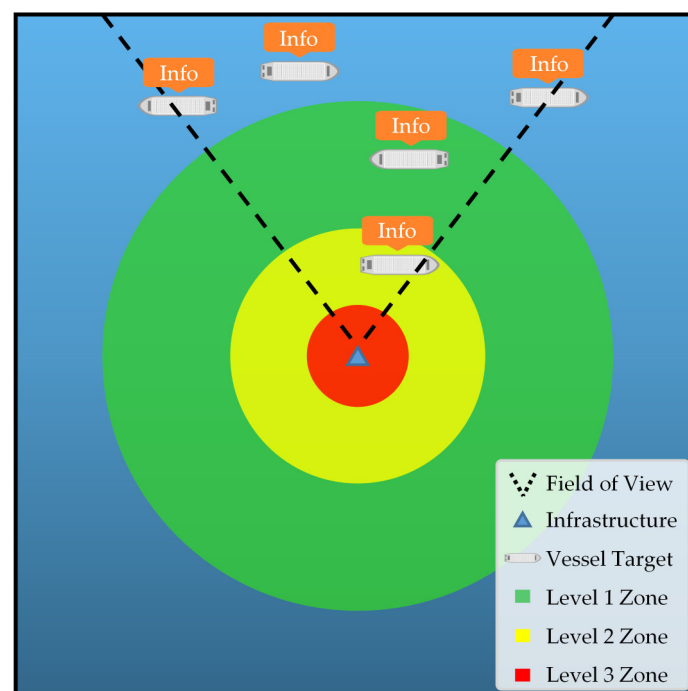


Figure 7. Digital maritime map for safety analysis.

3. Experimental Results and Discussion

In this section, the performance of the proposed sea horizon detection method and the vessel segmentation method are evaluated using public datasets to assess the practicality and reliability of the approach. Meanwhile, the monocular distance estimation method is evaluated through theoretical calculations in our case study. The experiments are conducted on a system featuring an Intel Core i5-12600KF CPU (Intel, Santa Clara, CA, USA), 32 GB of RAM, and an NVIDIA RTX 3070 GPU with 8 GB of VRAM (NVIDIA, Santa Clara, CA, USA).

3.1. Evaluation of Sea Horizon Detection

In this section, we describe the experiments that were conducted to evaluate the performance of the sea horizon detection algorithm. Considering the role of the sea horizon line in the self-calibration algorithm for camera extrinsic estimation, the evaluation focuses on comparing the detected sea horizon line with ground truth data in terms of angular and positional differences. The evaluation metrics include angular error, positional error, and detection accuracy.

3.1.1. Horizon Detection Dataset

The ground truth datasets for the horizon line used in the experiments come from the Manzoor-Umar: Sea Image Dataset (MU-SID) (<https://www.kaggle.com/datasets/umairatwork/manzoorumair-sea-image-dataset-musid>, accessed 3 January 2025) [47] and Singapore Maritime Dataset (SMD) (<https://sites.google.com/site/dilipprasad/home/singapore-maritime-dataset>, accessed 3 January 2025). The MU-SID contains an extensive collection of ocean images captured under various weather conditions and sea states. MU-SID includes a total of 2673 high-definition (1920×1080 pixels) RGB images, along with the positional information of the sea horizon line in each image. The SMD includes a diverse set of maritime images and videos collected in various maritime environments around Singapore. For our experiments, we utilized the visible-light video portion of the dataset, which contains 37 available videos. From these videos, we extracted 17,255 frames (1920×1080 pixels) that include the ground truth information for the sea horizon line. Compared to the MU-SID, the SMD presents additional challenges due to the presence of more obstructions that partially occlude the sea horizon line. These obstructions make the task of horizon extraction more complex and serve as a more stringent test for the accuracy and robustness of sea horizon detection algorithms.

3.1.2. Angular Error

In Section 2.4.2, we discussed the method for calculating the roll angle in camera extrinsic estimation, where the angle of the horizon line determines the result. Therefore, the angular accuracy of the detected horizon line directly impacts the subsequent estimation of camera extrinsic parameters.

To evaluate the angular accuracy of the horizon line detection, we calculated the angular error between the detected horizon line normal vector and the ground truth horizon line normal vector. The calculation formula is as follows:

$$E_{\theta} = \cos^{-1} \left(\frac{\vec{n}_{pred} \cdot \vec{n}_{gt}}{\|\vec{n}_{pred}\| \|\vec{n}_{gt}\|} \right) \quad (17)$$

where $\vec{n}_{pred}$ and $\vec{n}_{gt}$ represent the normal vectors of the detected and ground truth sea horizon lines, respectively.

3.1.3. Position Error

Similar to the sea horizon line angle, the position of the sea horizon line also directly impacts the pitch angle parameter in camera extrinsic estimation. To evaluate the positional accuracy of sea horizon line detection, we selected key points (e.g., the center point of the horizon line) in the image for calculations. The vertical position error for each key point is calculated as:

$$E_{pos} = \Delta v = |v - v_{gt}| \quad (18)$$

where v and v_{gt} denote the vertical coordinates of the predicted sea horizon line and the ground truth horizon line at the center point, respectively.

3.1.4. Comprehensive Evaluation

To assess the overall performance of the sea horizon line detection algorithm, we introduce a comprehensive error metric E_{all} , which combines accuracy with angular and positional errors, weighted by their relative contributions and normalized by maximum tolerances. The formula for E_{all} is defined as:

$$E_{all} = Acc \times \left(w_{\theta} \cdot \frac{\frac{1}{N_{Acc}} \sum_{i=1}^{N_{Acc}} E_{\theta_i}}{T_{\theta}} + w_{pos} \cdot \frac{\frac{1}{N_{Acc}} \sum_{i=1}^{N_{Acc}} E_{pos_i}}{T_{pos}} \right) + (1 - Acc) \quad (19)$$

Here, Acc represents the proportion of samples where both angular error E_{θ} and positional error E_{pos} are within their respective maximum tolerances T_{θ} (2°) and T_{pos} (5 pixels). N_{Acc} is the number of samples classified as accurate results. The weights w_{θ} and w_{pos} reflect the relative importance of angular and positional errors, which are both set to 0.5 in our experiments. The term $1 - Acc$ penalizes the proportion of samples that exceed the user-predefined tolerances, accounting for inaccuracies in detection. A smaller E value indicates better performance in horizon line extraction.

3.1.5. Horizon Detection Results

We compared our proposed horizon line detection method with Grounded SAM against the widely used traditional approaches, namely the Canny edge detection combined with a RANSAC-based line fitting without Grounded SAM, on the public MU-SID dataset. The experimental results are presented in Table 1.

Table 1. Sea horizon detection results of the traditional method (Canny and RANSAC without Grounded SAM) and our method (Canny and RANSAC with Grounded SAM) on MU-SID and SMD datasets. The best results are *italicized*.

Dataset	Method	↓Average E_{θ} ($^{\circ}$)	↓Average E_{pos} (pixels)	↑ Acc	↓ E_{all}
MU-SID	Traditional	0.077	0.786	0.10	0.91
	Ours	<i>0.053</i>	<i>0.995</i>	<i>0.75</i>	<i>0.33</i>
SMD	Traditional	0.871	3.147	0.14	0.94
	Ours	<i>0.342</i>	<i>1.599</i>	<i>0.53</i>	<i>0.60</i>

As shown in Table 1, our detection method improves Acc from 10% to 75% on MU-SID and from 14% to 53% on SMD compared to the traditional method, which proves that using Grounded SAM to extract ROIs for sea horizon regions is helpful in detecting correct sea horizon lines. This improvement prevents the sea horizon lines' edges from being more blurred in conditions such as bad weather, causing other potential line segments in the image to be mistakenly detected as sea horizon lines. Furthermore, the angular (E_{θ}) and positional (E_{pos}) errors remain low, which ensures accurate extrinsic parameter estimation,

including both roll and pitch. Since the same detection methods (Canny and RANSAC) are applied to both the correctly identified sea horizons in our method and the traditional method, it is a normal phenomenon that the E_{pos} value of the traditional method slightly surpasses ours in some cases. The overall error (E_{all}) is also reduced from 0.91 to 0.33 on MU-SID and from 0.94 to 0.60 on SMD, indicating our method's reliability in sea horizon detection and extrinsic parameter estimation.

3.2. Evaluation of Maritime Vessel Instance Segmentation

In this section, the performance of maritime vessel instance segmentation algorithms is evaluated, focusing on a comparison between Grounded SAM and the previously dominant segmentation method, Mask R-CNN [48], and its improved version, GM R-CNN [49]. Additionally, segmentation results are visually presented.

3.2.1. Vessel Segmentation Dataset

To evaluate the methods, we use the MariShipInsSeg dataset for marine vessel instance segmentation, annotated by Sun et al. [50]. This dataset contains 4000 visible-light marine vessel images with 8413 labeled instances. The dataset provides standard annotation files in the COCO format, making it suitable for benchmarking instance segmentation methods [50].

3.2.2. Metrics

The segmentation performance was evaluated using standard metrics derived from the COCO evaluation framework, including:

- AP (Average Precision): measures the mean precision across multiple Intersection-over-Union (IoU) thresholds (from 0.5 to 0.95, with increments of 0.05);
- AP_{50} and AP_{75} : precision at an IoU threshold of 0.50 and 0.75, representing lenient matching and stricter matching, respectively; and
- AP_S , AP_M , and AP_L : precision for small (less than 32^2 pixels), medium (between 32^2 pixels and 96^2 pixels), and large (greater than 96^2 pixels) instances, classified based on object size.

3.2.3. Vessel Segmentation Results

Based on the evaluation metrics described above, we performed quantitative comparisons with the instance segmentation performance of Grounded SAM, Mask R-CNN, and GM R-CNN. Since the authors did not release the source code of GM R-CNN, we directly compared our results to their experimental outcomes (conducted on the same MariShipInsSeg dataset) reported in [45]. The detailed results are presented in Table 2.

Table 2. Instance segmentation results on MariShipInsSeg dataset. The best results are *italicized*.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Size (MB)	Time (ms)
Mask R-CNN [48]	46.76	82.15	48.12	28.82	49.02	62.94	1100.0	163
GM R-CNN [49]	48.62	82.95	48.13	29.11	49.11	66.44	1189.9	165
Grounded SAM [42]	52.92	89.04	56.22	31.68	54.20	69.70	3340.3	961

From Table 2, it can be observed that Grounded SAM outperforms the other two methods, Mask R-CNN and GM R-CNN, across all evaluation metrics. Grounded SAM demonstrates superior results under both lenient and strict conditions, excelling at AP_{50} and AP_{75} . Additionally, the Grounded SAM also presents advantages in segmenting objects of various sizes, achieving the highest scores for small, medium, and large targets. While Grounded SAM has a larger model size and longer execution time for every frame, no

additional training is required to support its zero-shot capabilities. It ensures stronger overall performance with relatively higher computational costs and slower response times.

The segmentation results of Grounded SAM are illustrated in Figure 8, which shows its ability to delineate maritime vessels under diverse conditions accurately.

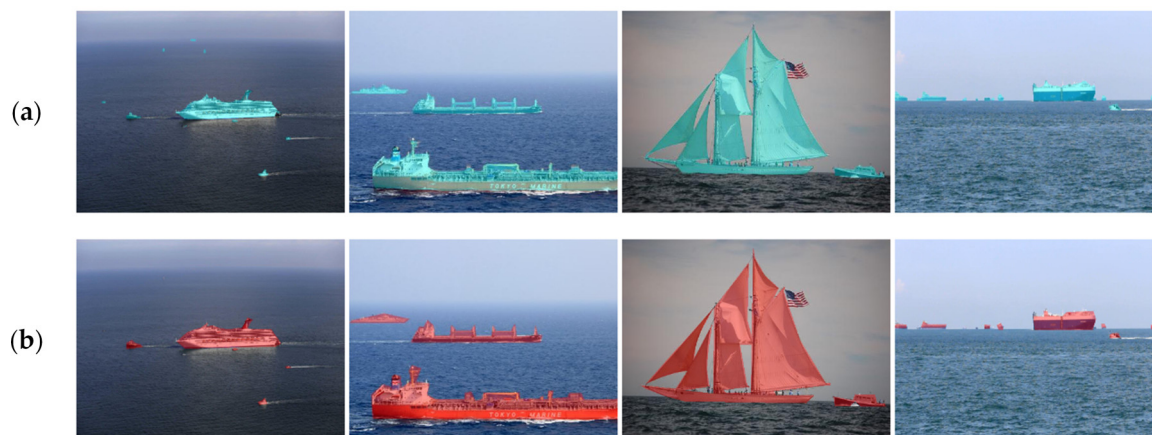


Figure 8. Segmentation results from Grounded SAM: (a) ground truth masks and (b) prediction masks.

3.3. Analysis of Monocular Distance Estimation

In the previous section, we derived the calculation method for the distance d from the pixel center (u, v) to the world coordinate $(X_w, 0, Z_w)$. When considering the range of the pixel boundary, the four corner points $(u \pm 0.5, v \pm 0.5)$ are substituted into the above method (Equations (5)–(10)) to calculate their respective distance values d_i ($i = 1, 2, 3, 4$). The maximum pixel-wise distance error Δd_{max} can then be defined as the maximum absolute difference between the distance at the pixel center and the d_i values, expressed as:

$$\Delta d_{max} = \max_{i=1,2,3,4} |d - d_i| \quad (20)$$

Based on the ratio of Δd_{max} to the distance value d , the maximum relative error (RE) for the pixel's distance estimation can be calculated as:

$$RE = \frac{\Delta d_{max}}{d} \quad (21)$$

In this study, the experimental conditions are meticulously established to ensure the accuracy and reliability of the monocular distance estimation method. The camera was calibrated using Zhang's calibration method, resulting in the intrinsic matrix K , as presented in Equation (22). Furthermore, based on the position of the sea horizon line in the case image and the extrinsic parameter estimation method (Equations (11)–(15)) detailed in Section 2.4.2, the camera's roll and pitch angles were accurately determined as $\theta = 2.012^\circ$ and $\phi = -2.326^\circ$, respectively. These intrinsic and extrinsic parameters serve as fundamental inputs for the subsequent error analysis.

$$K = \begin{pmatrix} 2362.73 & 0 & 1246.48 \\ 0 & 2363.08 & 1054.59 \\ 0 & 0 & 1 \end{pmatrix} \quad (22)$$

In Figure 9a, we present an error analysis of the distance estimation for individual pixel points within the sea surface region of the image. The purpose of this experiment is to analyze the error behavior at various distances under typical camera setup conditions, providing numerical references for the applicable monitoring range of the framework.

Moreover, under the same camera parameters, the error behavior across different images is entirely consistent. Therefore, we limit the discussion here to the relative error in distance estimation on a single image. The analysis shows that the single-pixel error progressively increases with distance, and the error growth rate accelerates as pixels approach the horizon line. Specifically, Figure 9b illustrates the RE trend for a column of pixels at $u = 1000$ with respect to pixel v . As shown in Table 3, at a distance of 500 m from the camera, the RE reaches 3.75%; at 1000 m, it escalates to 7.57%; and at 2000 m, it further increases to 18.21%, with the error continuing to grow as the distance extends beyond 2000 m. Based on this analysis, the optimal range of the monocular distance estimation method proposed in this study is confined to a 2000 m monitoring area for the accuracy requirements, which generally meet the demands of offshore infrastructure safety monitoring and warning scenarios in coastal waters.

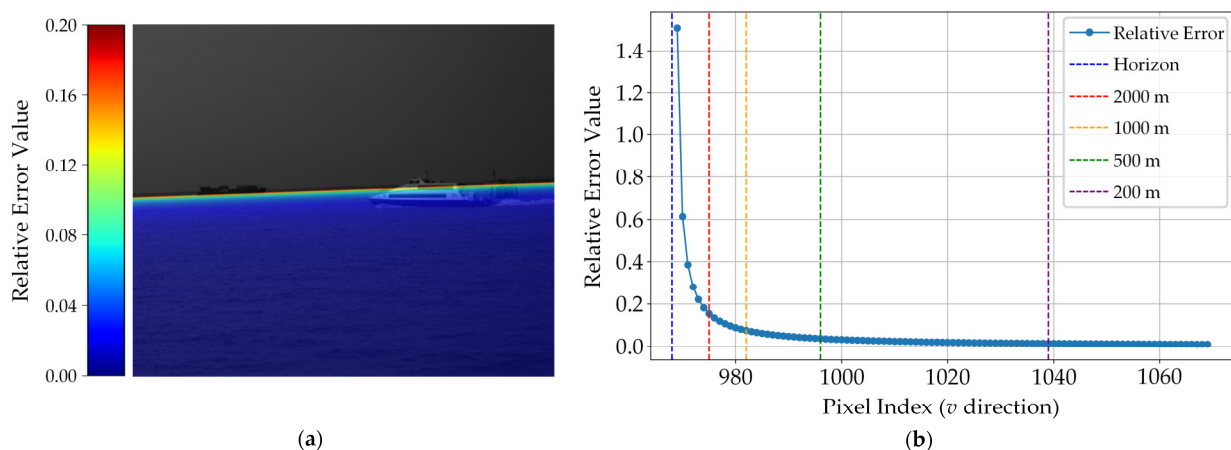


Figure 9. Relative error of distance in the monitoring area. (a) RE map ($RE < 0.20$): the color gradient represents relative error levels, with blue indicating more minor errors and red indicating more significant errors. (b) Relative error vs. pixel index v ($u = 1000$): the graph shows the trend of the relative error increasing as the target moves farther from the camera and closer to the horizon line. These results demonstrate that the relative error of monocular distance estimation grows with increasing distance and proximity to the horizon.

Table 3. Analysis of distance estimation ($u = 1000$).

Distance d (m)	Pixel Index v	Δd_{max} (m)	RE
100	1110	0.74	0.74%
200	1039	2.94	1.47%
500	996	18.73	3.75%
1000	982	75.70	7.57%
1500	977	178.85	11.92%
2000	974	364.20	18.21%
2500	973	552.35	22.09%
3000	972	842.55	28.09%

3.4. Application and Case Study

To better validate and deploy the proposed methods in practical applications, we designed and implemented a prototype of offshore infrastructure safety monitoring and warning system coupled with the proposed algorithms (i.e., sea horizon detection, vessel instance segmentation, and monocular distance estimation).

We collected video data at Baoyang Wharf in Shanghai, China, an area characterized by high vessel traffic and complex maritime conditions. A professional industrial-grade monocular camera (MV-CS050-10GC HIKRobot, Hangzhou, China) was installed

on a fixed platform to collect video streams, ensuring clear horizon visibility and stable data acquisition.

Considering that the monocular vision analysis is a compute-intensive task, all the proposed algorithms were implemented by standalone Python scripts running on the server side. Video streams were automatically processed through a unified pipeline that integrates the functions of horizon line detection, vessel segmentation, and distance estimation. Given the fact of potential camera shifts, it should be highlighted that periodic sea horizon re-detection is performed to ensure reliable parameter updates and minimize mis-detections. Processed frames and all extracted information are transmitted to the client side for real-time visualization and monitoring.

To ensure intuitive and efficient maritime monitoring for end users, we design a user-friendly front-end interface as a web-based application under the lightweight progressive Vue.js framework using HTML5, JavaScript, and cascade style sheet (CSS) in a Visual Studio Code environment, integrated with the Leaflet.js, and WebRTC libraries, as shown in Figure 10. The developed prototype provides the functions of the visualization of real-time video streaming, maritime target detection, safety evaluation and warning, and events statistics, enabling intuitive maritime surveillance and efficient decision-making.



Figure 10. User interface of the web client.

1. Real-time video streaming (WebRTC): the center of the interface displays live video feeds of the monitored area, streamed via WebRTC [51]. Alert zones are overlaid on the video to highlight the spatial relationship between detected vessels and restricted areas, offering operators clear and actionable insights.
2. Safety evaluation and warning: the top-left corner of the interface provides the warning area setting information, including the distance from the vessel to the offshore infrastructure that triggers the different levels of alarms. This part complements the alarm visualization overlay layer on the video stream, helping operators assess the level of risk quickly. In this case study, 200 m, 500 m, and 1000 m are set and can be modified by the users at any time and updated to the layers of the video stream automatically. Once a vessel enters the guarded area, different levels of alarms will be triggered, providing early warning and enabling users to make quick responses.

3. Maritime map visualization (Leaflet.js): on the right side of the interface, 2D maritime maps built with Leaflet.js display vessel positions and distances (calculated using monocular distance estimation) and restricted zones. The maps support panning and zooming, enabling operators to monitor geospatial data in real time.
4. Events statistics: the bottom-left corner logs historical alert data, providing operators with a timeline of past safety incidents for post-event analysis and strategy refinement.

4. Conclusions

This study presents a monocular vision-based safety monitoring framework designed for offshore infrastructures, integrating cutting-edge computer vision technologies for real-time maritime surveillance. The proposed framework utilizes the Grounded SAM for vessel detection and segmentation and a horizon-based self-calibration method. Evaluations of public datasets confirmed significant improvements in segmentation and horizon extraction accuracy, demonstrating the framework's reliability and practicality and the contribution of Grounded SAM in improving the existing methods. Moreover, the proposed framework and methodology can be extended to similar applications, providing valuable reference and inspiration for researchers in the field. Another notable feature of this framework is its cost-effectiveness, making it accessible for more applications: (1) lower hardware requirements: the framework uses cheaper hardware compared to many other solutions, such as radar or binocular vision systems; (2) reduced installation and maintenance costs: by utilizing a monocular optical camera in combination with self-calibration algorithms, the framework simplifies installation and minimizes maintenance effort compared to more complex sensor setups; and (3) lower data processing and labor costs: Grounded SAM's zero-shot capability eliminates the need for additional training specific to each application, saving both time and human resources.

Additionally, we develop a web-based monitoring platform that provides real-time video streaming, vessel geo-location visualization, and intrusion alerts, enabling intuitive and efficient monitoring of sea areas near offshore infrastructures. The platform was developed using a lightweight browser-based architecture with minimal hardware requirements on the client side to ensure convenience and usability. In summary, this research offers a practical, scalable, and supplemental solution for offshore infrastructure safety monitoring, contributing to improved situational awareness and safety in complex marine environments.

However, this research has some limitations, namely, the reduced accuracy of distance estimation at long distances and the large size and high computational requirements of the visual foundation model. Future work will focus on addressing these challenges by incorporating multi-sensor fusion to improve distance estimation accuracy and robustness and by optimizing models to reduce computational overhead while maintaining scalability and performance.

Author Contributions: Conceptualization, S.X. and R.Q.; methodology, S.X.; software, S.X.; validation, S.X.; writing—original draft preparation, S.X.; writing—review and editing, S.X., R.Q., L.M., Z.L. and Y.L.; supervision, Y.L. and R.Q.; project administration, R.Q.; and funding acquisition, R.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Interdisciplinary Project in Ocean Research of Tongji University (2023-1-YB-05).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are not publicly available due to confidentiality agreements associated with the project. Requests for access to the data should be directed to the corresponding author, who will require a detailed explanation of the intended use of the data.

Acknowledgments: The authors are grateful to Rufu Qin for his support with the theoretical aspects of this study. The authors thank Rufu Qin, Lianjiang Ma, Zhenghu Liu, and Yang Lu for their help in reviewing and editing this paper. We also thank the reviewers and editors for their suggestions to improve the quality of this paper.

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

References

1. Costanza, R. The ecological, economic, and social importance of the oceans. *Ecol. Econ.* **1999**, *31*, 199–213. [\[CrossRef\]](#)
2. Sala, E.; Mayorga, J.; Bradley, D.; Cabral, R.B.; Atwood, T.B.; Auber, A.; Cheung, W.; Costello, C.; Ferretti, F.; Friedlander, A.M.; et al. Protecting the global ocean for biodiversity, food and climate. *Nature* **2021**, *592*, 397–402. [\[CrossRef\]](#) [\[PubMed\]](#)
3. van Elden, S.; Meeuwig, J.J.; Hobbs, R.J.; Hemmi, J.M. Offshore Oil and Gas Platforms as Novel Ecosystems: A Global Perspective. *Front. Mar. Sci.* **2019**, *6*, 548. [\[CrossRef\]](#)
4. Su, X.; Wang, X.; Xu, W.; Yuan, L.; Xiong, C.; Chen, J. Offshore Wind Power: Progress of the Edge Tool, Which Can Promote Sustainable Energy Development. *Sustainability* **2024**, *16*, 7810. [\[CrossRef\]](#)
5. Zhou, X.; Zhao, X.; Zhang, S.; Lin, J. Marine Ranching Construction and Management in East China Sea: Programs for Sustainable Fishery and Aquaculture. *Water* **2019**, *11*, 1237. [\[CrossRef\]](#)
6. Buck, B.H.; Troell, M.F.; Krause, G.; Angel, D.L.; Grote, B.; Chopin, T. State of the Art and Challenges for Offshore Integrated Multi-Trophic Aquaculture (IMTA). *Front. Mar. Sci.* **2018**, *5*, 165. [\[CrossRef\]](#)
7. Bailey, K.; Steinberg, C.; Davies, C.; Galibert, G.; Hidas, M.; McManus, M.A.; Murphy, T.; Newton, J.; Roughan, M.; Schaeffer, A. Coastal Mooring Observing Networks and Their Data Products: Recommendations for the Next Decade. *Front. Mar. Sci.* **2019**, *6*, 180. [\[CrossRef\]](#)
8. Mendoza, J.; Nielsen, J.S.; Sørensen, J.D.; Köhler, J. Structural reliability analysis of offshore jackets for system-level fatigue design. *Struct. Saf.* **2022**, *97*, 102220. [\[CrossRef\]](#)
9. Syed Ahmad, S.Z.A.; Abu Husain, M.K.; Mohd Zaki, N.I.; Mukhlis, N.A.; Mat Soom, E.; Azman, N.U.; Najafian, G. Offshore Structural Reliability Assessment by Probabilistic Procedures—A Review. *J. Mar. Sci. Eng.* **2021**, *9*, 998. [\[CrossRef\]](#)
10. Dai, W.; Yu, Y. Data Analysis of Safety Assessment and Early Warning for Offshore Platform. *J. Coast. Res.* **2020**, *106*, 536–539. [\[CrossRef\]](#)
11. Ye, H.; Jiang, C.; Zu, F.; Li, S. Design of a Structural Health Monitoring System and Performance Evaluation for a Jacket Offshore Platform in East China Sea. *Appl. Sci.* **2022**, *12*, 12021. [\[CrossRef\]](#)
12. Mandal, S.; Nagarajan, V.; Sha, O.P. Navigational Safety and Traffic Pattern Analysis Using AIS Data on the Western Coast of India. *Curr. Sci.* **2018**, *114*, 2473–2481. [\[CrossRef\]](#)
13. Chen, P.; Li, M.; Mou, J. A Velocity Obstacle-Based Real-Time Regional Ship Collision Risk Analysis Method. *J. Mar. Sci. Eng.* **2021**, *9*, 428. [\[CrossRef\]](#)
14. Ma, W.; Zhu, Y.; Grifoll, M.; Liu, G.; Zheng, P. Evaluation of the Effectiveness of Active and Passive Safety Measures in Preventing Ship–Bridge Collision. *Sensors* **2022**, *22*, 2857. [\[CrossRef\]](#)
15. Kao, S.-L.; Chang, K.-Y.; Hsu, T.-W. A Marine GIS-based Alert System to Prevent Vessels Collision with Offshore Platforms. *J. Mar. Sci. Technol.* **2021**, *29*, 597–608. [\[CrossRef\]](#)
16. Zhang, X.; Fu, X.; Xiao, Z.; Xu, H.; Qin, Z. Vessel Trajectory Prediction in Maritime Transportation: Current Approaches and Beyond. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 19980–19998. [\[CrossRef\]](#)
17. Wang, X.; Zhang, S. Evaluation of multipath signal loss for AIS signals transmitted on the sea surface. *Ocean Eng.* **2017**, *146*, 9–20. [\[CrossRef\]](#)
18. Androjna, A.; Pavić, I.; Gucma, L.; Vidmar, P.; Perkovič, M. AIS Data Manipulation in the Illicit Global Oil Trade. *J. Mar. Sci. Eng.* **2023**, *12*, 6. [\[CrossRef\]](#)
19. Cope, S.; Tougher, B.; Zetterlind, V.; Gilfillan, L.; Aldana, A. Building a Practical Multi-Sensor Platform for Monitoring Vessel Activity near Marine Protected Areas: Case Studies from Urban and Remote Locations. *Remote Sens.* **2023**, *15*, 3216. [\[CrossRef\]](#)
20. Liu, R.W.; Guo, Y.; Nie, J.; Hu, Q.; Xiong, Z.; Yu, H.; Guizani, M. Intelligent Edge-Enabled Efficient Multi-Source Data Fusion for Autonomous Surface Vehicles in Maritime Internet of Things. *IEEE Trans. Green Commun. Netw.* **2022**, *6*, 1574–1587. [\[CrossRef\]](#)
21. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S. A review of video surveillance systems. *J. Vis. Commun. Image Represent.* **2021**, *77*, 103116. [\[CrossRef\]](#)

22. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. [[CrossRef](#)]
23. Gu, W.; Bai, S.; Kong, L. A review on 2D instance segmentation based on deep neural networks. *Image Vis. Comput.* **2022**, *120*, 104401. [[CrossRef](#)]
24. Wang, J.; Guan, Y.; Kang, Z.; Chen, P. A Robust Monocular and Binocular Visual Ranging Fusion Method Based on an Adaptive UKF. *Sensors* **2024**, *24*, 4178. [[CrossRef](#)]
25. Lyu, H.; Shao, Z.; Cheng, T.; Yin, Y.; Gao, X. Sea-Surface Object Detection Based on Electro-Optical Sensors: A Review. *IEEE Intell. Transp. Syst. Mag.* **2023**, *15*, 190–216. [[CrossRef](#)]
26. Zeng, F.; Xiao, C.; Wang, W.; Gan, L.; Yuan, H.; Li, Q. A Novel Ship-Height Measurement System Based on Optical Axes Collinear Stereo Vision. *IEEE Sens. J.* **2023**, *23*, 6282–6291. [[CrossRef](#)]
27. Liang, H.; Ma, Z.; Zhang, Q. Self-Supervised Object Distance Estimation Using a Monocular Camera. *Sensors* **2022**, *22*, 2936. [[CrossRef](#)]
28. Gladstone, R.; Moshe, Y.; Barel, A.; Shenhav, E. Distance estimation for marine vehicles using a monocular video camera. In Proceedings of the 2016 24th European Signal Processing Conference (EUSIPCO), Budapest, Hungary, 29 August–2 September 2016; pp. 2405–2409.
29. Kang, B.-S.; Jung, C.-H. Detecting Maritime Obstacles Using Camera Images. *J. Mar. Sci. Eng.* **2022**, *10*, 1528. [[CrossRef](#)]
30. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep Learning for Computer Vision: A Brief Review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [[CrossRef](#)]
31. 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](#)]
32. Liu, T.; Zhou, B.; Zhao, Y.; Yan, S. Ship Detection Algorithm based on Improved YOLO V5. In Proceedings of the 2021 6th International Conference on Automation, Control and Robotics Engineering (CACRE), Dalian, China, 15–17 July 2021; pp. 483–487.
33. Dablain, D.; Jacobson, K.N.; Bellinger, C.; Roberts, M.; Chawla, N.V. Understanding CNN fragility when learning with imbalanced data. *Mach. Learn.* **2024**, *113*, 4785–4810. [[CrossRef](#)]
34. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; pp. 3992–4003.
35. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *Computer Vision—ECCV 2024 Proceedings of the 18th European Conference, Milan, Italy, 29 September–4 October 2024*; Springer: Berlin, Germany, 2025; pp. 38–55.
36. Chen, F.; Chen, L.; Han, H.; Zhang, S.; Zhang, D.; Liao, H. The ability of Segmenting Anything Model (SAM) to segment ultrasound images. *BioSci. Trends* **2023**, *17*, 211–218. [[CrossRef](#)] [[PubMed](#)]
37. Mazurowski, M.A.; Dong, H.; Gu, H.; Yang, J.; Konz, N.; Zhang, Y. Segment anything model for medical image analysis: An experimental study. *Med. Image Anal.* **2023**, *89*, 102918. [[CrossRef](#)] [[PubMed](#)]
38. Yan, Z.; Li, J.; Li, X.; Zhou, R.; Zhang, W.; Feng, Y.; Diao, W.; Fu, K.; Sun, X. RingMo-SAM: A Foundation Model for Segment Anything in Multimodal Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5625716. [[CrossRef](#)]
39. Arbash, E.; De Lima Ribeiro, A.; Thiele, S.; Gnann, N.; Rasti, B.; Fuchs, M.; Ghamisi, P.; Gloaguen, R. Masking Hyperspectral Imaging Data with Pretrained Models. In Proceedings of the 2023 13th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), Athens, Greece, 31 October–2 November 2023; pp. 1–5.
40. Barakat, M.; Chung, G.C.; Lee, I.E.; Pang, W.L.; Chan, K.Y. Detection and Sizing of Durian using Zero-Shot Deep Learning Models. *Int. J. Technol.* **2023**, *14*, 1206–1215. [[CrossRef](#)]
41. Ma, B.; Xu, W. Efficient Fine Tuning for Fashion Object Detection. *Sensors* **2023**, *23*, 6083. [[CrossRef](#)]
42. Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. *arXiv* **2024**, arXiv:2401.14159. [[CrossRef](#)]
43. Shi, B.Q.; Zhao, C.G.; Hao, Z.J.; Xu, M.C.; Bian, C.T. Implementation of the Real-Time Parameters Monitoring System for Underground Equipments Based on Panoramic Images. *Appl. Mech. Mater.* **2012**, *155–156*, 825–830. [[CrossRef](#)]
44. Zhang, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [[CrossRef](#)]
45. Freeman, H. On the Encoding of Arbitrary Geometric Configurations. *IRE Trans. Electron. Comput.* **1961**, *EC-10*, 260–268. [[CrossRef](#)]
46. Zhao, L.; Shi, G. A method for simplifying ship trajectory based on improved Douglas–Peucker algorithm. *Ocean Eng.* **2018**, *166*, 37–46. [[CrossRef](#)]
47. Hashmani, M.A.; Umair, M. A Novel Visual-Range Sea Image Dataset for Sea Horizon Line Detection in Changing Maritime Scenes. *J. Mar. Sci. Eng.* **2022**, *10*, 193. [[CrossRef](#)]
48. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [[CrossRef](#)] [[PubMed](#)]

49. Sun, Y.; Su, L.; Luo, Y.; Meng, H.; Li, W.; Zhang, Z.; Wang, P.; Zhang, W. Global Mask R-CNN for marine ship instance segmentation. *Neurocomputing* **2022**, *480*, 257–270. [[CrossRef](#)]
50. 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 *Computer Vision—ECCV 2014 Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014, Proceedings, Part V*; Springer: Berlin, Germany, 2014; pp. 740–755.
51. Blum, N.; Lachapelle, S.; Alvestrand, H. WebRTC. *Commun. ACM* **2021**, *64*, 50–54. [[CrossRef](#)]

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