Journal of *Marine Science and Engineering*

MDPI

Article

Advanced Underwater Measurement System for ROVs: Integrating Sonar and Stereo Vision for Enhanced Subsea Infrastructure Maintenance

Jiawei Zhang [1], Fenglei Han [1,*], Duanfeng Han [1], Jianfeng Yang [1], Wangyuan Zhao [1] and Hansheng Li [2]

[1] College Of Shipbuilding Engineering, Harbin Engineering University, Harbin 150001, China; zhang950719@hrbeu.edu.cn (J.Z.); handuanfeng@hrbeu.edu.cn (D.H.); hertz@hrbeu.edu.cn (J.Y.); zhaozuai@hrbeu.edu.cn (W.Z.)

[2] Instituto Superior Técnico, University of Lisbon, 1049-001 Lisboa, Portugal; hansheng.li@centec.tecnico.ulisboa.pt

* Correspondence: fenglei_han@hrbeu.edu.cn

Abstract: In the realm of ocean engineering and maintenance of subsea structures, accurate underwater distance quantification plays a crucial role. However, the precision of such measurements is often compromised in underwater environments due to backward scattering and feature degradation, adversely affecting the accuracy of visual techniques. Addressing this challenge, our study introduces a groundbreaking method for underwater object measurement, innovatively combining image sonar with stereo vision. This approach aims to supplement the gaps in underwater visual feature detection with sonar data while leveraging the distance information from sonar for enhanced visual matching. Our methodology seamlessly integrates sonar data into the Semi-Global Block Matching (SGBM) algorithm used in stereo vision. This integration involves introducing a novel sonar-based cost term and refining the cost aggregation process, thereby both elevating the precision in depth estimations and enriching the texture details within the depth maps. This represents a substantial enhancement over existing methodologies, particularly in the texture augmentation of depth maps tailored for subaquatic environments. Through extensive comparative analyses, our approach demonstrates a substantial reduction in measurement errors by 1.6%, showing significant promise in challenging underwater scenarios. The adaptability and accuracy of our algorithm in generating detailed depth maps make it particularly relevant for underwater infrastructure maintenance, exploration, and inspection.

Keywords: marine structure measurement; underwater extreme environments; stereo matching; image sonar; marine engineering

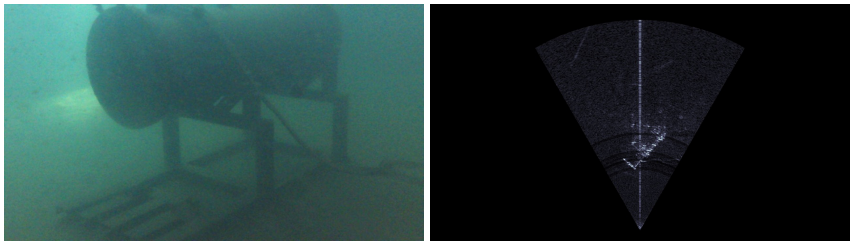
Citation: Zhang, J.; Han, F.; Han, D.; Yang, J.; Zhao, W.; Li, H. Advanced Underwater Measurement System for ROVs: Integrating Sonar and Stereo Vision for Enhanced Subsea Infrastructure Maintenance. *J. Mar. Sci. Eng.* **2024**, *12*, 306. https://doi.org/10.3390/jmse12020306

Received: 3 January 2024
Revised: 26 January 2024
Accepted: 7 February 2024
Published: 9 February 2024

1. Introduction

In recent years, ocean research and exploration have become increasingly important due to the expanding interest in underwater resources such as oil and gas, minerals, and seafood all over the world. To reach and study the depths of the ocean, specialized equipment such as submarines, ROVs (remotely operated vehicles), and diving suits are used for both underwater exploration and maintenance. These tools can be used for underwater maintenance, repairs, and inspections on submerged structures such as ships, oil platforms, pipelines, and cable networks, while high-precision underwater object measurement raises significant demand for the detection and tracking of marine organisms and structures. Whether it is for surveying the seafloor, locating lost, damaged structures, or laying down new pipelines, high-precision measurements are fundamental to ensuring safety, efficiency, and accuracy [1]. Consequently, high-precision underwater measurement is a rapidly growing field, with new technologies and techniques continually being developed and improved to keep abreast of the challenges of operating in the deep ocean [2].

Visual methods are commonly used for land-based applications. Most of the research in underwater environments has focused on restoration, reconstruction, and color correction. Recognition, depth, and shape recovery are also important but less researched areas, with limited datasets providing ground truth, depth information, and labeled data [3]. Stereo vision is an essential technology for many computer vision systems that require depth information. There are several commonly used algorithms for feature detection and matching in computer vision, such as the Semi-Global Block Matching (SGBM) algorithm and others like SGM [4], ORB [5], SIFT [6], etc. Among them, SGBM has been widely used for stereo matching tasks such as object detection, tracking, and scene reconstruction, but imaging conditions in underwater environments are challenging due to scattering, attenuation, limited visibility, and low light conditions, making it difficult to obtain accurate 3D maps of underwater objects [7]. Meanwhile, image acquisition by underwater platforms and divers can also introduce motion artifacts, blurring, and jitter, further degrading the image quality. As another commonly used underwater ranging device, image sonar is relatively robust and can be used regardless of water quality and lighting conditions. However, as shown in Figure 1, its resolution is low, and its imaging quality is often coarse, making it unsuitable for tasks that require color-related information, such as underwater pipeline corrosion measurement [8].



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

Figure 1. Comparative visualization of an underwater object. (**a**) captured through optical imaging, showing the object's appearance with ambient light, and (**b**) captured through image sonar, depicting acoustic reflections used for object detection and mapping in low-visibility underwater conditions.

To address these challenges, a promising approach is to combine the strengths of acoustic and optical sensors through multimodal data fusion [9]. Multimodal data fusion has emerged as a promising approach for integrating data from multiple sensors to provide more comprehensive and accurate measurements in various domains, including underwater ranging. By combining the strengths of acoustic and optical sensors, the resulting system can provide high-resolution images with color information while also accurately measuring distance. In such systems, the acoustic sensor provides accurate range measurements, while the optical sensor captures high-resolution images that complement the acoustic data. The two sets of data are then combined using sophisticated algorithms to produce a three-dimensional representation of the scene that is more complete and accurate than either sensor could achieve on its own. This integrated approach has the potential to significantly enhance our ability to detect and monitor underwater objects and environments and has applications in fields such as marine biology, ocean engineering, and underwater exploration. However, the use of the optical–acoustic method is still relatively unexplored.

In this paper, we introduce an enhanced version of the SGBM method. The proposed approach leverages the information obtained from image sonar to guide the cost aggregation and disparity calculation stages, thereby integrating more information into the SGBM algorithm for more accurate depth map reconstructions. The main objective of our research is to improve the accuracy and robustness of depth estimation in challenging environments where traditional stereo matching methods often fail to produce satisfactory results. By introducing the sonar data, we hope to overcome some of the limitations of traditional SGBM and enhance its performance in complex scenarios. Our experiments demonstrate

that the improved SGBM method with sonar data integration outperforms traditional SGBM in terms of both accuracy and computational efficiency.

The rest of this paper is structured as follows. In Section 2, we review relevant research on underwater stereo matching methods and optical–acoustic methods. In Section 3, we present our proposed method, which combines semi-global stereo matching with multibeam image sonar. In Section 4, we present the results of our experiments and provide the analysis. Finally, in Section 5, we offer our conclusions.

2. Related Work

2.1. Stereo Matching Methods

Stereo vision, a key technology in computer vision systems for depth information acquisition, has seen broad adoption across various fields. Its applications extend to marine engineering, notably in fish size measurement [10,11], icebreaking processes [12], and 3D reconstruction [13,14]. These applications highlight stereo vision's crucial role in enhancing our understanding and interaction with marine environments. However, the challenges posed by underwater settings, including issues of scattering, attenuation, limited visibility, and low light conditions, necessitate innovative approaches. Researchers have thus ventured into various methodologies for underwater stereo matching, employing specialized camera systems engineered to function effectively in these challenging conditions. For instance, Zhai et al. proposed an underwater ranging method based on the frequency comb laser [15]. Additionally, the adaptation of stereo matching algorithms has emerged as a vital strategy for addressing issues like light attenuation and limited visibility, ultimately bolstering the robustness and reliability of underwater stereovision. Xu et al. proposed an energy function suitable for underwater scenes. It designed primary data terms and smoothing terms and proposed bilateral filtering to collect initialization matching costs and fill operations to handle occlusion issues [16]. Skineer et al. proposed an unsupervised model called UWStereoNet based on deep learning, which addressed the problem of 3D perception and image processing in underwater scenes through image depth estimation and color correction [17].

However, most of the existing stereo matching algorithms for underwater environments still have limitations in terms of accuracy and robustness. Therefore, there is a need to explore new approaches for more accurate and reliable depth estimation in challenging underwater environments.

2.2. Optical–Acoustic Methods

In addition to the inherent challenges in underwater vision systems, the utilization of image sonar introduces its own set of restrictions, including a deficiency in height angle information resulting in an incomplete understanding of the three-dimensional scene and a relatively reduced spatial resolution. Many researchers have investigated these issues. Huang et al. proposed an innovative method, ASFM, for recovering 3D scene structure from multiple 2D sonar images without assuming a flat surface, and it can use information from many frames [18]. Karimanzira et al. described the development of a generalized solution for underwater object detection using AutoML principles by sonar [19]. Purser et al. presented a new towed camera platform that integrates additional acoustical devices to collect seafloor photo and video data at depths up to 6000 m [20]. McConnell et al. aimed to restore the elevation angle by using a stereo pair of orthogonally oriented imaging sonars [21]. Cho et al. proposed a 3D seafloor scanning method using sonar images obtained by an acoustic lens-based multibeam sonar (ALMS) [22]. While significant progress has been made through the application of deep learning methods and the augmentation of acoustic sensor arrays, it is crucial to acknowledge that these advancements, although promising, have not yet fully unlocked the innate potential of acoustic sensors.

Recognizing this, optical–acoustic approaches have surfaced as a highly promising avenue for underwater ranging and imaging. These innovative methodologies leverage the

complementary strengths of both acoustic and optical sensors to yield more comprehensive and precise measurements. Optical sensors can capture high-resolution images with color information, while acoustic sensors can provide accurate range measurements. The two sets of data are then combined using sophisticated algorithms to produce a three-dimensional representation of the scene that is more complete and accurate than either sensor could achieve on its own. For the fusion model, Negahdaripour et al. presented the epipolar geometry of an optical–acoustic stereo imaging system for underwater inspections [23]. Pecheux et al. used motion comparisons between images from the monocular camera and multibeam imaging sonar to compute the transformation matrix between the camera and the sonar and estimate the camera's focal length [24]. While significant progress has been made in the field of robot navigation through the integration of acoustic and optical sensors, there remains substantial untapped potential in the realm of precision ranging techniques. Terayama et al. proposed a system that successfully generates realistic daytime images from sonar and night camera images [25]. Raaj et al. presented a method for underwater object localization using a combination of cameras, sonars, and odometry information [26]. Cong et al. proposed a system that combines a stereo camera with a multi-beam echo sounder for underwater object detection and tracking [9]. Rahaman et al. proposed a method of fusing sonar data into the visual–inertial SLAM framework, specifically, the OKVIS [27].

Furthermore, optical sensors, while offering unique advantages in terms of high-resolution imaging, are affected by issues such as light attenuation and reduced visibility in turbid waters. This combination of optical and acoustic sensing has the potential to provide comprehensive and precise ranging information, thereby enhancing robot navigation capabilities.

The objective of this work is to address these challenges and unlock the full potential of precision ranging in the context of underwater robot navigation. By leveraging advanced algorithms, we aim to refine and expand the current state of the art in underwater ranging. Not only is our research academically significant, but it also holds the potential to revolutionize underwater robotics, enabling applications in fields such as marine science, resource exploration, and environmental monitoring. In summary, precision ranging techniques for acoustic and optical fusion represent a compelling avenue for future research and development in ocean engineering.

3. Proposed Method

Figure 2 illustrates the framework.

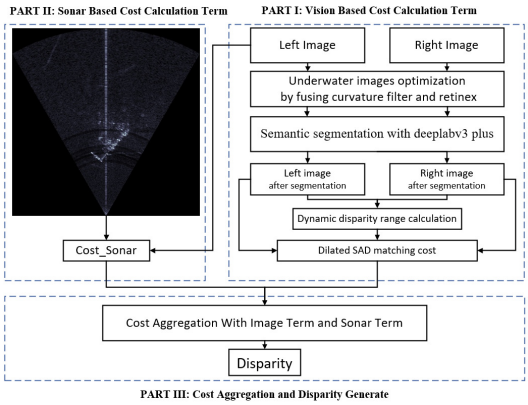


Figure 2. Framework of the proposed method. This framework shows a multi-stage approach to underwater stereo matching with a stereo camera and image sonar, starting with PART I: Vision-Based Cost Calculation utilizing stereo image optimization and semantic segmentation. PART II: Sonar-Based Cost Calculation, where sonar imagery informs cost metrics. PART III: Aggregation of costs from both visual and sonar data to compute disparity, illustrating the synergy between acoustic and optical sensing modalities for enhanced underwater perception.

3.1. Sensor Model and Camera Model

In a sonar system, the coordinate of a 3D point $\mathbf{p}$ can be denoted as $\mathbf{P}_S = (X_s, Y_s, Z_s)^T$. The Cartesian spherical coordinate (R, θ, ϕ) can be transformed as

$$\mathbf{P}_s = \begin{bmatrix} X_s \\ Y_s \\ Z_s \end{bmatrix} = R \begin{bmatrix} \cos\phi \sin\theta \\ \cos\phi \cos\theta \\ \sin\phi \end{bmatrix} \tag{1}$$

and

$$\begin{bmatrix} \mathcal{R} \\ \theta \\ \phi \end{bmatrix} = \begin{bmatrix} \sqrt{X_s^2 + Y_s^2 + Z_s^2} \\ \tan^{-1}(X_s / Y_s) \\ \tan^{-1}\left(Z_s / \sqrt{(X_s^2 + Y_s^2)}\right) \end{bmatrix}, \tag{2}$$

where θ and ϕ mean azimuth and elevation angles, respectively. The range R can be defined as

$$R = \frac{c\delta t}{2}, \tag{3}$$

where c is the sound velocity in the water and t is time factor, denoted by

$$\Delta t = 2\frac{R_{S\max} - R_{S\min}}{cN_t} = t_{i+1} - t_i, \quad i = 0, 1, \ldots, N_t - 1, \tag{4}$$

where N_t, the desired number of intervals (i.e., N_t = 512 for M1200D 2.1 [MHz]), fixes the range resolution, and i represents the bin number.

The sonar image, denoted by $I_s(R, \theta)$ and referred to as beam–bin data, is a 2D image that specifically conveys information in the range and azimuth directions. Notably, it does not include data related to elevation and, as such, it lacks elevation information. To utilize the beam–bin data $I_s(R, \theta)$, we transformed these polar coordinates to points $\mathbf{s} = [x_s, y_s] = R[sin\theta, cos\theta]$ on a zero-elevation plane.

For the camera model, the perspective camera model is the most commonly used. The perspective camera model assumes a pinhole projection system, where an image is formed as rays of light from objects pass through the center of the lens (the projection center) and intersect to create an image on a focal plane.

In the camera system, the coordinate of a 3D point $\mathbf{p}$ can be denoted as $\mathbf{P}_c = (X_c, Y_c, Z_c)^T$. Additionally, $\mathbf{p}_c = (u, v, 1)^T$ is the image coordinate result of using perspective projection to project a point from the camera coordinate system to the image coordinate system. The mapping from the 3D camera coordinate system to the 2D image coordinate system is described by the perspective projection equation:

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = KP = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix}, \tag{5}$$

where (f_x, f_y) are the camera's focal lengths, (c_x, c_y) is the optical center of the image, and λ is a scale factor, typically effectively equal to the distance Z_c from the object point to the camera's center.

This perspective projection equation projects 3D world points (X_c, Y_c, Z_c) onto 2D image points (u, v). The equation takes into account the focal lengths, optical center, and spatial location of the point, resulting in a perspective effect where objects farther from the camera appear smaller in the image, and objects closer to the camera appear larger.

Multi-beam image sonar is an acoustic sensor that emits multiple beams of sound waves and measures the time delay and intensity of the reflected echoes to generate an image of the scene. Compared to optical sensors, sonar is relatively robust and can be used regardless of water quality and lighting conditions. In order to fuse information from both sonar and

camera sensors, optical–acoustic epipolar geometry is a critical step. In the context of the combined binocular and sonar coordinate system, a point p that is visible in both the camera and sonar coordinate systems can be mathematically represented as:

$$\mathbf{P}_s = \mathbf{R}\mathbf{P}_c + \mathbf{T}, \tag{6}$$

where $\mathbf{P}_s$ represents the coordinates of point p in the sonar coordinate system, R is the rotation matrix that defines the transformation from the camera coordinate system to the sonar coordinate system, T is the translation vector that accounts for the displacement between the camera and sonar coordinate systems, and $\mathbf{P}_c$ represents the coordinates of point p as observed in the camera coordinate system. For the sake of simplifying calculations, we have:

$$T = \begin{bmatrix} t_x \\ t_y \\ 0 \end{bmatrix} \quad R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}, \tag{7}$$

which means there is no rotation between the camera and sonar coordinate systems, simplifying the transformation to a linear translation. Referring to Figure 3, the diagram illustrates the combined camera–sonar coordinate system. Given calibrated camera and sonar points, and knowing the rotation (**R**) and translation (**T**) parameters, we can obtain

$$\begin{aligned} X_s &= \mathbf{r}_1\mathbf{P}_c + t_x = Z_c\mathbf{r}_1\mathbf{p}_c + t_x, \\ Y_s &= \mathbf{r}_2\mathbf{P}_c + t_y = Z_c\mathbf{r}_2\mathbf{p}_c + t_y, \end{aligned} \tag{8}$$

where $\mathbf{r}_i (i = 1, 2, 3)$ represents the row vectors of the rotation matrix. Since the sonar's elevation angle is within the range of $\pm 6^\circ$, we have:

$$\begin{aligned} &\sqrt{X_s^2 + Y_s^2} = \mathcal{R}\cos\phi \approx \mathcal{R} = \sqrt{x_s^2 + y_s^2}, \\ &X_s = x_s, Y_s = y_s, \end{aligned} \tag{9}$$

so (8) can be written as:

$$\begin{aligned} x_s &= \mathbf{r}_1\mathbf{P}_c + t_x = Z_c\mathbf{r}_1K^{-1}\mathbf{p}_c + t_x, \\ y_s &= \mathbf{r}_2\mathbf{P}_c + t_y = Z_c\mathbf{r}_2K^{-1}\mathbf{p}_c + t_y. \end{aligned} \tag{10}$$

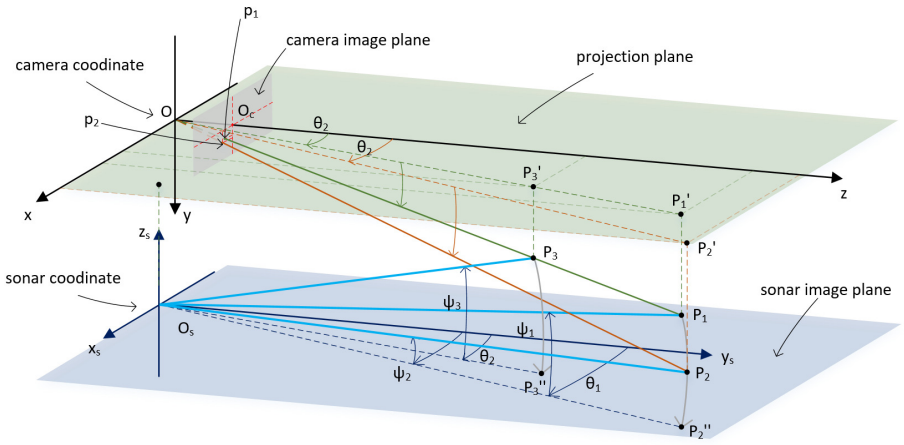


Figure 3. Joint coordinate system for left camera and sonar.

3.2. Vision-Based Cost Calculation Term

Stereo matching, which is the process of finding correspondences between points in a pair of stereo images, can typically be divided into the following four main steps: cost calculation, cost aggregation, disparity calculation, and disparity optimization. In

this section, we describe the algorithm we used in the cost calculation part of the SGBM for image data terms. In this step, a method using a dilated SAD window along with a dynamically determined disparity range derived through semantic segmentation is utilized to enhance the accuracy of the cost computation. To maintain the coherence of the paper, we provide a concise overview of this aspect within this section, considering it as the initial stage of stereo matching.

In practical operational scenarios of a real underwater environment, only the relevant information of the target objects is meaningful. Therefore, the extraction of these target objects from the images is of paramount importance. To achieve this objective, a semantic segmentation method is employed for the specific reason that it preserves the contour and shape of the target objects during extraction. In this method, the DeepLab V3 Plus model was employed for the purpose of conducting semantic segmentation. This approach allows for the optimization of image data utilization pertaining to the object of interest by harnessing semantic details about the target.

The purpose of matching cost calculation is to quantify the correlation between the pixel under consideration and the candidate pixel. Regardless of whether two pixels are corresponding points, their matching cost can be calculated using a matching cost function. A smaller cost indicates a higher degree of correlation, implying a greater probability of being corresponding points. The BT-SAD cost refers to the matching cost calculation in the SGBM algorithm. As shown in Figure 4, the scale of the cost volume is reduced. To improve the robustness of the BT-SAD cost under textureless conditions, we introduced an enhanced iteration of the BT-SAD algorithm, which incorporates a dilated SAD window tailored to the specific requirements of underwater measurement and image matching (see Figure 5). Before searching for corresponding pixels, it is common to specify a disparity search range for each pixel, limiting the range to within d, and the disparity d is treated as a fixed range, $d \in [D_{min}, D_{max}]$. However, a fixed range may not meet the requirements for all issues. To address this, we propose the disparity arrays of D_{min} and D_{max}, which are generated based on the object's outline present in the mask image. In row i, the disparity pair (D_{min}, D_{max}) is illustrated as follows:

$$\begin{aligned} \overline{D}_{max}(i) &= \max\{|I_r^+ - I_l^+|, |I_r^- - I_l^-|\} + \gamma \\ \overline{D}_{min}(i) &= \min\{|I_r^+ - I_l^+|, |I_r^- - I_l^-|\} + \gamma, \end{aligned} \tag{11}$$

where I_l^+ and I_r^+ mean the left boundary coordinates in the left image I_L and the right image I_R, respectively, and I_l^- and I_r^- mean the right boundary coordinates in the left image I_L and the right image I_R, respectively, while γ is a dynamic parameter that is fixed in the computation to fit the result. Points along the same epipolar line can possess a shared disparity $\overline{D}_{max}$ and $\overline{D}_{min}$.

To avoid redundancy, the aforementioned is only a concise overview. A comprehensive explanation of the theory, pertinent parameters, and experimental details can be found in [28].

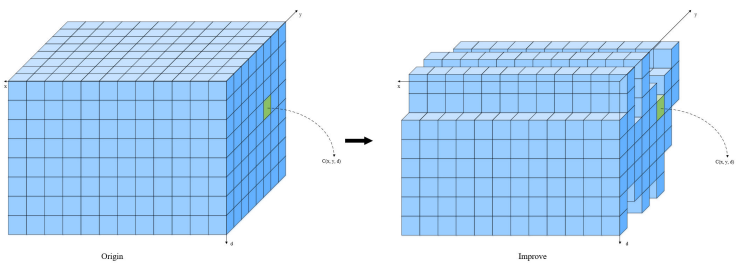


Figure 4. Cost volume. The left volume is the origin cost volume, and the right volume is the cost volume after improvement, where C means the cost.

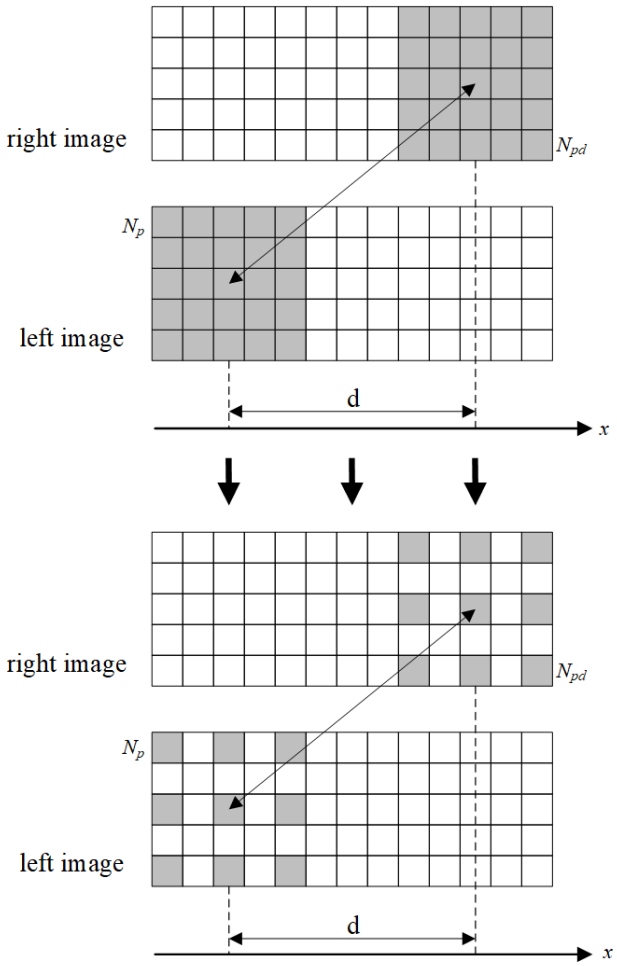


Figure 5. The upper portion of the illustration depicts the traditional process for calculating the SAD cost. In contrast, the lower portion illustrates the cost calculation process utilizing a dilated SAD window. N_P is the reference support window, and N_{pd} is the target support window.

3.3. Sonar-Based Cost Calculation Terms

In this section, we delve into the integration of sonar data into the cost calculation framework. This strategic incorporation of sonar data aims to significantly elevate the precision and robustness of stereo matching processes.

Sonar data bestow a unique advantage by furnishing essential depth information in the context of underwater scenarios. This information plays a pivotal role in achieving more precise determinations of object positions and spatial distances. The primary objective of fusing sonar cost calculations into the SGBM algorithm is to bolster the overall accuracy of stereo matching, a particularly challenging task in the underwater domain. Therefore, this synergy not only elevates the performance and dependability of underwater vision systems but also equips them with vital insights, thereby amplifying the accomplishment rate and operational efficiency of underwater missions.

In the process of matching the sonar image I_s, the primary reference is derived from the left image I_L. As depicted in Figure 3, two distinct points, P_1 and P_3, in the world coordinate system are projected to the same point, p_1, in the image plane. However, these two points project as distinct entities in the sonar plane. In the world coordinate system, their projection corresponds to a single point in the image plane, which, when transposed to the sonar coordinate system, results in two distinct points. A single point in the sonar plane can also correspond to two distinct points when projected onto the image plane. Hence,

we define the pixel dissimilarity for the sonar component by considering both the sonar echo intensity and the pixel values within the sonar image. This cost can be written as:

$$C_s(p_c, d) = 255 - I_s(x_s, y_s), \tag{12}$$

where d is the disparity between the matched points from the stereo image pair. The maximum pixel value for an 8-bit image is 255, which is why it is used as the factor. Using Equation (10), we can transmit the left image key point p_c to the sonar point p_s. For the known parameter Z_c, this can be calculated by the equation:

$$Z_c = \frac{fb}{d}, \tag{13}$$

where d means the baseline of the stereo camera.

Given the inherent lack of elevation angle information in sonar images, the vertical dimension tends to be overlooked during acoustic and optical matching. To address this limitation, we employ sub-pixel interpolation techniques to enhance the wealth of pixel information concerning anterior and posterior positions. By modifying Equation (12) to

$$\begin{aligned} \hat{C}_s(p_c, d) = 255 - \max(&|I_s(x_s + 1, y_s + 1) - I_s(x_s, y_s)|, \\ &|I_s(x_s - 1, y_s - 1) - I_s(x_s, y_s)|), \end{aligned} \tag{14}$$

it facilitates a more extensive integration of depth and positional information, thereby advancing the enhancement of underwater scene comprehension. This adaptation offers a holistic perspective on the fusion of data, leading to refined insights into the underwater environment.

3.4. Cost Aggregation with Vision and Sonar Terms

Since the cost calculation step considers only local correlations and is highly sensitive to noise, it cannot be directly employed to calculate the optimal disparity. Therefore, the cost aggregation step enhances the accuracy of aggregated cost values, which better reflect the inter-pixel correlations.

To integrate multi-beam image sonar data into the SGBM algorithm, we propose a modified cost aggregation stage that takes into account the sonar data. Specifically, we used the sonar data to guide the cost aggregation process by assigning higher weights to pixels that are similar in both the optical and sonar domains. This ensures that the cost aggregation process focuses on areas of the image where both sensors provide reliable information.

The modified cost aggregation stage can be formulated as follows:

$$C_{agg}(p_c, d) = wC(p_c, d) + (1 - w)\hat{C}_s(p_c, d), \tag{15}$$

where $C_{agg}(p_c, d)$ is the aggregated cost for pixel p and disparity d, $C(p_c, d)$ is the cost at pixel q and disparity d, and w is the weight assigned to the image cost and sonar cost. The energy function is:

$$\begin{aligned} E(D) = \sum_{\mathrm{p}} \Bigg(& Cagg(\mathbf{p}, D_{\mathrm{p}}) + \sum_{\mathrm{q} \in N_{\mathrm{p}}} P_1 \, \mathrm{T}\left[\left|D_{\mathrm{p}} - D_{\mathrm{q}}\right| = 1\right] \\ & + \sum_{\mathrm{q} \in N_{\mathrm{p}}} P_2 \, \mathrm{T}\left[\left|D_{\mathrm{p}} - D_{\mathrm{q}}\right| > 1\right] \\ & + \sum_{\mathrm{q} \in N_{\mathrm{p}}} P_3 \, \mathrm{T}\left[\left|D_{\mathrm{p}} - D_{\mathrm{q}}\right| \geq 1\right] \Bigg). \end{aligned} \tag{16}$$

The term $Cagg(\mathbf{p}, D_{\mathrm{p}})$ represents the loss associated with point $\mathbf{p}$ under disparity map D_{p}, and its construction is based on the results of the cost calculation. T serves as a function or

condition primarily utilized to assess whether a particular condition is satisfied. It evaluates $\left|D_{\mathrm{p}}-D_{\mathrm{q}}\right|$ to determine if it equals 1 or exceeds 1 and, based on this assessment, selects the corresponding penalty term, either P_1 or P_2. Specifically, the loss term $P_1\ \mathrm{T}\left[\left|D_{\mathrm{p}}-D_{\mathrm{q}}\right|=1\right]$ penalizes cases where adjacent pixel disparities vary only slightly, while the loss term $P_2\ \mathrm{T}\left[\left|D_{\mathrm{p}}-D_{\mathrm{q}}\right|>1\right]$ penalizes cases where adjacent pixel disparities change significantly. Typically, the penalty strength of P_2 is greater than P_1, signifying that our algorithm leans towards ensuring smooth disparity transitions rather than abrupt, drastic changes in the disparity map. In order to seamlessly integrate sonar data into our stereo matching process, we introduce a penalty function denoted as $P_3\ \mathrm{T}\left[\left|D_{\mathrm{p}}-D_{\mathrm{q}}\right|\geq 1\right]$. The primary objective of this function is to penalize disparities in the disparity map characterized by discontinuities and abrupt changes, aiming to yield smoother and more consistent depth estimates. This penalty function encourages smoothness in the disparity map, reducing errors and inconsistencies while improving the system's robustness. By appropriately tuning the parameters of the penalty function, we can strike a balance between smoothness and accuracy that is tailored to the specific requirements of different underwater applications.

The incorporation of three penalty terms into the loss function introduces the concept of global or semi-global optimization to the problem. In essence, the addition of these penalty terms transforms the task into a broader optimization challenge, where the goal is to achieve the best possible solution for all pixels, either globally or within a specific region. Specifically, $\mathrm{q}\in N_{\mathrm{p}}$ implies that determining the optimal disparity for a particular pixel requires knowledge of the optimal disparities of its neighboring pixels. This interdependence of disparities among adjacent pixels makes the problem more intricate, as it necessitates a holistic understanding of the entire image. Taking into account the multifaceted aspects and complexities of the matter at hand, a dynamic programming approach is adopted. Dynamic programming is a versatile technique commonly used to solve intricate optimization problems like this one. In this context, it is applied to find the best disparities on a global or semi-global scale, minimizing the overall loss function. To compute the matching cost for a pixel across all disparities, we implement a process that begins with the one-dimensional aggregation of matching costs for pixels situated along a specific path encircling the target pixel. These one-dimensional aggregated costs represent the path's aggregation cost. Subsequently, the aggregation costs from all paths are summed to yield the comprehensive matching cost for the pixel across all disparities. Here, we provide a qualitative framework for path cost aggregation, which can be expressed in the following formula:

$$L_{\mathbf{r}}(\mathbf{p},d)=wL_{\mathbf{r}}^{s}(\mathbf{p},d)+(1-w)L_{\mathbf{r}}^{c}(\mathbf{p},d). \tag{17}$$

This framework is designed to provide a flexible approach to path cost aggregation, allowing us to combine the sonar-based cost term $L_{\mathbf{r}}^{s}$ and the computer vision-based cost term $L_{\mathbf{r}}^{c}$. In this equation, $\mathbf{r}$ denotes the direction of the path. The path direction $\mathbf{r}$ is a critical factor in the cost aggregation process, determining how the costs from different pixels along the path are combined to estimate the disparity at a specific pixel $\mathbf{p}$. By adjusting the path direction, we can control the spatial relationships and interactions considered in the aggregation, allowing us to adapt the algorithm to the geometry of the scene and improve the quality of the disparity map.

Building upon the framework we have presented, we further elaborate on the role of the two components, $L_{\mathbf{r}}^{c}(\mathbf{p},d)$ and $L_{\mathbf{r}}^{s}(\mathbf{p},d)$, in the path cost aggregation process. Firstly, the $L_{\mathbf{r}}^{c}(\mathbf{p},d)$ term represents the conventional matching cost, denoted as:

$$\begin{aligned} L_{\mathbf{r}}^{c}(\mathbf{p},d)=&C(\mathbf{p},d)+\min(L_{\mathrm{r}}^{c}(\mathbf{p}-\mathbf{r},d),\\ &L_{\mathrm{r}}^{c}(\mathbf{p}-\mathbf{r},d-1)+P_1,\\ &L_{\mathrm{r}}^{c}(\mathbf{p}-\mathbf{r},d+1)+P_1,\\ &\min_{i}L_{\mathbf{r}}^{c}(\mathbf{p}-\mathbf{r},i)+P_2\Big)-\min_{k}L_{\mathbf{r}}^{c}(\mathbf{p}-\mathbf{r},k), \end{aligned} \tag{18}$$

influenced by the pixel-wise intensity differences. It follows a dynamic programming approach along the path direction $\mathbf{r}$, accumulating costs with respect to the disparities. This term considers the cost $C(\mathbf{p}, d)$, which reflects the intensity difference between the left and right images, as well as penalty terms P_1 and P_2. The role of P_1 is to penalize disparities with magnitude differences of 1, promoting smooth disparity maps. Conversely, P_2 is responsible for penalizing large discontinuities in disparities, which aims to reduce disparities' abrupt transitions. The aggregation over the path, through minimizing and subtracting, effectively integrates information along the path direction $\mathbf{r}$. On the other hand, the $L^s_{\mathbf{r}}(\mathbf{p}, d)$ term introduces the influence of sonar data, denoted as:

$$\begin{aligned} L^s_{\mathbf{r}}(\mathbf{p}, d) =& \hat{C}_s(\mathbf{p}, d) + \min(L^s_{\mathrm{r}}(\mathbf{p} - \mathbf{r}, d), \\ & L^s_{\mathrm{r}}(\mathbf{p} - \mathbf{r}, d-1) + P_3, \\ & L^s_{\mathrm{r}}(\mathbf{p} - \mathbf{r}, d+1) + P_3, \\ & \min_i L^s_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, i) + P_3\Big) - \min_k L^s_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, k). \end{aligned} \tag{19}$$

It encompasses the sonar matching cost term $\hat{C}_s(\mathbf{p}, d)$ and the similar penalty term P_3. This addition is crucial for underwater scenarios, as it utilizes sonar data to enhance the disparity estimation. The penalty term P_3 serves a similar purpose as P_1 and P_2 in the sonar context. It helps reduce disparities' discontinuities and promotes smoother depth estimations, considering the specific characteristics of sonar data.

Combining these components in the cost aggregation process allows us to balance the influence of traditional pixel-wise intensity matching (left image) with that of sonar data (sonar image) to obtain more accurate and robust disparity maps, particularly in challenging underwater conditions. This novel approach enables a more comprehensive fusion of depth and positional data, ultimately contributing to the refinement of underwater scene understanding. This combination of $L^c_{\mathbf{r}}(\mathbf{p}, d)$ and $L^s_{\mathbf{r}}(\mathbf{p}, d)$, tailored to the unique characteristics of the respective data sources, provides the foundation for our enhanced stereo vision system, contributing to more precise and reliable results in underwater environments.

4. Experimental Comparisons

The algorithm's performance was rigorously assessed via a battery of controlled tank experiments, wherein the target was deliberately positioned at the tank's submerged base. Detailed information regarding the specifications of the test tank, characteristics of the target, and the comprehensive details of the experimental setup is provided in Section 4.1.

4.1. Experimental Environment and Equipment

This section delves into the intricacies of the tank's dimensions, the composition of the test target, the specific sensor configurations, as well as the methodology employed in conducting these crucial experiments.

The tank test was meticulously conducted at the Ocean Underwater Engineering Science Research Institute of Shanghai Jiao Tong University, where a dedicated facility was utilized to ensure the precision and validity of the experiments. The test tank itself boasts extensive dimensions, measuring 25 m in length and 15 m in width. It features varying water depths, with a 6-m depth in the 15-m section and a 10-m depth in the 10-m section. This design allowed for a comprehensive simulation of different underwater conditions, ensuring that the testing environment closely mimicked real-world scenarios. To replicate the underwater environment with the utmost fidelity, target objects were deliberately positioned at a depth of 6 m within the tank. This depth aligns with the shallower section of the tank, facilitating a thorough evaluation of the algorithm's performance under specific underwater conditions.

Figure 6 illustrates the complexity of the target objects, and Table 1 shows the dimension of the target objects and ROV. Four distinct underwater structures were included in the test. These diverse structures were deliberately selected to represent typical underwater

objects, providing a rigorous assessment of the algorithm's ability to accurately measure their distance and dimensions.

Table 1. Dimensions of the target objects and ROV.

	Frame Structure	Pressure Tank	Platform Structure	Sphere Structure	ROV
height (cm)	165.4	134.5	66	64	40
width (cm)	53	113	80	64	35
depth (cm)	101.2	217	80	64	75

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

(**c**) (**d**)

Figure 6. Target objects used in the experiment: (**a**) frame-type structure; (**b**) pressure tank; (**c**) sphere structure; (**d**) platform structure.

To conduct the experiments, we deployed an underwater robotic system featuring an 8-thruster remotely operated vehicle (ROV) outfitted with a stereo camera and a multi-beam sonar, as illustrated in Figure 7. This carefully engineered setup provided a high degree of realism, closely mimicking the operational environment of an actual underwater robot. The system was built on the robust Robot Operating System (ROS) platform. In this setup, the ROV functioned as the host, while the upper computer acted as the slave, facilitating efficient communication and control. The upper computer was equipped with an NVIDIA Jetson Xavier (Nvidia, Santa Clara, CA, USA), a powerful processor adept at handling intensive tasks such as stereo matching and various image processing activities. This processing capacity was crucial for the successful execution of the algorithm's complex calculations, ensuring precise and real-time data processing. Additionally, the system was enhanced with an NVIDIA TX2 on the lower computer, which acted as the center for data collection and robot control. This setup allowed for effective data gathering and real-time management of the robot's functions. This comprehensive arrangement ensured that the

experiments were conducted in conditions that closely resembled real-world scenarios, thereby increasing the validity and applicability of the experimental results.

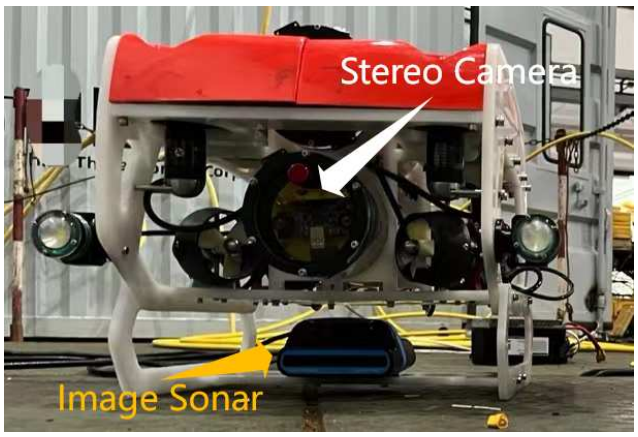


Figure 7. ROV system with stereo camera and image sonar.

Table 2 presents the calibration parameters for the stereo camera system. In this table, f_x and f_y indicate the focal lengths in the horizontal and vertical directions, respectively; u_0 and v_0 are the coordinates of the principal point, which is the intersection of the optical axis with the image plane. The distortion coefficients, which account for lens distortions in the camera, are denoted by k_c. Further, *R* represents the rotation matrix, defining the orientation of one camera in relation to the other, while *T* is the translation vector, indicating the position of the left camera in relation to the right camera. Lastly, *E* describes the extrinsic parameters, depicting the position of the left camera in relation to the sonar.

Table 2. the calibration parameter of the stereo camera

Parameters	Left Camera	Right Camera
(f_x, f_y)	(1542.96428, 1576.11235)	(1541.3057, 1571.6980)
(u_0, v_0)	(635.42367, 451.18523)	(628.82976, 471.85445)
k_c	(0.58524, −0.50156, 0.07526, −0.00169, 0)	(0.53044, −0.10992, 0.06693, −0.00207, 0)
R	(−0.01497, 0.00743, −0.00753)	
T (mm)	(−58.11, 0, 1.4)	
E (mm)	(−30, −70, −10)	

The stereo camera employed in this research is the PXYZ-S-H65-060 model (China), which offers a high-resolution output of 2560 × 720 pixels. This camera features a focal length of 3.5 mm, which is conducive to capturing detailed underwater imagery. Complementing the stereo camera, the Multibeam Sonar M1200D (Blueprint Oculus, Ulverston, UK) was used as the image sonar system. This setup enhances the overall capability of the system to provide precise and comprehensive underwater imaging and mapping.

4.2. Underwater Object Measurement Experiment

To provide a comprehensive illustration of our algorithm's effectiveness, this research employed a comparative analysis involving four distinct algorithms. These algorithms are thoughtfully chosen to encompass a spectrum of techniques, including both cutting-edge deep learning approaches and traditional methodologies. The selected comparative algorithms, detailed below, are pivotal to our evaluation:

- GA-Net: Guided Aggregation Net for End-to-End Stereo Matching [29];
- IGEV: Iterative Geometry Encoding Volume for Stereo Matching [30];
- BM: Block Match;
- SGBM: Semi-Global Block Matching.

These carefully chosen comparative algorithms facilitated a comprehensive evaluation, showcasing the strengths and limitations of our proposed algorithm within the context of stereo matching. Among the selected algorithms, both GA-Net and IGEV have garnered recognition as state-of-the-art (SOTA) methods. Their impressive performance is evident from their standings in the top 10 of the KITTI stereo benchmarks, solidifying their positions as leading algorithms in the field of stereo matching. This recognition at a highly competitive benchmark underscores their exceptional performance and establishes them as benchmarks in the field of stereo matching. These algorithms, lauded for their cutting-edge techniques, have proven their mettle by providing highly accurate depth estimations from stereo image pairs. In our comparative analysis, we aim not only to evaluate the effectiveness of our proposed algorithm but also to highlight how it measures up against these high-performing SOTA methods. By doing so, we can provide a comprehensive and insightful assessment of our algorithm's capabilities and contributions to the field of stereo matching.

The specific algorithmic performance is visually presented in Figure 8. While it is noteworthy that GA-Net (Figure 8c) and IGEV (Figure 8d) excel at KITTI benchmarks, achieving impressive results in controlled environments, their performance falters significantly in the challenging underwater scenarios explored in this study. These state-of-the-art deep learning methods seem to grapple with generalization issues when transitioning to the underwater domain, where complexities like scattering, attenuation, and low light conditions introduce new challenges. Their inability to yield reliable measurements in such conditions becomes evident. This limitation underscores a vital point: despite their achievements in specific contexts, the applicability of these methods is not universal and should be carefully assessed regarding the specific environment. This phenomenon highlights the importance of dedicated underwater stereo matching algorithms that are tailored to navigate the unique hurdles posed by underwater imaging. Figure 8e showcases the disparity map generated by the conventional SGBM (Semi-Global Block Matching) algorithm applied to an original underwater image. The disparity map clearly shows that effective contour matching is predominantly limited to areas of high contrast. This highlights a general limitation of visual algorithms in low-light conditions. Even when illumination is sufficient, only regions with distinct features, such as shelves, exhibit relatively better matching. However, the matching is not detailed enough to allow for clear differentiation between target objects and their surrounding environment. One of the primary challenges in processing underwater images is the degradation of features, which makes background matching highly challenging and can be likened to attempting to distinguish between two similar white walls. Moreover, it is noticeable that the original SGBM algorithm is particularly vulnerable to errors in areas lacking texture or contrast. This issue is exacerbated in underwater environments, where water's absorption and scattering of light significantly reduce the visibility and contrast of objects. Consequently, in such scenarios, visual algorithms face substantial difficulties in achieving reliable and accurate matches. Figure 8f shows the output generated by the Block Match (BM) algorithm. Upon a thorough examination of the results, it becomes evident that the BM algorithm, similar to the previously discussed methods, excels primarily in extracting depth information from high-contrast regions, particularly at object edges. However, it struggles to provide accurate depth information in regions with lower contrast, yielding mostly erroneous depth measurements in these areas. In a significant portion of the image, the BM algorithm's performance is marred by incorrect matching, further highlighting the challenges of underwater stereo vision. Consequently, conventional algorithms like BM face substantial difficulties in achieving precise depth estimations, particularly in scenarios with low-contrast textures or objects. Figure 8g shows the results obtained through the innovative method proposed in this paper. A meticulous examination of these results reveals the ability of our method to deliver continuous and effective depth information. It is noteworthy that our method excels in dealing with all four target objects, with notably uniform and precise depth estimations, especially in scenarios involving structures like the platform, sphere, and tank. The fusion of sonar data enhances the performance of underwater stereo vision, providing more reliable data for underwater applications.

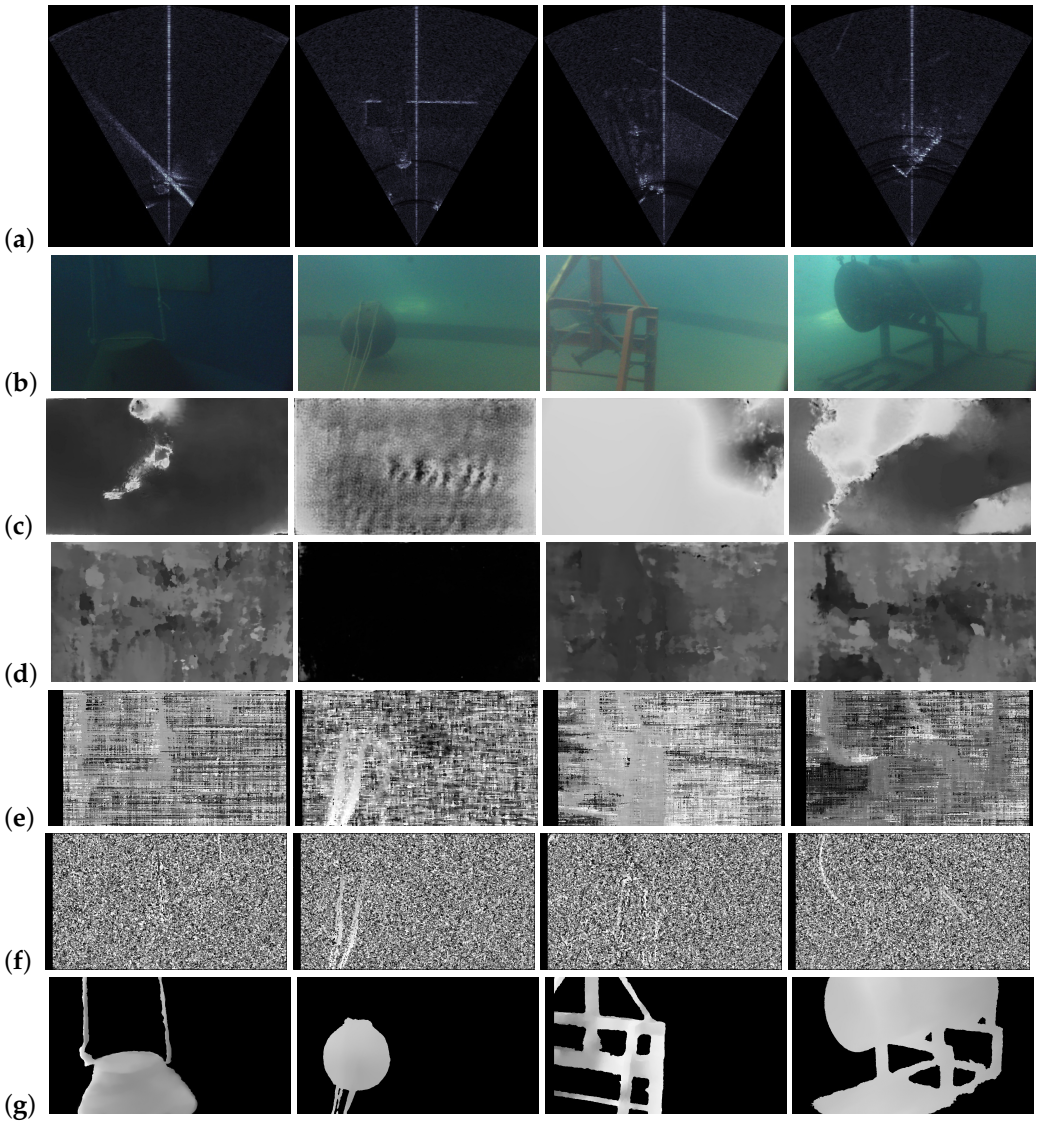


Figure 8. Comparisons of different stereo matching methods: (**a**) sonar image, (**b**) raw image, (**c**) IGEV, (**d**) GA-Net, (**e**) SGBM, (**f**) BM, (**g**) ours.

Given the inherent challenges of acquiring ground truth data in underwater environments, conventional evaluation metrics commonly used in terrestrial settings are often impractical. To address this issue, our evaluation method utilizes the length of underwater objects as a proxy metric for assessing the accuracy of the algorithm. Due to the difficulty in obtaining true distance values in underwater stereo matching, but given the regular size and shape of the target objects, the error can be defined by comparing the measured length of the target with its actual dimensions to obtain the error. In Table 3, we employ the width of the target object as the evaluation metric for accuracy. Notably, the GA-Net, IGEV, and BM algorithms are excluded from the table, as they failed to produce valid measurements. To ensure result precision, we conducted 50 separate measurements, and the final result is represented as the average of these measurements. This rigorous approach was taken to mitigate variability and enhance the reliability of our assessment.

Table 3. Average measurements for four object widths with different methods.

Methods	Shelf (mm)	Tank (mm)	Sphere (mm)	Platform (mm)	Error (%)
Ground Truth	530	1130	640	800	
SGBM	742	1576	831	1043	34.7
Ours (without sonar)	562	1251	663	843	6.2
Ours (with sonar)	542	1149	649	813	1.7

Table 3 reveals an interesting insight into the performance of the SGBM algorithm in underwater environments. Although it is functional in such conditions, the level of precision it offers falls short of practical standards. This is particularly evident when considering the error metrics, as illustrated in Figure 9, where the errors for each type of target object are relatively substantial. Building upon the analysis in Table 3, Figure 9 showcases the significant disparities in the algorithm's results when acoustic sonar data are integrated compared to when they are not (with a weight factor of 0). When we reference Figure 10, it becomes apparent that the inclusion of sonar data leads to substantial improvements in the accuracy of depth estimation for multiple target categories. Moreover, the depth maps generated with sonar data exhibit a noticeable enhancement in terms of fine-grained precision and continuity. These improvements are readily discernible to the human eye. This enhancement in performance, particularly in complex underwater environments, underscores the value of integrating acoustic sonar data into the stereo vision process. The visual improvements observed in Figure 9 correlate with the quantitative findings in Table 3, affirming that our approach can elevate the accuracy and continuity of depth information. In essence, the integration of sonar data offers a considerable boost in the overall performance of our algorithm, making it a robust solution for underwater depth estimation that significantly outperforms traditional methods in challenging underwater scenarios.

Furthermore, an in-depth analysis of the results presented in Figures 9 and 10 reveals a notable trend. Specifically, it is evident that the closed-form target objects such as sphere, tank, and platform exhibit significantly improved accuracy when compared to the frame-like structure of the shelf target. Referencing Figure 8a, a fascinating observation can be made regarding the acoustic sonar data, particularly from the profile-like imaging perspective it offers. The measurement process appears to benefit substantially from the inherent characteristics of closed shapes. These characteristics include reduced noise levels and higher image intensity, contributing to improved accuracy. Additionally, the sonar data demonstrate better continuity, further enhancing the measurement process. In essence, the observations suggest that, of the four target categories evaluated in this study, the shelf target shows the least improvement in accuracy. This is likely due to its inherent features, which do not align as seamlessly with the sonar imaging process as the more enclosed targets. These findings underscore the impact of target geometry on measurement accuracy and the significant advantages of integrating sonar data into the stereo vision process for underwater depth estimation. The improved performance, especially for closed-form objects, highlights the potential of our approach in enhancing the accuracy and continuity of depth information in challenging underwater scenarios.

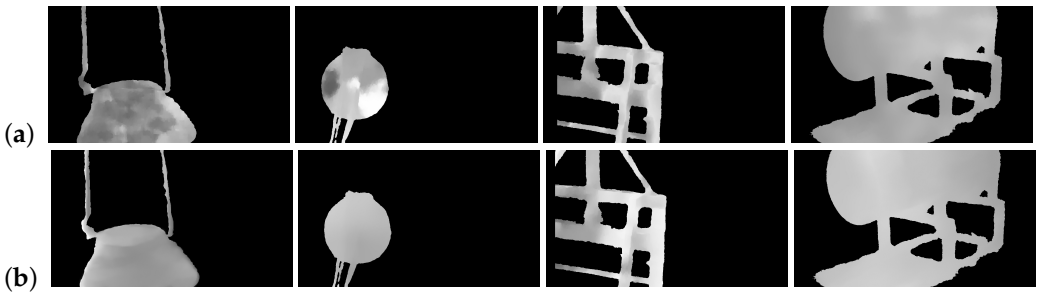


Figure 9. Algorithm performance (**a**) without sonar data and (**b**) with sonar data.

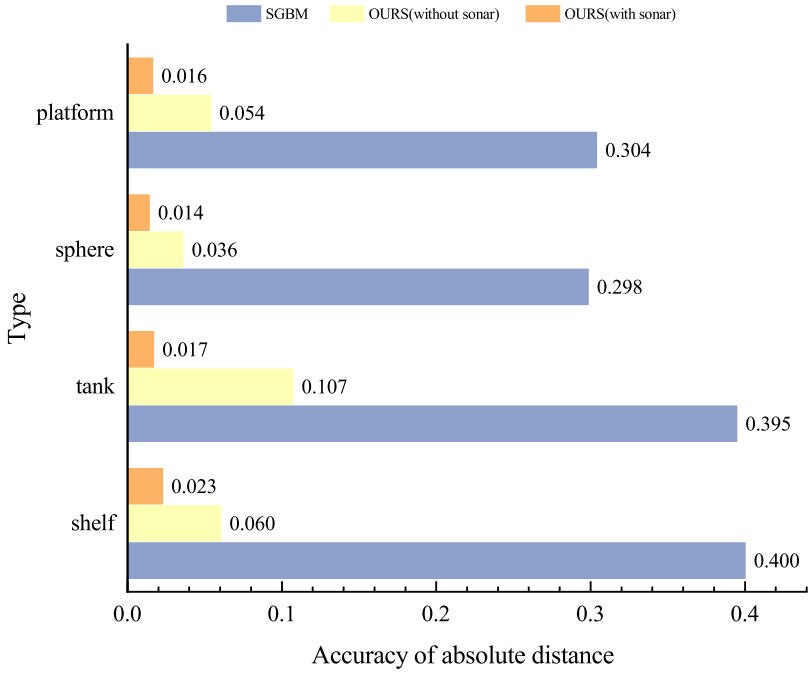


Figure 10. Histograms of the measurement accuracy.

Figure 11 offers an illuminating view of our method's performance under varying sonar data weightings. As depicted in the graph, it is evident that an excessively high weight assigned to sonar data results in its overwhelming dominance over the depth map generation process. This, in turn, leads to a loss of fine-grained detail within the depth map, ultimately sacrificing its accuracy and completeness. Conversely, when sonar data are assigned relatively lower weights, the depth map becomes more susceptible to the influence of image noise. Consequently, critical structural details of the target objects may be obscured, leading to an incomplete representation of the underwater scene. The crux of this observation underscores the necessity of striking a delicate balance in determining the optimal weight for sonar and image data. It is only through this fine-tuned calibration that a depth map with the desired level of accuracy and detail can be achieved. In summary, Figure 11 emphasizes the pivotal role played by the relative weighting of sonar and image data in the effectiveness of our method. Achieving the right balance is key to producing accurate and comprehensive depth maps, enabling a more reliable understanding of the underwater environment.

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

Figure 11. Results with different weights: (**a**) $w = 0.3$, (**b**) $w = 0.6$, (**c**) $w = 0.9$.

5. Conclusions

In this paper, we addressed the challenging issue of underwater target measurement in extreme conditions by incorporating sonar data into the SGBM cost computation and aggregation processes. Through comparative experiments, we have demonstrated that our approach can reduce errors to as low as 1.6%. This underscores the strong adaptability of our algorithm to extreme underwater environments, expanding the potential of underwater

applications by enhancing measurement accuracy. Our proposed algorithm offers several notable advantages. Firstly, it effectively leverages the complementary strengths of both sonar and image data in underwater depth mapping. By combining these two data sources, we have achieved a more comprehensive and accurate representation of the underwater scene. Secondly, our method demonstrates robustness in challenging underwater conditions. Unlike some existing algorithms that struggle in low-contrast or low-texture areas, our algorithm showcases improved performance even in such adverse scenarios. Moreover, the algorithm's versatility is a key asset. It can be applied across various underwater environments, making it adaptable for a wide range of underwater applications. The introduction of sonar data not only enhances the accuracy of depth mapping but also contributes to the visualization of underwater scenes in cases where image data alone fall short. Similarly, our algorithm has certain limitations. First, its adaptability to complex-shaped targets is relatively weak. Second, because we employ deep learning for target extraction, it requires the creation of target-specific datasets, limiting its applicability to specific target types. In summary, our algorithm excels in its fusion of sonar and image data, resilience in challenging conditions, adaptability, and improved depth mapping accuracy, making it a promising solution for underwater applications.

Our future research endeavors are poised to address several key areas. First and foremost, we aim to enhance our algorithm's adaptability to complex and irregularly shaped underwater targets, such as those with intricate structures. Moreover, we envision the integration of sonar data and stereo vision data into underwater robot Simultaneous Localization and Mapping (SLAM) systems as a primary focus. This endeavor holds the promise of improving underwater robot localization accuracy, thereby expanding the utility of such robots in various underwater applications, including underwater exploration, inspection, and maintenance tasks. By marrying the strengths of sonar and stereo vision in a seamless SLAM framework, we anticipate notable advancements in underwater robotics technology. In conclusion, our current research has paved the way for innovative applications of sonar–stereo vision fusion in underwater object measurement. As we move forward, we are committed to addressing the limitations and to evolving our techniques to achieve more accurate and robust underwater measurements while contributing to the broader field of underwater robotics.

Author Contributions: Conceptualization, J.Z., D.H. and F.H.; methodology, J.Z.; software, J.Y. and H.L.; validation, W.Z., F.H. and J.Z.; formal analysis, J.Y.; investigation, W.Z.; resources, F.H.; data curation, F.H.; writing—original draft preparation, J.Z.; writing—review and editing, J.Z.; visualization, H.L.; supervision, F.H.; project administration, F.H.; funding acquisition, F.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Heilongjiang Province of China, grant number LH2021E047, and National Key R&D Program of China, grant number 2022YFB3306200 .

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset available on request from the authors.

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

References

1. Massot-Campos, M.; Oliver-Codina, G. Optical sensors and methods for underwater 3D reconstruction. *Sensors* **2015**, *15*, 31525–31557. [CrossRef] [PubMed]
2. Henderson, J.; Pizarro, O.; Johnson-Roberson, M.; Mahon, I. Mapping submerged archaeological sites using stereo-vision photogrammetry. *Int. J. Naut. Archaeol.* **2013**, *42*, 243–256. [CrossRef]
3. Sahoo, A.; Dwivedy, S.K.; Robi, P. Advancements in the field of autonomous underwater vehicle. *Ocean. Eng.* **2019**, *181*, 145–160. [CrossRef]
4. Hirschmuller, H. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *30*, 328–341. [CrossRef] [PubMed]

5. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
6. Ng, P.C.; Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **2003**, *31*, 3812–3814. [CrossRef] [PubMed]
7. González-Sabbagh, S.P.; Robles-Kelly, A. A Survey on Underwater Computer Vision. *ACM Comput. Surv.* **2023**, *55*, 268. [CrossRef]
8. 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, Monterey, CA, USA, 19–23 September 2016; pp. 1–6.
9. Cong, Y.; Gu, C.; Zhang, T.; Gao, Y. Underwater robot sensing technology: A survey. *Fundam. Res.* **2021**, *1*, 337–345. [CrossRef]
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. Shi, C.; Wang, Q.; He, X.; Zhang, X.; Li, D. An automatic method of fish length estimation using underwater stereo system based on LabVIEW. *Comput. Electron. Agric.* **2020**, *173*, 105419. [CrossRef]
12. Kim, M.; Lee, S.; Hong, J.W. Empirical estimation of the breaker index using a stereo camera system. *Ocean. Eng.* **2022**, *265*, 112522. [CrossRef]
13. Aykin, M.D.; Negahdaripour, S. Forward-look 2-D sonar image formation and 3-D reconstruction. In Proceedings of the 2013 OCEANS-San Diego, San Diego, CA, USA, 23–27 September 2013; pp. 1–10.
14. Chung, D.; Kim, J. Underwater visual mapping of curved ship hull surface using stereo vision. *Auton. Robot.* **2023**, *47*, 109–120. [CrossRef]
15. Zhai, X.; Meng, Z.; Zhang, H.; Xu, X.; Qian, Z.; Xue, B.; Wu, H. Underwater distance measurement using frequency comb laser. *Opt. Express* **2019**, *27*, 6757–6769. [CrossRef]
16. Xu, Y.; Yu, D.; Ma, Y.; Li, Q.; Zhou, Y. Underwater stereo-matching algorithm based on belief propagation. *Signal Image Video Process.* **2023**, *17*, 891–897. [CrossRef]
17. Skinner, K.A.; Zhang, J.; Olson, E.A.; Johnson-Roberson, M. Uwstereonet: Unsupervised learning for depth estimation and color correction of underwater stereo imagery. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 7947–7954.
18. Huang, T.A.; Kaess, M. Towards acoustic structure from motion for imaging sonar. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 758–765.
19. Karimanzira, D.; Renkewitz, H.; Shea, D.; Albiez, J. Object detection in sonar images. *Electronics* **2020**, *9*, 1180. [CrossRef]
20. Purser, A.; Marcon, Y.; Dreutter, S.; Hoge, U.; Sablotny, B.; Hehemann, L.; Lemburg, J.; Dorschel, B.; Biebow, H.; Boetius, A. Ocean Floor Observation and Bathymetry System (OFOBS): A new towed camera/sonar system for deep-sea habitat surveys. *IEEE J. Ocean. Eng.* **2019**, *44*, 87–99. [CrossRef]
21. McConnell, J.; Englot, B. Predictive 3D Sonar Mapping of Underwater Environments via Object-specific Bayesian Inference. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 6761–6767. [CrossRef]
22. Cho, H.; Kim, B.; Yu, S.C. AUV-Based Underwater 3-D Point Cloud Generation Using Acoustic Lens-Based Multibeam Sonar. *IEEE J. Ocean. Eng.* **2018**, *43*, 856–872. [CrossRef]
23. Negahdaripour, S. Epipolar Geometry of Opti-Acoustic Stereo Imaging. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1776–1788. [CrossRef] [PubMed]
24. Pecheux, N.; Creuze, V.; Comby, F.; Tempier, O. Self Calibration of a Sonar–Vision System for Underwater Vehicles: A New Method and a Dataset. *Sensors* **2023**, *23*, 1700. [CrossRef] [PubMed]
25. Terayama, K.; Shin, K.; Mizuno, K.; Tsuda, K. Integration of sonar and optical camera images using deep neural network for fish monitoring. *Aquac. Eng.* **2019**, *86*, 102000. [CrossRef]
26. Raaj, Y.; John, A.; Jin, T. 3D Object Localization using Forward Looking Sonar (FLS) and Optical Camera via particle filter based calibration and fusion. In Proceedings of the OCEANS 2016 MTS/IEEE Monterey, Monterey, CA, USA, 19–23 September 2016; pp. 1–10.
27. Rahman, S.; Li, A.Q.; Rekleitis, I. Sonar visual inertial slam of underwater structures. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 5190–5196.
28. Zhang, J.; Han, F.; Han, D.; Su, Z.; Li, H.; Zhao, W.; Yang, J. Object measurement in real underwater environments using improved stereo matching with semantic segmentation. *Measurement* **2023**, *218*, 113147. [CrossRef]
29. Zhang, F.; Prisacariu, V.; Yang, R.; Torr, P.H. GA-Net: Guided Aggregation Net for End-to-end Stereo Matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 185–194.
30. Xu, G.; Wang, X.; Ding, X.; Yang, X. Iterative Geometry Encoding Volume for Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023.