Journal of
*Marine Science and Engineering*

MDPI

*Article*

# Research on Visual Perception for Coordinated Air–Sea through a Cooperative USV-UAV System

**Chen Cheng [1,*], Dong Liu [2], Jin-Hui Du [1] and Yong-Zheng Li [1]**

[1] School of Naval Architecture and Ocean Engineering, Jiangsu University of Science and Technology, Zhenjiang 212013, China; 212241801310@stu.just.edu.cn (J.-H.D.); justyzli@163.com (Y.-Z.L.)
[2] Yancheng Maritime Safety Administration of People's Republic of China, Yancheng 224008, China; ccccmaofei@163.com
* Correspondence: 202200000103@just.edu.cn

**Abstract:** The identification and classification of obstacles in navigable and non-navigable regions, as well as the measurement of distances, are crucial topics of investigation in the field of autonomous navigation for unmanned surface vehicles (USVs). Currently, USVs mostly rely on LiDAR and ultrasound technology for the purpose of detecting impediments that exist on water surfaces. However, it is worth noting that these approaches lack the capability to accurately discern the precise nature or classification of those obstacles. Nevertheless, the limited optical range of unmanned vessels hinders their ability to comprehensively perceive the entirety of the surrounding information. A cooperative USV-UAV system is proposed to ensure the visual perception ability of USVs. The multi-object recognition, semantic segmentation, and obstacle ranging through USV and unmanned aerial vehicle (UAV) perspectives are selected to validate the performance of a cooperative USV-UAV system. The you only look once-X (YOLOX) model, the proportional–integral–derivative-NET (PIDNet) model, and distance measurements based on a monocular camera are utilized to realize these problems. The results indicate that by integrating the viewpoints of USVs and UAVs, a collaborative USV-UAV system, employing the aforementioned methods, can successfully detect and classify different objects surrounding the USV. Additionally, it can differentiate between navigable and non-navigable regions for unmanned vessels through visual recognition, while accurately determining the distance between the USV and obstacles.

**Keywords:** visual perception; cooperative USV-UAV system; YOLOX; PIDNet; monocular camera vision



**Citation:** Cheng, C.; Liu, D.; Du, J.-H.; Li, Y.-Z. Research on Visual Perception for Coordinated Air–Sea through a Cooperative USV-UAV System. *J. Mar. Sci. Eng.* **2023**, *11*, 1978. https://doi.org/10.3390/jmse11101978

Academic Editor: Marco Cococcioni

Received: 18 September 2023
Revised: 8 October 2023
Accepted: 9 October 2023
Published: 12 October 2023



## 1. Introduction

As sensing technology, artificial intelligence algorithms, and intelligent control algorithms continue to advance, the development of intelligent spacecraft continues to advance. In recent years, the demand for unmanned ships has increased, and unmanned surface vehicles (USVs) have become research hotspots for numerous unmanned vehicles. USVs are small, intelligent ships that can navigate autonomously without the need for human operation and automatically complete specific water tasks [1].

The correct capture of surrounding environmental information is a crucial requirement for ensuring the safe and autonomous navigation of USVs in complex aquatic environments. USVs are able to effectively navigate through a dynamic environment while promptly avoiding obstacles and accurately identifying water surface target objects. USVs typically employ a range of sensors for various purposes, including radar navigation, millimeter-wave radars, LiDAR, sonar, and vision sensors. The initial sensors exhibit several limitations, including elevated costs, notable environmental ramifications, and a restricted capacity to perceive and gather comprehensive environmental data, hence impeding the acquisition of additional information [2,3]. Visual sensors have the potential to enhance the perception capabilities of USVs, allowing them to effectively observe a larger

expanse of water and gather valuable information about the aquatic environment. This, in turn, enables USVs to gain a thorough understanding of water navigation situations. The advancement of image processing technology and deep learning-based detection and recognition technology has led to a notable enhancement in the perceptual accuracy of vision systems for unmanned ships. Visual sensors have found extensive applications in diverse mobile intelligent platforms, assuming a crucial function in the detection and recognition of water targets, monitoring aquatic environments, and mitigating potential collisions with unmanned vessels [4].

The limited installation height of vision sensors on USVs poses challenges in achieving a comprehensive perception of the surrounding environment, particularly in detecting obstructed areas ahead. Consequently, this creates a blind spot inside the field of vision. Simultaneously, the close proximity of visual sensors to the water surface introduces a susceptibility to environmental influences such as water ripples, reflections, and illumination, hence presenting challenges in image processing and target recognition. Unmanned Aerial Vehicles (UAVs) possess a notable advantage in terms of their elevated flying height, which significantly enhances their visual perception range. Consequently, the images captured by UAVs are subjected to lesser influence from the surrounding water environment [5]. The limited cargo capacity of UAVs necessitates a reliance on the battery module for flying power, resulting in a relatively short flight duration that restricts its ability to carry out prolonged, complex tasks. In addition, it should be noted that UAVs currently possess limited processing and computing capabilities, rendering them inadequate for executing intricate visual processing tasks within UAV systems.

Consequently, numerous researchers have directed their focus towards investigating the collaborative systems of UAVs and USVs, aiming to leverage their individual strengths in order to address the challenges posed by the limited endurance of UAVs and the restricted perception range of USVs. This paper aims to investigate the research pertaining to visual-based environment perception in collaborative UAV and USV systems. The findings of this study will offer fundamental technological assistance for the advancement and implementation of autonomous navigation, maritime supervision, and cruise control in unmanned maritime vessels.

The organization of this paper can be summarized as follows: Section 2 lists the recent related works. In Section 3, the several methods are introduced. Section 4 demonstrates the experimental setup and data process. The results are discussed in Section 5. Finally, the conclusion is summarized in Section 6.

## 2. Literature Review

The perception of autonomous ships can be categorized into two distinct aspects based on their perceived content: self-perception and the perception of the external environment. The accurate determination of USVs' own state can be achieved by utilizing GPS positioning sensors and IMU inertial units. The stability and precision of this determination are often dependent on the device's performance [6]. The process of perceiving the external environment primarily relies on the use of diverse sensory mechanisms. The heightened unpredictability of the environment is a significant obstacle in accurately recognizing the exterior surroundings of USVs, hence creating difficulties in ensuring their safe and autonomous navigation. The perception of the aquatic navigation environment can be categorized into two groups based on the various operating methods of sensors: active perception and passive perception [7].

Active perception refers to the act of transmitting signals to the external environment using sensing devices, and subsequently acquiring information about the surrounding environment by receiving the returned signal information. Examples of such sensing equipment are radar navigation and LiDAR sensors [8]. Carlos et al. (2009) incorporated radar technology into the ROAZ USV system in order to facilitate the identification of obstacles and the prevention of collisions [9]. Zhang et al. (2011) employed the Gaussian particle filtering technique to effectively analyze maritime radar data and successfully

accomplish dynamic target tracking [10]. Han et al. (2019) presented a novel technique that combines radar technology with simultaneous location and map building for unmanned ship systems. This algorithm aims to overcome the issue of GPS signal loss in difficult surroundings, ultimately enabling accurate positioning in coastal areas [11]. Esposito et al. (2014) employed LiDAR technology for the purpose of automatically identifying docks, hence facilitating the autonomous docking of USVs [12].

Passive perception mostly pertains to the utilization of visual sensors for the acquisition of information concerning the surrounding navigation environment. The fundamental premise involves capturing visual data of the surrounding environment using visual sensors, followed by the interpretation of the environmental information based on the color and texture characteristics of the captured images [13,14]. Kristan et al. (2014) first utilized either monocular or stereo vision techniques to gather real-time data on the water surface environment. Subsequently, they used a water antenna recognition algorithm to ascertain the precise location of the water antenna. Next, they conducted a search for potential targets in close proximity to the water antenna in order to successfully detect water surface targets [15]. Wang et al. (2015) employed both monocular and binocular vision techniques to achieve the real-time and efficient identification of obstacles on the sea surface. Their approach enables the detection and localization of multiple objects across a distance range spanning from 30 to 100 m [16].

Following this, the integration of deep learning techniques was used in the domain of USV vision with the aim of enhancing the robustness and precision of algorithms pertaining to USV vision. Shi et al. (2019) made enhancements to the single-shot multi-box detector (SSD) algorithm in order to effectively identify and localize impediments and targets in the vicinity of unmanned vessels [17]. Song et al. (2019) put forth an algorithm for real-time obstacle identification. This approach utilized the Kalman filtering method to combine the SSD and Faster RCNN models. The objective of this algorithm was to detect obstacles on the sea surface for USVs [18]. Zhan et al. (2019) introduced a novel network segmentation algorithm that utilized self-learning techniques to identify and classify water and non-water surface regions in visual pictures captured by USVs. This approach aims to enable autonomous collision avoidance capabilities in USVs, hence ensuring the safety of their navigation [19].

The installation height of visual sensors on USVs presents a constraint that not only diminishes the sensing range of these ships, but also renders their vision vulnerable to the effects of water waves and reflections. Consequently, this low installation height poses challenges for the visual processing capabilities of USVs. In recent years, there has been significant advancement in UAV technology, leading to their widespread utilization across several domains. UAVs possess notable maneuverability capabilities and have an extensive perception range that is attributable to their elevated flight altitudes. Due to this rationale, numerous academics have endeavored to engage in collaborative research pertaining to USVS and UAVs in order to accomplish intricate aquatic assignments. Xu and Chen (2022) presented a comprehensive analysis of a multi-agent reinforcement learning (MARL) methodology, specifically designed for UAV clusters. The primary challenges were the assembly and formation maintenance in UAV cluster formation control [20]. Zhang et al. (2019) employed a distributed consistency technique in order to develop and simulate the control algorithm. In relation to the issue of cooperative path-following control between USVs and UAVs [21]. Li et al. (2022) introduced a novel conceptual framework for establishing a coherent and efficient connection between USVs and UAVs. This framework, referred to as the logical virtual ship–logical virtual aircraft guidance principle, aims to facilitate an effective association between these two types of unmanned vehicles [22].

## 3. Methodology

### *3.1. YOLOX*

The predominant techniques employed for target recognition at present encompass SSD, CenterNet, and YOLO. The SSD algorithm is influenced by the anchor notion intro-

duced in Faster R-CNN, wherein individual units establish previous boxes with varying scales or aspect ratios. The anticipated boundary boxes are derived from the past boxes, hence mitigating the challenges encountered during training. CenterNet, alternatively referred to as Objects as Points, has garnered significant attention from users, owing to its very straightforward and refined architecture, robust capability for handling diverse tasks, rapid inference speed, commendable accuracy, and the absence of the necessity for non-maximum suppression (NMS) post-processing. The YOLO algorithm addresses object detection by formulating it as a regression problem. Utilizing an independent end-to-end network, the task involves processing the input data that originate from the original image and generating the corresponding output, which comprises the positions and categories of objects. In comparison to these networks, YOLOX demonstrates superior performance in accomplishing the identical task, while concurrently preserving a highly competitive inference speed. The YOLOX object detection network comprises four components: the input terminal of the model, the Darket53 backbone network, the feature enhancement network neck, and the model prediction [23].

The YOLOX network Incorporates two data augmentation techniques, namely Mosaic and MixUp, at its input end. Additionally, it establishes a Focus structure. Mosaic data, firstly proposed in YOLO4, aim to improve the background of an image through the application of random scaling, cropping, and the arrangement of many photos [24]. MixUp is a supplementary augmentation method that is implemented in conjunction with Mosaic. This strategy significantly improves the quality of photos by merging two images together using a specific fusion coefficient, while minimizing the computational overhead [25]. The Focus structure is designed to extract four distinct feature layers from an image by selecting alternate pixels. These layers are subsequently combined to consolidate width and height information into channel information. This process results in a concatenation of feature layers, increasing the number of channels from three to twelve, thereby quadrupling the channel count.

The backbone serves as the primary architectural framework of YOLOX. Within YOLOX, the prominent feature extraction network employed is CSPDarknet53. CSPDarknet53 is composed of 72 convolutional layers, each possessing a dimension of $3 \times 3$ and a stride of 2. This configuration enables the network to effectively extract features and progressively down-sample the input data. The neck feature fusion structure employed in YOLOX is founded upon three fundamental elements: Feature Pyramid Networks (FPNs), Spatial Pyramid Pooling (SPP), and Path Aggregation Networks (PANs) [26]. The primary components of prediction encompass the decoupled head, anchor-free, label assignment, and loss calculation, which facilitate the execution of classification and regression tasks inside the model.

The primary filtering technique employed in YOLOX is SimOTA. Initially, the anchor boxes undergo a screening process to extract the position IoU matrix [27]. Subsequently, the loss function is computed for the chosen candidate detection boxes and ground truth. The cost function is then determined through a weighted summation of the resulting loss functions, as demonstrated below:

$$C_{ij} = L_{ij}^{clsloss} + \gamma \times L_{ij}^{regloss} \quad (1)$$

where $C_{ij}$ denotes the total loss for a specific bounding box; $L_{ij}^{clsloss}$ is the classification loss, measuring the difference between the predicted and true class labels; and $L_{ij}^{regloss}$ is the regression loss, evaluating the disparity between the predicted and actual bounding box positions.

This paper employs the CBAM attention mechanism to enhance the conventional YOLOX network [28]. In contrast to the original network, the present network exhibits a heightened focus on pertinent characteristics of diminutive entities, hence minimizing the risk of detection oversight. Identifying little objects poses a greater challenge because of their low resolution and limited visual information, in contrast to larger objects. Conse-

quently, the CBAM module is incorporated into the Dark3 module of the shallow network. Attention weights are then derived from both the spatial and channel dimensions. These weights are subsequently multiplied by the feature map ratio of 80 × 80, resulting in an enhanced feature response specifically for small objects. The purpose of the CBAM is accomplished by the utilization of two distinct attention modules: the channel attention module (CAM) for assessing the correlation among channels, and the spatial attention module (SAM) for evaluating the correlation among positions. The structure of the CBAM module can be viewed in Figure 1.

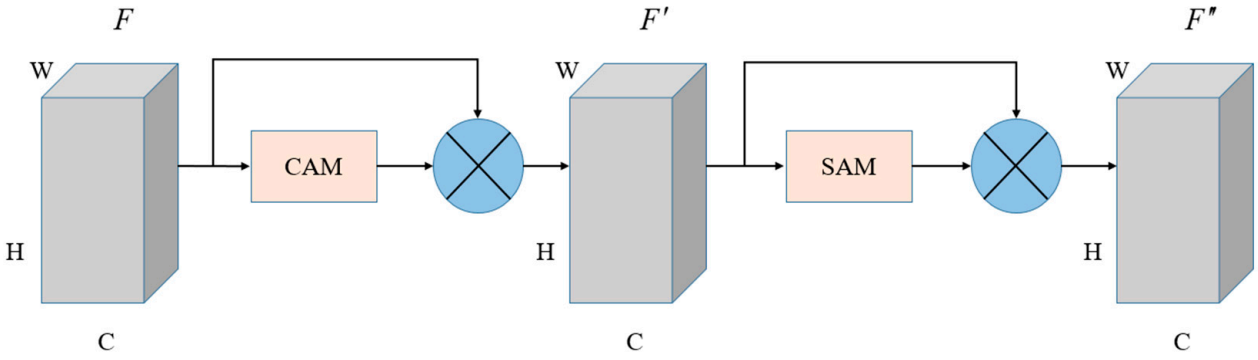


**Figure 1.** The structure of CAM module.

The convolutional attention module initially conducts spatial domain operations, specifically maximum pooling and average pooling, on the input feature map ($F$) with dimensions H × W × C. This process generates two channel information vectors of size 1 × 1 × C. These vectors are subsequently fed into a multi-layer perceptron (MLP) and individually summed. The application of the sigmoid activation function is the final step in obtaining the weight coefficient, $M_c$. This coefficient is then multiplied by the original feature map to derive the channel attention feature map, as shown in Equations (2) and (3).

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{2}$$

$$F' = M_c(F) \otimes F \tag{3}$$

The IoU (Intersection over Union) loss function is employed in YOLOX, which is a widely utilized metric within the domain of object detection. The computation methodology for the IoU is as follows:

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} \tag{4}$$

### 3.2. PIDNet

A proportional (P) controller, an integral (I) controller, and a derivative (D) controller comprise a PID controller, which can be viewed in Figure 2. The PI controller implementation might be expressed as:

$$c_{out}[n] = k_p e[n] + k_i \sum_{i=0}^{n} e[i] \tag{5}$$

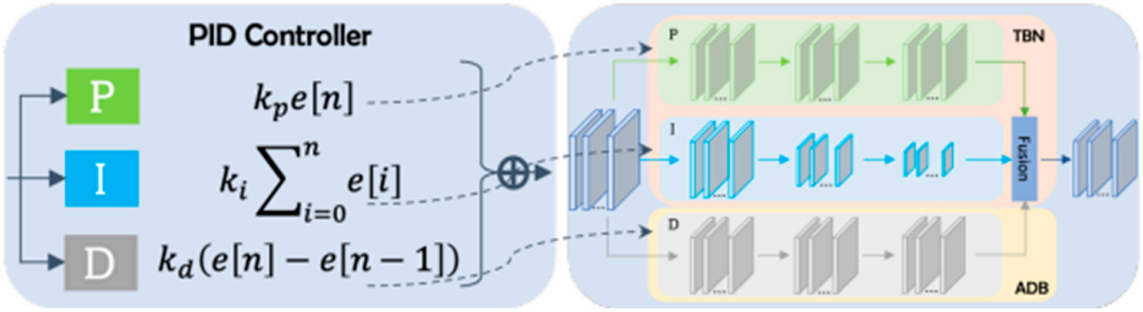


**Figure 2.** The structure of CBAM module.

The *P* controller concentrates on the present signal, whereas the *I* controller gathers all previous signals. Subsequently, the introduction of the *D* controller is implemented, wherein the *D* component assumes a negative value when the signal decreases, acting as a dampening mechanism to mitigate overshooting. In a similar manner, two-branch networks (TBNs) analyze the contextual and intricate information via the utilization of several convolutional layers, both with and without strides. In this particular one-dimensional example, it is worth noting that both the detailed and contextual branches comprise three layers, without the inclusion of batch normalization (BN) and rectified linear units (ReLUs). The output maps can be calculated as

$$O_D[i] = K^D_{i-3} I[i-3] + \ldots + K^D_i I[i] + \ldots + K^D_{i+3} I[i+3] \quad (6)$$

$$O_C[i] = K^C_{i-7} I[i-7] + \ldots + K^C_i I[i] + \ldots + K^C_{i+7} I[i+7] \quad (7)$$

where

$$K^D_i = k_{31}k_{22}k_{13} + k_{31}k_{23}k_{12} + k_{32}k_{21}k_{13} + k_{32}k_{22}k_{12} + k_{32}k_{23}k_{13} + k_{33}k_{21}k_{12} + k_{33}k_{22}k_{11} \quad (8)$$

$$K^C_i = k_{32}k_{22}k_{12} \quad (9)$$

and where $k_{mn}$ refers to the *n-th* value of the kernel in layer *m*.

PIDNet is composed of three branches that have distinct roles: the proportional (*P*) branch is responsible for parsing and preserving detailed information in feature maps with high resolution; the integral (I) branch aggregates context information at both local and global levels to parse long-range dependencies; and the derivative (*D*) branch extracts high-frequency features to predict boundary regions. A semantic head is positioned at the output of the initial Pag module in order to generate an additional semantic loss, denoted as $l_0$, with the aim of enhancing the optimization process of the entire network. Instead of using dice loss, we employ weighted binary cross entropy loss, $l_1$, to address the issue of imbalanced boundary detection. This is because emphasizing the coarse border is desired in order to emphasize the boundary region and increase the characteristics for smaller items. The variables $l_2$ and $l_3$ are used to denote the cross-entropy (CE) loss in our study. Specifically, for $l_3$, we employ the boundary awareness CE loss (Towaki. 2019), which leverages the output of the boundary head to effectively coordinate the tasks of semantic segmentation and boundary detection. This approach enhances the functionality of the Bag module. Therefore, the final loss for PIDNet can be calculated as [29]:

$$Loss = \lambda_0 l_0 + \lambda_1 l_1 + \lambda_2 l_2 + \lambda_3 l_3 \quad (10)$$

### 3.3. Monocular Vision Scale–Distance by USVs

The objective here is to gather images of water surfaces, analyze them to determine the specific area where the target is situated within each image, and afterwards compute the greatest pixel ordinate value of that area, together with its related mean horizontal

coordinate, utilizing the aggregated coordinates as the pixel coordinates for the points of observation. The objective is to determine the depth information of the observation point within the camera coordinate system by utilizing the camera's internal parameters and the geometric relationship of perspective projection. Next, the three-dimensional data of the observation point in the coordinate system of the USV attachment will be determined using rigid body transformation. Subsequently, the distance between the observation point and the USV will be calculated. The range-measuring model utilized in this paper is depicted in Figure 3. This model can be seen as a process that maps the items present in the three-dimensional (3D) scene onto two-dimensional (2D) images using a pinhole camera.

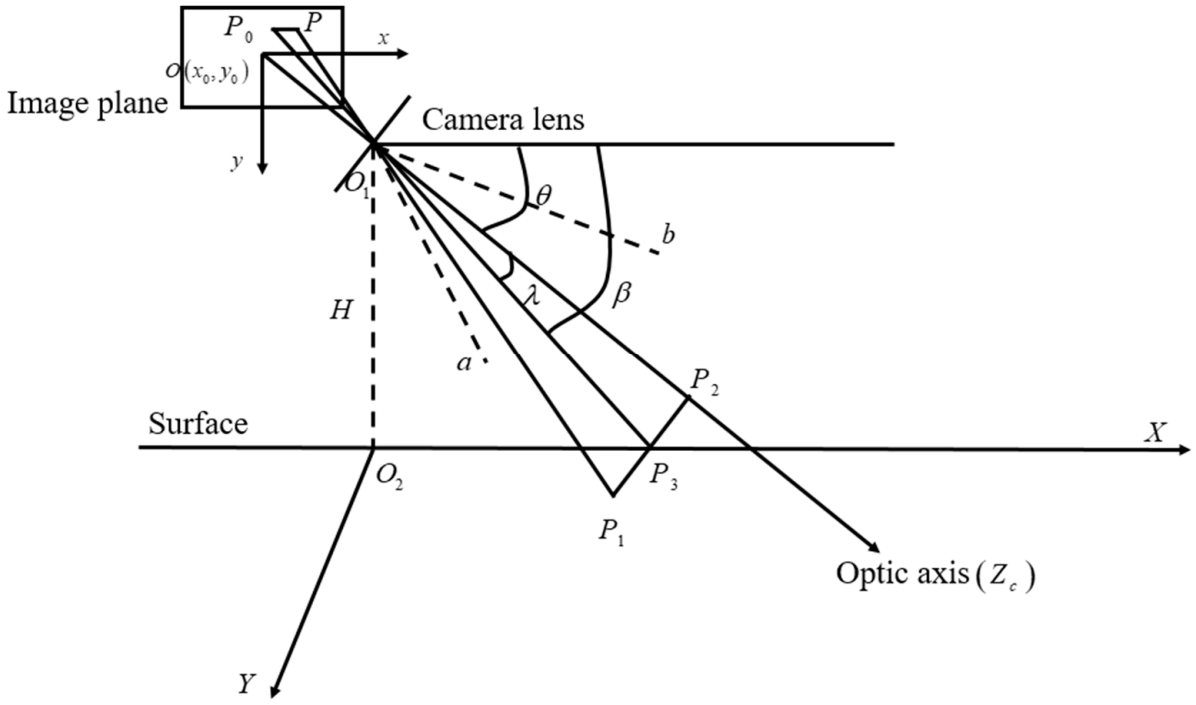


**Figure 3.** Ranging model.

In Figure 3, $xoy$ denotes the image coordinate system while $Z_c$ denotes the $Z$-axis and optical axis of the camera coordinate system; $xo_2y$ represents the $Z = 0$ plane and water surface in the USV coordinate system; $O_1$ is the camera lens; the two dashed lines, $a$ and $b$, represent visual field range; $\theta$ indicates the camera pitch angle. The observation point $P_1$ is positioned at the imaging point $P$ on the image plane, the projection point $P_2$ on the optical axis, and the projection point $P_3$ on the $X$-axis; the point $P_3$ corresponds to the point $P_0$ when projected onto the picture plane.

The distance from the observation point to the center of the USV can be calculated as follows:

$$O_2P_1 = \sqrt{x_s^2 + y_s^2} \tag{11}$$

where $x_s$ and $y_s$ denote the coordinate values of $P_s(x_s, y_s, z_s)$ at the observation point in the USV fitted coordinate system, which can be estimated based on the link between the coordinates of the observation point in the pixel coordinate system and the coordinates in the USV fitted coordinate system:

$$z_c \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} f_x & 0 & u_0 & 0 \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{R} & \mathbf{T} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_s \\ y_s \\ z_s \\ 1 \end{pmatrix} \tag{12}$$

where $f_x$, $f_y$, $u_0$, and $v_0$ indicate the intrinsic camera parameters, which can be calibrated based on Pei (2015) [30]. **R** and **T** denote the rotation matrix and translation matrix from the USV fitted coordinate system to the camera coordinate system, respectively. $z_c$ is the depth coordinate value of the observation point in the camera coordinate system $P_c(x_c, y_c, z_c)$, which is the distance of $O_1P_2$ and can be calculated as

$$z_c = O_1P_2 = O_1P_3 \times \cos\lambda \tag{13}$$

$$O_1P_3 = \frac{H}{\sin(\theta + \lambda)} \tag{14}$$

where $H$ denotes the height of camera; $\lambda$ is the angle between the light at point $P_0$ and the camera's optical axis, which can be shown as follows:

$$\lambda = \arctan\frac{y - y_0}{f} \tag{15}$$

where $y$ denotes the ordinate of point $P$ in the image coordinate system; $f$ is the focal length of the camera. In the pixel coordinate system, $\lambda$ can be calculated as follows:

$$\lambda = \arctan\frac{(v - v_0) \times dy}{f_y \times dy} = \arctan\frac{v - v_0}{f_y} \tag{16}$$

where $v$ denotes the ordinate of point $P$ in the pixel coordinate system; $v_0$ is the vertical axis of the image center; $dy$ is the unit pixel length in the y-direction; and $f_y$ is the normalized focal length.

In order to mitigate the potential influence of measurement errors pertaining to camera height and pitch angle on the accuracy of ranging in situations characterized by uncertainty, this research study presents a calibration technique for the aforementioned camera parameters.

In the event that the sea surface is generally tranquil, it is feasible to approximate it as a flat plane. In Figure 4, $A(x_1, y_1, 0)$ is the intersection point between the plane $z = 0$ in the grid coordinate system and the camera optical axis (the Z-axis of the camera coordinate system), and the coordinate of point $A$ in the camera coordinate system is $C_a(0, 0, z_1)$. $B(x_2, y_2, a)$ is the intersection point between the plane $z = 0$ in the grid coordinate system and the camera optical axis; the coordinate of point $B$ in the camera coordinate system is $C_b(0, 0, z_2)$. The coordinate of the vertical projection point $B'$ on the $z = 0$ plane of intersection $B$ in the grid coordinate system is $(x_2, y_2, 0)$. The pitch angle of the camera is $\angle BAB'$, which can be shown as

$$\angle BAB' = a\cos\frac{AB \cdot AB'}{|AB||AB'|} \tag{17}$$

$$H = z_1 \times \sin\angle BAB' \tag{18}$$

Camera calibration is a process that yields the grid coordinate system, which lies on the calibration plate and is coplanar with the water surface. Additionally, it provides the rotation matrix ($\mathbf{R_1}$) and translation matrix ($\mathbf{T_1}$) of the camera coordinate system. The relationship between the coordinate $\boldsymbol{XX}$ in the grid coordinate system and the corresponding coordinate $\boldsymbol{XX_c}$ in the camera coordinate system can be expressed as follows:

$$XX_c = \mathbf{R}_1^* XX + \mathbf{T}_1 \tag{19}$$


**Figure 4.** Camera height and pitch angle model.

Substitute the coordinate variables of $A$ and $C_a$ into Equation (19):

$$\begin{pmatrix} 0 \\ 0 \\ z_1 \end{pmatrix} = \begin{pmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ y_1 \\ 0 \end{pmatrix} + \begin{pmatrix} T_1 \\ T_2 \\ T_3 \end{pmatrix} \tag{20}$$

Solve Equation (20) to obtain the coordinate values of $A$ and $C_a$. Similarly, the coordinate values of $B$ and $C_b$ can also be calculated.

## 4. Experimental Setup

### *4.1. Data Processing*

Pre-processing actions for images are deemed crucial in accordance with the stipulations of model application settings. The images within the experimental dataset were obtained from a total of 60 videos, each of which had a minimum duration of 60 s. During the cropping process, a proportional comparison is conducted on the target instance. If the resulting area of the target instance is equal to or more than 60% of the original instance area, the instance is retained. Otherwise, it is removed. The dataset has a total of 8588 pictures, with the training set and test set accounting for 80% and 20% of the dataset, respectively. The objective of this study is to examine the feasibility of the cooperative USV and UAV platform architecture. This is achieved by initially categorizing the dataset into five distinct groups, namely ship (representing various types of ships), USV, buoy, building, and people, with an equal proportion among each category.

In addition, 8588 images are also utilized to train and test the PIDNet model. The labelme software is employed for the purpose of annotating all photographs, with a primary focus on labelling distinct regions such as navigable water surfaces, non-navigable skies, and diverse barriers that are present on the water. In order to address intricate marine barriers, a multi-point framing approach is employed to accurately delineate the desired area. The detailed annotation method is shown in Figure 5.

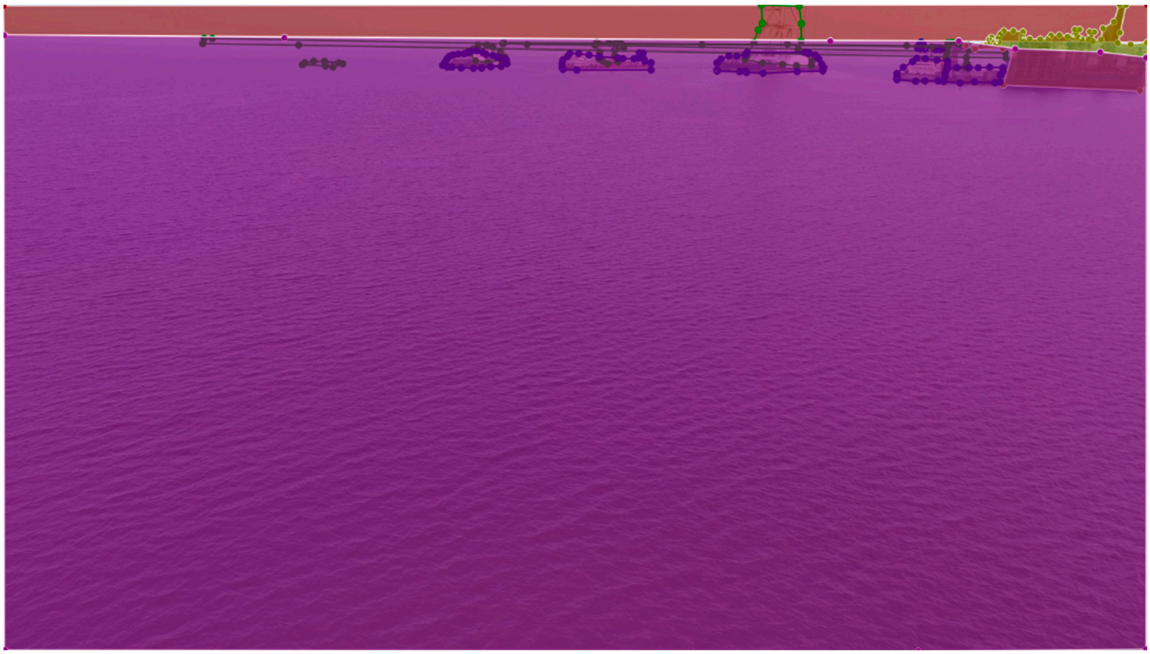

**Figure 5.** Annotation diagram.

### 4.2. Experimental Platform

Jiangsu University of Science and Technology invented and developed the cooperative USV-UAV platform, including an unmanned catamaran and a quadrotor, which can be viewed in Figure 6. Furthermore, this platform is equipped with a USV, with two lithium batteries housed within the hull. Additionally, a satellite communication module is located within the hull compartment, while four cameras are fitted atop the USV.

**Figure 6.** Cooperative USV-UAV platform.

The experimental training computer is configured with the Windows 10 operating system, with an NVIDIA GTX2080Ti graphics processing unit (GPU). The deep learning framework is Pytorch 1.5.0. Additional information can be observed in Table 1.

**Table 1.** Experimental computer environment.

| Name | Versions |
|---|---|
| System Environment | Windows 10 64-bit |
| CPU | Intel(R) Core (TM) i9-9980XE |
| GPU | NVIDIA GTX 2080Ti |
| Python | 3.6.0 |
| Pytorch | 1.5.0 |

### 4.3. Evaluation Criteria

Various occupations are evaluated using diverse metrics. This study introduces assessment criteria, including frames per second (FPS), precision (P), recall (R), and average precision (AP), which are developed by considering relevant demand variables. Frames per second (FPS) is a quantitative metric, utilized to gauge the rate at which images are processed within a given time frame of one second. Recall (R) is employed as a metric for evaluating the comprehensiveness of target detection. Conversely, precision (P) is utilized to ascertain the accuracy of recognition precision, which may be calculated as follows:

$$\begin{cases} Precision = \frac{N_{TP}}{N_{TP}+N_{FP}} \\ [l]\text{Re}call = \frac{N_{TP}}{N_{TP}+N_{FN}} \end{cases} \tag{21}$$

where $N_{TP}$, $N_{FP}$, $N_{FN}$ indicate the number of successfully detected targets, the number of wrongly detected targets, and the correct number of targets missed by the model, respectively. The average precision (AP) is computed as follows:

$$Averageprecision = \int_0^1 P(R)dR \tag{22}$$

## 5. Results Analysis

### 5.1. Multi-Target Recognition

The field experiments were carried out on the Huanghai Sea in Yancheng city, China. The YOLOX model was selected to track the USV on calm water. Figure 7 demonstrates the test results. The model demonstrates a high level of effectiveness in accurately identifying the USV during its operation on undisturbed water surfaces. Additionally, the UAV is strategically positioned above the USV at this particular instance, which is considered the most favorable condition for optimal recognition. Moreover, even when the USV is not positioned in the center of the screen, the model is still capable of accurately detecting it and generating reliable detection outcomes.

This article focuses on the YOLO model and YOLOX is a fundamental model in the project, and is compared with other models (e.g., SSD, CenterNet, and other YOLO versions). According to the data presented in Table 2, it is evident that YOLOX exhibits superior speed performance compared to alternative models (baseline: FPS). The YOLOX model shows a recognition accuracy that surpasses the YOLO V4 model by 6.2 percent, and demonstrates superior performance compared to other existing models. The YOLOX model displays suboptimal recognition outcomes for small targets due to its tendency to prioritize the prediction of larger targets at higher levels while neglecting the accurate prediction of smaller targets. Despite the somewhat lower FPS achieved with the YOLOX model, it remains a highly promising approach. The YOLOX algorithm presents several benefits and holds potential for practical implementations due to its straightforward architecture, rapid processing capabilities, high precision, and efficient memory utilization.

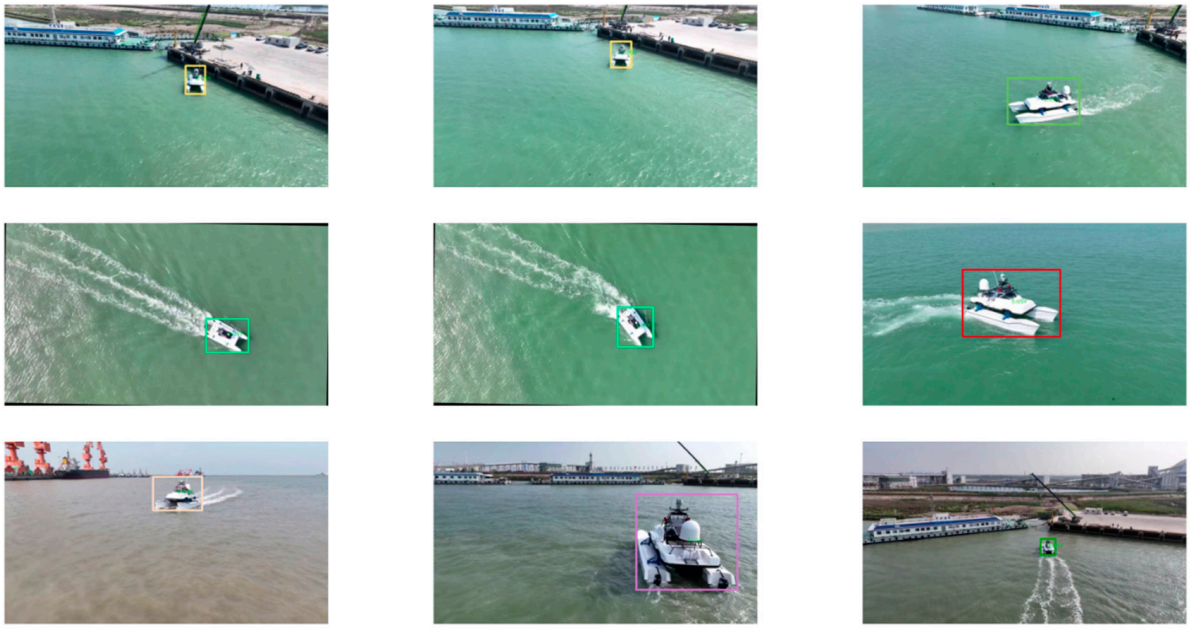

**Figure 7.** Results of the field test.

**Table 2.** Model comparison.

| Model | FPS | $AP_{50}$(%) |
|---|---|---|
| SSD | 52 | 79.9 |
| CenterNet | 57 | 81.5 |
| YOLO V3 | 51 | 83.6 |
| YOLO V4 | 46 | 84.1 |
| YOLO V5 | 47 | 82.7 |
| YOLOX | 42 | 90.3 |

In this article, a UAV is employed to capture the testing photographs with the purpose of assessing the viability of the platform migration application. Figure 8 shows the multi-object recognition based on the UAV's perspective. From the standpoint of a UAV, it is possible to precisely monitor the movements of USVs from various vantage points. Additionally, the presence of other ships in motion near the unmanned vessel, as well as stationary obstructions like buoys, quayside barges, and moored ships, may be reliably detected and identified. The cooperative USV-UAV's perception system is highly successful in identifying even the smallest pixels in the distance of the image, including buoys and ships. This ensures the system's effectiveness in covering the entire water region. Figure 9 demonstrates the multi-object recognition from the USV's perspective. From this standpoint, the USV possesses the capability to effectively detect and classify diverse forms of impediments in its vicinity. Furthermore, it can precisely discern the movements of individuals situated beside the vessel. Nevertheless, the current perspective of the USV lacks sufficient breadth to adequately monitor diverse barrier conditions in every direction. Hence, the approach that relies on the cooperative USV-UAV system can effectively facilitate the comprehensive detection of expansive maritime regions by unmanned ships at sea. This approach holds resemblance to the concept of bird's-eye view (BEV) technology, employed in autonomous driving systems.

**Figure 8.** Multi-object recognition based on UAV's perspective ((**a**–**i**) represent different angles of view from UAV).

**Figure 9.** Multi-object recognition based on USV's perspective ((**a**–**d**) represent different angles of view from USV).

### 5.2. Semantic Segmentation

The reliable identification and classification of water surfaces play a crucial role in facilitating the autonomous movement of USVs. This is because any area that is not composed of water is highly likely to be an obstacle, hence presenting a possible hazard to the USV's navigation. The proposed PIDNet is based on the graph-based segmentation algorithm.

Figures 10 and 11 illustrate the process of semantic segmentation, specifically focusing on the identification and differentiation of water, surface obstacles, land, and sky. This segmentation is achieved through the utilization of both USV and UAV perspectives. The initial row showcases the unaltered input image, while the subsequent row presents the corresponding ground truth. The subsequent row exhibits the segmentation output that is predicted by PIDNet. The red region depicted in the diagram denotes the expanse of the water surface, which is designated as a navigable zone. The black region depicted

in the illustration represents the celestial expanse known as the sky, which is deemed impassable for navigation purposes. Other colors are used to symbolize different barriers found on both water and land, namely locations that are not suitable for navigation. The results suggest that PIDNet demonstrates a high level of efficacy in discerning navigable and non-navigable regions on intricate water surfaces. This offers dependable technical assistance for the autonomous navigation of USVs, relying on visual inputs.

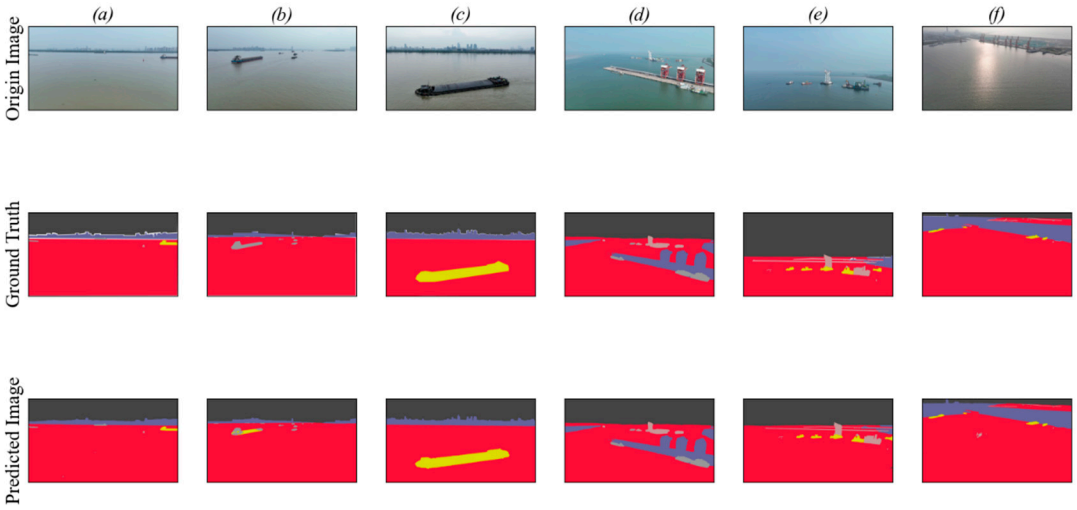


**Figure 10.** Semantic segmentation based on UAV perspective ((**a**–**f**) represent different angles of view from UAV).

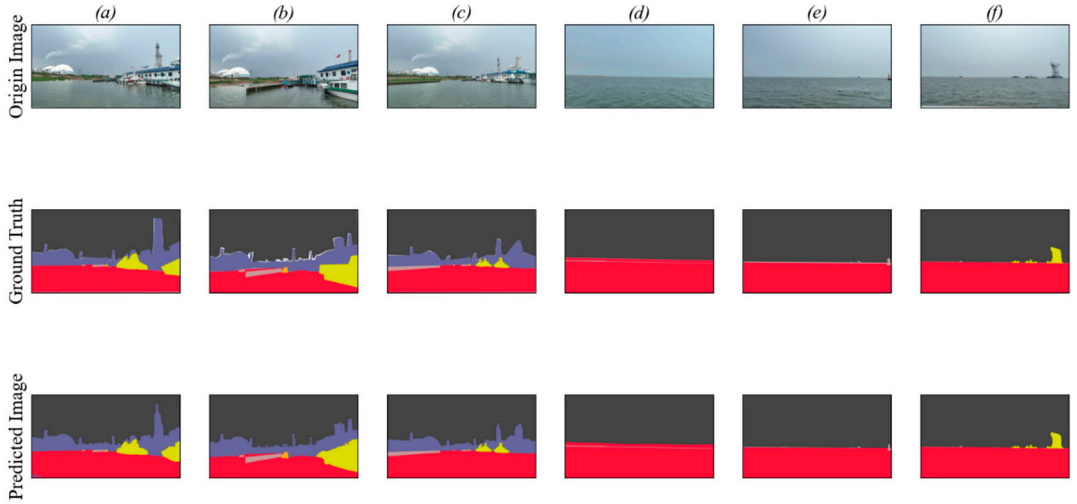


**Figure 11.** Semantic segmentation based on USV perspective ((**a**–**f**) represent different angles of view from USV).

In order to showcase the resilience of the PIDNet algorithm, we conducted a comparative analysis between our approach and other cutting-edge methodologies, specifically focusing on graph-based segmentation algorithms (U-Net, Refine-Net, and DeepLab). The performance evaluation of several models was conducted by quantitatively assessing their accuracy in semantic segmentation. This assessment involved the use of metrics such as mean intersection over union (MIoU), pixel accuracy (PA), and frames per second (FPS). It was observed that the accuracy of semantic segmentation was comparable to that of water surface segmentation. The results can be viewed in Table 3. The networks that are being compared in this study were subjected to retraining using our dataset, incorporating the most optimal hyperparameters. The results demonstrate that the PIDNet has the most

accurate prediction ability, and the network's processing speed is exceptionally high as a result of its efficient architecture.

**Table 3.** Segmentation algorithms comparison.

| Networks | Params (M) | MIOU (%) | PA (%) | FPS |
|---|---|---|---|---|
| U-Net | 34.0 | 79.82 | 80.81 | 9 |
| Refine-Net | 55.1 | 81.63 | 84.26 | 15 |
| DeepLab | 44.3 | 87.22 | 89.13 | 30 |
| PIDNet | 29.5 | 91.08 | 94.32 | 40 |

Figure 12 shows the sea–skyline detection results based on the PIDNet. This part focuses on evaluating the effectiveness of PIDNet in detecting the sea–skyline in various water surface situations, including sunny, foggy, rainy, evening, and reflective conditions. It is evident that, despite the presence of reflection interference issues in the dataset, the alignment achieved using our technique roughly corresponds to the reference sea–skyline in both situations. The obtained outcome serves as evidence that our sea–skyline detection approach possesses the ability to effectively adapt to diverse environmental conditions.

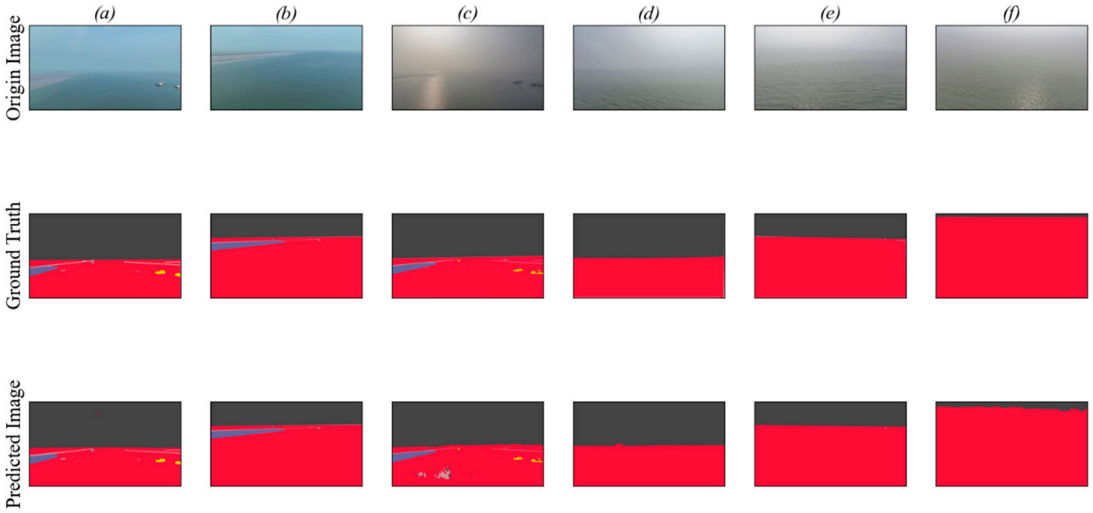


**Figure 12.** Sea–skyline detection algorithm based on PIDNet (**a**–**f**) represent different angles of view from UAV).

### *5.3. Stereo Distance Measurement*

Based on the aforementioned distance measurement model, it is evident that the outcomes of distance measurement are subject to the influence of factors such as the camera height, pitch angle, and pixel coordinates of the observation site. The gathered single frame image is presented in Figure 13, extracting the area where the sea surface target (black box) is located. We determined the uppermost vertical coordinate value among the pixel values within the designated area, and computed the mean value of the related horizontal coordinates. The pixel coordinates obtained by combining the coordinates serve as the basis for calculating the distance between the observation point and the USV.

**Figure 13.** Target on water surface.

Through the manipulation of the camera height and pitch angle to collect photos of a consistent area, this study aims to examine the influence of camera height on the accuracy of range measurements. The findings of this investigation are presented in Table 4.

**Table 4.** The influence of different camera heights on ranging results.

| Real Distance (m) | Camera Height (m) | Pitch Angle (Degree) | Test Distance (m) | Relative Error (%) |
|---|---|---|---|---|
| 5.08 | 2.13 | 18 | 4.92 | −3.1 |
| | 2.32 | 20 | 5.05 | −0.5 |
| | 2.51 | 22 | 5.11 | 0.5 |
| 10.25 | 2.13 | 18 | 11.12 | 8.4 |
| | 2.32 | 20 | 10.94 | 6.7 |
| | 2.51 | 22 | 10.44 | 1.8 |
| 15.18 | 2.13 | 18 | 16.12 | 6.1 |
| | 2.32 | 20 | 15.88 | 4.6 |
| | 2.51 | 22 | 16.30 | 7.3 |
| 22.41 | 2.13 | 18 | 24.18 | 7.8 |
| | 2.32 | 20 | 25.02 | 11.6 |
| | 2.51 | 22 | 23.44 | 4.5 |
| 35.20 | 2.13 | 18 | 39.14 | 11.9 |
| | 2.32 | 20 | 38.52 | 9.4 |
| | 2.51 | 22 | 37.80 | 7.3 |

It is verified that the pitch angle has little impact on the ranging results [31]. Therefore, the camera height serves as the primary variable for evaluating the precision of range measurements. The distance values presented in Table 4 were obtained using radar technology. Subsequently, experiments were conducted to validate these measurements. The findings revealed that the relative error ((test distance − real distance)/real distance) diminishes as the camera height increases, but it increases with greater measurement distances.

## 6. Conclusions

A visual perception technology for coordinated air–sea via a cooperative USV-UAV system is proposed in this paper. The utilization of UAVs can serve as a means to address the limited visibility of unmanned maritime vessels. The primary purpose of this technology is to offer technological assistance in the realm of visual perception for USVs in complex sea

regions. The research areas of USV visual perception encompass multi-object recognition, semantic segmentation, and obstacle recognition, which are regarded as highly significant. The main contributions of this paper can be summarized as follows:

1. The cooperative platform utilizes the YOLOX model to carry out a range of sea detection tasks, including ship recognition, various obstacle detection, and the identification of individuals. The findings of the YOLOX study demonstrate the versatility and effectiveness of the collaborative USV-UAV system, and provides improved detection accuracy and increased detection speed compared to other mainstream methods;
2. The PIDNet model is firstly used to handle the semantic segmentation of sea and air. Compared to other approaches, the results indicate that PIDNet has a significant degree of effectiveness in distinguishing between areas that can be navigated and those that cannot be navigated on complex water surfaces. This offers dependable technical assistance for the autonomous navigation of USVs, relying on visual inputs. The PIDNet model also has a strong ability to detect the sea–skyline in different environmental conditions;
3. The application of distance measurements based on monocular camera vision is used to range the distance between the USV and its targets. The results show that this method can effectively estimate the distance of obstacles. Nevertheless, the findings also suggest that, as the distance from the obstruction rises, the precision of the anticipated outcomes will correspondingly deteriorate. Hence, in instances where USVs exhibit high velocities, the utilization of visual ranging technology in isolation is inadequate for ensuring the safety of these USVs.

**Author Contributions:** C.C. finished the paper, designed the code and performed the experiments; D.L. provided the experimental fields and revised the paper; J.-H.D. carried out the experiments; Y.-Z.L. revised the paper. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** No data were used for the research described in the article.

**Acknowledgments:** We express our gratitude to the China Maritime Safety Administration for their provision of a sea test range.

**Conflicts of Interest:** The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

## References

1. Shao, G.; Ma, Y.; Malekian, R.; Yan, X.; Li, Z. A novel cooperative platform design for coupled USV-UAV systems. *IEEE Trans. Ind. Inf.* **2019**, *15*, 4913–4922. [CrossRef]
2. Anderson, K.; Gaston, K.J. Lightweight unmanned aerial vehicles will revolutionize spatial ecology. *Front. Ecol. Environ.* **2013**, *11*, 138–146. [CrossRef] [PubMed]
3. Woellner, R.; Wagner, T.C. Saving species, time, and money: Application of unmanned aerial vehicles (UAVs) for monitoring of an endangered alpine river specialist in a small nature reserve. *Biol. Conserv.* **2019**, *233*, 162–175. [CrossRef]
4. Campbell, S.; Naeem, W.; Irwin, G.W. A review on improving the autonomy of unmanned surface vehicles through intelligent collision avoidance maneuvers. *Ann. Rev. Control* **2012**, *36*, 267–283. [CrossRef]
5. Murphy, P.P.; Steimle, E.; Griffin, C.; Cullins, C.; Hall, M.; Pratt, K. Cooperative use of unmanned sea surface and micro aerial vehicles at Hurricane Wilma. *J. Field Robot.* **2008**, *25*, 164–180. [CrossRef]
6. Mostafa, M.Z.; Khater, H.A.; Rizk, M.R.; Bahasan, A.M. GPS/DVL/MEMS-INS smartphone sensors integrated method to enhance USV navigation system based on adaptive DSFCF. *IET Radar Sonar Navig.* **2018**, *13*, 1616–1627. [CrossRef]
7. Han, J.; Cho, Y.; Kim, J.; Kim, J.; Son, N.-S.; Kim, S.Y. Autonomous collision detection and avoidance for ARAGON USV: Development and field tests. *J. Field Robot.* **2020**, *37*, 987–1002. [CrossRef]
8. Ma, H.; Smart, E.; Ahmed, A.; Brown, D. Radar Image-Based Positioning for USV Under GPS Denial Environment. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 72–80. [CrossRef]

9. Almeida, C.; Franco, T.; Ferreira, H.; Martins, A.; Santos, R.; Almeida, J.M.; Carvalho, J.; Silva, E. Radar based collision detection developments on USV ROAZ II. In Proceedings of the OCEANS 2009—EUROPE, Bremen, Germany, 11–14 May 2009; pp. 1–6.
10. Zhang, J.Y.; Su, Y.M.; Liao, Y.L. Unmanned surface vehicle target tracking based on marine radar. In Proceedings of the 2011 International Conference on Computer Science and Service System (CSSS), Nanjing, China, 27–29 June 2011; pp. 1872–1875.
11. Han, J.; Cho, Y.; Kim, J. Coastal SLAM with marine radar for USV operation in GPS-restricted situations. *IEEE J. Ocean. Eng.* **2019**, *44*, 300–309. [CrossRef]
12. Esposito, J.M.; Graves, M. An algorithm to identify docking locations for autonomous surface vessels from 3-D Li DAR scans. In Proceedings of the IEEE International Conference on Technologies for Practical Robot Applications, Woburn, MA, USA, 14–15 April 2014; pp. 1–6.
13. Su, L.; Yin, Y.; Liu, Z. Small surface targets detection based on omnidirectional sea-sky-line extraction. In Proceedings of the 33rd Chinese Control Conference, Nanjing, China, 28–30 September 2014; pp. 4732–4736.
14. Tao, M.; Jie, M. A sea-sky line detection method based on line segment detector and Hough transform. In Proceedings of the 2nd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, 14–17 October 2016; pp. 700–703.
15. Kristan, M.; Perš, J.; Sulič, V.; Kovačič, S. A graphical model for rapid obstacle image-map estimation from unmanned surface vehicles. In Proceedings of the 12th Asian Conference on Computer Vision, Singapore, 1–5 November 2014; pp. 391–406.
16. Wang, H.; Mou, X.; Mou, W.; Yuan, S.; Ulun, S.; Yang, S.; Shin, B.-S. Vision based long range object detection and tracking for unmanned surface vehicle. In Proceedings of the 2015 IEEE 7th International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), Siem Reap, Cambodia, 15–17 July 2015; pp. 101–105.
17. Shi, B.; Zhou, H. Marine Object Recognition Based on Deep Learning. In Proceedings of the International Conference on Computer, Network, Communication and Information Systems (CNCI 2019), Qingdao, China, 27–29 March 2019.
18. Song, X.; Jiang, P.; Zhu, H. Research on Unmanned Vessel Surface Object Detection Based on Fusion of SSD and Faster-RCNN. In Proceedings of the Chinese Automation Congress (CAC), Hangzhou, China, 22–24 November 2019; pp. 3784–3788.
19. Zhan, W.; Xiao, C.; Wen, Y.; Zhou, C.; Yuan, H.; Xiu, S.; Zhang, Y.; Zou, X.; Liu, X.; Li, Q. Autonomous visual perception for unmanned surface vehicle navigation in an unknown environment. *Sensors* **2019**, *19*, 2216. [CrossRef] [PubMed]
20. Xu, D.; Chen, G. Autonomous and cooperative control of UAV cluster with multi agent reinforcement learning. *Aeronaut. J.* **2020**, *126*, 932–951. [CrossRef]
21. Zhang, J.; Wang, W.; Zhang, Z.; Luo, K.; Liu, J. Cooperative Control of UAV Cluster Formation Based on Distributed Consensus. In Proceedings of the 2019 IEEE 15th International Conference on Control and Automation (ICCA), Edinburgh, UK, 16–19 July 2019; pp. 788–793.
22. Li, J.; Zhang, G.; Li, B. Robust adaptive neural cooperative control for the USV-UAV based on the LVS-LVA guidance principle. *J. Mar. Sci. Eng.* **2022**, *10*, 51. [CrossRef]
23. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
24. Bochkovskiy, A.; Wang, C.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
25. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
26. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
27. Zheng, Z.; Wang, P.; Liu, W.; Lu, T.; Ye, R.; Ren, D. Distance-iou loss: Faster and better learning for bounding box regression. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12993–13000. [CrossRef]
28. Woo, S.; Park, J.; Lee, J.K.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
29. Towaki, T.; David, A.; Varun, J.; Sanja, F. Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5229–5238.
30. Pei, X.Y. *Autonomous Navigation Technology of Unmanned Surface Vehicle*; Shanghai Maritime University: Shanghai, China, 2015.
31. Zhao, M.H.; Wang, J.H.; Zheng, X.; Zhang, S.J.; Zhang, C. Monocular vision based water-surface target distance measurement method for unmanned surface vehicles. *Transducer Microsyst. Technol.* **2021**, *40*, 47–54.