sensors

MDPI

Article

A Two-Mode Underwater Smart Sensor Object for Precision Aquaculture Based on AIoT Technology

Chin-Chun Chang [1], Naomi A. Ubina [1,2], Shyi-Chyi Cheng [1,*], Hsun-Yu Lan [3], Kuan-Chu Chen [1] and Chin-Chao Huang [1]

[1] Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung City 202, Taiwan
[2] College of Computing Studies, Information and Communications Technology, Isabela State University, PM9Q+C9F, Echague 3309, Philippines
[3] Department of Aquaculture, National Taiwan Ocean University, Keelung City 202, Taiwan
* Correspondence: csc@mail.ntou.edu.tw

Abstract: Monitoring the status of culture fish is an essential task for precision aquaculture using a smart underwater imaging device as a non-intrusive way of sensing to monitor freely swimming fish even in turbid or low-ambient-light waters. This paper developed a two-mode underwater surveillance camera system consisting of a sonar imaging device and a stereo camera. The sonar imaging device has two cloud-based Artificial Intelligence (AI) functions that estimate the quantity and the distribution of the length and weight of fish in a crowded fish school. Because sonar images can be noisy and fish instances of an overcrowded fish school are often overlapped, machine learning technologies, such as Mask R-CNN, Gaussian mixture models, convolutional neural networks, and semantic segmentation networks were employed to address the difficulty in the analysis of fish in sonar images. Furthermore, the sonar and stereo RGB images were aligned in the 3D space, offering an additional AI function for fish annotation based on RGB images. The proposed two-mode surveillance camera was tested to collect data from aquaculture tanks and off-shore net cages using a cloud-based AIoT system. The accuracy of the proposed AI functions based on human-annotated fish metric data sets were tested to verify the feasibility and suitability of the smart camera for the estimation of remote underwater fish metrics.

Keywords: sonar images; stereo RGB images; Mask R-CNN; gaussian mixture models; convolutional neural networks; semantic segmentation networks; object detection CNN

check for updates

Citation: Chang, C.-C.; Ubina, N.A.; Cheng, S.-C.; Lan, H.-Y.; Chen, K.-C.; Huang, C.-C. A Two-Mode Underwater Smart Sensor Object for Precision Aquaculture Based on AIoT Technology. *Sensors* **2022**, *22*, 7603. https://doi.org/10.3390/s22197603

Academic Editors: Nunzio Alberto Borghese and Matteo Luperto

Received: 31 August 2022
Accepted: 4 October 2022
Published: 7 October 2022

1. Introduction

Aquaculture, with the aim of the farmed production of fish or shellfish, has been one of the great contributors to supplying fish or seafood products for human consumption. In 1974, aquaculture contributed 7% of the fish supply and reached 50% in 2020 [1]. With this vast increase in production and to cope with the supply–demand due to the increasing population, many aquaculture farms are expanding rapidly. However, these expansions require effective fish farm management, which is much needed to solve relevant aquaculture issues, including environmental degradation, disease and parasite outbreaks, labor shortage, and productivity maximization by efficiently managing its resources. These challenges can be addressed by integrating fish farm monitoring-based technology with Artificial Intelligence-based Internet of Things (AIoT). At the same time, machine learning and big-data analytics make it possible to collect, process, and analyze large volumes of heterogeneous datasets. When combined, these powerful technologies craft a precision aquaculture framework that uses sensors, cloud, and analytics to enable real-time, evidence-based decision-making to optimize operations [2].

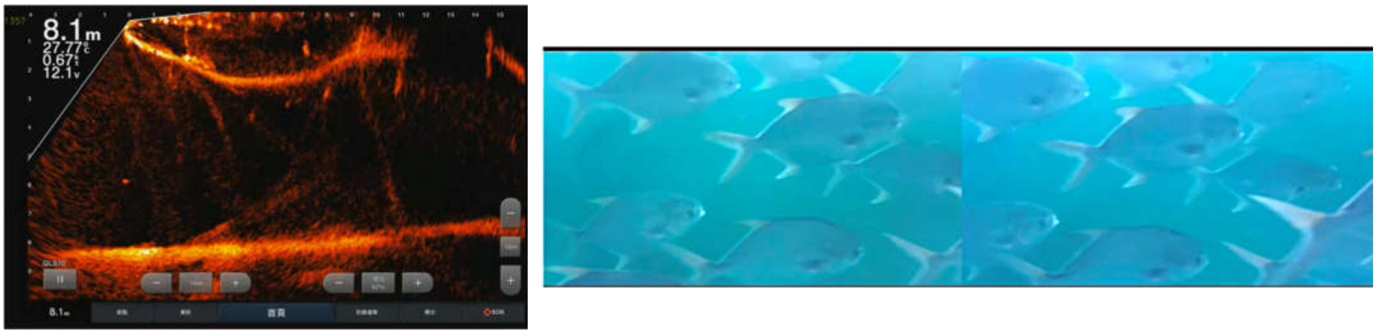
Precision aquaculture requires adopting technologies such as information-based management with big data and models to guide the production process [3] to fully understand the environmental and fish conditions in the cage. Its goal is to enable farmers to make intelligent decisions by providing objective information to improve their capability to monitor and control factors that involve fish production; thus, farming decisions are adjusted to improve fish health and maximize farm production. Large and modern aquaculture farms must incorporate technological innovation to automate their processes, minimize workforce requirements, and maximize their fish feeding process. It enables farmers to integrate technology and data-driven decisions making, enabling efficient aquaculture farm management and remote monitoring, especially for farms situated in the open sea. The use of machine learning and computer vision in Artificial Intelligence (AI), together with sensors and Internet of Things (IoT) technologies, have been widely used to monitor fish feeding behavior, disease, and growth as a non-invasive method, thereby enabling objective observation of the fish farm. Such a mechanism also allows data collection and real-time image acquisition using reliable wireless communication channels [4] without relying so much on human intervention [5].

Various sensors such as temperature, position, humidity, flow, and photo optic or camera sensors have changed how the world accesses data from remote locations. These devices have bridged the gap in collecting data from the physical environment and transmitting wirelessly to a platform with a network of remote servers for storage, management, and data processing [6]. Data collection from physical environments can be carried out by other means such as underwater vehicles [7]. The advancement of cloud computing and IoT has brought tremendous innovation and improvement to aquaculture farming. Cloud computing services enable the collection and storage of big data for processing using AI methodologies capable of predictive analysis to provide informed decision-making mechanisms for precise aquaculture. It enables a brand-new farming approach [8] that eases the burden of the farming industry in terms of monitoring.

For aquaculture farms, it is vital to monitor the fish growth and population as an essential parameter to approximate fish food and assess the overall wellness of the fish species. To achieve the goal of smart aquaculture, fish counting and body length estimation using underwater images are essential to estimate the fish growth curve [9,10]. Cameras as sensors can now be used to capture underwater fish images in an off-shore cage in a non-intrusive manner that reduces the manual handling of the fish, thus reducing direct contact that can cause stress, injury, and growth disturbance to the fish species in the cage. In addition, sonar and stereo cameras for data collection and computer vision can estimate the fish's biological information. Sonar and RGB cameras, such as stereo systems, are just one of the most widely used and studied systems for underwater environment monitoring.

In the underwater environment where the lighting condition is poor or low, RGB cameras are limited. In contrast, sonar cameras are more robust concerning the issue of light attenuation and water turbidity that severely affects optical sensors. In terms of area to cover in capturing the underwater environment, sonar cameras have more scope and a higher range than stereo cameras, as shown in Figure 1. In addition, a sonar camera provides a depth reference value for 3D images, further improving the length or size estimation accuracy. Various studies also dealt with using sonar systems and their applicability to fish length estimation [11–13]. A 3D sonar camera allows direct representation of a scene's 3D information, drastically reducing or no longer requiring the 3D reconstruction process from 2D views, making it more viable for real-time data processing [14]. However, the cost of a high-resolution 3D sonar camera is expensive. To meet the cost concerns of sensors for aquaculture management, in this study, we proposed a fusion of using a low-cost sonar imaging device and a stereo camera system for aquaculture fish monitoring.

Figure 2 shows the framework of our AIoT system where the two camera sensors (sonar and stereo camera) were deployed and installed in the aquaculture farm to collect images/videos from the site. In addition, these sensors were equipped with wireless transmission capabilities to send the collected data to the AI cloud services, where each of the trained deep learning and machine learning models performed the necessary AI function for a specific application.



(**a**) Sonar image (**b**) Stereo image pair

Figure 1. Data captured from the sensor devices is transmitted to the cloud for storage and big data analytics.

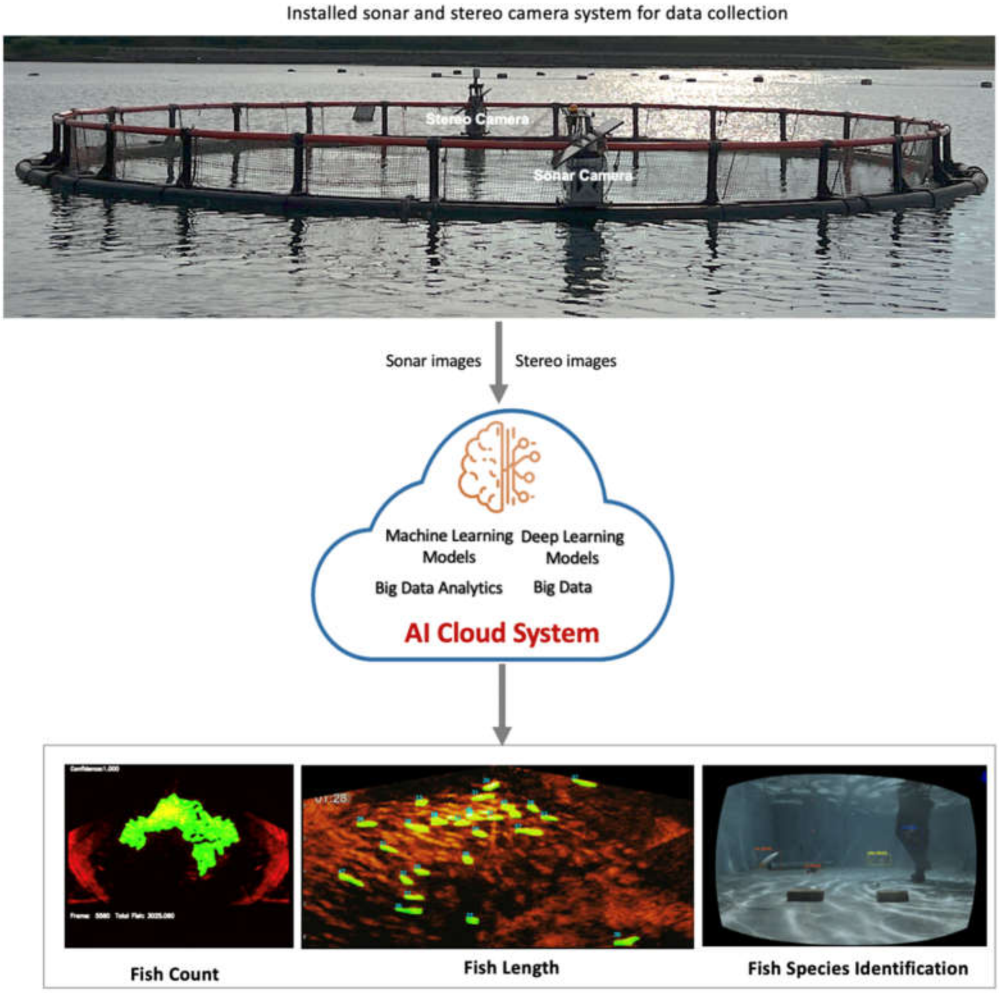


Figure 2. The framework of our proposed AIoT technology for smart Underwater surveillance for precision aquaculture.

Our proposed sensor fusion comprises four steps: data collection from the aquaculture sites using our sonar and stereo camera system sensors, 3D-point cloud estimation, overlapping detection, and object detection to integrate AI functions; the details of these are discussed in the subsequent section. In this work, we used the sonar camera system as the primary sensor device for collecting depth information from the target fish objects to perform fish metric estimation, specifically for fish length and fish count. It uses its beam to collect fish information by sending multiple sound waves to the scene. Thus, it can capture the real environment's depth information. Sonars use depth information to form an image object much different from an optical image.

Although sonar devices, as mentioned earlier, provide bigger coverage, they do not have texture and color information since they just provide depth information. Due to refraction, the shape of the captured fish sonar images is affected. Thus, they can only map macro-features due to their limited resolution [15]. The stereo camera system addresses this concern or limitations of the sonar camera system. These two devices can work together to provide a clearer picture of the underwater fish object in the aquaculture cages or ponds. Since sonar images lack color information, we used the RGB images captured from the low-cost sonar camera to provide additional functions for fish-type annotation.

One of the challenges of sensor fusion is to detect the common area of each sensor or their corresponding images, and considering the environment is underwater, more problems arise. Additionally, the target objects in the sonar and stereo camera systems have different positions, so a mechanism should be devised to project the same target image into the same plane. Incorporating transformation in their corresponding rotation and translation vectors will map the features from sonar to optical coordinate system using the extrinsic sensor calibration method. Since each sensor's range is different, the target object and its corresponding shape should be recognized by both [14]. In this work, we proposed a method to determine the sensor's overlapping areas using their corresponding 3D point cloud information so that the sonar images are projected to the stereo images. We used four markers (four pixels) or points with their corresponding 3D point sets. Both the sonar and stereo camera systems should be able to detect or distinguish these markers. We had two phases to integrate plane projection conversion. First, the learning phase obtained the transformation matrix in rotation and translation based on the marker's information. Second, each pixel in the sonar image was transformed into its corresponding pixel in the stereo image (left image). Then, the result of the transformation was projected into the 2D space to identify the pixel correspondence for both camera systems. In the testing phase, for each pixel in the frame of the sonar camera, the transformation matrices were then used to locate the corresponding pixel in the synchronized frame of the optical camera. Thus, the common area covered by the sonar and RGB cameras was detected.

The contributions of our paper are the following:

- We proposed an AIoT system that provides sonar and stereo camera fusion that supports automatic data collection from aquaculture farms and performs artificial intelligence functions such as fish type detection, fish count, and fish length estimation. To our knowledge, combining a low-cost sonar and stereo camera system tested in various aquaculture environments with different AI monitoring functions is a novel work.
- We designed a methodology to perform sonar and stereo camera system fusion. However, deploying IoT can be expensive, and to limit the cost of its implementation, we employed low-cost sensors that do not entail high additional expenses for aquaculture farmers.
- Using a sonar camera system, we developed our mechanism to estimate the fish's length and weight. Additional plugin AI functions can also be deployed in the cloud to meet the emerging requirements of decision-making for aquaculture management based on the collected big data sets. Agile development realizes the design of learnable digital agents to achieve the goal of precision aquaculture.

The paper is structured as follows: Section 2 provides the related works, and Section 3 contains the materials and methods, which detail our approach to addressing the issue discussed earlier. Sections 4 and 5 include the experimental results and discussion, respectively. Finally, the last section outlines our conclusions and recommendations for future works.

2. Related Works

IoT and AI have gained popularity in the past few years due to their efficiency and promising results in various fields. For example, in aquaculture production, they have been widely used to improve the accuracy and precision of farming operations, facilitate autonomous and continuous monitoring, provide a reliable decision support system, and limit manual labor demand and subjective assessment of fish conditions [16]. In addition, vision cameras such as stereo and sonar systems are popular for computer vision-based problems with the capability of image processing, object detection and classification, and image segmentation.

Imaging sonar systems have been applied to many aquaculture applications, such as fish counting [17,18], estimation of fish length [11], analysis of fish population [19–21], fish tracking [22], fish detection [23], monitoring of fish behavior [24], and control of fish feeding [25]. Hightower et al. [12] used multibeam sonar to determine the reliability of length estimates of some fish species. The sonar was positioned approximately 0.5 m above the bottom, and the beam was aimed slightly off the bottom. Additionally, using sonar image analysis and statistical methodologies, a non-invasive procedure using multibeam sonar was used to count and size the fish in a pond. Simulation software was developed to calculate the abundance correction factor, which depends on the transducer beam size based on the pond size [26]. DIDSON acoustic system with an ultra-high-resolution lens was used to evaluate the accuracy and precision of estimating length from images of tethered fish insonified at the side aspect. The device used has a good potential for discriminating sizes among different species [27]. Lagarde et al. [28] integrated an ARIS acoustic camera to perform counts and size estimates for European eels. Count estimates were performed using 58 videos. The acoustic camera was installed in a channel that links a lagoon to the Mediterranean Sea. It was positioned in the narrowest part of the channel and came 53 m wide and 3.5 m deep. A real-time system for scientific fishery biomass estimator was proposed by Sthapit et al. [29] using a compact single beam advanced echosounder. The device, composed of a transducer, processing unit, a keypad, and a display unit, analyzes ping data continuously. In real-time, it calculates various parameters and simultaneously displays the echogram results on the screen.

Convolutional Neural Networks, or CNNs, have been applied to process sonar images for many applications, such as detecting objects on the sea floor [30] and counting fish [31]. As illustrated in Figure 3, some characteristics of the fish schools in sonar images are as follows:

- Fish schools swim in the three-dimensional space;
- In the sonar image, fish close to the sonar system can often be incomplete, and fish away from the sonar system become blurrier;
- In sonar images, fish are often overlapped, and the location difference of fish in the direction perpendicular to the sonar beam is indistinguishable [32];
- Annotators are often required to examine successive sonar images to identify fish in sonar images because they find fish by the change of the pattern and strength of echoes.

The stereo camera system has also been extensively used in computer vision. For example, a DeepVision stereo camera system was used by Rosen et al. [33] for continuous data collection of fish color images passing inside the extension of the trawl. Out of 1729 fish captured while trawling, 98% were identified in terms of species. Such a mechanism increases the scope of the information collected specifically on documenting the fine-scale distribution of individual fish and species overlap. The information that can be drawn from this can help interpret acoustic data. The underwater stereo video was also used to determine population counts and spatial and temporal frequencies, incorporating detection and identification [34]. Stereo vision is also integrated for video-based tracking [35], fish volume monitoring [36] or abundance [37], and 3D tracking of free-swimming fish [38].

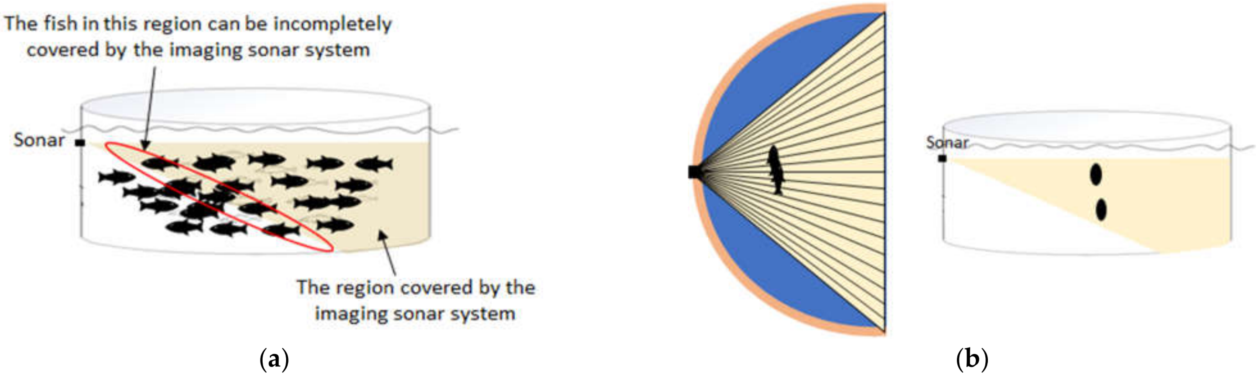

The fish in this region can be incompletely covered by the imaging sonar system
Sonar
The region covered by the imaging sonar system
(a)
Sonar
(b)

Figure 3. An illustration of the region covered by an imaging sonar system where (**a**) shows that the imaging sonar system can partly cover the fish, and (**b**) shows that the fish at different locations in the direction perpendicular to the sonar beam can be overlapped in the sonar image.

Stereo camera systems also have been widely used in fish length estimations [39–42], using disparity information to provide 3D information about an object [43–45]. In aquaculture, many are now putting their interest and efforts into integrating stereo cameras for fish length and biomass estimations [40,41,46]. In our previous work, we integrated a low-cost stereo camera system to perform fish metrics estimation. We used a reliable object-based matching using sub-pixel disparity computation with video interpolation CNN and tracked and computed the fish length in each video frame [10].

Through the years, interest in combining various sensors to achieve higher accuracy and efficiency has been widespread. Many studies regarding sensor fusions have been successfully integrated and applied in multiple fields, such as camera-lidar integration for semantic mapping [47], driver aid systems for intelligent vehicles [48,49], target tracking for robotic fish [50], activity detection of sound sources [51] and avian monitoring [52]. An underwater acoustic-optic image matching was proposed by Zhou et al. [53]. Their work combined the advantages of CNN depth features extraction to determine the image visual attribute conversion; the difference between the acousto-optic images was discarded. Their matching technique used current advanced learned descriptions in the generated target image (acoustic) and the original image (optical). The data aggregation method was utilized in displaying the calibrated matching correspondence between the two types of images.

3. Materials and Methods

3.1. Devices Used and Experimental Environments

Figure 4 shows the sonar equipment we used for the image capture with GARMIN Panoptix LiveScope System (Garmin Ltd., Taiwan), which includes a sonar screen, a processor, and a sonar transducer probe. The sonar system uses an Intel NUC minicomputer (Intel Corporation, Santa Clara, CA, USA) to collect and analyze the sonar images enclosed with the sonar block box. Meanwhile, we used a low-cost camera to set up our stereo camera system using two Go Pro Hero 8 devices (GoPro, San Mateo, CA, USA). The two cameras were mounted in a fixed relative position, as shown in Figure 5a, with a baseline or camera distance of 11 cm. A waterproof case was used to cover the two Go Pro cameras to protect them from water damage since they would be submerged in the water during the data capturing. Next, we calibrated the two stereo cameras using the popular checkboard-based method, as shown in Figure 5b, since the patterns are distinct and easy to detect. The calibration checkboard has an A4 paper size with a 2.5 cm grid size. The first step of the sensor fusion relies on the stereo image rectification process of the left and right images of the low-cost stereo cameras. One potential problem of a low-cost stereo image camera system is that it is incomplete or incorrectly synchronized, causing the object's pose to be different in the left and right images. As seen in Figure 5c, the checkboard corners serve as the corresponding points of the left and right images.

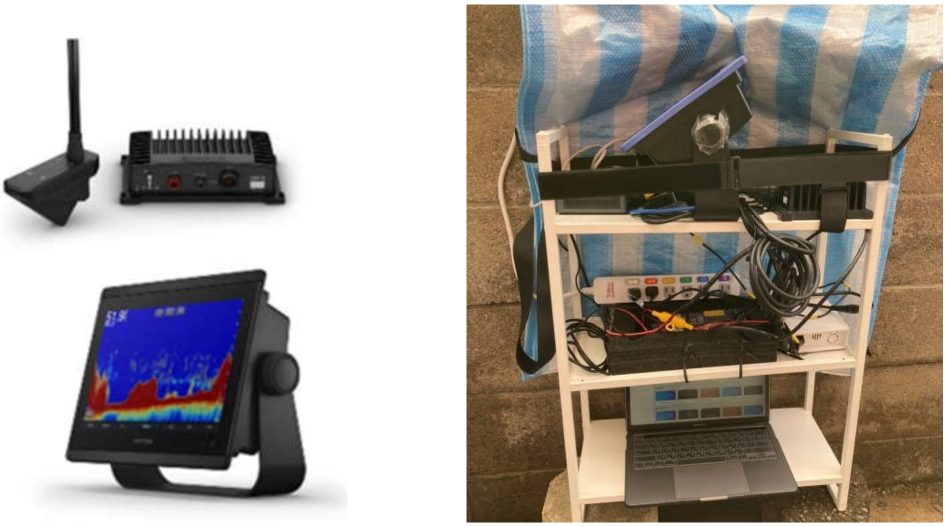
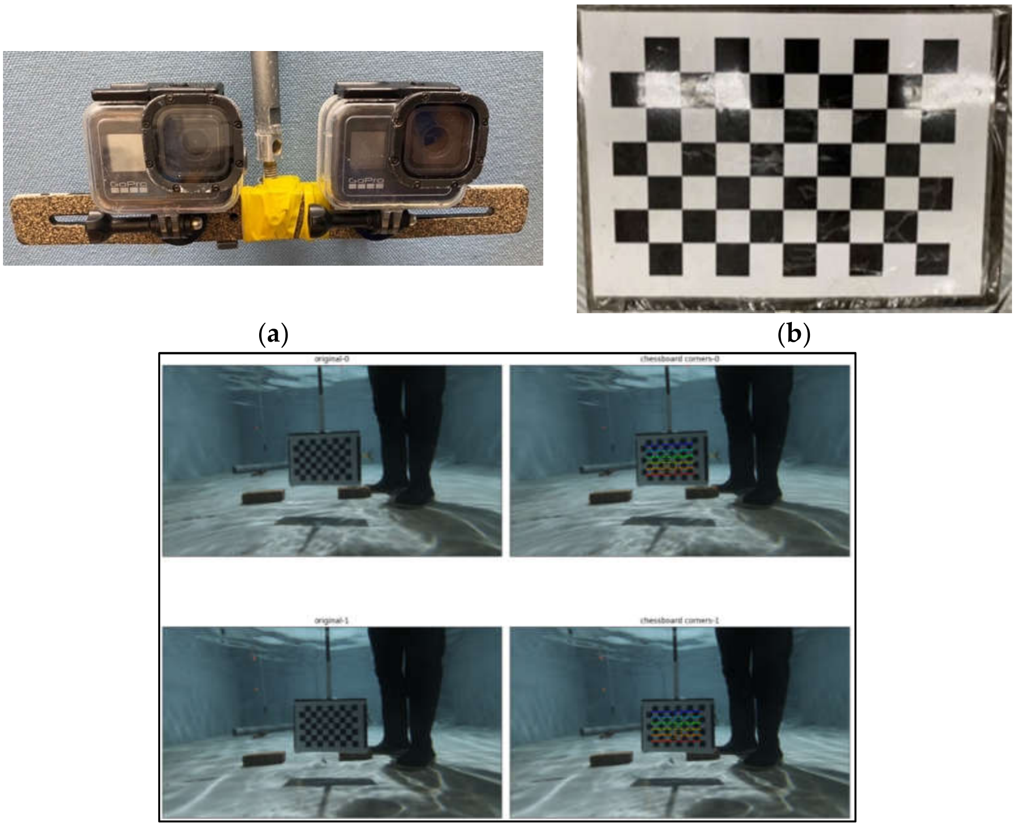


Figure 4. Sonar camera device used for data gathering.

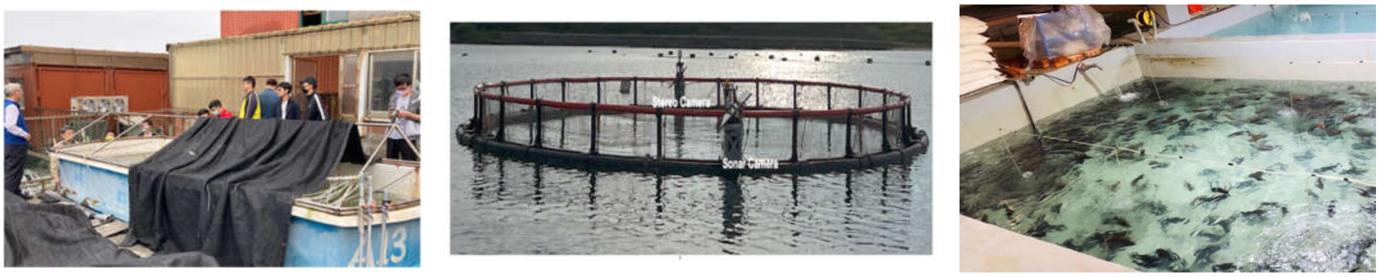


(**a**) (**b**)

(**c**)

Figure 5. Set-up of the low-cost stereo camera system: (**a**) the stereo camera; (**b**) the correction checkboard for calibrating; (**c**) the stereo camera calibration based on the warping of the check-board map.

The experimental site has three locations representing indoor and outdoor environments and less dense and highly dense fish populations. Figure 6 shows the set-up of the environment and locations with its corresponding fish species. Fish tank A is 4 m in length, 1 m in width, and 0.8 m in depth, and the fish instances are small. Fish tank B is an off-shore cage with 50 m in circumference and 25 m in depth. Lastly, tank C with crowded fish instances is 5.3 m in length, 4 m in width, and 0.8 m in depth.



(**a**) A: AAC-A13, Keelung
Species: Oplegnathus punctatus

(**b**) B: Offshore Cage, Penghu
Species: Trachinotus blochii

(**c**) C: LongDann-C10, Pingtung
Species: Cephalopholis sonnerati

Figure 6. Experimental environments utilized for training various deep learning models for fish length estimation, fish count estimation, and fish type annotation.

For simultaneous data capture using the two sensors for calibration, we used a laptop computer as the sonar recording. In contrast, the data captured by the stereo camera were saved in their respective storage devices. The distance between the sonar and stereo camera was 50 cm. The Go Pro camera, which acts as the stereo camera system, used the rg174 signal line to confirm that the target object was captured. In contrast, the sonar was confirmed by the human operators. The computer device that trains the neural networks had an Intel i7-107000k 3.8. GHz CPU, NVIDIA GeForce RTX 3090 GPU, and 48 GB memory.

3.2. Sonar and Stereo Camera Fusion

Figure 7 shows the scheme for the sensor fusion that captures underwater images from a fish pond or a net cage. The 3D point clouds of the common object in the left image and the sonar image can be used to calculate the transformation matrix using a 3D affine transformation algorithm [54]. In the training phase, the common object was detected by applying the pre-trained object detection CNN, e.g., YOLOv4 [55] to both the synchronized RGB image and the sonar image. As shown in Figure 8, the bounding boxes of the common object (the brick) were detected from the stereo images and the synchronized sonar image using YOLOv4.

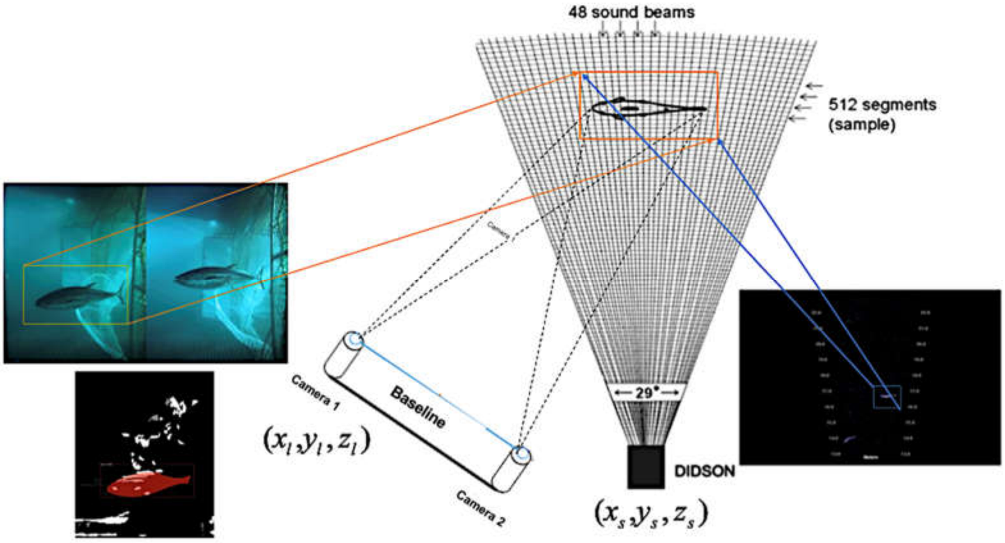
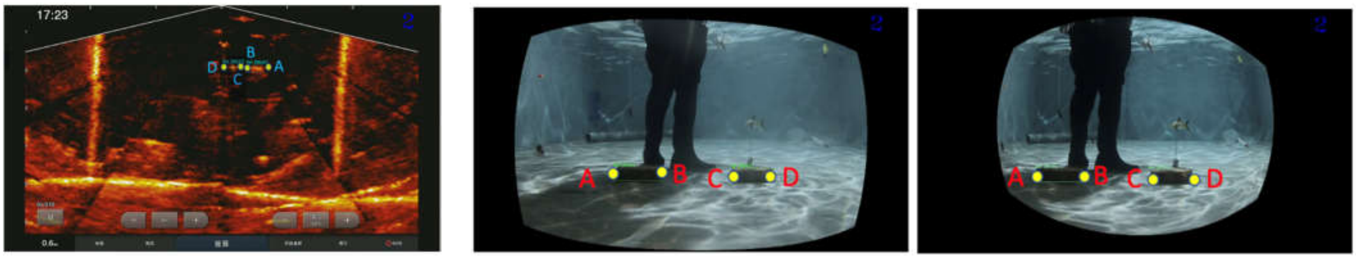


Figure 7. The sensor fusion consists of a stereo camera, and a sonar imaging device captures the fish images from a pond or a net cage. The 3D point clouds of the common object in the left image and the sonar image can be used to calculate the transformation matrix using a 3D affine transformation algorithm [54].



(**a**) Sonar image

(**b**) Stereo images

Figure 8. Brick detection and its corresponding four-point marks using YOLOv4 [55] in (**a**) sonar image and (**b**) stereo image pair.

Four markers (A, B, C, D) exist in both sonar and stereo images. Each marked point has an image coordinate of (u, v). To combine sonar and stereoscopic images using camera calibration, we used two bricks as the target object and integrated YOLOv4 to mark or capture the point coordinates provided by the bounding box of the disparity conversion. For the stereo image, the left and the right images were captured and underwent an image rectification process to obtain the correct intrinsic and extrinsic parameters using camera calibration. To find the corresponding points between a stereo pair and plot them into the 3D space, given a point in the left image, its corresponding point in the right image lies on its epipolar line. Using a stereo-image-based disparity matching algorithm in finding the correspondence of the left image to the right image, take a pixel in the left image and search on the epipolar line for that pixel in the right image. The pixel with the minimum cost is selected, and the disparity can now be computed. The point is located on the epipolar line, which would only require a one-dimensional search where cameras need to be aligned along the same axis. To obtain the depth of a stereo image pair, the disparity information is the difference in the image location of the same 3D point projected using two different cameras. The disparity of a pixel $\overrightarrow{x} = (x, y)$ in the left image can be computed by obtaining the difference between

$$d = x - x' \tag{1}$$

where x' is the x-coordinate of the corresponding pixel in the right image. Once the disparity value of the pixel $\overrightarrow{x}$ in the left image has been computed, the depth value from disparity d and its 3D coordinates $X_O = (x_O, y_O, z_O)$ can be determined using triangulation:

$$\begin{cases} z_O = f * {}^{b}/_{d} \\ x_O = {}^{(x - c_x) \, * \, z_O}/_{f} \\ y_O = {}^{(y - c_y) \, * \, z_O}/_{f} \end{cases} \tag{2}$$

where f is the focal length of the camera, b is the baseline, defined as the distance between the centers of the left and right cameras, and (c_x, c_y) is the center of the projected 2D plane.

Similarly, the 3D point P3D (r, θ, φ) of the spherical coordinates where θ is the azimuth direction, and φ is the spread in the elevation direction and can be expressed in Cartesian coordinates as follows:

$$X = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} r \cos\theta \ \cos\varnothing \\ r \sin\theta \ \cos\varnothing \\ r \sin\varnothing \end{bmatrix} \tag{3}$$

The 2D point $\overrightarrow{x} = (u, v)$ projected on the sonar image plane is expressed as follows:

$$\overrightarrow{x} = \begin{bmatrix} u \\ v \end{bmatrix} = \frac{1}{\cos\varnothing} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} r \cos\theta \\ r \sin\theta \end{bmatrix} \tag{4}$$

Thus, in the 2D sonar images, the information of the elevation angle and, therefore, the height of information of the target fish objects cannot be identified. For the target object in the underwater environment, critical points refer to the shortest distance points, where the sonar's acoustic beams reflect on the object. The critical position in the *j*th acoustic beam is expressed as $r_{cp}(j)$ and $\theta_{cp}(j)$ using the sonar system's local coordinate, which can be calculated using:

$$\begin{bmatrix} x_{cp}^{Local}(j) \\ y_{cp}^{Local}(j) \\ z_{cp}^{Local}(j) \end{bmatrix} = \begin{bmatrix} r_{cp}(j)\ \sqrt{1 - sin^2\left(t + \frac{s}{2}\right) - sin^2\left(\theta_{cp}(j)\right)} \\ r_{cp}(j)\ sin\left(\theta_{cp}(j)\right) \\ r_{cp}(j)\ sin\ \left(t + \frac{s}{2}\right) \end{bmatrix} \tag{5}$$

where t and s are the imaging sonar's tilt angle and spreading angle, respectively. Since the imaging sonar is tilted by an angle t, azimuth angle $\theta_{cp}(j)$, it differs from the azimuth angle of the spherical coordinate. Hence, $x_{cp}^{Local}(j)$ can be calculated using $y_{cp}^{Local}(j)$ and $z_{cp}^{Local}(j)$. The critical point's position in the global coordinates can be expressed using the rotation matrix $R = R_z R_y R_x$ and the position of the imaging sonar $(x_{S,W}, y_{S,W}, z_{S,W})$ in terms of the world coordinate system is represented using:

$$\begin{bmatrix} x_W(j) \\ y_W(j) \\ z_W(j) \end{bmatrix} = \begin{bmatrix} x_{S,W} \\ y_{S,W} \\ z_{S,W} \end{bmatrix} + R \begin{bmatrix} x_{cp}^{Local}(j) \\ y_{cp}^{Local}(j) \\ z_{cp}^{Local}(j) \end{bmatrix} \tag{6}$$

where R represents a 3D rotation transformation matrix to determine the roll angle, pitch angle, and yaw angle of the imaging sonar. The 3D point cloud of the sonar scene can be generated by accumulating the calculated coordinates while scanning [56].

Once the corresponding pixel pairs between the sonar image and the stereo images are detected, the 3D coordinates of the matched points between sonar and stereo images are computed based on the above 3D coordinates computing scheme. Let $X_{O,A} = (x_{O,A}, y_{O,A}, z_{O,A})$ and $X_{S,A} = (x_{S,A}, y_{S,A}, z_{S,A})$ be the 3D coordinates of the common point A generated from the stereo images and the sonar image, respectively. Obviously, $X_{O,A} \neq X_{S,A}$ since they locate point A in different 3D coordinate systems. Let, $X_{W,A} = (x_A, y_A, z_A)$ be the 3D coordinates of the point A in the world coordinate system. Then we can apply the following 3D transformation to transform $X_{S,A}$ or $X_{O,A}$ into $X_{W,A}$:

$$X_{W,A} = R_{O\to W} X_{O,A} + T_O = R_{S\to W} X_{S,A} + T_S \tag{7}$$

where $R_{O\to W}$ ($R_{S\to W}$) is the 3D rotation matrix that aligns the z-axis of the optical camera (sonar camera) coordinate system with the z-axis of the world coordinate system; T_O (T_S) is the translation vector that locates the center of the optical camera (sonar camera) in the world coordinate system. Equation (7) can be rewritten as

$$X_{O,A} = R_{S\to O} X_{S,A} + T_{S\to O} \tag{8}$$

where $R_{S\to O} = R_{O\to W}^{-1} R_{S\to W}$ and $T_{S\to O} = R_{O\to W}^{-1}(T_S - T_O)$. Equation (8) facilitates the computation of the transformation matrices $R_{S\to O}$ and $T_{S\to O}$ by directly matching the set of 3D points, i.e., A, B, C, D points, of the common object in the optical image against that in the sonar image.

Figure 9 shows an example of the 1080 × 1092 sonar images captured from a 630 cm × 600 cm fish pond. Thus, each pixel in the sonar image would occupy an area of 0.583 cm × 0.3125 cm in the fish pond. If a sonar pixel occupies $w \times h$ cm^2 in the fish pond, the 3D coordinates of the pixel $\vec{x} = (u, v)$ can be calculated as

$$X_{S,\vec{x}} = \left[x_S - (u * w),\quad y_S,\quad v * h\right] \tag{9}$$

where x_S and y_S are the x-coordinate and the y-coordinate of the position of the sonar device in the world coordinate system. The value of y_S equals to 0 when we set the origin of the world coordinate system to be $[0,\ d,\ 0]$ where d is the depth of the sonar device. In practice, we can obtain the value of d by putting a depth sensor in the center of the sonar camera.

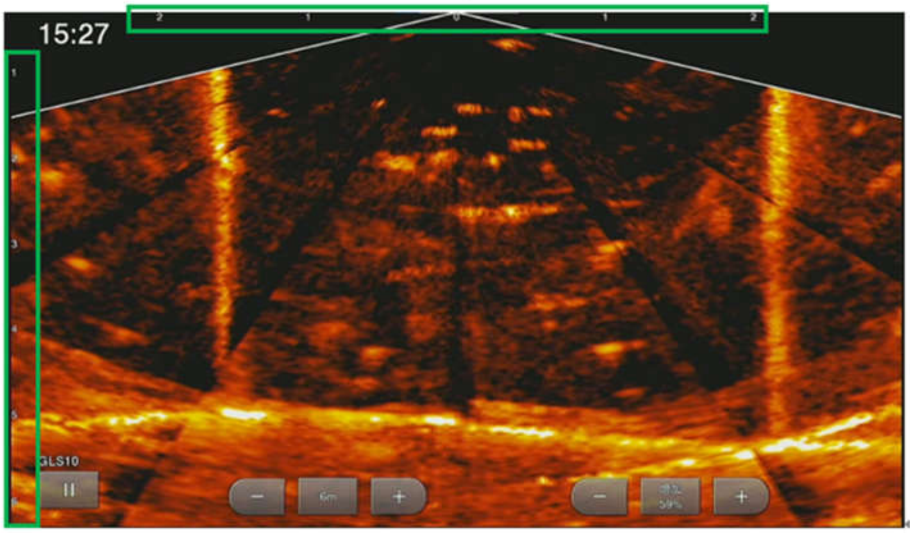


Figure 9. An example of the 1080 × 1092 sonar images captured from a 630 cm × 600 cm fish pond.

Given the pixel coordinates of A, B, C, D in the sonar image, Equation (9) thus generates the 3D point set $[X_{S,A}, X_{S,B}, X_{S,C}, X_{S,D}]$. Similarly, the corresponding 3D point set $[X_{O,A}, X_{O,B}, X_{O,C}, X_{O,D}]$ can be computed based on the four pixels in the stereo images using the disparity computing algorithm mentioned above. Equation (8) can now be written as

$$\begin{bmatrix} X_{O,A} \\ X_{O,B} \\ X_{O,C} \\ X_{O,D} \end{bmatrix} = R_{S\to O} \begin{bmatrix} X_{S,A} \\ X_{S,B} \\ X_{S,C} \\ X_{S,D} \end{bmatrix} + T_{S\to O} \tag{10}$$

To solve the unknown parameters $\theta = (R_{S\to O},\ T_{S\to O})$, we can apply any optimization scheme to minimize the following loss function:

$$L_\theta = \frac{1}{4} \sum_{i\in[A,B,C,D]} ||X_{O,i} - R_{S\to O}X_{S,i} + T_{S\to O}||_2 \tag{11}$$

where $||X||_2$ is the L2 norm of the vector X. Although the object detection CNN for the common object detection has been proved to be accurate, the 2D coordinates of the detected matched points in both the sonar image and the stereo images still contain some errors. This error implies that the resulting 3D point pairs contain noises that reduce the reliability of the parameters θ. To further improve the quality of the learned parameters, the loss function for minimization can be rewritten as

$$L_\theta = \frac{1}{4N} \sum_{i=1}^{N} \sum_{j\in[A_i,B_i,C_i,D_i]} \big|\big|X_{O,ij} - R_{S\to O}X_{S,ij} + T_{S\to O}\big|\big|_2 \tag{12}$$

where N is the number of frames used for parameter training.

3.3. Sonar and Stereo Camera Fusion for Fish Metrics Estimation

Figure 10 shows the block diagram of the proposed fish metrics estimation using the sonar and stereo camera fusion and the cloud-based AI functions. First, as mentioned above, we can compute the 3D point clouds P_S and P_O for each captured sonar image I_S and its synchronized stereo image pair I_O, respectively. Next, the overlapping detection module is applied to detect the area of the monitored fish pond or cage both cameras watch. Finally, the two 3D point clouds are inputted simultaneously into the overlapping detection module to identify their correspondence. Figure 11 shows the overlapping area of each camera system that was converted into the 3D point cloud discussed in the previous subsection. The two-mode fish count estimation could be an added feature to support the sonar camera for type-specific fish count estimation if the monitored fish pond or cage contains multiple types of fish.

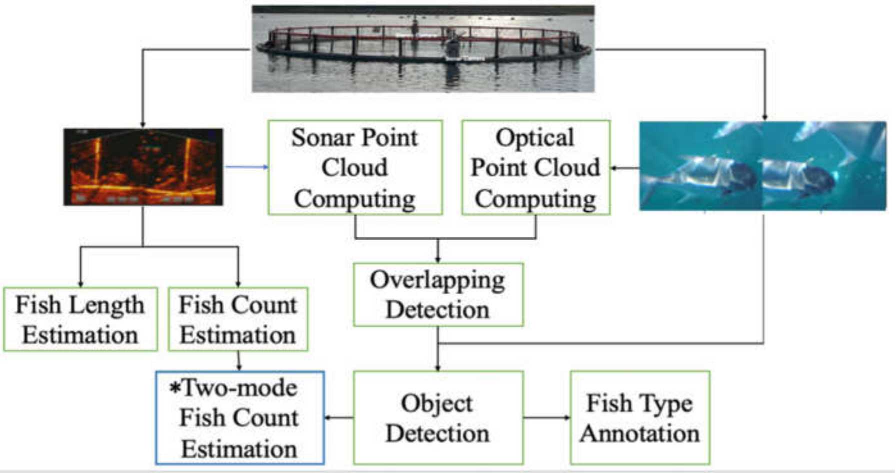


Figure 10. The fish metrics estimation uses sonar and stereo camera fusion and cloud-based AI functions.

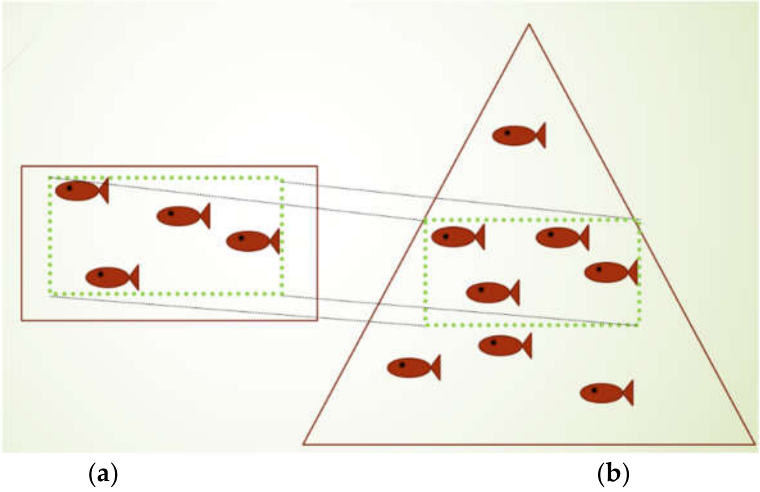


Figure 11. Schematic diagram of overlapping area detection.

The area covered by the optical camera is contained by the sonar device. Once the transformation parameters $\theta = (R_{S\to O},\ T_{S\to O})$ are obtained, the first step of our overlapping area detection is to compute the transformed point cloud:

$$\dddot{P}_S = R_{S\to O} P_S + T_{S\to O} \tag{13}$$

Next, we compute the overlapped point cloud:

$$\hat{P}_S = P_S \wedge \dddot{P}_S \tag{14}$$

For each 3D point $\hat{X}_{S,\vec{x}} = (x, y, z)$ in $\hat{P}_S$, we can then estimate the 2D coordinates of the pixel $\vec{x} = (u, v)$ as:

$$\vec{x} = [(x - x_S)/w, z/h] \tag{15}$$

where w and h are the width and the height of a pixel in the sonar image, respectively; x_S is the x-coordinate of the position of the sonar device in the world coordinate system. Finally, the bounding box B_S to crop the sonar image is defined by the two corner pixels:

$$\begin{cases} \vec{x}_{lu} = \left[\min\limits_{\vec{x}_i \in \hat{P}_S} u_i, \min\limits_{\vec{x}_i \in \hat{P}_S} v_i \right] \\ \vec{x}_{rb} = \left[\max\limits_{\vec{x}_i \in \hat{P}_S} u_i, \max\limits_{\vec{x}_i \in \hat{P}_S} v_i \right] \end{cases} \tag{16}$$

3.3.1. Estimation of Fish Standard Length and Weight Using Sonar Image

The distributions of the standard length and weight of fish are essential to assessing the health and growth of the fish culture. As Figure 12 shows, there are four main steps for estimating those two distributions:

- Apply Mask R-CNN to identify fish instances in each frame of the input sonar video. The standard length of an identified fish instance is estimated by the distance between the two farthest points on this instance.
- Apply the EM algorithm [57] to learn a GMM for the distribution of the length of the identified fish instance. The GMM for the distribution of the length x can be expressed as follows:

$$p\,(x) = \sum_{i=1}^{c} w_i\, N\left(x;\, \mu_i,\, \sigma_i^2\right) \tag{17}$$

where c denotes the weight of the ith Gaussian components, w_i denotes the weight of the ith Gaussian component, and $N\left(x;\, \mu_i,\, \sigma_i^2\right)$ denotes the probability density of the Gaussian distribution with the mean of μ_i and variance σ_i^2. The probability of sample x from the ith Gaussian components, denoted by $p(G_i|x)$, can be estimated by:

$$p(G_i|x) = \frac{w_i\, N\left(x;\, \mu_i,\, \sigma_i^2\right)}{\sum_{i=1}^{c} w_i\, N\left(x;\, \mu_i,\, \sigma_i^2\right)} \tag{18}$$

The sample x belongs to the ith Gaussian component G_i if $p(G_i|x)$ is the largest among $p(G_i|x)$, $i = 1, \ldots, c$. Given the number of Gaussian components c, the EM algorithm can find the parameters $w_{i,}$ μ_i and σ_i^2 for each of the c components through maximum-likelihood estimation. In this paper, a non-Gaussianity criterion $\Phi\,(c)$ was defined in terms of the standardized skewness, and kurtosis [58] was adopted to determine the number c of Gaussian components:

$$\Phi\,(c) = \frac{1}{c}\sum_{i=1}^{c}\left(\left|\frac{\frac{1}{|G_i|}\sum_{x\in G_i}(x-\mu_i)^3}{\sigma_i^3}\right| + \left|\frac{\frac{1}{|G_i|}\sum_{x\in G_i}(x-\mu_i)^4}{\sigma_i^4} - 3\right|\right) \tag{19}$$

The EM algorithm was applied with the number of Gaussian components ranging from one to five. Then, the GMM with the least non-Gaussianity criterion $\Phi\,(c)$ was selected for the subsequent analysis.

- Select the Gaussian component G_{i^*} with the largest component weight as the component comprising a single fish instance. Then, output the statistics of the fish length in G_{i^*}.
- Apply K-nearest neighbor regression with the training set, where the length and weight of the fish are measured manually to estimate the weight using the fish length in G_{i^*}. This paper set parameter K for the K-nearest neighbor regression to 5.

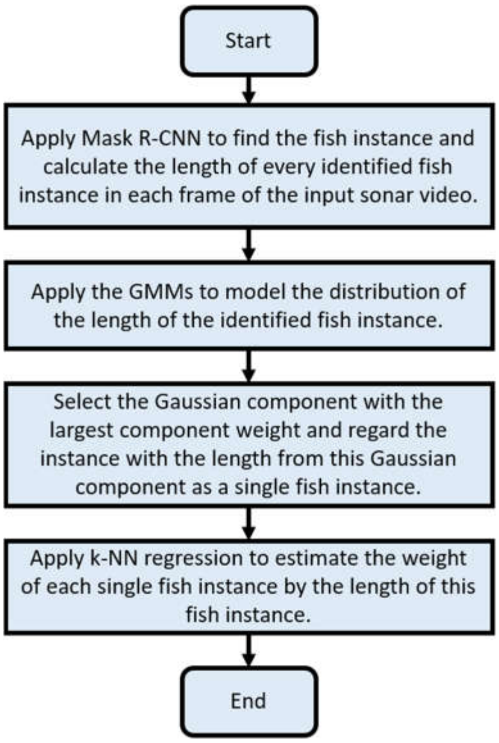

Start
Apply Mask R-CNN to find the fish instance and calculate the length of every identified fish instance in each frame of the input sonar video.
Apply the GMMs to model the distribution of the length of the identified fish instance.
Select the Gaussian component with the largest component weight and regard the instance with the length from this Gaussian component as a single fish instance.
Apply k-NN regression to estimate the weight of each single fish instance by the length of this fish instance.
End

Figure 12. The flowchart to estimate the fish length and weight distribution.

3.3.2. Estimation of the Quantity of Fish in an Off-Shore Net Cage Using Sonar Image

The quantity n_{fish} of fish in an off-shore net cage is estimated using the volume of the fish school that is swimming on the water surface and grabbing food pellets as follows:

$$n_{fish} = \frac{V \times \delta}{V_{fish}} \tag{20}$$

where δ denotes the average fish density of the fish school, and V and V_{fish} represent the volume of the fish school and the volume of the space occupied by a fish instance, respectively. In this paper, V_{fish} is roughly estimated by $V_{fish} = l_{fish} \times \frac{l_{fish}}{2} \times \frac{l_{fish}}{2}$, which is the volume of the cuboid covering the space of a fish instance, where l_{fish} denotes the average length of the fish instance and is measured beforehand. The volume V and the average length fish density d of the fish school are estimated by the average normalized pixel value of the fish region $\mathcal{F}$ in the sonar image as follows:

$$\delta = \frac{1}{g_{max} \times |\mathcal{F}|} \sum_{\mathrm{x} \in \mathcal{F}} g(x) \tag{21}$$

where g_{max} is the maximum pixel value in the fish region and $|\mathcal{F}|$ denotes the number of pixels in $\mathcal{F}$.

There are several ways to scan the fish school to estimate their volume using the sonar system. For example, it can rotate and sideway scan the fish school. In this paper, for simplicity purposes, the fish school was analyzed without rotating the sonar beam. The sonar beam in Figure 13a passes through the fish school in a slantwise position, where the angle between the sonar beam and the seaplane is θ. Meanwhile, in Figure 13b, the space of the fish school when the fish is grabbing pellets is enclosed by an irregular prism, and the volume of the fish school is estimated by the volume of its irregular prism and can be expressed as:

$$V = A \times d \tag{22}$$

where A is the area of the fish regions in the sonar image projected onto the sea plane and d is the depth information of the fish school.

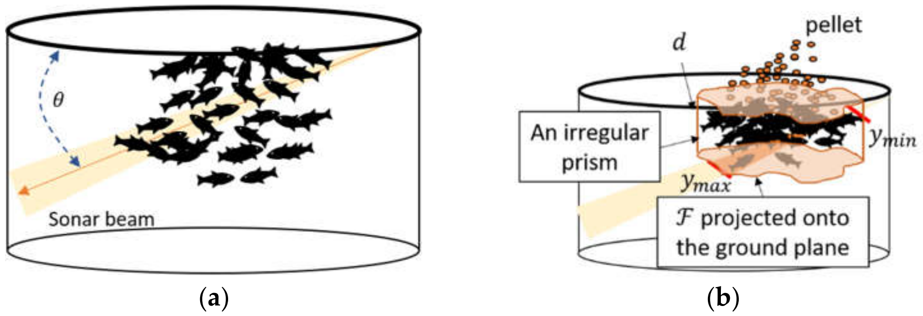


Figure 13. Illustration of the net cage and the fish school with the imaging sonar system, where (**a**) is the angle between the water and the plane of the sonar beam is θ; and (**b**) is the feeding fish school showing is grabbing pellets during feeding which is enclosed by an irregular prism.

The pattern of the fish school in the sonar image when the fish gathers and swims toward the fish surface to grab the pellets is different compared with when the fish disperses, as shown in Figure 14. In this work, a CNN is first applied to find the frame in the sonar video where the fish school gathers and grabs the pellets. Then, the fish region $\mathcal{F}$ in the said frame is identified using a semantic segmentation network; the details of these two neural networks will be presented later. The area A of the bottom of the prism and the depth of the fish school can be estimated by:

$$\begin{aligned} A &= |\mathcal{F}| \times \Delta_x \times \Delta_y \times \cos(\theta), \\ d &= y_{max} \times \Delta_y \times \sin(\theta) \end{aligned} \tag{23}$$

where y_{max} denotes the bottom row of this region and Δ_x and Δ_y denote the width and height of a pixel in centimeters, respectively.

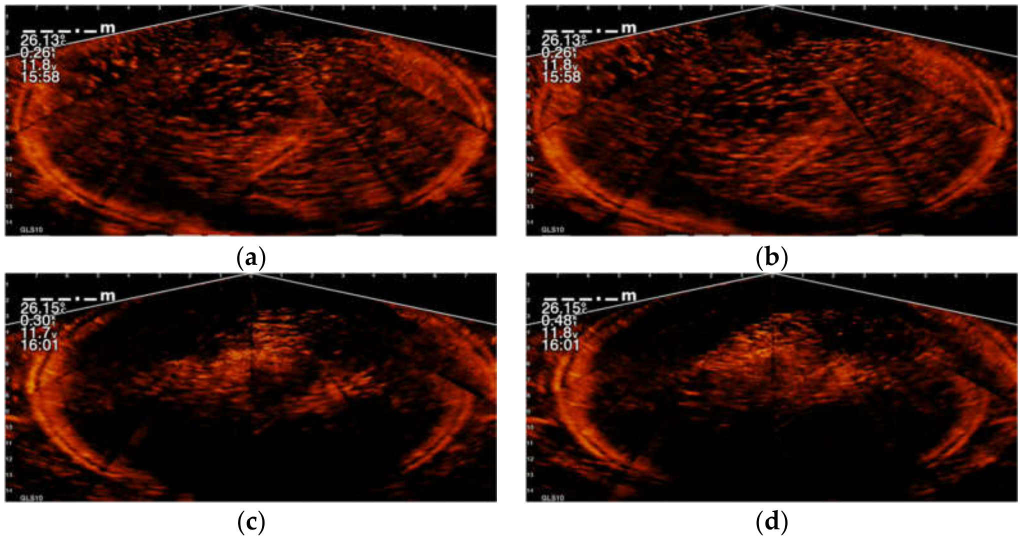


Figure 14. Sonar images of a fish school in an off-shore net cage, where (**a**) and (**b**) show the fish dispersing; and (**c**) and (**d**) show fish swimming toward the water surface to grab feed pellets.

Figure 15 shows how to estimate the quantity of fish in an off-shore net cage. For the first step, it constructs the input for the subsequent two CNNs by stacking five successive frames to form a five-channel image. These five frames are the target frame, two preceding, and two succeeding frames of the target frame. Next, the CNN presented in Section 3.3.3 is applied to determine if the given frame is a fish-gathering frame. If the target frame is classified as a fish-gathering frame, the CNN in Section 3.3.4 is applied to segment the fish region in the target frame. Equations (20), (21), and (22) are then used to estimate the fish quantity. The neural network architectures of the two CNNs are described in the succeeding subsections.

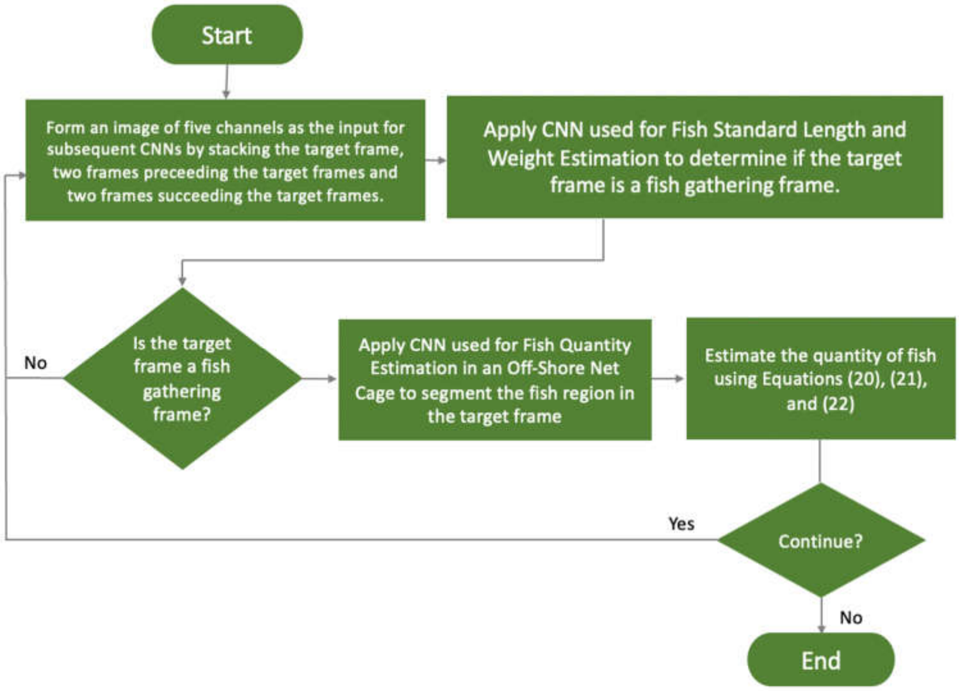

Start
Form an image of five channels as the input for subsequent CNNs by stacking the target frame, two frames preceeding the target frames and two frames succeeding the target frames.
Apply CNN used for Fish Standard Length and Weight Estimation to determine if the target frame is a fish gathering frame.
Is the target frame a fish gathering frame?
No
Apply CNN used for Fish Quantity Estimation in an Off-Shore Net Cage to segment the fish region in the target frame
Estimate the quantity of fish using Equations (20), (21), and (22)
Continue?
Yes
No
End

Figure 15. The flowchart of estimating the quantity of fish in an off-shore net cage.

3.3.3. CNN for Detecting the Fish-Gathering Frame

Figure 16 shows the neural network architecture of the CNN for detecting the fish-gathering frames. The input for the CNN is a five-channel image comprising five successful sonar image frames. The kernel size of the first ten convolutional layers of the CNN is all of size 3 × 3 with a corresponding activation function ReLu. Meanwhile, the last three layers also have a ReLu activation function, and a sigmoid was incorporated into the last 1 × 1 convolution layer.

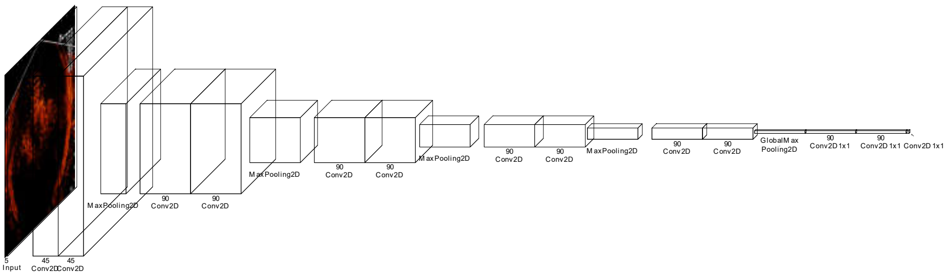


Figure 16. The neural network architecture for detecting fish-gathering frame.

3.3.4. Semantic Segmentation Network for Segmenting Fish Regions

The semantic segmentation networks' neural network architecture to segment fish regions in the sonar image is shown in Figure 17. The neural network is based on the U-Net architecture [59]. The input of this CNN is a five-channel image comprising five successive sonar image frames. In the semantic segmentation network, the transposed convolutional layer with strides (2,2) is adopted for up-sampling. The kernel size of the convolutional layer and the transposed convolutional layer is 3 × 3. The activation of the last later is softmax, and the activation function of the other layers ReLu.

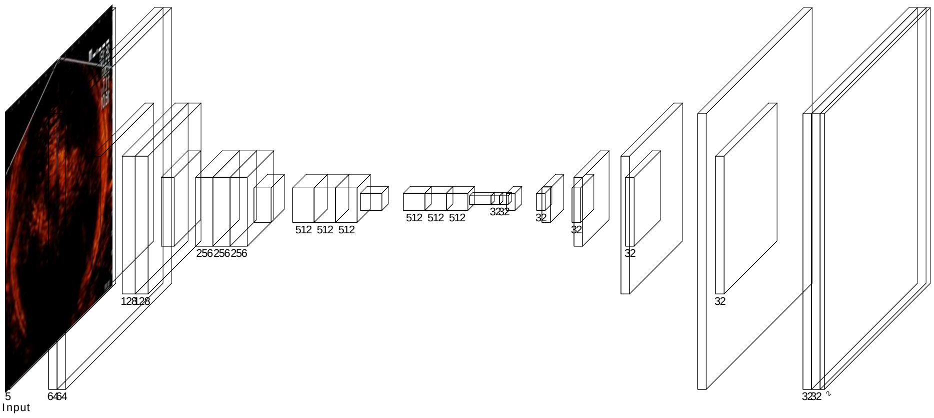


Figure 17. The neural network architecture for segmenting fish regions in the sonar image.

3.4. Object Detection for Fish Type Identification and Two-Mode Fish Counting

The fish of the left image of the input stereo image pair is detected by any object detection CNN, e.g., the YOLOv4. The object detection results can be used to annotate the fish type since the types of fish are given in the training dataset to train the object detection CNN. Let c_i and c_{total} be the fish count of the *i*-th type and the total count of fish detected based on the RGB image. As mentioned above, the sonar image estimates the number of fish without information on fish types. To deal with the difficulty, our two-mode fish counting algorithm estimates the count of the *i*-th type fish as:

$$C_i = C_{sonar} \times c_i / c_{total} \tag{24}$$

In this study, we focused on the design of the two-mode smart sensor, which consists of a sonar scanning device and a stereo optical camera. The captured images are sent to the cloud using a wireless communication network. Although the object detection CNN is not new, we can design a new CNN architecture for underwater object detection to improve the accuracy of fish distributions using (24). Note that the functionality of the smart sensor is incremental since we can add a new AI function into the cloud to provide a new service for sensor fusion.

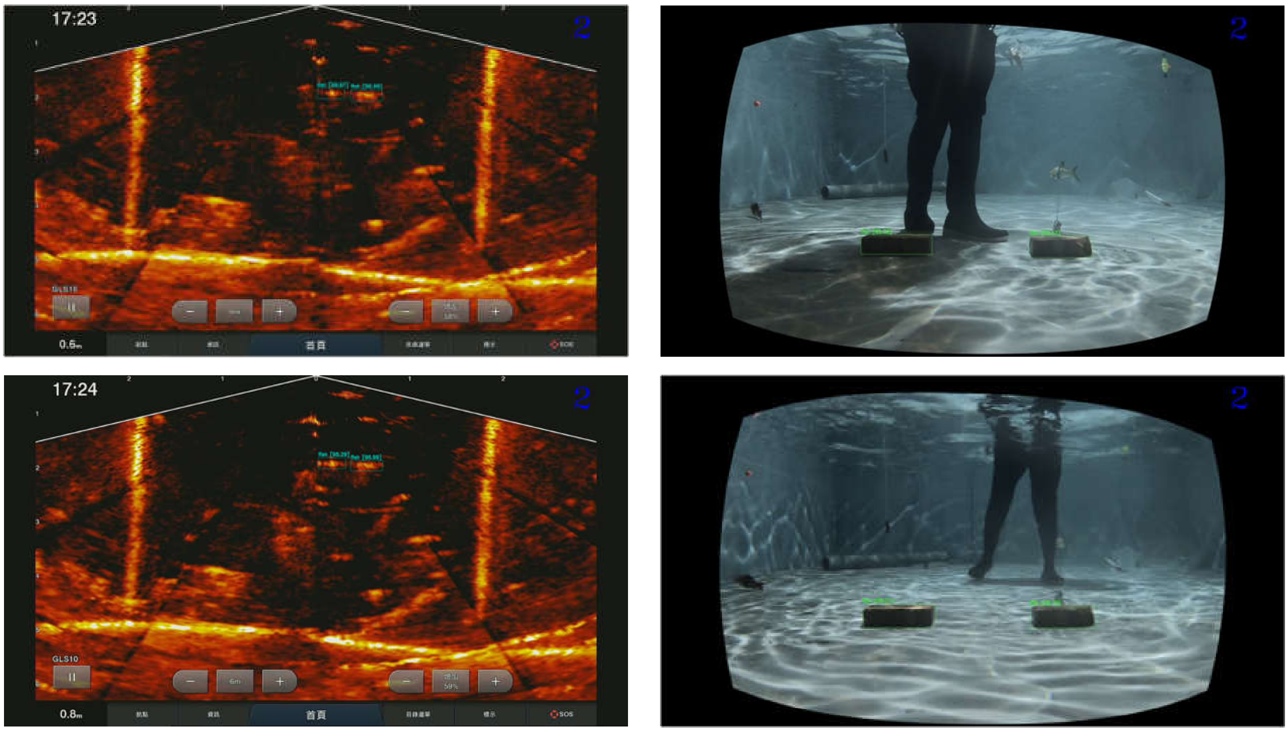
4. Experimental Results

4.1. Sonar and Stereo Camera Fusion Results

The fish objects detected in the sonar images are mapped into the stereo camera image as an area of interest. In Figure 18, we conducted an experiment using the two bricks as the target object representing the fish to determine the object detection capability of our proposed method. Of course, sonar images have a wider range than stereo images, and the positions of the target objects are entirely different. However, based on the detection results for both camera systems, our approach identified the target objects using different sonar and stereo image frames.

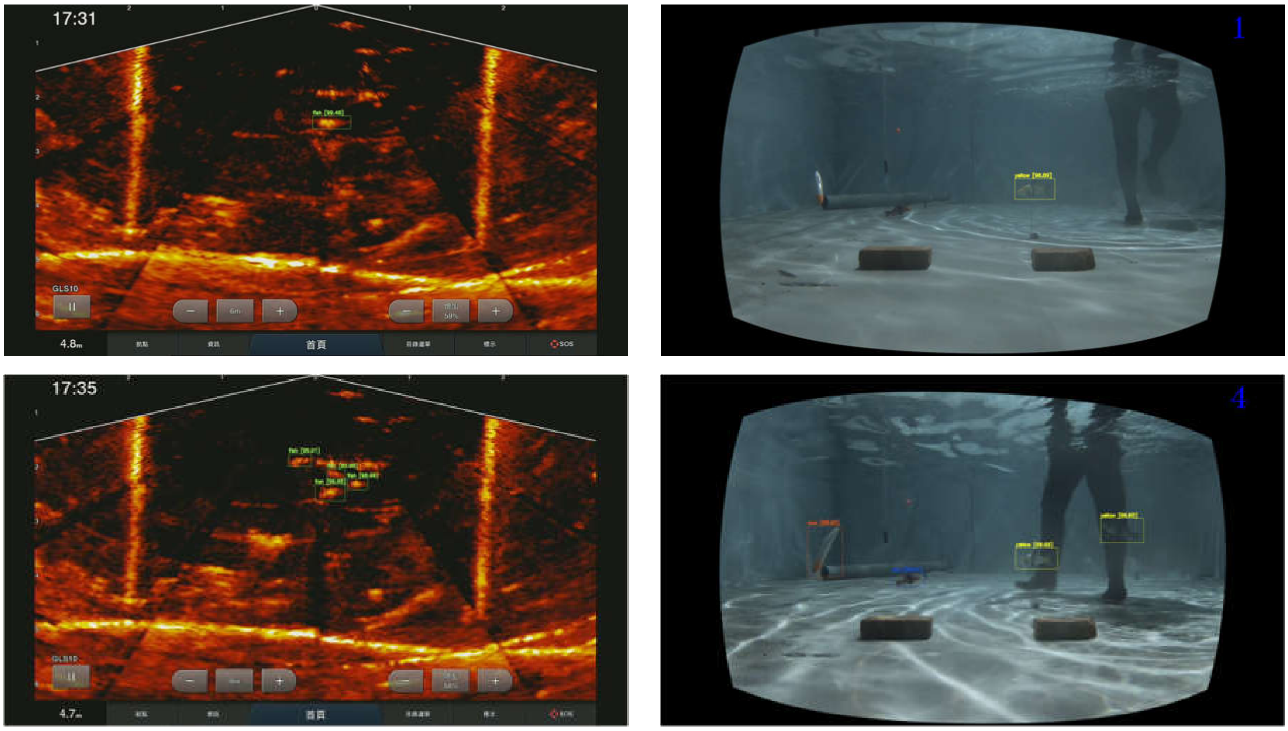
On the other hand, the fish detection in Figure 19a identified the same number of fish objects in the sonar image and are all correctly mapped and detected with fish annotated in the stereo images in Figure 19b.

To ensure that our mechanism detects a correct object in both sonar and stereo images, we integrated a bounding box that shows the area covered by the stereo image in the sonar image in Figure 20, where a shows the sonar image with its corresponding area covered in the stereo image while b shows the entire stereo image area that appears in some part of the sonar images. The images were taken from our various aquaculture locations.



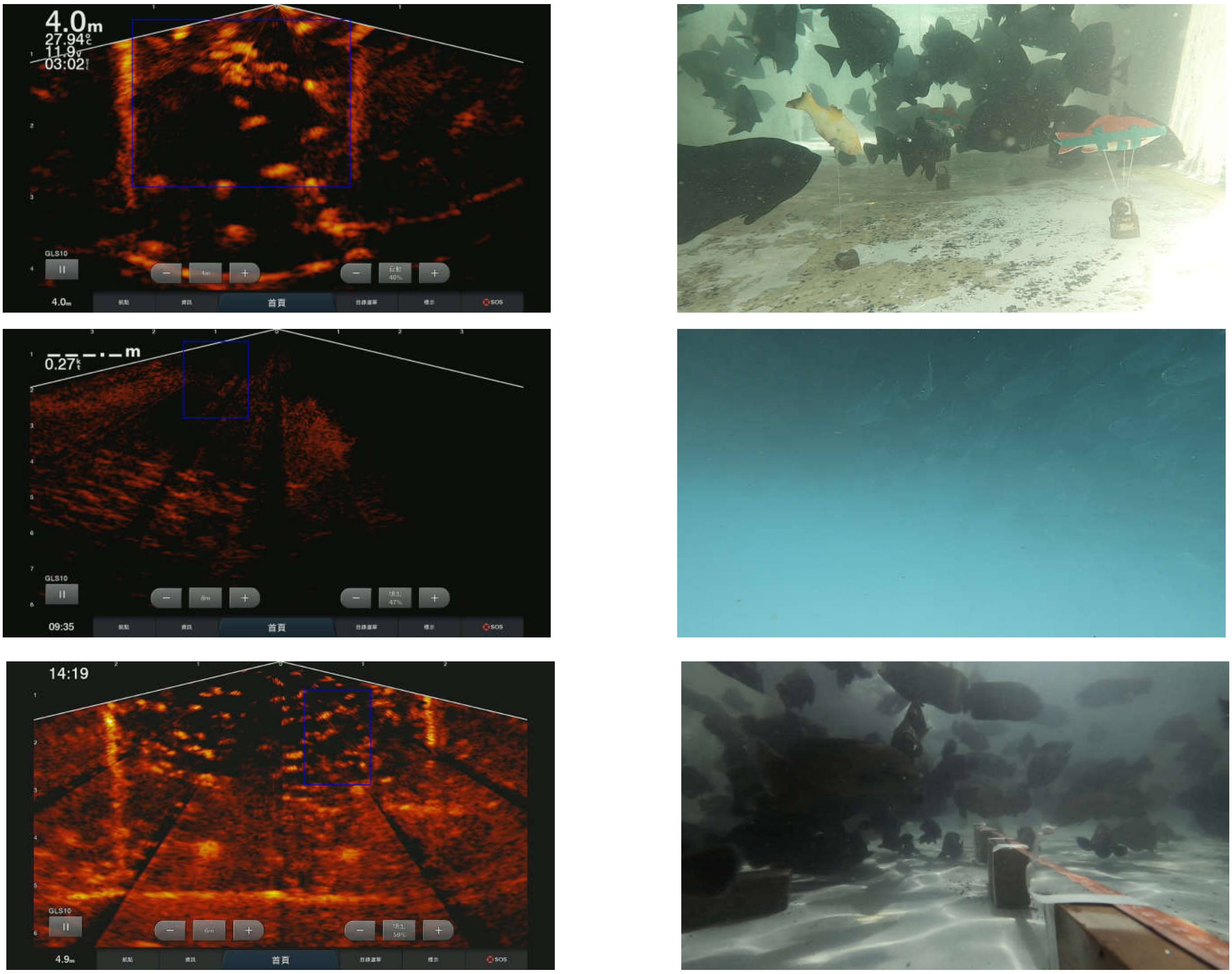
(**a**) (**b**)

Figure 18. Detection results (marker) of the target object from (**a**) sonar and (**b**) stereo images.



(**a**) (**b**)

Figure 19. Fish classification results using (**a**) sonar and (**b**) stereo images.



(**a**) sonar images

(**b**) stereo images

Figure 20. Bounding box detection results from the sensor fusion.

4.2. Estimation of Fish Standard Length and Weight Using Sonar Images

Figure 21 shows the fish instances of a sonar image detected by Mask R-CNN. Table 1 shows that the true positive rates of Mask R-CNN for the three experimental environments were approximately 85, 90, and 75%, respectively. Environment C incurred the lowest positive rate, which was affected by the crowded environment of the fish cage. The number values in the image represent the fish length estimation.

Table 1. The true positive rate of Mask R-CNN for different experimental environments.

Environment	True Positive Rates
A	85%
B	90%
C	75%

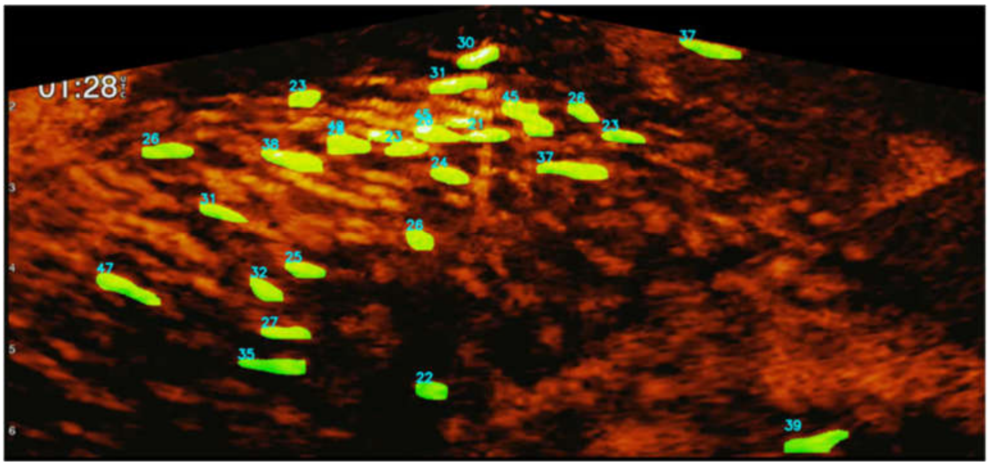


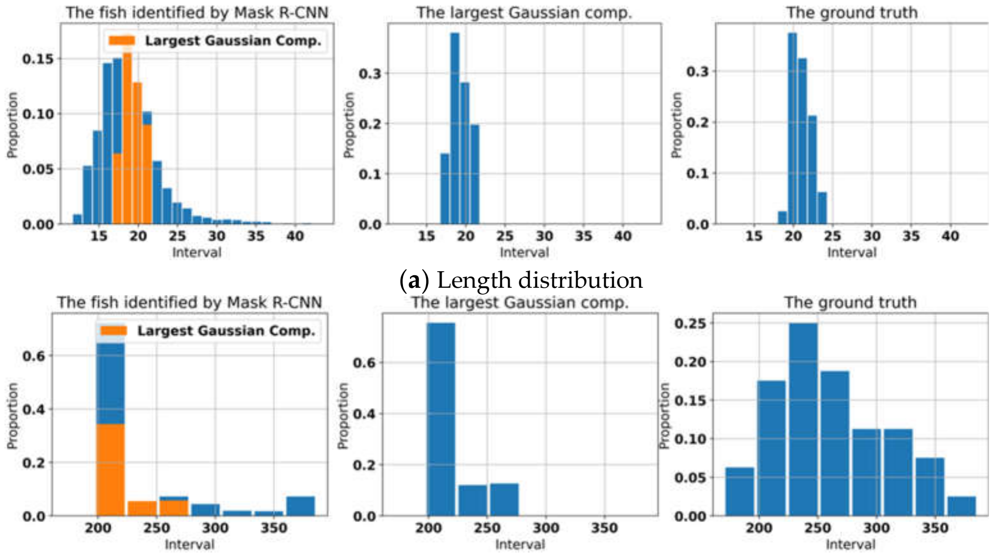
Figure 21. Fish detection using Mask R-CNN and length estimation results using the sonar camera system.

Table 2, on the other hand, shows the experimental results where the relative errors of the average length and weight can be reduced by applying GMMs. The length and weight of each fish in the tank were measured in all environments. We compared the distributions of all estimated data, the estimated data incorporating GMM, and the ground truth (Figure 22), all presented in Table 2. The t-test and Bartlett test were used to determine if the distributions of the two independent samples were significantly different or not in terms of means and variances, respectively. The comparison result showed that the length distributions of the three data were different in means. The p-values for the ground truth vs. GMM, and the ground truth vs. the distribution of fish length identified by Mask R-CNN were 1.11×10^{-41} and 2.57×10^{-8}, respectively. However, the variance of the fish length processed by the GMM and the ground truth was similar (the p-value is 0.61 and the p-value for the other pair is 1.12×10^{-24}). In manually measuring the fish length, we used the fork length.

Table 2. The relative error of the estimated average standard length and weight of fish, where the ground truth for the average standard length and weight of fish is measured manually, and ε, N, NG, and c denotes the relative error, the number of fish instances identified by Mask R-CNN, the number of the instance in the largest Gaussian component, and the number of Gaussian components, respectively.

Environment	Length (cm)					Weight (g)					N	N_G	c
	Manual	w/o GMM		w/GMM		Manual	w/o GMM		w/GMM				
		ℓ	ε	ℓ	ε		w	ε	w	ε			
A (month A)	21.98	8.82	0.14	19.29	0.12	286.37	236.45	0.17	220.32	0.23	4, 456	2, 020	4
A (month B)	21.98 †	27.13	0.24	23.27	0.06	286.37 †	337.39	0.18	315.49	0.10	1, 911	1, 319	2
A (month C)	21.98 †	23.77	0.08	22.31	0.01	286.37 †	315.44	0.10	297.75	0.04	210	168	2
A (month D)	21.98 †	25.16	0.14	22.63	0.03	286.37 †	335.13	0.17	305.94	0.07	33	21	2
A (month E)	21.98 †	24.38	0.11	22.21	0.01	286.37 †	316.57	0.11	296.78	0.04	2, 413	1, 879	2
B (month A)	18.01 †	20.18	0.12	18.90	0.05	180.12 †	210.93	0.17	201.03	0.12	1, 201	1, 034	2
B (month B)	18.01 †	20.30	0.13	19.26	0.07	180.12 †	218.01	0.21	210.78	0.17	3, 279	1, 448	5
B (month C)	18.01 †	21.16	0.18	21.16	0.18	180.12 †	228.60	0.27	228.60	0.27	1, 1228	1, 228	1
C (month A)	15.46	19.78	0.28	17.98	0.16	171.66	272.13	0.59	249.82	0.46	12, 101	9, 781	2
C (month B)	15.50	19.79	0.28	17.51	0.13	173.90	262.86	0.51	236.27	0.36	14, 488	11, 165	2
C (month C)	16.13	16.13	0.22	17.58	0.09	189.36	262.84	0.39	235.98	0.25	10, 121	7, 814	2
C (month D)	16.33	16.33	0.20	18.03	0.10	211.43	290.57	0.37	269.56	0.27	8, 480	7, 013	2

Note. †: no measurement of data available; data was based on the previous months.



(**a**) Length distribution

(**b**) Weight distribution

Figure 22. The distribution of fish length and weight, identified by Mask R-CNN, largest Gaussian component, and the distribution for data manually measured or ground truth.

4.3. Estimation of Fish Quantity in a Net Cage Using Sonar Images

Figure 23 shows the off-shore net cage environment in Penghu, Taiwan, for the fish quantity estimation with Trachinotous blochii as the fish species. The net cage is 15 m in diameter and 5 m in depth, with approximately 2200 fish instances during the experiment. The average standard length of the fish was 20 cm, and the sonar beam was positioned at a slant angle θ of 20°.

(**a**) (**b**)

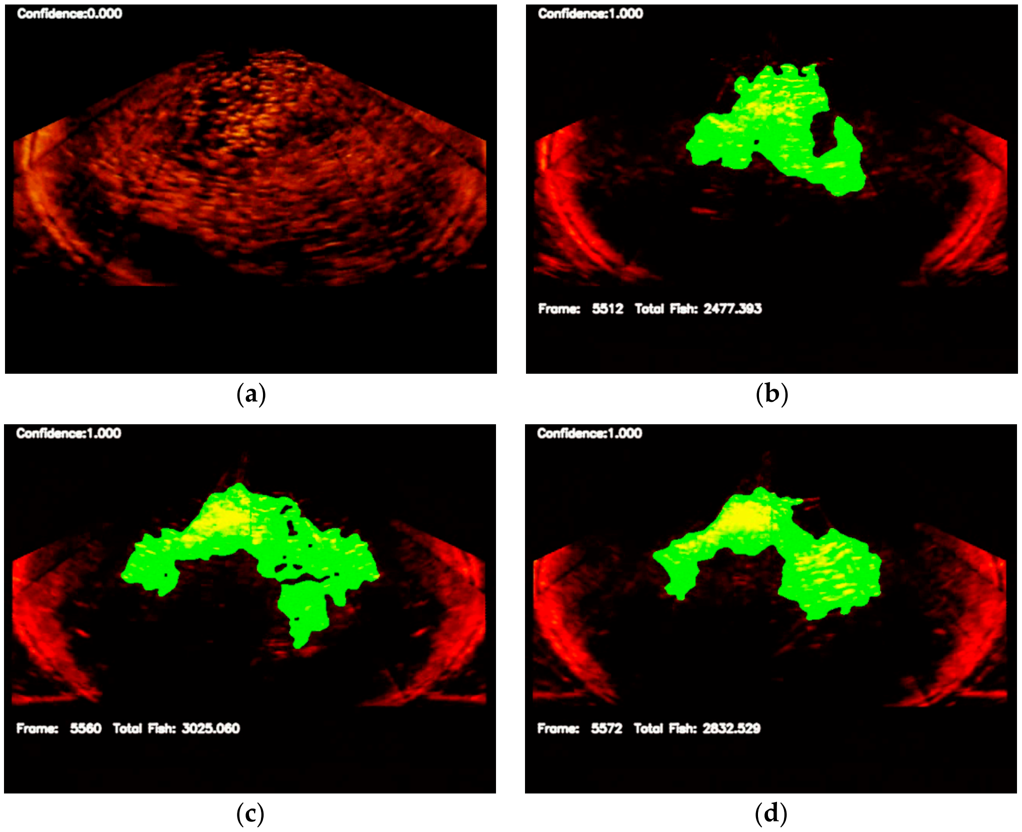
Figure 23. Off-shore cages in Penghu, Taiwan, where (**a**) shows the landscape of the off-shore net cages; and (**b**) is the environment of the net cage used for the experiment.

The dataset used to train the CNN in detecting fish-gathering frames consisted of 58 fish-gathering images and 116 fish-dispersing images. The CNN was evaluated using 10-fold cross-validation and obtained an accuracy of 0.98. Table 3 shows the confusion matrix results. The intersection over union (IoU) was adopted to represent the performance index for the semantic segmentation network to segment the fish region. This network was evaluated using 10-fold cross-validation, and the average IoU was 0.77 ± 0.66.

Figure 24 shows the results of applying the procedures in Figure 15 for fish quantity estimation. The quantity of the fish was estimated using 105 fish gathering frames. Figure 25 shows the distribution of the estimated fish quantity with a mean and standard deviation of 2578.72 and 569.099, respectively. Therefore, the manual estimation of the quantity of fish in the net cage was 2200, within the estimate's 68% and 95% confidence intervals with values of [2112.32, 3045.21] and [1659.33, 3498.11], respectively.

Table 3. The confusion matrix for the CNNs' detection of fish gathering frame.

Actual	**Predicted**	
	Gathering	**Dispersing**
Gathering	58	0
Dispersing	3	113


Confidence:0.000
Confidence:1.000
Frame: 5512 Total Fish: 2477.393
Confidence:1.000
Frame: 5560 Total Fish: 3025.060
Confidence:1.000
Frame: 5572 Total Fish: 2832.529

(**a**) (**b**) (**c**) (**d**)

Figure 24. Experimental results of estimating fish quantity where (**a**) shows the non-gathering characteristics of the fish school; and (**b**–**d**) show the gathering feature of the fish school and the estimated fish quantity based on the fish region identified by the semantic segmentation network.

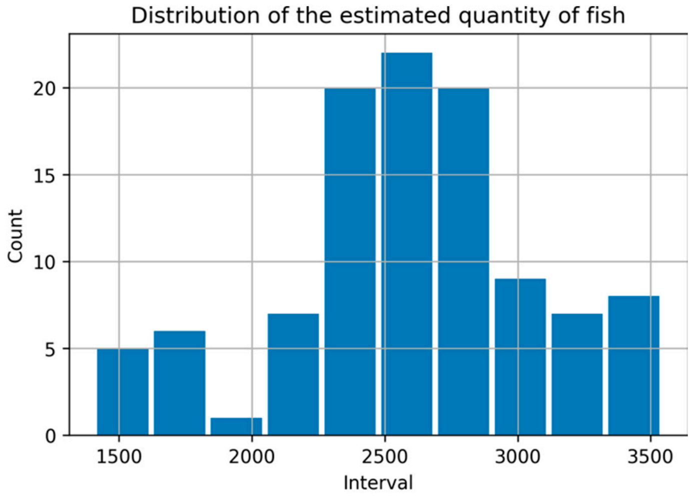
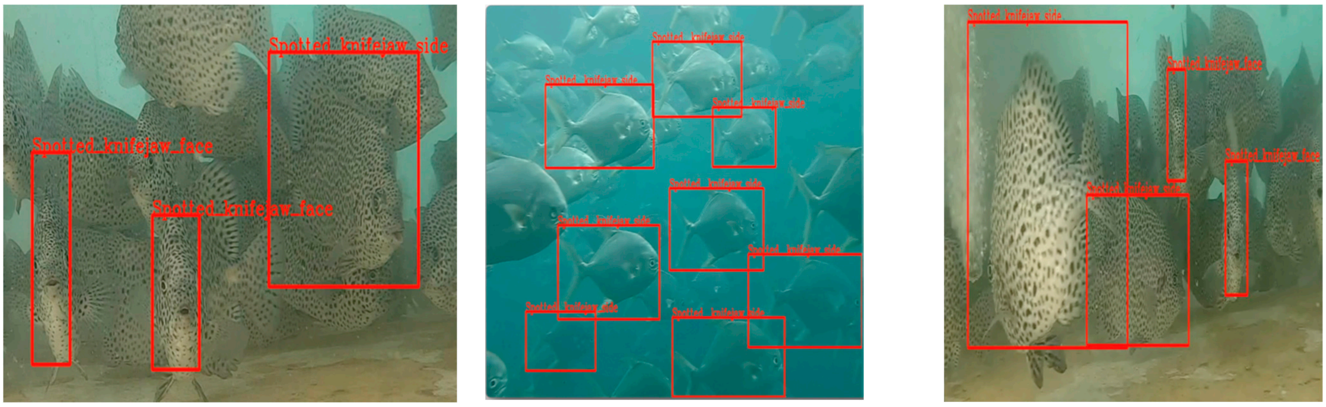


Figure 25. The distribution of the estimated fish quantity.

4.4. Object Detection for Fish Type Identification and Two-Mode Fish Count

The object detection model utilized the pre-trained YOLOv4 with the COCO dataset [60] was used for the object detection model. The object detection results representing three different aquaculture environments (Keelung, Penghu, and Pingtung locations) are shown in Figure 26. The experimental result for the fake fish experiment is in Figure 27, where three fish species were detected.



(**a**) AAC-A13, Keelung

(**b**) Offshore Cage, Penghu

(**c**) LongDann-C10, Pingtung

Figure 26. Fish target object detection results using YOLOv4 in the different aquaculture sites.

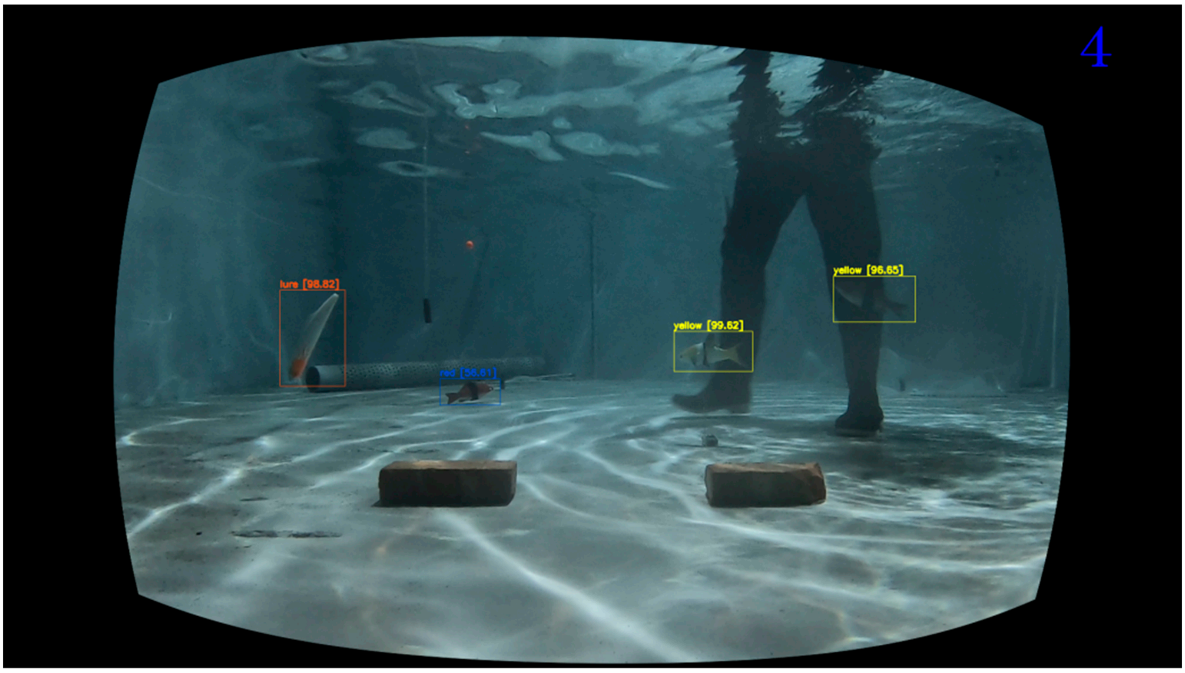


Figure 27. Fish target object detection results using YOLOv4 using the fake fish experiment.

Figure 28 is the result of the fish count estimation, which shows the actual Trachinotus blochii species detected from the images taken from the Penghu off-shore cage. Since the low-cost stereo camera range is short, it cannot see other fish objects beyond its reliable area coverage; thus, only 55 fish objects were detected and counted.

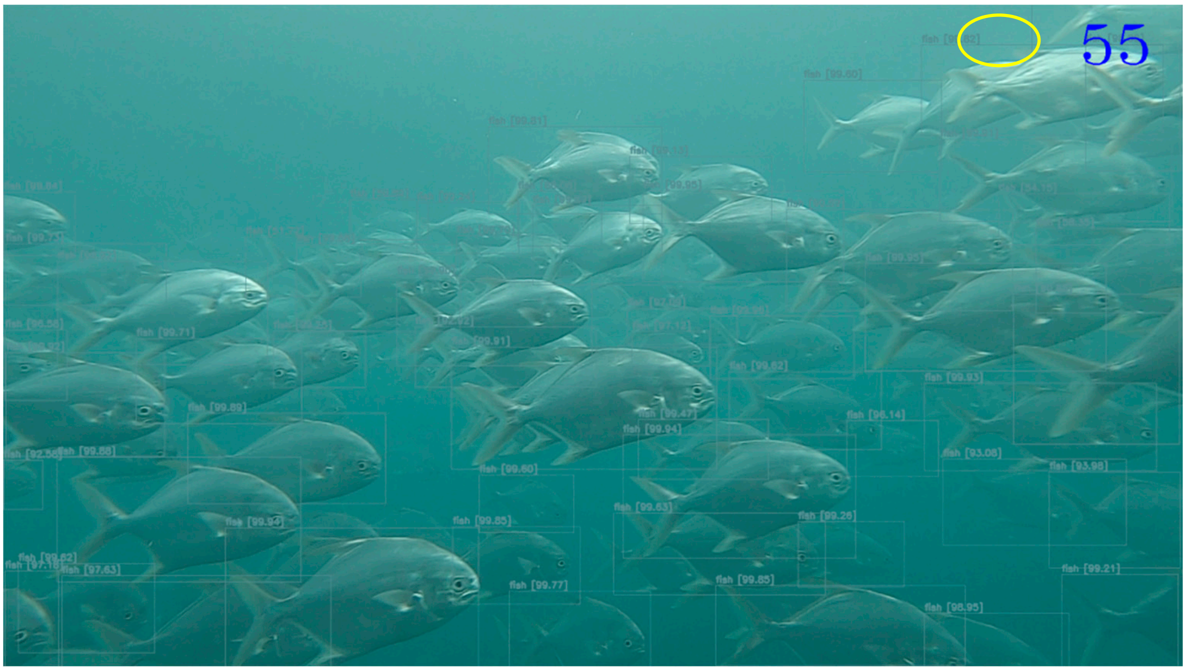


Figure 28. Stereo camera system trachinotus blochii species count estimation.

5. Discussion

Our sensor-based fusion mechanism was applied to aquaculture monitoring using optical (stereo camera) and sonar images. Each sensor system has its strength and limitations, and we took advantage of its capabilities to address the issues of the other sensors. For example, it would be difficult for an RGB camera to accurately estimate fish metrics due to poor underwater conditions addressed by the sonar camera system. Therefore, we took advantage of the texture information of the optical images to provide fish species annotation to sonar images. In addition, detecting the common area of each sensor poses a significant challenge considering the quality of images in the underwater environment.

Additionally, sonar cameras have a larger area covered when compared with stereo cameras. Thus, the target objects will be in different positions or locations. We also must consider that sensors vary in terms of errors, the origin of the coordinate axis, and the types of data received. Two essential issues need to be addressed for sensor fusion, namely, opti-acoustic extrinsic calibration and opto-acoustic feature matching [14]. To deal with this and improve the quality of our data fusion, we performed camera calibration to enable both sensors to be in a common world frame or coordinates. Our approach integrated the 3D point cloud information of both sensors to identify the overlapping areas by using markers (4 pixels) to project sonar images to stereo images as part of the learning phase. The integration of the transformation matrix made it possible to locate the corresponding pixels in both camera systems. Camera calibration performed a significant function in our sensor fusion by transforming their corresponding rotation matrix and translation vectors to match the features from sonar to optical coordinate system, thus taking advantage of the epipolar geometry for the multi-modal feature association [14]. For opti-acoustic feature matching, the 3D information was utilized to identify the same features of both sensor modalities.

One of the AI functions is to estimate the length and weight distribution of the fish in an indoor aquaculture tank. The main challenge of the first method is that the fish in the aquaculture tank are often crowded and overlapped in sonar images. Besides, sonar images of an aquaculture tank are usually noisy due to the echo from the air pumper and the bottom and wall of the tank. The Mask R-CNN [61] identified single fish instances in sonar images. Therefore, it may locate overlapped and incomplete fish for crowded aquaculture tanks as single fish instances. Because Mask R-CNN determines the fish instance, which looks like a single fish instance, an assumption that the length distribution of identified instances is a mixture of Gaussians and the length distribution of the valid single fish instance is the largest Gaussian component of this mixture of Gaussians. Based on that assumption, the first method employs Gaussian mixture models (GMMs) to model the length distribution of the fish instance identified by Mask R-CNN. Then, the proposed method regards the fish instance with the length from the Gaussian component with the largest mixture component weight as a single fish instance and estimates the weight of the instance by the *k*-nearest-neighbor regression. Since we manually measured the fish fork length as the basis for the length estimation, our proposed estimation method is biased since the fork length is usually more significant than the standard length.

Furthermore, the response of the caudal fish fin in the sonar image is usually weak, which makes the fish length measured by our proposed method close to the standard length of the fish. On the other hand, the estimated weight distribution and the ground truth were significantly different. This result could be attributed to the fish's weight being affected by other factors, such as the thickness of the fish. In practice, it is difficult to observe both the thickness and length of the fish instance from the view of the imaging sonar system. Overall, the relative error of the estimated average fish standard length was approximately 15%, while the relative error for the estimated average weight was less than 50%.

The second AI function for sonar images estimates fish quantity in an off-shore net cage, and we identified two main challenges. First, the fish considered in this paper, which were Trachinotus blochii, are often widely distributed in the net cage. Second, the view of an imaging sonar system only covers a small portion of the net cage. Since the fish of the target species can gather and swim close to the water surface to grab food pellets, the proposed method only estimates the quantity of fish in the net cage when feeding fish. The number of fish is assessed on the average fish volume and the fish school's estimated volume and density. In this paper, a convolutional neural network was developed to determine if fishes gathered and were grabbing food pellets. Then, we supposed that the fishes gathered and were grabbing feed. In that case, a semantic segmentation network was applied to segment the fish school in the sonar image, and the volume of the fish school was estimated on the segmentation result. The visible imaging has a short imaging distance underwater due to the light attenuation caused by water absorption and scattering. Therefore, the image was more blurred, and the quantity of the image decreased as the shooting distance increased. However, the sound wave can travel far through water without attenuation. Consequently, counting based on acoustics can still work when visual counting is inappropriate [62].

For object detection, this work used two images from the two sensors to capture a common target. The YOLOv4 [55], with an efficient and powerful object detection model, which makes it possible to achieve real-time object detection, was used to identify or detect the target object in the sonar and stereo images. In the two-mode fish counting estimation, we detected the type of fish found in the cage and provided an assessment of the number of populations for each species. Since we used a low-cost camera type, its range is minimal, so it cannot detect fish out of its range even with robust object detection deep learning models such as YOLOv4. Therefore, we only counted detected fish images within the reliable range, which is why we had a lower fish count when compared with the estimated number of fish in the cage, unlike the sonar camera system that covers a larger and broader range. Thus, we will still rely on the sonar camera system for the final fish population count to perform such estimation. The two-mode fish counting estimation now serves as a sampling device to support and assist the sonar camera system in providing information about the fish species distribution as added sonar image analytics. However, at the moment, the dataset available is only one species per cage/pond. However, we tested our mechanism to perform annotation functions using a well-known object detection CNN to check if our proposed method can detect various fish types. We can replace the YOLOv4 with a state-of-the-art object detection CNN because YOLOv4 does not work well in the underwater environment. This shortcoming will be one of our future works to improve the performance of our smart sensor fusion.

The success of the data collection procedure for this study poses a difficult task. First, the assistance of the aquaculture operators is much needed during the data collection, and they must be present in every data collection activity. Second, the data can only be obtained when the aquaculture operators feed the fish, usually once a day. Third, the weather is another major factor since the net cages are in the open sea. Finally, it is essential to consider the sea current to make sure that the transducer of the imaging sonar system is steady since it greatly affects the quality of the data. After several attempts to collect data, only a few were collected due to difficulties.

6. Conclusions

The 3D point clouds for each camera system were separately obtained, extracted, and matched to find each correspondence. Our sensor fusion approach detected the corresponding pixels or the bounding box of the common area. Thus, it could also detect the fish objects in the images of both sensors to be utilized for fish type annotation. In this paper, two methods were developed to estimate the quantity and the distribution of the standard length and weight of fish using a sonar imaging system. The first method was developed for estimating the distribution of the standard length and weight of fish. Using GMMs to find the distribution of the standard length of single fish instances and employing K-nearest-neighbor regression to estimate the weight of fish by the length, the relative errors of the estimated average fish standard length and weight were approximately less than 15 and 50%, respectively. Those errors can be reduced if the fish length manually measured is based on the standard length instead of the fork length. Therefore, the proposed method can be applied to monitoring the growth of culture fish. The second method estimates the fish quantity in an off-shore net cage. The preliminary experimental result showed that the quantity of fish could be within the estimate's 68 and 95% confidence intervals. The 68 and 95% confidence interval widths were approximately 900 and 1800, respectively. Preliminary experimental results showed that the proposed method is feasible. Lastly, the fish target object detection provides an additional function to annotate fish species and offers additional information to the sonar system. For our future works, we plan to incorporate Generative Adversarial Networks to convert the optical image of the target object into a sonar image. Additionally, we will integrate sonar and stereo camera fusion for fish length and weight estimation.

Author Contributions: Conceptualization, C.-C.C. and S.-C.C.; methodology, N.A.U.; software, H.-Y.L., K.-C.C., and C.-C.H.; validation, H.-Y.L., N.A.U. and S.-C.C.; formal analysis, C.-C.C., S.-C.C. and H.-Y.L.; investigation, C.-C.C., S.-C.C. and N.A.U.; resources, H.-Y.L., K.-C.C. and C.-C.H.; data curation, H.-Y.L., K.-C.C. and C.-C.H.; writing—original draft preparation, C.-C.C., K.-C.C., C.-C.H. and N.A.U.; writing—review and editing, H.-Y.L., N.A.U. and S.-C.C.; visualization, H.-Y.L., K.-C.C., C.-C.H. and N.A.U.; supervision, C.-C.C. and S.-C.C.; project administration, S.-C.C.; funding acquisition, S.-C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Ministry of Science and Technology, Taiwan, under grant number MOST 111-2221-E-019-068-MY3 and by Fisheries Agency, Council of Agriculture, Taiwan, under grant number 111AS-6.2.1-FA-F6.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to acknowledge Hoang-Yang Lu and Chih-Yung Cheng, both at the Department of Electrical Engineering of National Taiwan Ocean University, for their valuable contributions during the early stage of this research work. Additionally, the authors would like to acknowledge Shih-Syun Lin, Yi-Zeng Hsieh and Fan-Hua Nan for their valuable comments on the project results. Finally, the authors acknowledge Tang-Yuan Lee and Kai-Xiang Zhang in helping complete the experiments.

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

Abbreviations

AIoT	Artificial Intelligence-based Internet of Things
CNNs	Convolutional Neural Networks
EM	Expectation Maximization
GMM	Gaussian Mixture Model
IoT	Internet of Things
K-NN	K-Nearest Neighbors
ReLU	Rectified Linear Unit
RGB	Red, green, blue
YOLOv4	You only look once version 4

References

1. Food and Agriculture Organizations of the United Nations. *State of the World and Aquaculture*; FAO: Rome, Italy, 2020.
2. O'Donncha, F.; Grant, J. Precision Aquaculture. *IEEE Internet Things Mag.* **2019**, *2*, 26–30. [CrossRef]
3. O'Donncha, F.; Stockwell, C.; Planellas, S.; Micallef, G.; Palmes, P.; Webb, C.; Filgueira, R.; Grant, J. Data Driven Insight Into Fish Behaviour and Their Use for Precision Aquaculture. *Front. Anim. Sci.* **2021**, *2*, 695054. [CrossRef]
4. Antonucci, F.; Costa, C. Precision aquaculture: A short review on engineering innovations. *Aquac. Int.* **2019**, *28*, 41–57. [CrossRef]
5. Gupta, S.; Gupta, A.; Hasija, Y. Transforming IoT in aquaculture: A cloud solution in AI. In *Edge and IoT-based Smart Agriculture A Volume in Intelligent Data-Centric Systems*; Academic Press: Cambridge, MA, USA, 2022; pp. 517–531.
6. Mustapha, U.F.; Alhassan, A.-W.; Jiang, D.-N.; Li, G.-L. Sustainable aquaculture development: A review on the roles of cloud computing, internet of things and artificial intelligence (CIA). *Rev. Aquac.* **2021**, *3*, 2076–2091. [CrossRef]
7. Petritoli, E.; Leccese, F. Albacore: A Sub Drone for Shallow Waters A Preliminary Study. In Proceedings of the MetroSea 2020–TC19 International Workshop on Metrology for the Sea, Naples, Italy, 5–7 October 2020.
8. Acar, U.; Kane, F.; Vlacheas, P.; Foteinos, V.; Demestichas, P.; Yuceturk, G.; Drigkopoulou, I.; Vargün, A. Designing An IoT Cloud Solution for Aquaculture. In Proceedings of the 2019 Global IoT Summit (GIoTS), Aarhus, Denmark, 17–21 June 2019.
9. Chang, C.-C.; Wang, Y.-P.; Cheng, S.-C. Fish Segmentation in Sonar Images by Mask R-CNN on Feature Maps of Conditional Random Fields. *Sensors* **2021**, *21*, 7625. [CrossRef] [PubMed]
10. Ubina, N.A.; Cheng, S.-C.; Chang, C.-C.; Cai, S.-Y.; Lan, H.-Y.; Lu, H.-Y. Intelligent Underwater Stereo Camera Design for Fish Metric Estimation Using Reliable Object Matching. *IEEE Access* **2022**, *10*, 74605–74619. [CrossRef]
11. Cook, D.; Middlemiss, K.; Jaksons, P.; Davison, W.; Jerrett, A. Validation of fish length estimations from a high frequency multi-beam sonar (ARIS) and its utilisation as a field-based measurement technique. *Fish. Res.* **2019**, *218*, 56–98. [CrossRef]
12. Hightower, J.; Magowan, K.; Brown, L.; Fox, D. Reliability of Fish Size Estimates Obtained From Multibeam Imaging Sonar. *J. Fish Wildl. Manag.* **2013**, *4*, 86–96. [CrossRef]
13. Puig-Pons, V.; Muñoz-Benavent, P.; Espinosa, V.; Andreu-García, G.; Valiente-González, J.; Estruch, V.; Ordóñez, P.; Pérez-Arjona, I.; Atienza, V.; Mèlich, B.; et al. Automatic Bluefin Tuna (Thunnus thynnus) biomass estimation during transfers using acoustic and computer vision techniques. *Aquac. Eng.* **2019**, *85*, 22–31. [CrossRef]
14. Ferreira, F.; Machado, D.; Ferri, G.; Dugelay, S.; Potter, J. Underwater optical and acoustic imaging: A time for fusion? A brief overview of the state-of-the-art. In Proceedings of the OCEANS 2016 MTS/IEEE, Monterey, CA, USA, 19–23 September 2016.
15. Servos, J.; Smart, M.; Waslander, S.L. Underwater stereo SLAM with refraction correction. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013.
16. Føre, M.; Frank, K.; Norton, T.; Svendsen, E.; Alfredsen, J.; Dempster, T.; Eguiraun, H.; Watson, W.; Stahl, A.; Sunde, L.; et al. Precision fish farming: A new framework to improve production in aquaculture. *Biosyst. Eng.* **2018**, *173*, 176–193. [CrossRef]
17. Hughes, J.B.; Hightower, J.E. Combining split-beam and dual-frequency identification sonars to estimate abundance of anadromous fishes in the roanoke river, North Carolina. *N. Am. J. Fish. Manag.* **2015**, *35*, 229–240. [CrossRef]
18. Jing, D.; Han, J.; Wang, X.; Wang, G.; Tong, J.; Shen, W.; Zhang, J. A method to estimate the abundance of fish based on dual-frequency identification sonar (DIDSON) imaging. *Fish. Sci.* **2017**, *35*, 229–240. [CrossRef]
19. Martignac, F.; Daroux, A.; Bagliniere, J.-L.; Ombredane, D.; Guillard, J. The use of acoustic cameras in shallow waters: New hydroacoustic tools for monitoring migratory fish population. a review of DIDSON technology. *Fish Fish.* **2015**, *16*, 486–510. [CrossRef]
20. Baumann, J.R.; Oakley, N.C.; McRae, B.J. Evaluating the effectiveness of artificial fish habitat designs in turbid reservoirs using sonar imagery. *N. Am. J. Fish. Manag.* **2016**, *36*, 1437–1444. [CrossRef]
21. Shahrestani, S.; Bi, H.; Lyubchich, V.; Boswell, K.M. Detecting a nearshore fish parade using the adaptive resolution imaging sonar (ARIS): An automated procedure for data analysis. *Fish. Res.* **2017**, *191*, 190–199. [CrossRef]
22. Jing, D.; Han, J.; Wang, G.; Wang, X.; Wu, J.; Chen, G. Dense multiple-target tracking based on dual frequency identification sonar (DIDSON) image. In Proceedings of the OCEANS 2016, Shanghai, China, 10–13 April 2016.
23. Wolff, L.M.; Badri-Hoeher, S. Imaging sonar- based fish detection in shallow waters. In Proceedings of the 2014 Oceans, St. John's, NL, Canada, 14–19 September 2014.

24. Handegard, N.O. An overview of underwater acoustics applied to observe fish behaviour at the institute of marine research. In Proceedings of the 2013 MTS/IEEE OCEANS, Bergen, Norway, 23–26 September 2013.
25. Llorens, S.; Pérez-Arjona, I.; Soliveres, E.; Espinosa, V. Detection and target strength measurements of uneaten feed pellets with a single beam echosounder. *Aquac. Eng.* **2017**, *78*, 216–220. [CrossRef]
26. Estrada, J.; Pulido-Calvo, I.; Castro-Gutiérrez, J.; Peregrín, A.; López, S.; Gómez-Bravo, F.; Garrocho-Cruz, A.; De La Rosa, I. Fish abundance estimation with imaging sonar in semi-intensive aquaculture ponds. *Aquac. Eng.* **2022**, *97*, 102235. [CrossRef]
27. Burwen, D.; Fleischman, S.; Miller, J. Accuracy and Precision of Salmon Length Estimates Taken from DIDSON Sonar Images. *Trans. Am. Fish. Soc.* **2010**, *139*, 1306–1314. [CrossRef]
28. Lagarde, R.; Peyre, J.; Amilhat, E.; Mercader, M.; Prellwitz, F.; Gael, S.; Elisabeth, F. In situ evaluation of European eel counts and length estimates accuracy from an acoustic camera (ARIS). *Knowl. Manag. Aquat. Ecosyst.* **2020**, *421*, 44. [CrossRef]
29. Sthapit, P.; Kim, M.; Kang, D.; Kim, K. Development of Scientific Fishery Biomass Estimator: System Design and Prototyping. *Sensors* **2020**, *20*, 6095. [CrossRef]
30. Valdenegro-Toro, M. End-to-end object detection and recognition in forward-looking sonar images with convolutional neural networks. In Proceedings of the 2016 IEEE/ OES Autonomous Underwater Vehicles (AUV), Tokyo, Japan, 6–9 November 2016.
31. Liu, L.; Lu, H.; Cao, Z.; Xiao, Y. Counting fish in sonar images. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018.
32. Christ, R.D.; Wernli, R.L. Chapter 15-Sonar. In *The ROV Manual*; Butterworth-Heinemann: Oxford, UK, 2014; pp. 387–424.
33. Rosen, S.; Jørgensen, T.; Hammersland-White, D.; Holst, J.; Grant, J. DeepVision: A stereo camera system provides highly accurate counts and lengths of fish passing inside a trawl. *Can. J. Fish. Aquat. Sci.* **2013**, *70*, 1456–1467. [CrossRef]
34. Shortis, M.; Ravanbakskh, M.; Shaifat, F.; Harvey, E.; Mian, A.; Seager, J.; Culverhouse, P.; Cline, D.; Edgington, D. A review of techniques for the identification and measurement of fish in underwater stereo-video image sequences. In Proceedings of the Videometrics, Range Imaging, and Applications XII; and Automated Visual Inspection, Munich, Germany, 14–16 May 2013.
35. Huang, T.-W.; Hwang, J.-N.; Romain, S.; Wallace, F. Fish Tracking and Segmentation From Stereo Videos on the Wild Sea Surface for Electronic Monitoring of Rail Fishing. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 3146–3158. [CrossRef]
36. Vale, R.; Ueda, E.; Takimoto, R.; Martins, T. Fish Volume Monitoring Using Stereo Vision for Fish Farms. *IFAC-PapersOnLine* **2020**, *53*, 15824–15828. [CrossRef]
37. Williams, K.; Rooper, C.; Towler, R. Use of stereo camera systems for assessment of rockfish abundance in untrawlable areas and for recording pollock behavior during midwater trawls. *Fish. Bull.-Natl. Ocean. Atmos. Adm.* **2010**, *108*, 352–365.
38. Torisawa, S.; Kadota, M.; Komeyama, K.; Suzuki, K.; Takagi, T. A digital stereo-video camera system for three-dimensional monitoring of free-swimming Pacific bluefin tuna, Thunnus orientalis, cultured in a net cage. *Aquat. Living Resour.* **2011**, *24*, 107–112. [CrossRef]
39. Cheng, R.; Zhang, C.; Xu, Q.; Liu, G.; Song, Y.; Yuan, X.; Sun, J. Underwater Fish Body Length Estimation Based on Binocular Image Processing. *Information* **2020**, *11*, 476. [CrossRef]
40. Voskakis, D.; Makris, A.; Papandroulakis, N. Deep learning based fish length estimation. An application for the Mediterranean aquaculture. In Proceedings of the OCEANS 2021, San Diego, CA, USA, 20–23 September 2021.
41. Shi, C.; Wang, Q.; He, X.; Xiaoshuan, Z.; Li, D. An automatic method of fish length estimation using underwater stereo system based on LabVIEW. *Comput. Electron. Agric.* **2020**, *173*, 105419. [CrossRef]
42. Garner, S.B.; Olsen, A.M.; Caillouet, R.; Campbell, M.D.; Patterson, W.F. Estimating reef fish size distributions with a mini remotely operated vehicle-integrated stereo camera system. *PLoS ONE* **2021**, *16*, e0247985. [CrossRef]
43. Kadambi, A.; Bhandari, A.; Raskar, R. 3D Depth Cameras in Vision: Benefits and Limitations of the Hardware. In *Computer Vision and Pattern Recognition*; Springer International Publishing: Cham, Switzerland, 2014; pp. 1–26.
44. Harvey, E.; Shortis, M.; Stadler, M. A Comparison of the Accuracy and Precision of Measurements from Single and Stereo-Video Systems. *Mar. Technol. Soc. J.* **2002**, *36*, 38–49. [CrossRef]
45. Bertels, M.; Jutzi, B.; Ulrich, M. Automatic Real-Time Pose Estimation of Machinery from Images. *Sensors* **2022**, *22*, 2627. [CrossRef]
46. Boldt, J.; Williams, K.; Rooper, C.; Towler, R.; Gauthier, S. Development of stereo camera methodologies to improve pelagic fish biomass estimates and inform ecosystem management in marine waters. *Fish. Res.* **2017**, *198*, 66–77. [CrossRef]
47. Berrio, J.S.; Shan, M.; Worrall, S.; Nebot, E. Camera-LIDAR Integration: Probabilistic Sensor Fusion for Semantic Mapping. *IEEE Trans. Intell. Transp. Syst.* **2022**, *7*, 7637–7652. [CrossRef]
48. John, V.; Long, Q.; Liu, Z.; Mita, S. Automatic calibration and registration of lidar and stereo camera without calibration objects. In Proceedings of the 2015 IEEE International Conference on Vehicular Electronics and Safety (ICVES), Yokohama, Japan, 5–7 November 2015.
49. Roche, V.D.-S.J.; Kondoz, A. A Multi-modal Perception-Driven Self Evolving Autonomous Ground Vehicle. *IEEE Trans. Cybern.* **2021**, 1–11. [CrossRef]
50. Zhong, Y.; Chen, Y.; Wang, C.; Wang, Q.; Yang, J. Research on Target Tracking for Robotic Fish Based on Low-Cost Scarce Sensing Information Fusion. *IEEE Robot. Autom. Lett.* **2022**, *7*, 6044–6051. [CrossRef]
51. Dov, D.; Talmon, R.; Cohen, I. Multimodal Kernel Method for Activity Detection of Sound Sources. *EEE/ACM Trans. Audio Speech Lang. Processing* **2017**, *25*, 1322–1334. [CrossRef]
52. Mirzaei, G.; Jamali, M.M.; Ross, J.; Gorsevski, P.V.; Bingman, V.P. Data Fusion of Acoustics, Infrared, and Marine Radar for Avian Study. *IEEE Sens. J.* **2015**, *15*, 6625–6632. [CrossRef]

53. Zhou, X.; Yu, C.; Yuan, X.; Luo, C. A Matching Algorithm for Underwater Acoustic and Optical Images Based on Image Attribute Transfer and Local Features. *Sensors* **2021**, *21*, 7043. [CrossRef]
54. Andrei, C.-O. 3D Affine Coordinate Transformations. Master's Thesis, School of Architecture and the Built Environment Royal Institute of Technology (KTH), Stockholm, Sweden, 2006.
55. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
56. Kim, B.; Joe, H.; Yu, S.-C. High-precision Underwater 3D Mapping Using Imaging Sonar for Navigation of Autonomous Underwater Vehicle. *Int. J. Control. Autom. Syst.* **2021**, *19*, 3199–3208. [CrossRef]
57. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **1977**, *39*, 1–38.
58. Gkalelis, N.; Mezaris, V.; Kompatsiaris, I. Mixture subclass discriminant analysis. *IEEE Signal Processing Lett.* **2011**, *18*, 319–322. [CrossRef]
59. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Munich, Germany, 5–9 October 2015.
60. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
61. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef]
62. Li, D.; Miao, Z.; Peng, F.; Wang, L.; Hao, Y.; Wang, Z.; Chen, T.; Li, H.; Zheng, Y. Automatic counting methods in aquaculture: A review. *J. World Aquac. Soc.* **2020**, *52*, 269–283. [CrossRef]