Effective marine monitoring with multimodal sensing and improved underwater robotic perception towards environmental protection and smart energy transition.

FARHADI TOLIE, H., REN, J., HASAN, M.J., MA, P, KENNAN, S. and LI, Y.

2024

Journal of Geodesy and Geoinformation Science
http://jggs.chinasmp.com

Article Open Access

Effective Marine Monitoring with Multimodal Sensing and Improved Underwater Robotic Perception towards Environmental Protection and Smart Energy Transition

Hamidreza FARHADI TOLIE[1,2], REN Jinchang[1,2], Md Junayed HASAN[1,2], MA Ping[1,2], Somasundar KANNAN[1], LI Yinhe[1,2]

1 School of Computing, Engineering, and Technology, Robert Gordon University, Aberdeen AB10 7AQ, U. K. ;
2 National Subsea Centre, Aberdeen AB21 0BH, U. K.

ABSTRACT

Effective underwater sensing is crucial for environmental protection and sustainable energy transitions, particularly as we face growing challenges in marine ecosystem monitoring, resource management, and the need for efficient energy infrastructure. To support these efforts, we propose a multimodal sensing approach that enhances underwater detection and distance estimation by combining affordable sonar technology with stereo vision-based depth cameras. Our method integrates the Ping 360 single-beam sonar for target detection and distance measurement with depth refinement from the Intel RealSense D455 camera. A promptable segmentation model automates sonar target detection, overcoming challenges such as acoustic noise and shadowing without requiring large labeled datasets. Depth images from the stereo camera are enhanced using a Depth-Anything model, addressing underwater-specific issues like noise, missing regions, and light attenuation, achieving accurate depth maps for distances up to 1.2 meters underwater. By leveraging multimodal sensing, this approach not only improves underwater robotics for navigation, manipulation, and exploration but also plays a key role in monitoring and maintaining energy infrastructure, such as offshore wind farms and underwater pipelines. Accurate, real-time sensing of these installations ensures more efficient operations, minimizes the environmental impact, and aids in the sustainable management of ocean resources. This enables better energy production and resource utilization, which are essential for a smarter and more sustainable energy transition.

Key words: underwater multi-modal sensing; image fusion; depth refinement; sonar image analysis

Citation: Hamidreza FARHADI TOLIE, REN Jinchang, Md Junayed HASAN, et al. Effective Marine Monitoring with Multimodal Sensing and Improved Underwater Robotic Perception towards Environmental Protection and Smart Energy Transition[J]. Journal of Geodesy and Geoinformation Science, 2024, 7(4): 19-35. DOI: 10.11947/j.JGGS. 2024. 0403.

1 Introduction

The ocean plays a vital role in controlling the global climate, supporting diverse life forms, and providing renewable resources. However, increasing challenges like climate change, overfishing, and pollution are threatening the stability of these ecosystems[1]. To better understand and protect the marine environment, it is essential to use advanced monitoring technologies. Multimodal sensing systems, incorporating both sonar and optical imaging, enable comprehensive underwater mapping and as-

* Foundation support: SeaSense project (No. SPARK-2294).
First author: Hamidreza FARHADI TOLIE.
E-mail:h.farhadi-tolie@rgu.ac.uk.
Corresponding author: REN Jinchang.
E-mail: jinchang.ren@ieee.org.

sessment. These technologies provide valuable insights into ocean features, water quality, and the overall health of marine ecosystems, facilitating improved environmental protection and resource management. Additionally, advancements in underwater robotics have transformed our ability to monitor vast and remote oceanic regions. Equipped with cutting-edge sensing technologies, these robots can navigate and collect data from previously inaccessible areas, significantly enhancing the accuracy of marine assessments.

Leveraging artificial intelligence for data processing, enables robotic systems to detect changes in marine environments, which is vital for protecting biodiversity and fostering sustainable seabed energy projects. Together, multimodal sensing and enhanced robotic capabilities are essential for developing smart energy solutions and ensuring the responsible use of ocean resources. This includes managing fisheries[2] to safeguard vulnerable species and advancing renewable energy development[3] including but not limited to automated inspection and condition monitoring.

However, underwater exploration and task execution have long posed significant challenges due to the complex nature of the subsea environment. Historically, human divers have been tasked with conducting inspections, maintenance, and object retrieval, often at great personal risk[4]. However, with advances in technology, Remotely Operated Vehicles (ROVs) and Autonomous Underwater Vehicles (AUVs) have taken over many of these responsibilities[5]. These vehicles rely heavily on optical sensors for navigation and perception. Unfortunately, the underwater world presents unique obstacles such as poor visibility, light scattering, and water turbidity, which severely limit the effectiveness of traditional optical sensors[6].

To overcome these challenges, the integration of multiple sensing modalities has emerged as a promising approach for enhancing underwater perception[7]. One such advancement is the use of depth cameras, which provide accurate 3D measurements of the environment, crucial for understanding object shapes, distances, and dimensions. Despite their benefits, depth cameras face considerable limitations in underwater environments, particularly in conditions of low visibility where optical sensors struggle to perform effectively.

In these scenarios, sonar-based systems, which rely on acoustic signals rather than light, offer a reliable alternative. Sonar can provide precise distance measurements even in murky or dark environments, making it an invaluable tool for underwater object detection and obstacle avoidance[8]. Multi-beam sonar systems, for example, can generate detailed 3D maps of underwater surroundings, but their high cost often limits their use. In contrast, Single-Beam Sonar (SBS) systems are more cost-effective, though they can suffer from noise and shadowing effects that hinder their ability to provide detailed object-specific information.

Recognising the strengths and limitations of both optical and acoustic sensing systems, our research explores the potential of multi-modal sensing for underwater applications. By integrating stereo camera modules with sonar systems, we aim to leverage the complementary strengths of these two modalities. This combination allows for improved perception, particularly in scenarios where one modality alone might struggle—such as when optical sensors are affected by water conditions or when sonar systems provide insufficient detail.

Rather than focusing solely on the development of a prototype, this work emphasises the potential of this multimodal approach to improve underwater object detection, depth and range measurement, and control. Through the collection and analysis of data from both the stereo camera and sonar modules, we demonstrate how combining these modalities can enhance perception in underwater environments. While we outline a potential prototype for future demonstration, the core of our work lies in proving the efficacy and advantages of this integrated sensing system.

Herein, we aim to develop more robust and reliable underwater sensing technologies, paving the way for improved haptic sensing and robotic control in challenging subsea conditions.

1.1 Current research gap

Despite significant advancements in underwater sensing technologies, several gaps remain in the effective utilisation of multimodal systems, particularly in stereo vision and sonar integration for underwater object detection and distance estimation.

1) Limitations of depth image refinement techniques: Most depth image refinement techniques currently rely on color image- or adjacent pixel-based refinements, which are particularly ineffective in underwater environments where light scattering, colour distortion, and shadow regions degrade the quality of captured data. Techniques such as those proposed by CHEN et al.[9] and MATSUO et al.[10], though useful in air-based scenarios, struggle underwater due to the more pronounced effects of noise, missing regions, and non-uniform colour distributions. Moreover, deep learning-based approaches such as those by ZHANG and WU[11] that aim to refine depth images often fail to address the underwater-specific noise characteristics, especially when large portion of data is missing. Existing methodologies also tend to focus on enhancing edge detection and noise reduction without fully addressing the unique challenges posed by the underwater environment, leaving a gap in robust and reliable depth estimation.

2) Challenges in sonar-based target detection: While sonar systems like the Ping 360 offer a cost-effective solution for underwater sensing, current research primarily focuses on navigation and obstacle avoidance. Target detection in single-beam sonar systems remains under-explored due to inherent limitations such as acoustic shadowing, speckle noise, and low resolution[9]. Most existing object detection methods, as seen in KIM et al.[12] and MCKAY et al.[13], are tailored to imaging sonar systems, which have better spatial resolution as well as shape information helping to distinguish between various objects. However, these methods often oversimplify sonar data when converting it into image-like outputs, ignoring the complexities of acoustic data and reducing their real-world applicability. Additionally, machine learning models for sonar data are typically trained on high-quality data, which may not reflect the noisy and low-resolution reality of single-beam systems like the Ping 360. This results in performance degradation when applied to practical underwater environments, leaving a significant research gap in single beam sonar-based target detection.

(3) Absence of multimodal integration: Although there has been considerable progress in stereo and sonar technologies individually, the integration of these sensing modalities for underwater applications remains largely unaddressed. Existing stereo vision models, including StereoNet[14] and PSMNet[15], demonstrate accurate depth estimation in controlled environments but struggle in underwater scenarios where the reliance on RGB-based refinements fails due to colour distortion and scattering. On the other hand, sonar systems excel at depth perception but lack detailed object information. The absence of a cohesive system that leverages the strengths of both stereo vision and sonar for underwater object detection and distance estimation represents a significant gap in current research.

1.2 Contributions

To address the aforementioned gaps, our research proposes an integrated approach that combines stereo vision with single-beam sonar data to enhance underwater object detection, distance estimation, and noise reduction.

1) Multimodal sensing approach: We propose a novel framework that integrates stereo camera modules with the Ping 360 sonar, leveraging the complementary strengths of optical and acoustic sensors. By combining the 3D spatial awareness of stereo vision with the reliable range detection of sonar, we enhance the accuracy of underwater object detection, especially in scenarios where either modality alone would fail. This integrated approach addresses challenges in murky water and limited object detail in sonar, enabling more reliable monitoring of marine ecosystems. Such improved sensing

is essential for environmental protection and sustainable management of ocean resources, supporting tasks like habitat preservation and seabed assessments for renewable energy projects.

2) Advanced depth image refinement: To overcome the limitations of traditional pixel-based depth refinements, we propose a method that combines RGB-based relative depth images with recorded depth data for more accurate measurements in underwater environments. Unlike previous techniques that fail in these conditions, our approach leverages both modalities to generate more complete and reliable depth maps. This refinement supports better environmental assessments, helping detect changes in oceanic features and ecosystems, which is critical for protecting biodiversity and facilitating sustainable energy infrastructure, such as offshore wind or tidal energy installations.

3) AI-enhanced sonar target detection: We address the challenges of single-beam sonar-based target detection by proposing an AI-based method for automating the distance estimation of multiple objects in real-time sonar data. Instead of relying on conventional object detection methods, we employ a promptable segmentation model that can detect target objects based on their presence, even in noisy and shadowed environments. This AI-driven approach enhances the ability to monitor vulnerable marine species and track changes in underwater habitats, ensuring more responsible resource management and aiding the sustainable development of subsea energy projects.

4) Prototype and real-world validation: While our primary focus is on demonstrating the potential of multimodal sensing, we aim to develop a simplified prototype that integrates stereo vision and sonar modules. This prototype will serve as a practical proof-of-concept, showcasing the viability of multimodal sensing for underwater object detection and manipulation tasks. By validating our approach in real-world scenarios, we contribute to the development of more effective and affordable underwater monitoring systems, which are crucial for both environmental conservation and supporting the transition to smarter, sustainable energy solutions.

2 Relevant Backgrounds

In this section, we will discuss about the necessary backgrounds of the selected sensor models, and relevant technical details for our proposed methodologies.

2.1 Sensor details

2.1.1 Intel RealSense D455 camera

The Intel RealSense D455 camera integrates stereoscopic depth sensors, an RGB sensor with a resolution of 1280 × 800 at a frame rate of 30 frames per second, and an infrared (IR) projector. By utilising these three sensors, the camera generates a depth map by detecting IR light reflected from objects in the scene. Compared to traditional stereo vision systems, the inclusion of IR light allows the RealSense camera to operate effectively even in low-light conditions, which can be beneficial for underwater applications.

Technically, depth images are formed using stereo vision algorithms. The IR projector emits invisible structured IR rays into the scene, which helps to enhance depth perception. Stereo vision algorithms then compute the correlation between each pixel from the left and right cameras using an onboard processor, generating a depth image. It is important to note that the depth image records the distance from the camera plane to the object plane, not the diagonal distance to the objects[16]. As demonstrated in Fig. 1, although the diagonal distances to objects vary, the depth distance remains the same for all four objects.

While the RealSense D455 camera is known for its accuracy in both indoor and outdoor environments, its performance significantly declines in underwater settings. This is primarily due to the refraction of IR light in water, which causes distortions. Moreover, the camera itself is not waterproof, requiring the use of an external waterproof housing, which introduces additional refraction. These factors lead to noisy depth images with significant missing regions[17]. As an example, Fig. 2 illustrates

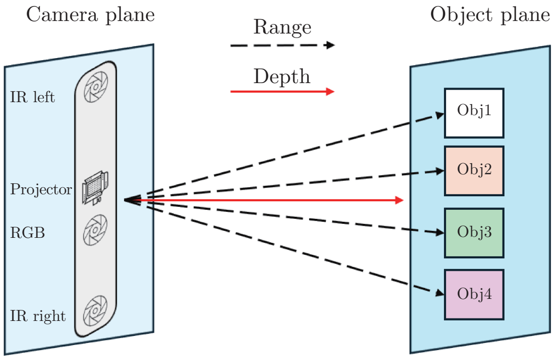


Figure 1 Schematic diagram of the depth and range (diagonal distance) from the camera plane to the objects

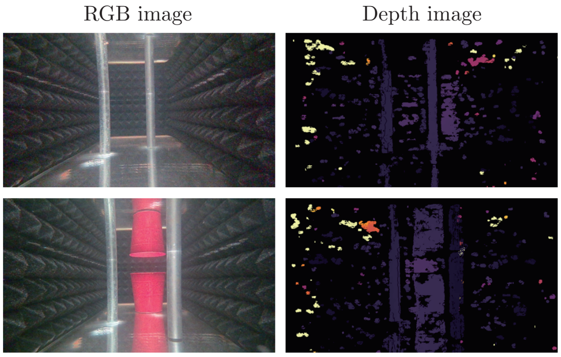


Figure 2 Sample RGB and depth images captured within the water tank

depth images captured in our experimental water tank. As seen, the depth images suffer from outliers, noise, and missing regions, making it challenging to accurately determine the shape and dimensions of objects.

2.1.2 Blue-robotics Ping 360 sonar

For sonar data acquisition, we utilised the Ping 360 sonar, a mechanical SBS designed for localising targets and inspecting underwater structures by detecting sound wave reflections[18]. The sonar operates through a publicly available API[19], providing flexibility in data collection. The Ping 360 offers a scanning range between 0.75 m and 50 m, with a full 360° scanning sector, and adjustable voltage gain levels (low, medium, and high), making it a versatile tool for underwater exploration and tracking.

However, during operation, the Ping 360 generates substantial noise near the sonar head due to the rotational motion of the transducer[18]. This results in a consistent noise zone extending approximately 0.25 m around the sonar head, which must be accounted for when analysing the sonar data.

3 Methodology

3.1 Promptable segmentation for distance measurement from sonar

The Ping 360 sonar system includes a graphical interface that allows users to establish connections, monitor real-time data, and record sonar readings. Additionally, it offers a distance axis for estimating the proximity of objects. However, for more precise distance measurements and to eliminate the need for manual interpretation, we propose integrating Artificial Intelligence (AI) to automate the process. By employing AI, the system can not only identify objects within the sonar image but also measure distances to multiple objects simultaneously.

Traditionally, object detection methods such as those in Literatures [20]—[23] use bounding boxes to identify objects, while segmentation models[24-26] aim to pinpoint exact object locations. However, these methods assume distinct object shapes, making them less effective for SBS data, which typically captures object presence without detailed shape information. In addition, the availability of training data is limited, and the noise and shadowing zones prevalent in underwater environments pose further challenges. As a result, we propose the use of advanced promptable segmentation methods to improve detection accuracy.

The proposed methodology is outlined in Fig. 4. Initially, statistical properties and region labeling of the sonar image are used to generate segmentation prompts, which are then fed into the state-of-the-art Segment-Anything Model (SAM)[28] along with the recorded sonar data to localise objects or targets. Based on the generated mask, the system

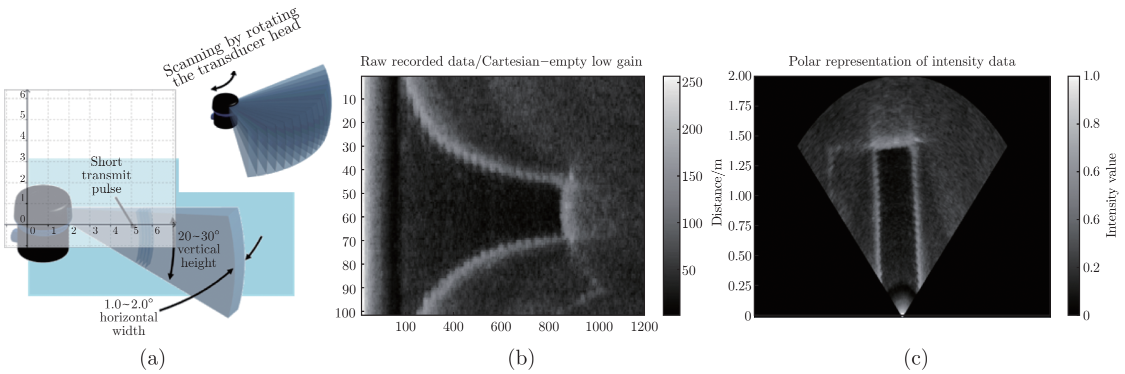


Figure 3 (a) Illustration of the sonar sensor positioned at 0° horizontally, with the transducer head aligned along the x-axis. The expected vertical and horizontal coverage of the transmitted pulse is depicted, along with the scan area (represented by the blue rectangular box) in a hypothetical tank[18]; (b) Raw sonar data from a 102° scan of an empty tank, with a boundary detected at approximately 1.50 meters. The left axis represents intensity values (I) ranging from 0 to 255, displayed in a polar coordinate system; (c) Converted polar representation of the raw data in (b), where the tank boundary is clearly visible at 1.50 meters. This format is typically used by the sonar interface software and is commonly employed in AI-driven research for analysing sonar data. All figures are provided for demonstration purposes.

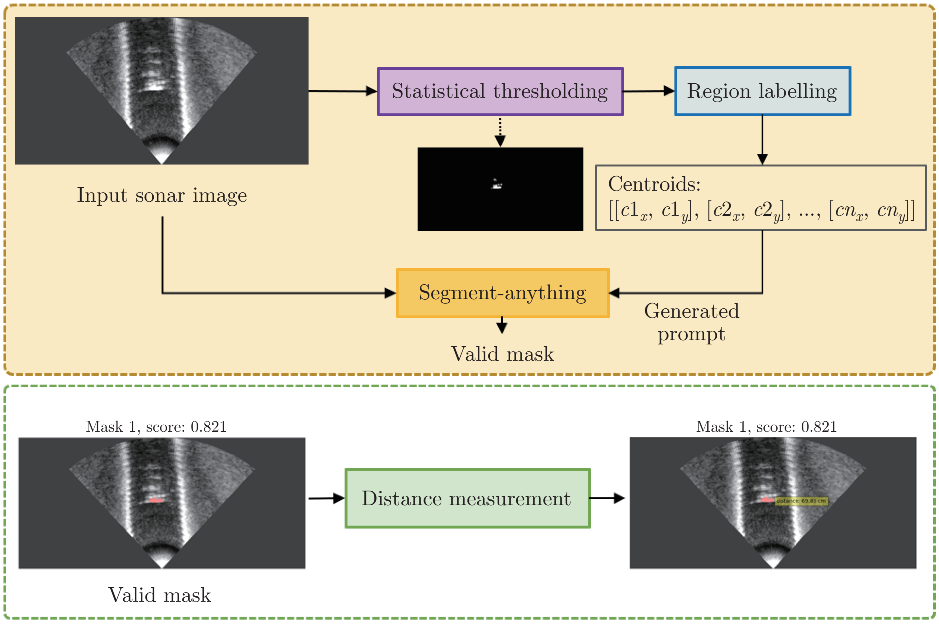


Figure 4 General framework of the proposed methodology using SAM[27]

calculates the diagonal distance to the detected object, taking the sonar's characteristics into account. The SAM model is preferred over simple region labeling due to its ability to more accurately capture complete object boundaries. For example, during our trials, single objects with partial colour coatings were sometimes detected as separate objects through basic region labeling. SAM, however, overcomes such issues by accounting for regional connectivity, leading to more accurate object identification and distance measurements. The subsequent subsections explain the prompt generation and distance measurement processes in detail.

3.1.1 Prompt generation

In the experimental setup, data points from each angle were recorded as intensity values I ranging from 0 to 255, with each point corresponding to a segment of the total scanned distance. For instance, with a maximum distance $D_{\max} = 2$ meters and 1200 samples, each sample point represents approximately 0.001 67 meters. Data within 0.25 meters of the sensor head and beyond 1.40 meters (based on the tank setup) were excluded as unreliable. Statistical thresholding was then applied to the Region Of Interest (ROI), retaining intensity values where I was greater than or equal to $2 \times \mu + \sigma$ (empirically determined), thereby filtering out noise. The mean intensity μ and standard deviation σ for each angle were calculated using the following equations

$$\mu = \frac{1}{N'} \sum_{i=1}^{N'} I_i \tag{1}$$

$$\sigma = \sqrt{\frac{1}{N'-1} \sum_{i=1}^{N'} (I_i - \mu)^2} \tag{2}$$

Where N' represents the number of samples within the ROI.

The denoised data was then transformed from Cartesian to polar coordinates for further analysis, using the following equations

$$r = \sqrt{x^2 + y^2} \tag{3}$$

$$\theta = \arctan\left(\frac{y}{x}\right) \tag{4}$$

To generate the input prompts for the SAM algorithm, we employed Python's scikit-image regionprops function, which identifies potential regions in the filtered image. Regions smaller than 600 pixels (based on empirical testing) were considered noise or shadowing zones and discarded. The central points of the remaining regions were used as input prompts for the SAM algorithm.

3.1.2 Sonar based distance measurement

To compute the distance between the sonar and detected objects, we used the identified object masks within the sonar image. First, a bounding box was drawn around each object based on the mask coordinates. The center point of the closest edge (bottom-most along the x-axis) was then located, and the distance d was calculated using the following equation

$$d = \sqrt{(X_c^o - X_c^s)^2 + (Y_c^o - Y_c^s)^2} \tag{5}$$

Where (X_c^o, Y_c^o) and (X_c^s, Y_c^s) represent the center points of the object and the sonar sensor, respectively.

Once d was computed in terms of pixel distance, we converted the value into centimeters. Given that each centimeter in the sonar image corresponds to 6 pixels (based on a maximum range of 200 centimeters across 1200 samples per angle), the distance d was divided by 6 to obtain the actual distance measurement in centimeters.

3.2 Stereo depth refinement using depth-anything

The framework for the proposed depth refinement strategy is illustrated in Fig. 5. The left and right infrared (IR) cameras of the RealSense system are used to capture the scene, and a depth map is generated based on RealSense's stereo vision technology. Additionally, we incorporate the recently introduced Depth-Anything method to produce a pseudo disparity image, which provides relative disparity information. Since our goal is to refine the depth data generated by the RealSense camera, we utilise the camera's baseline and focal length to convert the pseudo disparity into relative depth values as follows

$$\text{relative depth} = \frac{\text{baseline} \times \text{focal length}}{\text{pseudo disparity}} \tag{6}$$

For the RealSense D455 camera, the baseline and focal length are 95 mm and 1.88 mm, respectively.

As shown in Fig. 5, the generated relative depth image is clear of noise, and object shapes are well defined. To estimate accurate absolute depth values, we map the relative depth values to the RealSense-generated absolute depth data. Sample points with varying intensity values are selected from the relative depth image, and their corresponding absolute depth values are obtained from the RealSense depth image. To minimise the error

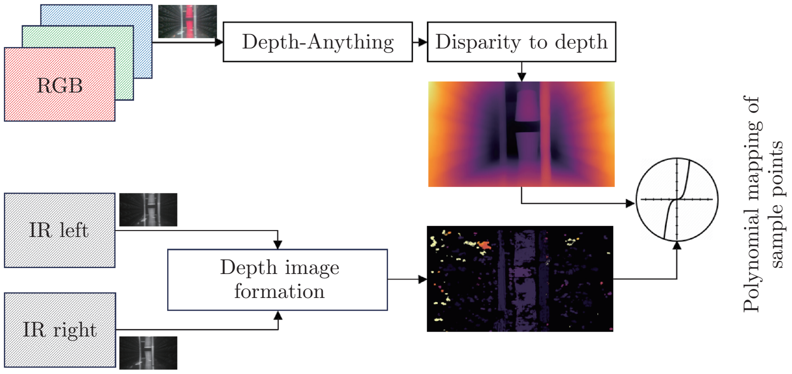


Figure 5 Framework of the proposed refinement strategy[29]

between the estimated and actual depth values, we fit a polynomial curve to the sample points. The curve fitting is described by the following polynomial equation

$$f(x) = a_6x^6 + a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0 \tag{7}$$

Where $\sum_{i=1}^{6} a_i$ are the coefficients determining the influence of each power of x, while a_0 represents the curve's constant term at $x = 0$. These coefficients are determined by solving a least-squares problem using data points from the sample images.

3.3 Stereo depth refinement using BasNet

In addition to the SAM, we have also tried to refine the depth images using saliency maps as they assist in identifying visually interested object regions[30]. To this end, we have utilized the Boundary-Aware Salient Network (BasNet)[31], that excels at detecting salient objects within RGB images by focusing on object boundaries, making it particularly effective for this task. By generating saliency maps, BasNet identifies and highlights the most prominent objects in a scene. In this paper, the obtained saliency maps are applied as masks to the corresponding depth images, effectively filtering out background noise. This process allows us to better distinguish the depth information of the objects of interest, ensuring that the refined depth images are not only more accurate but also more focused on the relevant parts of the scene. The flowchart for depth refinement using BasNet is depicted in Fig. 6.

To enhance the performance of saliency detection, we first preprocess the RGB images by applying denoising techniques to generate higher-quality inputs for the BasNet model. Specifically, we utilize Principal Component Analysis (PCA)[32] to extract the first principal component from RGB images. This approach effectively captures the most significant features while filtering out noise. By reducing the noise in RGB images, PCA allows saliency detection to concentrate on the objects of interest more effectively. The first principal component acts as a denoised representation of the original RGB image, preserving essential information while enhancing the clarity and focus of the image for subsequent processing. Next, these denoised RGB images are fed into BasNet to generate the corresponding saliency maps, which highlight the most visually prominent regions in the images. To further refine the saliency detection results, we apply the OTSU thresholding method[33] to the saliency maps. OTSU is an effective technique for automatically determining an optimal threshold by maximizing the variance between foreground and background regions. Additionally, OTSU is adaptive, making it suitable for handling images with varying illumination and contrast, ensuring consistent performance across different visual conditions. This step converts the continuous saliency values into

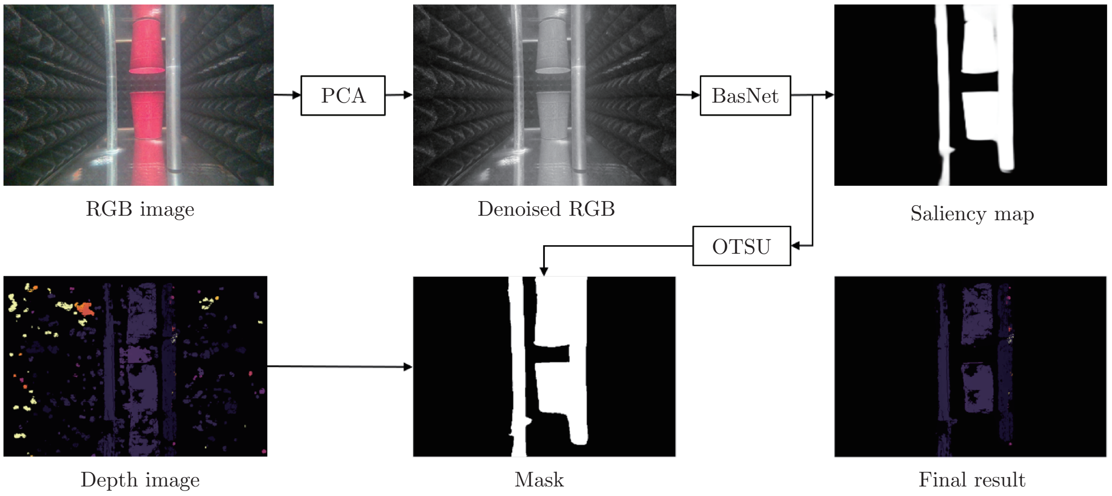
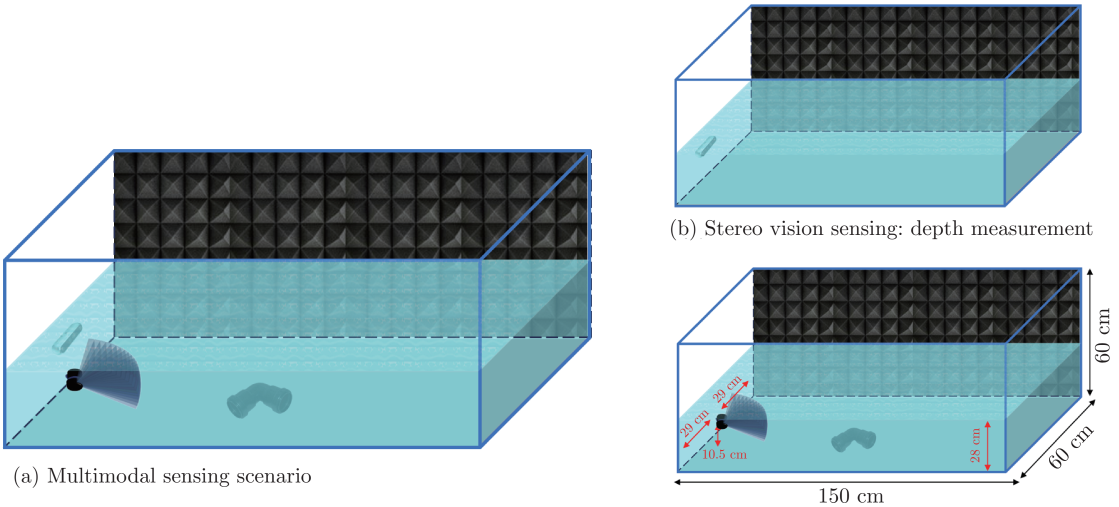


Figure 6 Framework of the depth refinement strategy using BasNet



(a) Multimodal sensing scenario

(b) Stereo vision sensing: depth measurement

(c) Sonar sensing: range measurement

Figure 7 Schematic diagrams of the testing environments

binary masks that precisely segment the objects of interest from the background. The sample generated results from denoising the RGB images to generating object masks, are illustrated in Fig. 6.

4 Further Discussions and Analysis

4.1 Experimental setup

To integrate the sonar and stereo vision-based sensing, we have placed both of these sensors in a water tank shown in Fig. 7(a). The experiments took place within a glass water tank measuring $60 \times 60 \times 150$ cm^3 (height × width × length), with tank walls 1 cm thick. The water level reached 28 cm in height. To enhance reflection quality attributable to the tank's glass structure, acoustic foams were affixed to its interior walls and the camera was attached to the exterior wall. Then, the stereo camera is attached to the exterior wall and the sonar was places inside the tank in an altitude of 10.5 cm from the tank's floor. However, to test each of the sensing modalities separately, we have conducted experiments with each of them individually as shown in Fig. 7(b) and (c).

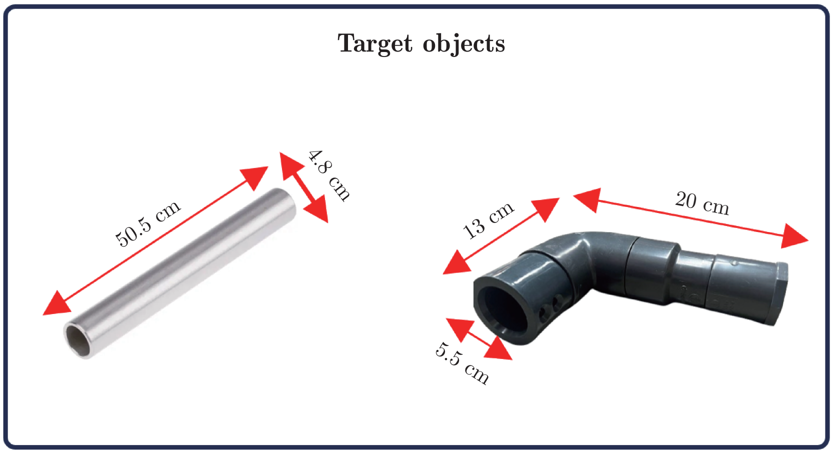


Figure 8 Target objects used for the experiments along with their corresponding dimensions (The uniform pipe on the left is galvanized, while the bent pipe is made of PVC fabric)

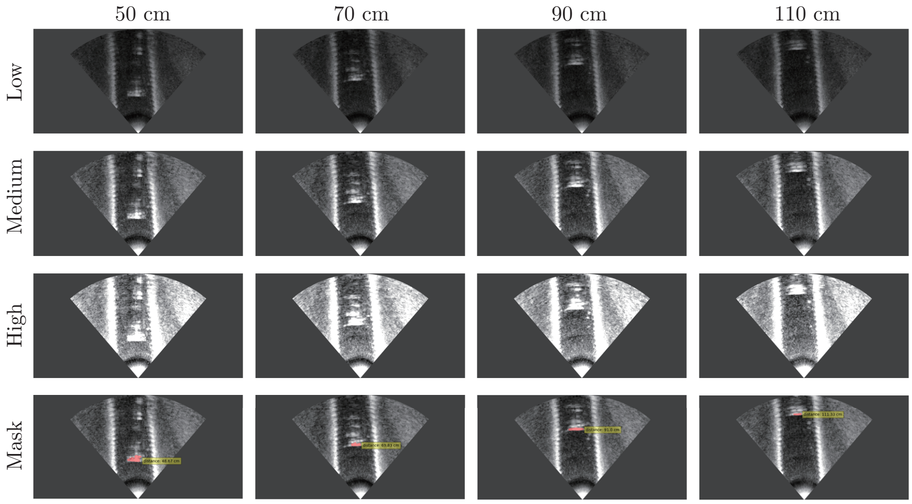


Figure 9 Sonar images obtained with low, medium, and high gain settings for a bent pipe placed at distances of 50, 70, 90, and 110 cm (The generated mask image is based on SAM, with the medium gain image as input)

To validate our depth refinement strategy, we conducted experiments in our water tank using two metal pipes and a bucket to capture depth images by placing these objects at the distances of 60 cm to 120 cm. Depth images were then used to evaluate the accuracy of our distance predictions compared to ground-truth distances. The reported distance values represent depth, which corresponds to the straight-line distance from the camera plane to the object plane. For the sonar, we have conducted experiments using two objects, a uniform and a bent pipe shown in Fig. 8, to evaluate the performance in diagonal distance measurement using the proposed SAM-based approach.

4.2 Sonar-related experimental details

To assess the performance and accuracy of the proposed approach, we first conducted a set of experiments with a bent pipe, as illustrated in Fig. 8. The goal was to identify the optimal gain setting for the Ping 360 sonar. The bent pipe was positioned at diagonal distances of 50, 70, 90, and 110 cm from the sensor, and data were collected at three different gain levels: low, medium, and high. The resulting sonar images are shown in Fig. 9. Based on the observations from the experiments, the following conclusions were made:

1) Low gain produced images with less noise, but the reflections from the object were too weak for clear identification.

2) High gain produced strong reflections from the object, but the images became excessively noisy, and the shadowing zone was exaggerated, leading to potential object misidentification.

3) Medium gain provided the best results, yielding clearer object identification and more accurate distance measurements across the test cases.

4) The appearance of shadow zones was influenced by the sensor's altitude and its viewing angle (whether straight or tilted). At a lower altitude of 10 cm with a straight field of view, an object placed 50 cm away produced more shadowing compared to other distances.

Next, to evaluate the performance of SAM in terms of object identification and distance measurement, we measured the distances of the bent pipe using the medium gain setting and placed it at various distances. The measured results are presented in Tab. 1. Based on the data, the measurement error varied between 0.17 and 1.33 cm, demonstrating high accuracy in object localization and distance estimation. Similarly, experiments were conducted using a uniform pipe (shown in Fig. 8), which was placed at the same distances. The corresponding results are shown in Tab. 2, with an absolute error range of 0.33 to 1.0 cm.

The following general observations were made during the experiments:

1) Although the Ping 360's optimal range is between 0.75 and 50 m, it still provides reliable data for objects located between 0.5 and 0.75 m.

2) A single Ping 360 sensor is insufficient to determine the precise shape or dimensions of an object.

3) With the proposed approach applied to Ping 360 data, diagonal object distances can be automatically measured with a maximum error of 1.5 cm.

4) Using a Tesla T4 GPU on Google Colab, the average processing time per sonar image was approximately 2.5 seconds.

4.3 Stereo vision-related experimental results

To validate the depth refinement strategy, experiments were carried out in a water tank using two metal pipes and a bucket to capture depth images. These images were subsequently used to assess the accuracy of the predicted distances relative to the ground-truth distances. The reported distances represent depth, defined as the straight-line distance from the camera plane to the object plane, as shown in Fig. 1. The key observations from the experiments are as follows:

1) The infrared technology used in the RealSense camera allows it to capture depth images effectively, even in low-light conditions.

2) The operational distance range of the Intel RealSense camera in the underwater setup is between 60 and 120 cm.

3) Depth measurement accuracy decreased by approximately 30% in the underwater environment, requiring a compensation factor of 1.30.

4) The error rate decreased as objects were positioned farther away, dropping to less than 1% of the ground-truth distance for distant objects.

5) The measured distances were influenced by the object's shape and the camera's viewpoint.

6) The refined depth images allowed for the re-

Table 1 Distance measurement results for the bent pipe at different distances

Ground truth/cm	50	70	90	110
Measured distance/cm	48.67	69.83	91.0	111.33
Absolute Error/cm	1.33	0.17	1.0	1.33

Table 2 Distance measurement results for the uniform pipe at different distances

Ground truth/cm	50	70	90	110
Measured distance/cm	51.0	70.67	90.33	111.0
Absolute Error/cm	1.0	0.67	0.33	1.0

construction of more visually accurate 3D images of the underwater scene.

Fig. 10 presents the RGB image, the depth image captured by the RealSense camera, the refined depth image generated using the Depth-Anything model, and the refined depth image using the BasNet-based approach. While the depth image from the RealSense camera lacked detailed information about the environment, the refined depth images effectively highlighted the target objects. As seen in Fig. 10, although the BasNet-based refinement approach produces prominent results by removing noise, it fails to identify objects in distance such as the left pipe of the depth image in the second row of the Fig. 10. Hence, we have conducted depth measurement only by applying the Depth-Anything-based refinement approach.

By applying the proposed Depth-Anything-based refinement strategy—mapping relative depth values to absolute depth data—and taking into account that depth refers to the straight-line distance from the camera to the object plane, we were able to provide detailed depth information for the objects of interest. The measured ground-truth and compensated depth values for each experimental setup are reported in Tab. 3.

From the results, we observed that the Intel RealSense D455 camera provided depth measurements with an absolute error of approximately 1 cm in the given experimental setup. Fig. 11 presents the relative and absolute error values. The minimum error was found for objects placed between 65 and 100 cm from the camera, suggesting that positioning the camera within this range on a robotic arm would minimize error during robotic manipulation. Furthermore, the relative error decreased as the object distance increased, demonstrating the camera's potential for accurate distance measurement at longer

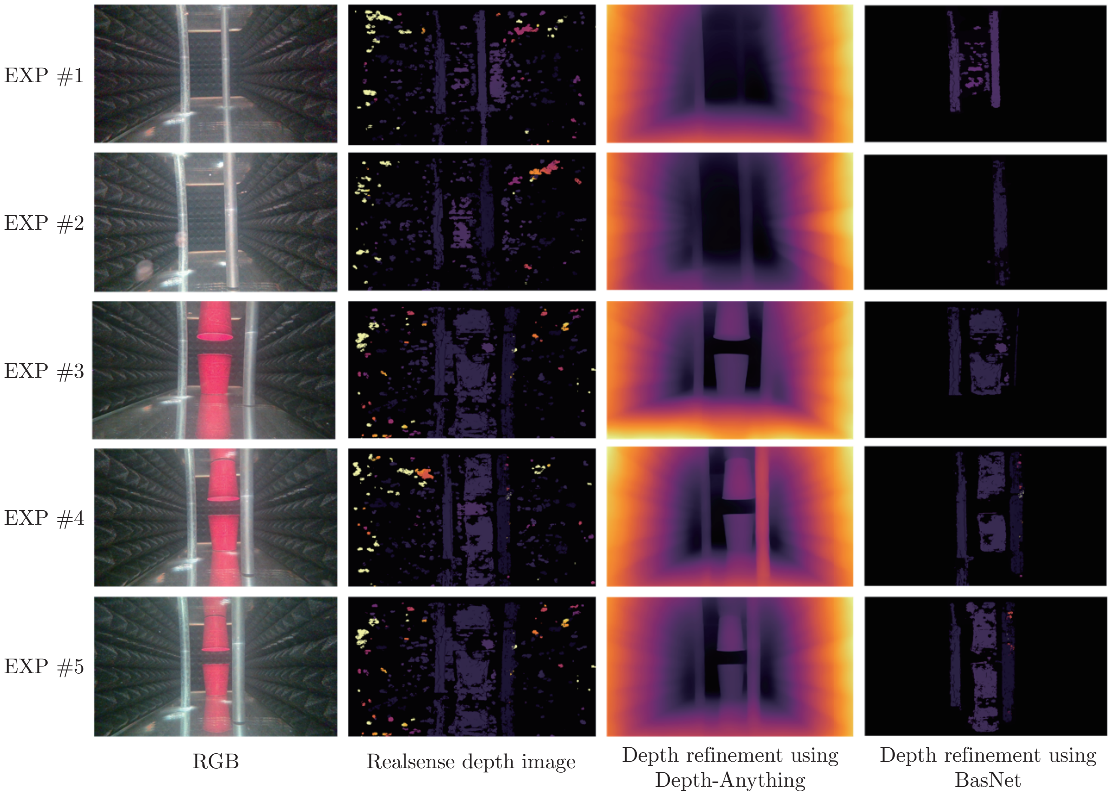

EXP #1
EXP #2
EXP #3
EXP #4
EXP #5
RGB
Realsense depth image
Depth refinement using Depth-Anything
Depth refinement using BasNet

Figure 10 RGB, RealSense depth, and relative depth images captured under five different experimental settings with objects placed at varying distances

Table 3 Distance measurement results for objects in various experiments

Experiment	Object	Ground-truth depth/cm	Compensated measured distance/cm	Absolute error/cm
EXP #1	Left pipe	84.0	83.7	0.3
EXP #1	Right pipe	97.5	98.2	0.7
EXP #2	Left pipe	84.0	83.7	0.3
EXP #2	Right pipe	61.5	62.5	1.0
EXP #3	Left pipe	84.0	83.7	0.3
EXP #3	Right pipe	64.8	63.8	1.0
EXP #3	Bucket	92.5	93.0	0.5
EXP #4	Left pipe	84.0	85.0	1.0
EXP #4	Right pipe	60.7	59.9	0.8
EXP #4	Bucket	104.3	105.3	1.0
EXP #5	Left pipe	84.0	83.7	0.3
EXP #5	Right pipe	60.7	59.9	0.8
EXP #5	Bucket	120.5	119.6	0.9

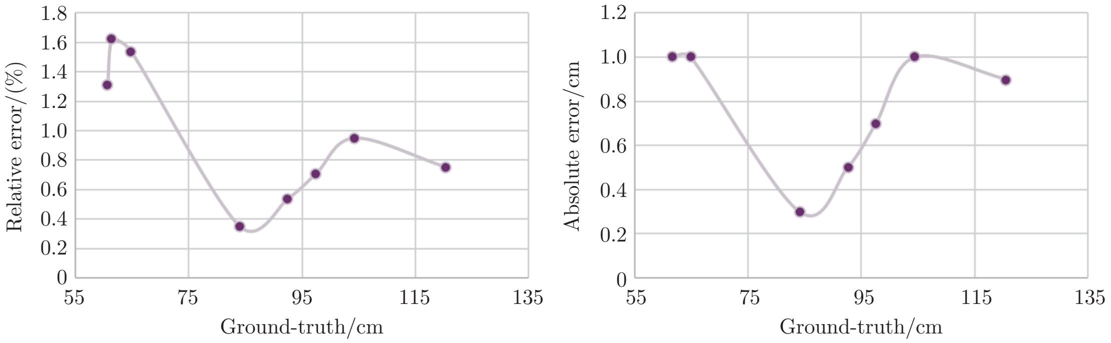


Figure 11 Diagrams showing the relative and absolute depth measurement error

ranges.

5 Contribution Analysis

This section analyzes the proposed contributions of our research and evaluates the extent to which these goals have been achieved based on the experimental results, with a focus on how these advancements contribute to the protection of marine ecosystems.

5.1 Multimodal sensing approach

This contribution has been successfully achieved. The experimental results demonstrate that the multimodal sensing approach effectively compensates for the limitations of each sensor. The sonar provided reliable distance measurements in underwater conditions where stereo cameras struggled, while stereo vision contributed detailed spatial awareness. By integrating AI for automating target detection within sonar images, this approach offers enhanced accuracy and robustness, significantly improving underwater sensing[27]. These advancements are crucial for ecosystem protection, as they enable more precise monitoring of marine habitats, facilitating the detection of changes or threats to biodiversity, such as habitat degradation or overfishing.

5.2 Advanced depth image refinement

The proposed depth refinement strategy has been successfully implemented and validated through experiments. The results indicate that the Depth-Anything-based[34] refinement approach effectively generates accurate and reliable depth images, addressing common challenges such as noise and missing regions. The refined depth images significantly improved measurement accuracy, particularly when objects were positioned at optimal distances from the camera, as reflected in the reported error margins[29]. This improvement in depth accuracy is essential for monitoring underwater ecosystems, as it allows for more detailed assessments of seafloor topography and habitat structure, which are key to maintaining healthy marine environments.

5.3 AI-enhanced sonar target detection

The implementation of a promptable segmentation model, along with statistical signal thresholding, allowed for precise target detection and distance estimation within sonar data. The experimental findings show that the medium gain setting provided the optimal balance between noise reduction and object detection accuracy. The AI-driven models enhanced the precision of these measurements, achieving a maximum error of only 1.5 cm in diagonal distance measurements, validating the effectiveness of this approach. By improving the ability to detect and track objects in complex underwater environments, this contribution aids in the monitoring of endangered species, habitat restoration efforts, and tracking human activities that could harm sensitive ecosystems, such as illegal fishing or underwater construction.

5.4 Prototype and real-world validation

The research successfully demonstrated the potential of the multimodal approach in controlled experimental settings. The experiments with both the Ping 360 sonar and Intel RealSense D455 camera provided tangible evidence of the system's capabilities. However, the description of a fully functional prototype and its testing in diverse real-world scenarios requires further elaboration. While the experimental results are promising, additional testing in more varied and challenging underwater conditions is necessary to ensure the broader applicability and robustness of the proposed system. By extending this validation to real-world environments, the system could play a critical role in environmental monitoring, enabling more accurate data collection in remote or sensitive regions, which is key to protecting marine ecosystems from further degradation.

5.5 Summary

In summary, the research has successfully addressed most of the proposed contributions. The integration of stereo vision with sonar and the advanced depth image refinement techniques represent significant progress in underwater sensing. The AI-enhanced sonar target detection further highlights the innovative nature of this work. These improvements in sensing accuracy and object detection directly support ecosystem protection by enabling more effective monitoring and management of marine environments. Nevertheless, while the experimental results validate the core concepts, further development and testing in real-world environments are essential to fully realize the potential of the proposed system in supporting both environmental conservation and sustainable resource management.

6 Conclusion

In conclusion, we have evaluated the usability of stereo vision-based depth measurement in underwater environments and explored the potential of the Intel RealSense D455 camera for depth estimation. Recognizing the limitations of stereo vision underwater, we proposed a multimodal sensing approach that integrates sonar and depth cameras to enhance underwater target detection and distance estimation. Specifically, we introduced a promptable image segmentation method applied to single-beam sonar images using the Ping 360 sonar for target detection and distance measurement. Our experiments demonstrated the effectiveness of this approach, particularly with a medium gain setting, in accurately identifying objects and measuring distances in complex underwater conditions.

To further improve depth estimation, we utilized the Depth-Anything model to refine the depth images captured by the stereo camera, addressing underwater-specific challenges such as noise, missing regions, and light attenuation. The refinement process, which included polynomial curve fitting, enabled the RealSense camera to produce high-quality depth maps for robotic operations within a range of 60 to 120 cm underwater. Our results validate the effectiveness of this integrated strategy, paving the way for more precise and reliable operations in subsea environments, with applications in underwater navigation, manipulation, and exploration.

The achieved improvements in sensing and detection also have broader implications for environmental conservation and sustainable resource management. By enhancing the precision of underwater sensing, our approach supports better monitoring of marine ecosystems, aiding in the protection of biodiversity and the assessment of underwater energy infrastructure. This is crucial for advancing sustainable energy projects, such as offshore wind farms and tidal energy, ensuring they operate efficiently while minimizing their environmental impact.

For future work, we aim to further improve depth accuracy by enhancing RGB image through extracting noise free features[35] for segmentation and assigning consistent depth values to object pixels using recorded data. This can be achieved through existing segmentation methods[36] or by applying change detection techniques[37-39] that incorporate temporal information and directional guided filters[40-41]. Expanding this multimodal framework in real-world, dynamic underwater conditions will be critical for advancing both environmental protection and the smart energy transition.

References

[1] HOEGH-GULDBERG O, BRUNO J F. The impact of climate change on the world’s marine ecosystems[J]. Science, 2010, 328(5985): 1523-1528.

[2] ZHANG Fengwei, TAO Wenxin. Application of a multimodal model optimized by multi-head-attention mechanism in the classification of fishery resource remote sensing data files[C]//Proceedings of the Third International Conference on Advanced Algorithms and Neural Networks (AANN 2023). Qingdao: SPIE, 2023: 549-554.

[3] LIU Hengguang, XIA Shaohong, FAN Chaoyan, et al. 3D geo-modeling framework for multisource heterogeneous data fusion based on multimodal deep learning and multipoint statistics: a case study in South China Sea[J/OL]. EGUsphere, 2024: 1-44. https://egusphere.copernicus.org/preprints/2024/egusphere-2024-684/.

[4] BARRATT D M, HARCH P G, METER K V. Decompression illness in divers: a review of the literature[J]. The Neurologist, 2002, 8(3): 186-202.

[5] RAY J P. Development of underwater robots for under water inspection and cleaning applications[D]. Lappeenranta: LUT University, 2023.

[6] TOLIE H F, REN Jinchang, ELYAN E. DICAM: deep inception and channel-wise attention modules for underwater image enhancement[J]. Neurocomputing, 2024, 584: 127585.

[7] CAREY N, WERFEL J, NAGPAL R. Fast, accurate, small-scale 3D scene capture using a low-cost depth sensor[C]//Proceedings of 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). Santa Rosa, CA: IEEE, 2017: 1268-1276.

[8] FERREIRA F, MACHADO D, FERRI G, et al. Underwater optical and acoustic imaging: a time for fusion? A brief overview of the state-of-the-art[C]// Proceedings of OCEANS 2016 MTS/IEEE Monterey. Monterey, CA: IEEE, 2016: 1-6.

[9] CHEN Li, LIN Hui, LI Shutao. Depth image enhancement for Kinect using region growing and bilateral filter[C]//Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). Tsukuba, Japan: IEEE, 2012: 3070-3073.

[10] MATSUO K, AOKI Y. Depth image enhancement using local tangent plane approximations[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

Boston, MA: IEEE, 2015: 3574-3583.

[11] ZHANG Xin, WU Ruiyuan. Fast depth image denoising and enhancement using a deep convolutional network[C]//Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China: IEEE, 2016: 2499-2503. DOI: 10.1109/ICASSP.2016.7472127.

[12] KIM B, YU S C. Imaging sonar based real-time underwater object detection utilizing AdaBoost method[C]//Proceedings of 2017 IEEE Underwater Technology (UT). Busan, Republic of Korea: IEEE, 2017: 1-5. DOI: 10.1109/UT.2017.7890300.

[13] MCKAY J, GERG I, MONGA V, et al. What' s mine is yours: pretrained CNNs for limited training sonar ATR[C]//Proceedings of OCEANS 2017-Anchorage. Anchorage, AK: IEEE, 2017: 1-7.

[14] KHAMIS S, FANELLO S, RHEMANN C, et al. StereoNet: guided hierarchical refinement for real-time edge-aware depth prediction[C]//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 573-590.

[15] CHANG Jiaren, CHEN Yongsheng. Pyramid stereo matching network[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE, 2018: 5410-5418.

[16] TADIC V, ODRY A, KECSKES I, et al. Application of intel realSense cameras for depth image generation in robotics[J]. WSEAS Transactions on Computers, 2019, 18: 107-112.

[17] DIGUMARTI S T, CHAURASIA G, TANEJA A, et al. Underwater 3D capture using a low-cost commercial depth camera[C]//Proceedings of 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Placid, NY: IEEE, 2016: 1-9.

[18] HWANG J, BOSE N, ROBINSON B, et al. Sonar based delineation of oil plume proxies using an AUV[J]. International Journal of Mechanical Engineering and Robotics Research, 2022, 11(4): 207-214.

[19] Blue Robotics. Ping-python[EB/OL]. [2024-11-03]. https://github.com/bluerobotics/ping-python.

[20] GIRSHICK R. Fast R-CNN[C]//Proceedings of IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 1440-1448.

[21] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV: IEEE, 2016: 779-788.

[22] NEVES G, RUIZ M, FONTINELE J, et al. Rotated object detection with forward-looking sonar in underwater applications[J]. Expert Systems with Applications, 2020, 140: 112870.

[23] FANG Zhenyu, REN Jinchang, ZHENG Jiangbin, et al. Dual teacher: improving the reliability of pseudo labels for semi-supervised oriented object detection[J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2024. https://ieeexplore.ieee.org/document/10804848.

[24] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Proceedings of the 13th European Conference on Computer Vision–ECCV 2014. Zurich, Switzerland: Springer, 2014: 740-755.

[25] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2961-2969.

[26] CHEN Zhe, WANG Yue, TIAN Wei, et al. Underwater sonar image segmentation combining pixel-level and region-level information[J]. Computers and Electrical Engineering, 2022, 100: 107853.

[27] TOLIE H F, REN Jinchang, HASAN M J, et al. Promptable sonar image segmentation for distance measurement using SAM[C]//Proceedings of 2024 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea). Portorose, Slovenia: IEEE, 2024: 229-233.

[28] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023: 4015-4026.

[29] TOLIE H F, REN Jinchang, HASAN M J, et al. Enhancing underwater situational awareness: realSense camera integration with deep learning for improved depth perception and distance measurement[C]//Proceedings of Artificial Intelligence for Security and Defence Applications II. Edinburgh, United Kingdom: SPIE, 2024: 34-42.

[30] DAI Yuchao, ZHANG Jing, HE Mingyi, et al. Salient object detection from multi-spectral remote sensing images with deep residual network[J]. Journal of Geodesy and Geoinformation Science, 2019, 2(2): 101-110.

[31] QIN Xuebin, ZHANG Zichen, HUANG Chenyang, et al. BASNet: boundary-aware salient object detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recogni-

tion. Long Beach, CA: IEEE, 2019: 7479-7489.

[32] KURITA T. Principal component analysis (PCA)[M]//IKEUCHI K. Computer Vision: a Reference Guide. Cham: Springer, 2019: 1-4.

[33] YANG Xiaolu, SHEN Xuanjing, LONG Jianwu, et al. An improved median-based Otsu image thresholding algorithm[J]. AASRI Procedia, 2012, 3: 468-473.

[34] YANG Lihe, KANG Bingyi, HUANG Zilong, et al. Depth anything: unleashing the power of large-scale unlabeled data[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA: IEEE, 2024: 10371-10381.

[35] MA Ping, REN Jinchang, SUN Genyun, et al. Multiscale superpixelwise prophet model for noise-robust feature extraction in hyperspectral images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5508912.

[36] HASAN M J, ELYAN E, YAN Yijun, et al. Segmentation framework for heat loss identification in thermal images: empowering Scottish retrofitting and thermographic survey companies[C]// Proceedings of the 13th International Conference on Brain Inspired Cognitive Systems. Kuala Lumpur, Malaysia: Springer, 2023: 220-228.

[37] LI Yinhe, REN Jinchang, YAN Yijun, et al. CBANet: an end-to-end cross-band 2-D attention network for hyperspectral change detection in remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5513011.

[38] ZHANG Erlei, ZONG He, LI Xinyu, et al. ICSF: integrating inter-modal and cross-modal learning framework for self-supervised heterogeneous change detection[J/OL]. IEEE Transactions on Geoscience and Remote Sensing, 2024. https://ieeexplore.ieee.org/document/10807406.

[39] YAN Yijun, REN Jinchang, SUN He, Robert WILLIAMS. Nondestructive quantitative measurement for precision quality control in additive manufacturing using hyperspectral imagery and machine learning[J]. IEEE Transactions on Industrial Informatics, 2024, 20(8): 9963-9975. DOI: 10.1109/TII.2024.3384609.

[40] TOLIE H F, FARAJI M R, QI Xiaojun. Blind quality assessment of screen content images via edge histogram descriptor and statistical moments[J]. The Visual Computer, 2024, 40(8): 5341-5356.

[41] TOLIE H F, REN Jinchang, CHEN Rongjun, et al. Blind sonar image quality assessment via machine learning: leveraging micro- and macro-scale texture and contour features in the wavelet domain[J]. Engineering Applications of Artificial Intelligence, 2025, 141: 109730.