remote sensing

IMPACT FACTOR 4.2

CITESCORE 8.3

Article

An Improved Underwater Visual SLAM through Image Enhancement and Sonar Fusion

Haiyang Qiu, Yijie Tang, Hui Wang, Lei Wang, Dan Xiang and Mingming Xiao

Special Issue

Advances of Underwater Remote Sensing of Methane: Spatiotemporal Distribution

Edited by

Prof. Dr. Changhui Jiang, Dr. Zuoya Liu, Dr. Yue Yu and Dr. Yuwei Chen

MDPI

https://doi.org/10.3390/rs16142512

remote sensing

MDPI

Article

An Improved Underwater Visual SLAM through Image Enhancement and Sonar Fusion

Haiyang Qiu [1,*], Yijie Tang [2], Hui Wang [1], Lei Wang [3], Dan Xiang [1] and Mingming Xiao [1]

[1] School of Naval Architecture and Ocean Engineering, Guangzhou Maritime University, Guangzhou 510725, China; wanghui@gzmtu.edu.cn (H.W.); gsxd@gpnu.edu.cn (D.X.); xmingm@gzmtu.edu.cn (M.X.)
[2] School of Automation, Jiangsu University of Science and Technology, Zhenjiang 212003, China; 211210301215@stu.just.edu.cn
[3] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China; lei.wang@whu.edu.cn
* Correspondence: hy.qiu@gzmtu.edu.cn

Abstract: To enhance the performance of visual SLAM in underwater environments, this paper presents an enhanced front-end method based on visual feature enhancement. The method comprises three modules aimed at optimizing and improving the matching capability of visual features from different perspectives. Firstly, to address issues related to insufficient underwater illumination and uneven distribution of artificial light sources, a brightness-consistency recovery method is proposed. This method employs an adaptive histogram equalization algorithm to balance the brightness of images. Secondly, a method for denoising underwater suspended particulates is introduced to filter out noise from images. After image-level processing, a combined underwater acousto–optic feature-association method is proposed, which associates acoustic features from sonar with visual features, thereby providing distance information for visual features. Finally, utilizing the AFRL dataset, the improved system incorporating the proposed enhancement methods is evaluated for its performance against the OKVIS framework. The system achieves a better trajectory estimation accuracy compared to OKVIS and demonstrates robustness in underwater environments.

Keywords: underwater SLAM; underwater image processing; visual feature point enhancement; sonar features; visual features

Citation: Qiu, H.; Tang, Y.; Wang, H.; Wang, L.; Xiang, D.; Xiao, M. An Improved Underwater Visual SLAM through Image Enhancement and Sonar Fusion. *Remote Sens.* **2024**, *16*, 2512. https://doi.org/10.3390/rs16142512

Academic Editor: Fabio Menna

Received: 17 May 2024
Revised: 29 June 2024
Accepted: 5 July 2024
Published: 9 July 2024

1. Introduction

Underwater environments present unique challenges for autonomous navigation and mapping due to their complex, dynamic, and often unstructured nature [1]. The utilization of robot-assisted technology for underwater exploration can alleviate the cognitive burden on divers and enhance work efficiency. A critical capability for these underwater robots is simultaneous localization and mapping (SLAM), which allows them to build a map of the unknown environment while simultaneously keeping track of their own location within it [2]. In recent years, numerous camera-based SLAM frameworks have emerged, capable of generating reliable state estimation results in both indoor and outdoor settings [3]. However, SLAM in underwater environments presents unique challenges that differ significantly from terrestrial or aerial applications. The underwater domain is characterized by poor visibility, limited lighting, and complex, dynamic conditions. Traditional vision-based SLAM techniques, which rely on optical cameras, often struggle due to issues such as turbidity, light absorption, and scattering. These conditions severely limit the range and quality of visual data, making reliable localization and mapping difficult [4].

Specifically, in underwater environments, the uneven attenuation of natural light within water results in color deviation and contrast degradation in images. When natural light is insufficient, scenes are often illuminated with artificial light sources. These artificial lights, typically single-point sources with limited power, cast numerous shadows in the

irregular underwater landscape [5]. Furthermore, suspended particles in the water cause diffuse reflection, disrupting the normal propagation of light. These conditions make cameras more susceptible to issues such as limited visibility, fogging, and fluctuations in light intensity, leading to blurred images compared to those taken in terrestrial settings [6].

As a result, the effectiveness of optical cameras, which rely on visual information, is significantly restricted in underwater environments. When using a vision-based state estimation system with continuous underwater image capture, the aforementioned adverse conditions can severely impact the extraction of stable feature points for motion estimation. Various noise disturbances may lead to the generation of numerous anomalous feature points, reducing estimation accuracy or causing tracking failures [7].

To enhance the feasibility of visual odometry in underwater environments, this paper proposes an improved VIO front-end method based on visual feature enhancement. The main contributions are as follows:

1. Image-Level Enhancements: The proposed method integrates image brightness enhancement and suspended particulate removal techniques. This significantly increases the probability of the successful detection of visual feature points after application.
2. Geometric Feature Association: From a spatial geometry perspective, a feature-association method that integrates sonar acoustic features with visual features is proposed, enabling visual features to obtain depth information.
3. Benchmarking with AFRL Dataset: A comparative analysis is conducted using the AFRL dataset against the classical OKVIS visual SLAM framework. This tests the limitations of traditional frameworks in underwater datasets and demonstrates the feasibility of the proposed method.

2. Related Work

In recent years, image registration has garnered widespread attention in the kinematic estimation research field. This method is applicable for positioning and navigation using data from monocular, binocular, and RGBD cameras [8]. Visual odometry (VO), obtained through continuous image registration, is a fundamental component of the front end of visual SLAM systems. However, visual motion estimation is easily influenced by environmental conditions, scale, and motion state. To enhance performance, SLAM systems often integrate other types of sensors, especially inertial sensors. OKVIS is a tightly coupled visual–inertial SLAM fusion framework that introduces a keyframe-based optimization method [8]. The VINS-MONO framework employs optical flow tracking at the front end and uses inertial sensors to recover scale, mitigating position drift and improving robustness in feature-poor conditions [9]. OpenVINS employs a multistate constraint Kalman filter (MSCKF) as its back-end information-processing method, which can support various feature point representations. The use of a filtering framework provides an advantage in computational speed [10].

As visual SLAM frameworks continue to develop and improve, their applications have begun to extend to underwater environments. Under adequate illumination and with abundant scene features, ORB-SLAM based on ORB features achieved promising results [11]. However, variations in lighting and object motion also impacted its performance. Compared to using feature-based methods for front-end motion tracking, an optical flow mechanism based on retracking has achieved better results in dynamic environments [12], albeit at the cost of additional computational demand. In structured scenes such as ship hull inspection, utilizing geometric models to select different image registration methods and combining them with pose-graph visual SLAM techniques enables the construction of a texture-mapped 3D model of the ship hull [13]. However, owing to the specific characteristics of underwater environments, conventional visual SLAM methods cannot be readily adapted for underwater use. Visual odometry systems are still notably sensitive to fluctuations in lighting conditions, and the uneven absorption of light in underwater environments can significantly impact the extraction and matching of feature points [14].

In addition to inertial sensors, visual SLAM systems also integrate more environmental perception sensors. In underwater environments, utilizing various types of sonar sensors can enhance the accuracy of localization and mapping. Multibeam sonar is an advanced acoustic imaging device that uses multiple sound beams to simultaneously scan underwater areas, generating high-resolution 2D or 3D images.

The data structure of sonar is quite similar to the point clouds in terrestrial SLAM using LIDAR, but it is sparser and contains more noise. Therefore, a variant of the iterative closest point (ICP) method based on probabilistic registration is applied to multibeam underwater SLAM systems, resulting in 2.5D bathymetric data [15]. By using multibeam sonar and employing active volumetric exploration and revisitation, the sonar data of keyframe submaps can be converted into a metric of submap saliency to reduce loop-closure matching errors [16,17].

Synthetic aperture sonar (SAS) is a powerful technique used to enhance spatial resolution for underwater imaging by extending the receiver array in time. This technique effectively increases the array length and aperture, resulting in high-resolution images. By combining SAS with deep learning techniques, it is possible to achieve underwater target detection, recognition, and segmentation [18,19]. The image information from SAS can be compared with targets in optical images for similarity, providing more accurate distance and location information for targets, and this SLAM method can aid in AUV navigation [20]. Yet, a significant drawback of SAS is its reliance on precise receiver position estimates. This requirement is critical as inaccuracies can notably diminish the quality of the resultant images [21].

Mechanical scanning imaging sonar (MSIS) is an acoustic imaging device that uses rotating mechanical components to scan underwater scenes. It captures underwater acoustic reflections by rotating transmitters and receivers. Typically, it has only one beam but offers lower costs. Due to the time required for one full rotation scan, which introduces motion distortion, motion estimation is essential for compensating MSIS's images. In structured environments, SLAM methods utilizing sonar image matching have shown promising results in localization and mapping [22]. In comparative experiments, integrating sonar into SLAM systems demonstrated significant performance advantages underwater compared to using only visual and inertial data [23]. The SVIN-2 underwater SLAM framework, an extension of OKVIS, supports robust initialization and relocalization, enhancing the reliability of underwater SLAM operations [24].

Lighting conditions are a key factor in the quality of underwater images. In scenarios where water depth surpasses 30 m, natural light becomes nearly nonexistent, prompting the necessity of employing active lighting methods for underwater optical imaging systems. To enhance the quality of underwater images, Ancuti [25] introduced a systematic processing method for enhancing underwater images, based on the principle of minimizing information loss to improve color and visibility. They proposed a dark channel a priori algorithm, which mitigates the influence of the red channel while accounting for the effects of optical radiation absorption and scattering on image degradation, thereby enhancing the visual quality of the image. Similarly, Barros [26] proposed a light-propagation model based on visual-quality perception. Building upon existing physical models, they integrated the physics of light propagation to mitigate the impact of optical radiation attenuation, further enhancing the quality of underwater images.

Water in natural environments also frequently harbors a substantial quantity of suspended matter, encompassing sediments, sand, and dust particles produced by diverse planktonic organisms. The irregular morphology and surface roughness of these suspended objects pose challenges in maintaining consistent observations, as the perceived information varies with viewing angles [27]. Consequently, image fidelity diminishes, contours become indistinct, and the signal-to-noise ratio declines. These suspended materials can be considered noise in image feature extraction, significantly impacting the extraction, matching, and tracking of visual feature points in images. As a result, the operational efficiency of underwater feature-based visual odometry is markedly reduced compared to

terrestrial scenarios [28]. Therefore, it is essential to filter underwater suspended particles from images.

The null domain denoising method partially mitigates noise by eliminating components at specific frequencies. However, when noise in underwater images intersects with their structural texture in the frequency domain, it leads to blurred textures and unclear edges. This issue can be addressed by a nonlinear median filter enhanced through the weighted median method. Linear filters in the wavelet domain are exemplified by the Wiener filter [29]. However, the degradation process of the actual signal may not conform to a Gaussian distribution, making this type of filter potentially detrimental to the visual quality of the denoised image. To address this, Celebi [30] introduced a wavelet domain spatially adaptive Wiener filter image-denoising algorithm to enhance the visual quality of the image after noise reduction. Additionally, C.J. Prabhakar [31] proposed an adaptive wavelet band thresholding method for reducing noise in underwater images. This method aims to filter out additive noises in the image, including scattering and absorption effects, as well as suspended particles visible to the naked eye, resulting from sand and dust on the seabed.

Based on a review of pioneers' work, the back-end estimation methods in underwater SLAM roughly converge with traditional SLAM approaches. However, due to the optical challenges in underwater environments, significant differences exist in the front-end visual odometry. Current efforts primarily focus on the precise identification and extraction of underwater visual features, like employing machine learning for feature detection and robust image-descriptor usage. In terms of underwater image processing, evaluations typically assess numerical metrics like denoising and contrast enhancement, lacking subsequent comparisons for practical image-application effects. In the domain of visual SLAM integrating sonar, approaches mostly utilize sonar-ranging information without considering the correspondence between sonar echoes and spatial geometric structures. Therefore, this study aims to enhance the front-end of underwater visual SLAM systems by integrating improved image-enhancement techniques and fusing sonar information. Real-world data will be employed to evaluate the efficacy of these methods.

3. System Overview

The main focus of this paper is to improve the front-end of visual SLAM systems for underwater environments. Therefore, the overall system design follows the traditional VIO (visual–inertial odometry) architecture while incorporating bundle adjustment (BA) optimization as the back-end. However, unlike traditional VIO frameworks, in this framework, the front-end images undergo an image-enhancement module specifically developed for underwater environments before proceeding to feature point extraction and matching, similar to SLAM systems. Some of the extracted feature points are fused with sonar features, generating a new type of sonar-camera feature. In the back-end optimization part, the pose is optimized by jointly minimizing the reprojection error of the new features, the visual features, and the predicted error from the IMU.

The overall flow of the system is depicted in the system block diagram as shown in Figure 1. Initially, data from each sensor undergo preprocessing to yield the camera image, the inertial measurement unit (IMU) preintegration term, and the position information of the sonar features. The original images are processed through brightness recovery and suspended matter removal modules, followed by the extraction and matching of visual feature points. The IMU data are used to correct aberrations caused by motion during the sonar-scanning cycle. Next, sonar features are matched with camera features, and the distance information from sonar detection is utilized to refine the depth estimation of visual feature points. The sonar feature information corresponding to the camera features is then used to enhance the accuracy of the camera features on the image plane, thereby reducing the reprojection error. Finally, a joint error optimization is conducted, incorporating the

reprojection error of the camera features, IMU error, and sonar-camera feature point to estimate landmark positions and the robot's state, as illustrated in Equation (1):

$$J(x) = \sum_{k=1}^{K} \sum_{j \in \tau(k)} e_r^{j,kT} \mathbf{P}_r^k e_r^{j,k} + \sum_{k=1}^{K-1} e_s^{kT} \mathbf{P}_s^k e_s^k + \sum_{k=1}^{K-1} e_I^{kT} \mathbf{P}_I^k e_I^k \quad (1)$$

where k is the index of all the keyframes K, and j is the landmark index belonging to set τ of keyframe k. $\mathbf{P}_r^k$, $\mathbf{P}_s^k$, and $\mathbf{P}_I^k$ are the information matrices of the visual landmark reprojection, sonar-camera feature matching, and IMU preintegration. Since the matching of the sonar-camera features and the IMU error is based on two adjacent keyframes, the superscript of the summation is $k-1$.

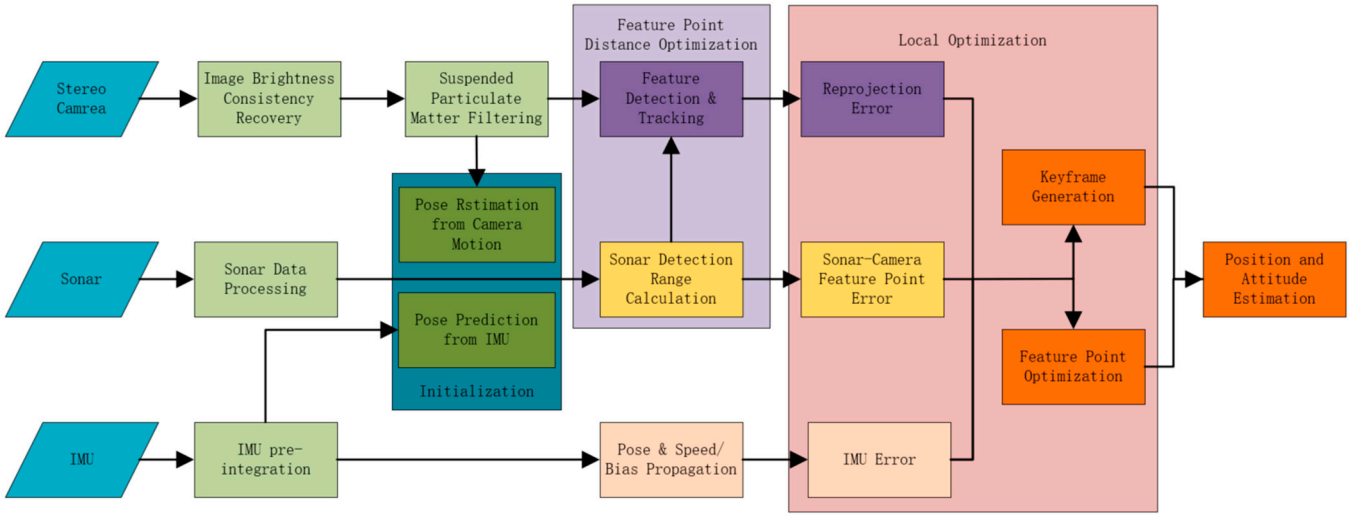


Figure 1. System architecture diagram for the improved visual SLAM with modules of image enhancement and sonar fusion.

4. Proposed Method

4.1. Underwater Image Brightness-Consistency Recovery

Visual odometry is highly sensitive to changes in light, and the uneven absorption of light in water significantly affects the extraction and matching of visual feature points. Different wavelengths of light attenuate at different rates in water: longer wavelengths like red light have a higher attenuation rate and weaker penetration, usually only reaching 3–4 m. In contrast, shorter wavelengths like blue and green light can travel further. This uneven light attenuation causes underwater optical image distortion, often resulting in images having a bluish or greenish tint. This color distortion reduces image contrast and increases the difficulty of extracting feature points.

In an underwater cave environment, the main source of light is the searchlight carried by the exploration platform, which is an artificial light source with strong directionality, limited by the power of the light source; the brightness difference between the inside and outside of the artificial light source illumination range is large. Using searchlights to illuminate a designated target can create shadow areas due to obstructions in the light path. These effects result in an uneven distribution of brightness in the image, typically characterized by high brightness in the central area and low brightness in the surrounding areas. Additionally, parts of protruding objects that face away from the light source appear as low-brightness black areas.

To improve the contrast of underwater images, image contrast enhancement processing is usually performed using the HE (histogram equalization) method. However, directly applying the algorithm to process underwater images can result in increased brightness in already high-brightness regions and decreased brightness in low-brightness regions,

exacerbating overexposure and underexposure issues. To mitigate these problems, it is essential to restore the light intensity of the underwater image before enhancement. This helps to reduce the brightness differences in various image regions caused by the uneven illumination of artificial light sources.

To establish an underwater light model, the image information recorded by the camera is considered as the superposition of the reflected light from the scene and the scattered light in the water. The light intensity at each position in the image can be expressed by the following equation:

$$p_{i,j} = \left[q_{i,j}\omega + a_{i,j}(1-\omega)\right]\alpha_{i,j} \tag{2}$$

where $p_{i,j}$ represents the light intensity of the image at the (i,j) position, the gray value of the image at that position. $q_{i,j}$ is the reflected light intensity at the location, $a_{i,j}$ is the scattered light intensity, and ω is the component weight; the setting of the weight parameter varies depending on factors such as water temperature, depth, salinity, light intensity, or incident angle. Due to the limitation of the irradiation range of the artificial light source resulting in different light intensities at different locations, the light range attenuation coefficient $\alpha_{i,j}$ is introduced to represent the attenuation coefficient at the image location (i,j).

The underwater image to be processed, centered at pixel $p_{i,j}$, with a window $I_{i,j}$ of size 20×20 covering pixels with a gray-scale mean value of $\mu_{i,j}$ and a standard deviation of $\sigma_{i,j}$ can be represented as follows:

$$\mu_{i,j} = \sum_{m=-9}^{10}\sum_{n=-9}^{10} p_{i+m,j+n}/400 \tag{3}$$

$$\sigma_{i,j}^2 = \sum_{m=-9}^{10}\sum_{n=-9}^{10} \left(p_{i+m,j+n} - \mu_{i,j}\right)^2/399 \tag{4}$$

The maximum value $\mu_{\max}$ of the mean gray value of the pixel points in the coverage range of each window is selected as the base brightness, and the range attenuation coefficient at this position is considered to be 1. According to the invariance of the light distribution, the mean and the standard deviation of the pixel distribution of each window should be roughly the same in the case of sufficient light. Therefore, the light attenuation coefficients at different positions can be obtained from the difference of each window $\mu_{i,j}$:

$$\alpha_{i,j} = \mu_{i,j}/\mu_{\max} \tag{5}$$

Since the range of the window $I_{i,j}$ is small, it can be approximated that the scattered light intensity a within the range of $I_{i,j}$ is unchanged; at any position within the window, the scattered light intensity is constant, so the variance $\sigma_{i,j}^2$ can be deduced as the following equation:

$$\begin{aligned}
\sigma_{i,j}^2 &= \sum_{m=-9}^{10}\sum_{n=-9}^{10} \left(p_{i+m,j+n} - \mu_{i,j}\right)^2/399 \\
&= \sum_{m=-9}^{10}\sum_{n=-9}^{10} \left[q_{i+m,j+n}\omega_{i,j} + a_{i+m,j+n}\left(1-\omega_{i,j}\right) - \overline{q}_{i,j}\omega_{i,j} - \overline{a}_{i,j}\left(1-\omega_{i,j}\right)\right]^2\alpha_{i,j}^2 \\
&= \omega_{i,j}^2\alpha_{i,j}^2 \sum_{m=-9}^{10}\sum_{n=-9}^{10} \left(q_{i+m,j+n} - \overline{q}_{i,j}\right)^2/399 \\
&= \omega_{i,j}^2\alpha_{i,j}^2\sigma_{q_i,j}^2
\end{aligned} \tag{6}$$

where $\sigma_{q\_i,j}^2$ is the variance of the reflected light distribution in window $I_{i,j}$, and $\overline{q}_{i,j}$ is the mean of the reflected light in that window. Next, find the maximum value of the standard

deviation $\sigma_{i,j}/\alpha_{i,j}$ after removing the effect of the attenuation coefficient in all windows, denoted as $\sigma_{\alpha\_\max}$:

$$\sigma_{\alpha\_\max} = \max \frac{\sigma_{i,j}\mu_{\max}}{\mu_{i,j}} \tag{7}$$

Approximating the weight of the scattered light at this position as 0, i.e., $\omega = 1$, based on the standard deviation invariance assumption $\sigma^2_{q\_i,j} = \sigma^2_{\alpha\_\max}$, the weight of the receivable reflected light $\omega_{i,j}$ satisfies the following equation:

$$\omega_{i,j} = \frac{\sigma_{i,j}}{\alpha_{i,j}\sigma_{a\_\max}} = \frac{\sigma_{i,j}\mu_{\max}}{\mu_{i,j}\sigma_{a\_\max}} \tag{8}$$

In the absence of natural light interference, the minimum value of the reflected light pixel gray value q in each window $I_{i,j}$ is close to 0:

$$q^I_{i,j\_\min} = 0 \tag{9}$$

Under these conditions:

$$\begin{aligned} p^I_{i,j\_\min} &= q^I_{i,j\_\min}\omega_{i,j}\alpha_{i,j} + \alpha_{i,j}(1-\omega_{i,j})\alpha_{i,j} \\ &= \alpha_{i,j}(1-\omega_{i,j})\alpha_{i,j} \end{aligned} \tag{10}$$

where $p^I_{i,j\_\min}$ denotes the minimum value of the light intensity of the image in window $I_{i,j}$. So, the light attenuation coefficient $\alpha_{i,j}$ at each position in the image can be expressed as follows:

$$a_{i,j} = \frac{p^I_{i,j\_\min}}{(1-\omega_{i,j})\alpha_{i,j}} \tag{11}$$

Based on the attenuation coefficient $\alpha_{i,j}$ and the scaling coefficient $\omega_{i,j}$, the pixel q can be calculated:

$$q_{i,j} = \left(p_{i,j} - p^I_{i,j\_\min}\right)\frac{\sigma_{\alpha\_\max}}{\sigma_{i,j}} \tag{12}$$

After performing the necessary calculations on the pixel points, it is possible to restore the image pixel values to reflect condition q of uniform and sufficient lighting. This approach helps to mitigate the problem of insufficient contrast enhancement caused by uneven illumination to a certain extent.

Then, the underwater images are processed using an adaptive histogram equalization (AHE) algorithm based on illumination consistency reduction. Initially, the images are divided into numerous small regions, and each region undergoes histogram equalization (HE) tailored to its local characteristics. For darker regions, the brightness is increased to enhance contrast and visual effect, while for brighter regions, the brightness is reduced to prevent overexposure or distortion.

Figure 2 shows the results of the original image, the HE-processed image, and the AHE-processed image. From an image perspective, it is evident that the image processed by HE exhibits a larger area of overexposure and white noise. This occurs because the HE algorithm adjusts the gray-level distribution across the entire image globally. In areas of higher brightness in the original image, enhancing overall contrast amplifies the brightness values of these areas, leading to overexposure and noise.

Conversely, darker regions have their gray levels reduced, which diminishes contrast and can result in the loss of fine image details and a reduction in feature points. In contrast, after AHE processing, the contrast near the rock surface is improved, and the contours of objects in the distant background become clearer. AHE adapts histogram equalization locally to each region of the image, avoiding the issues of brightness anomalies and white noise seen with global HE processing.

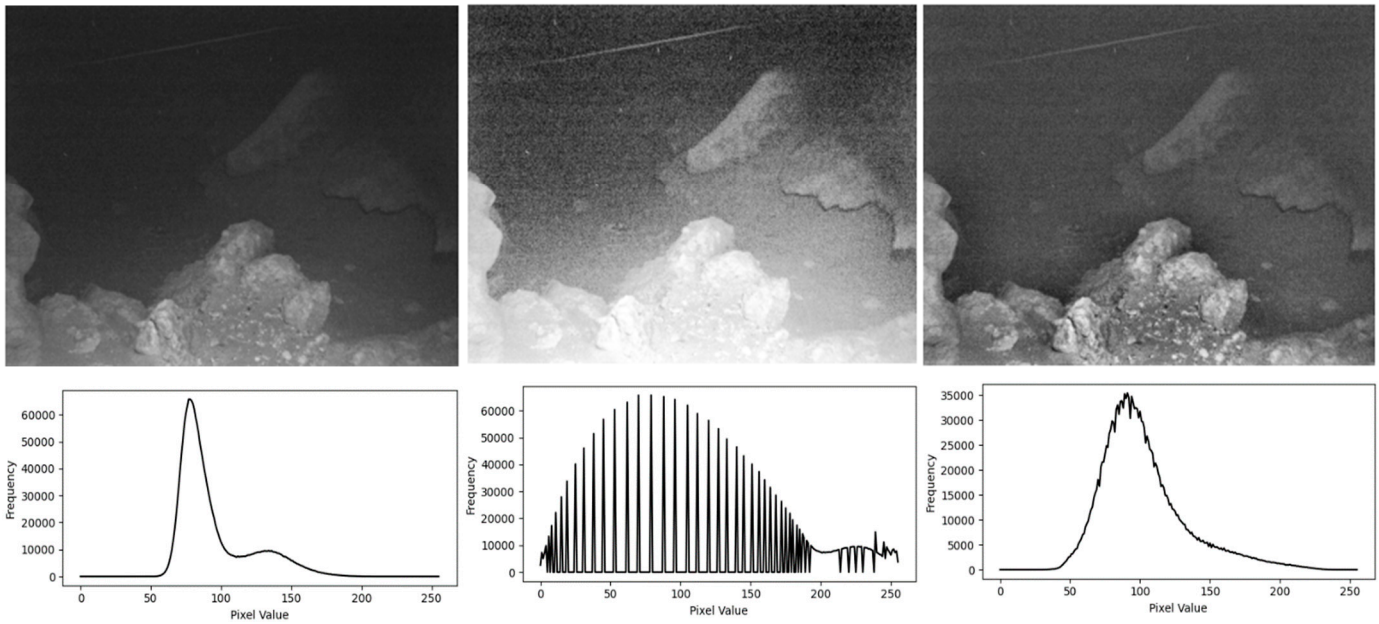


Figure 2. Gray-scale image, HE-processed image, illuminance-consistent AHE-processed image, and corresponding histograms.

From a quantitative perspective, it is clear that histograms play a crucial role in assessing image quality. An image with uniformly distributed gray levels across its pixels and spanning all possible gray levels signifies high contrast and diverse gray tones. Ideally, in a randomly distributed gray-value image, the histogram would resemble a normal distribution. To perform pixel-value statistics on three images, the horizontal axis represents pixel values ranging from 0 to 255, while the vertical axis indicates the frequency of each pixel value appearing in the images.

After applying the HE algorithm, the resulting histogram tends to show spikes. These spikes indicate an excess of pixels with certain gray levels that were expanded during the equalization process, as for pixel values greater than 150 and less than 50, highlighting noise or specific textures and details. This demonstrates the challenge of global histogram equalization in uniformly managing diverse image regions.

Conversely, in the AHE-processed image, the histogram distribution approaches a normal curve more closely compared to the original image. This reduction in peak gray values suggests effective brightness uniformity processing, as the occurrence frequency of pixel peaks has decreased from around 65,000 to 35,000. By addressing local brightness issues, AHE helps prevent excessive brightness in specific image areas, resulting in a more balanced and visually coherent image.

4.2. Underwater Suspended Particulate Filtration

Image blurring caused by suspended particulate matter underwater is akin to the haze-induced blurring observed in terrestrial environments. Therefore, these suspended particles can be considered analogous to noise and addressed using image-defogging algorithms. The underwater environment is highly dynamic and complex. Factors such as seasonal changes, water temperature variations, and lighting conditions can significantly impact the quality of images captured by cameras, even within the same body of water. Hence, it is crucial to assess the necessity of filtering suspended particulate matter based on specific image conditions. The presence of suspended particulate matter creates a grayish haze that can be detected using blurring-detection techniques.

By detecting the image gradient, the degree of blurring in an image can be assessed. The image gradient is calculated by determining the rate of change along the x-axis and y-axis of the image, thereby obtaining the relative changes in these axs. In image processing,

the gradient of an image can be approximated as the difference between neighboring pixels, using the following equation:

$$\frac{\partial f(x,y)}{\partial x} = f(x+1,y) - f(x,y) \tag{13}$$

$$\frac{\partial f(x,y)}{\partial y} = f(x,y+1) - f(x,y) \tag{14}$$

A Laplacian operator with rotational invariance can be used as a filter template for computing the partial derivatives of the gradient. The Laplacian operator is defined as the inner product of the first-order derivatives of the two directions, denoted as Δ:

$$\Delta = \nabla^2 f(x,y) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{15}$$

In a two-dimensional function $f(x,y)$, the second-order differences in the x and y directions are as follows:

$$\frac{\partial^2 f}{\partial x^2} = f(x+1,y) + f(x-1,y) - 2f(x,y) \tag{16}$$

$$\frac{\partial^2 f}{\partial y^2} = f(x,y+1) + f(x,y-1) - 2f(x,y) \tag{17}$$

The equation is expressed in a discrete form to be applicable in digital image processing:

$$\begin{array}{l} \nabla^2 f(x,y) = f(x+1,y) + f(x-1,y) \\ +f(x,y+1) + f(x,y-1) - 4f(x,y) \end{array} \tag{18}$$

If the pixels have high variance, the image exhibits a wide frequency-response range, indicating a normal, accurately focused image. Conversely, if the pixels have low variance, the image has a narrower frequency-response range, suggesting a limited number of edges. Therefore, the average gradient, which represents the sharpness and texture variation of the image, is used as a measure: a larger average gradient corresponds to a sharper image. Abnormal images are detected by setting an appropriate threshold value to determine the acceptable range of sharpness. When the calculated result falls below the threshold, the image is considered blurred, indicating that the concentration of suspended particulate matter is unacceptable and requires particulate-matter filtering. If the result exceeds the threshold, the image is deemed to be within the acceptable range of clarity, allowing for the next step of image processing to proceed directly.

After determining whether the image is blurry, the issue of blurring in underwater camera images resulting from suspended particulate matter can be addressed by drawing parallels with haze conditions on the ground. Viewing suspended particulate matter in water as a form of noise, an image-defogging algorithm (DCP) can be employed to mitigate the blurring effect.

It is hypothesized that in a clear image devoid of suspended particulate matter, certain pixels within non-water regions, such as rocks, consistently exhibit very low intensity values:

$$J^{dark}(x) = \min_{y \in \Omega(x)} \left(\min_{c \in \{r,g,b\}} J^c(y) \right) \tag{19}$$

where J^c denotes each channel of the color image, $\Omega(x)$ denotes a window centered on pixel x, and c denotes one channel of R, G, and B color. Dark channels in underwater images stem from three primary sources: shadows cast by elements within the underwater environment, such as aquatic organisms; brightly colored objects or surfaces, like aquatic plants and fish; and darkly colored objects or surfaces, such as rocks.

Before applying the DCP method to filter suspended particles from underwater images, it is crucial to acknowledge a significant disparity between underwater images and foggy images. The selective absorption of light by the water body results in a reduced red component in the image, which can potentially interfere with the selection of the dark channel. To effectively extract the dark channel of the underwater image, the influence of the red channel must be mitigated. Therefore, the blue–green channel is selected for dark channel extraction.

The imaging model of a foggy image is expressed as

$$I(x) = J(x)t(x) + A[1 - t(x)] \tag{20}$$

where $I(x)$ is the image to be defogged, $J(x)$ is the fog-free image to be recovered, the parameter A is the optical component, which is a constant value, and $t(x)$ is the transmittance. The two sides of the equation are deformed by assuming that the transmittance $t(x)$ is constant within each window and defining it as $\widetilde{t}(x)$, and then two minimum operations are performed on both sides to obtain the following equation.

The imaging model with the presence of more suspended particulate matter is expressed as

$$\min_{y\in\Omega(x)}\left[\min_{c}\frac{I^c(y)}{A^c}\right] = \widetilde{t}\min_{y\in\Omega(x)}\left[\min_{c}\frac{J^c(y)}{A^c}\right] + 1 - \widetilde{t}(x) \tag{21}$$

According to the dark primary color theory, the intensity of the dark channel of the fog-free image tends to zero. It means the intensity of the dark channel in the fogged image is greater than that of the fog-free image Because in foggy environments, the light is subjected to scattering by particles, which results in additional light, and the intensity of the fogged image is higher than that of the fog-free image:

$$J^{dark}(x) = \min_{y\in\Omega(x)}\left[\min_{c} J^c(y)\right] = 0 \tag{22}$$

and it can be deduced that

$$\min_{y\in\Omega(x)}\left[\min_{c}\frac{I^c(y)}{A^c}\right] = 0 \tag{23}$$

The intensity of the dark channel of a fogged image is used to approximate the concentration of fog, which is expressed as the density of suspended particulate matter in the underwater image. Considering the situation in an actual underwater environment, retaining a certain degree of suspended particulate matter, the transmissivity can be set as

$$\widetilde{t}(x) = 1 - \omega\min_{y\in\Omega(x)}\left[\min_{c}\frac{I^c(y)}{A^c}\right] \tag{24}$$

In some cases, extreme values of the transmittance can occur. In order to prevent the value J from being abnormally large when the value t is very small, leading to the overall overexposure of the image screen, a threshold value t_0 is set. By empirical judgment and experimental measurement, the value is considered as $\omega = 0.9$, and $t_0 = 0.1$. The final image-recovery formula is as follows:

$$J(x) = \frac{I(x) - A}{\max[t(x), t_0]} + A \tag{25}$$

To verify the effectiveness of the defogging algorithm on underwater images, a photo was taken from a dataset near an artificial shipwreck on a sediment-covered rocky seabed. The overall image has a greenish tint, and due to the presence of numerous particles in the water, there is severe scattering, making the image appear whitish. As shown in Figure 3, after applying the image-defogging algorithm, the suspended particles in the water are

effectively removed, resulting in a clearer image. The improvement in image quality allows for a more detailed observation of the seabed terrain.

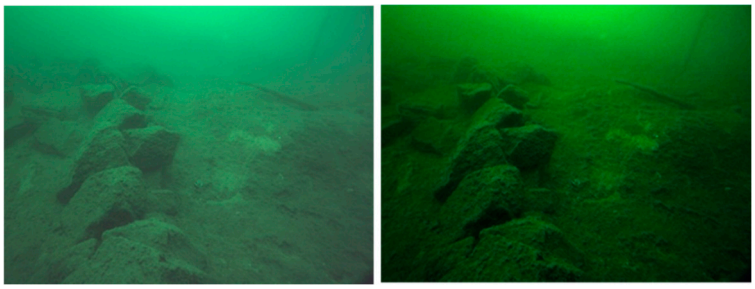


Figure 3. Original image and processed image using DCP.

4.3. Acoustic and Visual Feature Association for Depth Recovery

The spatial detection range of a sonar is typically visualized as a spherical configuration with the sonar device at its center. When targeting a specific direction, the detection area is effectively confined to a prism-shaped region. Leveraging the sonar's horizontal resolution, individual beams are associated with a fan-shaped ring in cross-section. In terms of data structure, each angle in the horizontal plane corresponds to a beam. Each beam consists of many bin values, where each bin corresponds to a distance extending outward from the center of the sonar's emission. For example, if an object is located 20 m from the sonar, and assuming each bin has a resolution of 0.1 m, the position at 20 m would correspond to the 200th bin value. Echoes caused by the object would increase the 200th bin value, while bin values without echoes would remain at 0.

The uncertainty in sonar detection range increases with the distance between targets. As targets move farther away, a single sonar beam covers a broader area, especially noticeable in the vertical dimension where the sonar beam aperture widens. In previous processing methods, a common approach involved extracting the point with the highest bin value within a beam and then calculating the spatial distance to its centroid, which was considered the spatial feature point for the sonar. However, the uncertainty in sonar features arises from two main factors. Firstly, the sparse resolution of sonar, combined with an underwater environment's impact on bin-value distortion, hampers accurate distance measurements to targets. Secondly, as one moves farther from the center of the sonar, a bin value represents a spatial region rather than an exact point, as shown in Figure 4.

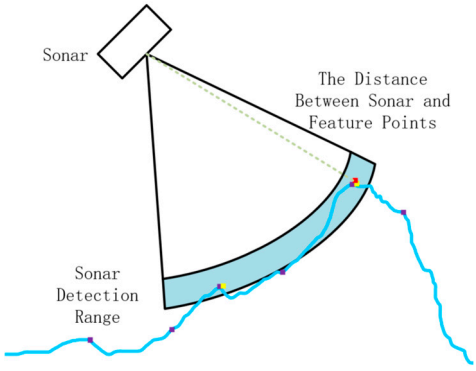


Figure 4. Bin coverage increases with growing distance, and all objects in this range correspond to only one value of the bin, which reduces the probability of obtaining an accurate distance and direction of the object through sonar echoes. The blue curve represents the edge of the target, such as the seabed or rock. The sector represents a beam, and the blue arc-shaped area represents a bin value in the beam. Feature points in the area, such as red marked point, will cause acoustic echoes and increase the bin value.

In the coverage area of a bin, there may be multiple objects, and their edges may correspond to many visual feature points. Therefore, accurately correlating sonar feature points with visual feature points is challenging. This paper explores an alternative method. Firstly, it computes the spatial positions of coarse visual feature points within the scanning area of the sonar. Then, for each beam (corresponding to an angle in space), it identifies the maximum bin value. Subsequently, it considers the two values before and after this maximum bin value, totaling five consecutive bin values, as illustrated by the five consecutive black dots in Figure 5. Based on the geometric structure of the sonar beam, the spatial region corresponding to this set of bins is calculated, as shown by the middle section between the two green sector areas. The true object that makes the maximum bin should be in this region, as the orange rectangle. The visual feature points that fall within this spatial region are referred to as candidate visual feature points, as shown with green points.

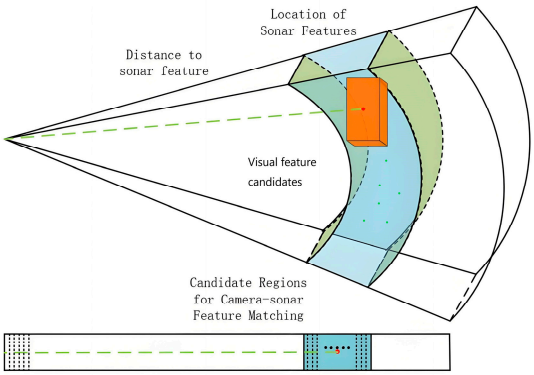


Figure 5. Candidate matching of visual feature points with one sonar feature point.

Typically, the distribution of feature points in an image is quite random, with some local areas containing more edges and corners, resulting in a higher density of extracted feature points. To improve computational efficiency, a quadtree method is often used to achieve a more uniform distribution of feature points as shown in Figure 6. In traditional quadtree homogenization, feature points in the image are recursively partitioned into four equally divided regions. This process continues until a predetermined condition related to the number of feature points is met. Ultimately, only one feature point is retained in each final segmented region after the equalization process is completed.

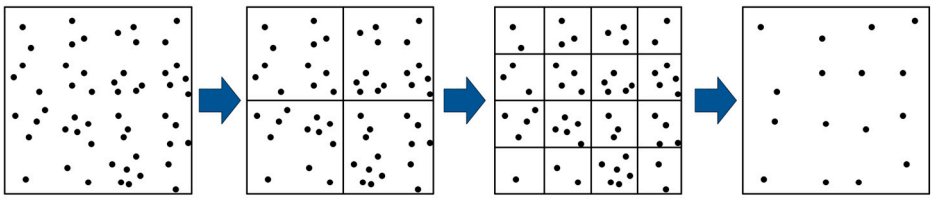


Figure 6. Using a quadtree for feature unification.

After applying the association method described earlier, a single sonar feature may be associated with one or multiple visual features. When multiple visual features are involved, these features may extend across multiple image regions following quadtree segmentation. During this partitioning process, when a group of mutually correlated points is connected to form a polygon, the quadtree's partitioning region for this group of points should be larger than its minimum bounding rectangle (MBR).

As depicted in Figure 7, in the partitioning process shown in the bottom-left corner, each small area ensures it contains visual feature points. However, the visual feature points associated with a sonar feature are divided into different child-node regions. Contrastingly, in the top-right corner, all points of a node's region are contained within a single set.

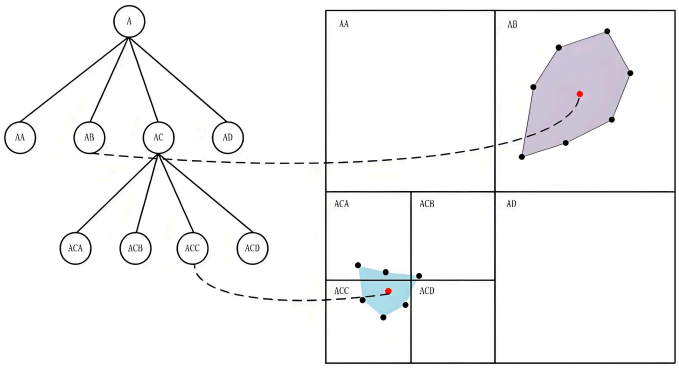


Figure 7. Different polygenes in the partitions of the quadtree.

In traditional quadtrees, only leaf nodes can be assigned an object (one polygon), hence an object may be assigned to more than one leaf node; it means these leaf nodes share the same depth from the sonar feature. While in the proposed quadtree segmentation, if the range matrix of a node contains the MBR of an indexed object and the range matrices of its four child nodes do not contain the MBR of that indexed object (intersecting or diverging), the object is added to that node. In this way, the root-node intermediate nodes are able to be assigned indexed objects, and the objects assigned to each node are not duplicated.

The termination condition for quadtree recursion in this method does not solely rely on the number of feature points but must also consider the resolution capabilities of the sonar. By incorporating the detection distance and resolution of the sonar, the maximum detection range of the sonar opening can be calculated. When the size of the divided area in the quadtree becomes smaller than the area of the polygenes, simple image segmentation becomes ineffective in providing optimal information for feature matching between the sonar and the camera. Thus, image segmentation is terminated to conserve computational resources.

5. Results and Experiments

The system proposed in this paper for underwater SLAM will undergo testing utilizing an AFRL dataset [32], which encompasses a binocular camera, an inertial measurement unit, a sonar, and a water-pressure sensor. The detailed specifications of these sensors are listed in Table 1.

Table 1. Sensor specifications of the underwater dataset.

Sensor	Specifications
Cameras (IDS UI-3251LE)	15 frames per second Resolution: 1600 × 1400
Sonar (IMAGENEX 831L)	Angular resolution: 0.9° Maximum detection distance: 6 m Scanning period: 4 s Effective intensity range: 6 to 255
IMU (Microstrain 3DM-GX4-15)	Frequency: 100 Hz Noise density: $0.005^\circ/s/\sqrt{Hz}$ Drift: $10^\circ/h$
Pressure Sensor (Bluerobotics Bar30)	Frequency: 15 Hz Maxdepth: 300 M

All acquired and processed data are recorded in a package file using ROS. The detailed ROS topics for the sensors are shown in Table 2. The dataset was collected from a cave in Ginnie Spring, FL, USA. The sonar was configured with a higher rate to accommodate the underwater environmental scene. As natural lighting was lacking in the cave, an

underwater searchlight was utilized to supplement the scene during video recording. However, due to constraints related to the searchlight's light angle and power, it could only illuminate within a certain angle corresponding to the direction of the underwater robot's travel. This led to significant differences in illumination between the center and edge areas, resulting in poorer camera light conditions compared to shallow water environments. In this dataset, the presence of dynamic obstacles such as fish and crabs, coupled with substantial amounts of suspended particulate matter in the underwater environment, will significantly affect both acoustic-signal-based sonar and optical-based underwater cameras.

Table 2. Topic information of the different sensors in the dataset.

Sensor	Ros Topic	Data
Camera	/slave1/image_raw/compressed /slave2/image_raw/compressed	Left camera image Right camera image
IMU	/imu/imu	Angular velocity and acceleration data
Sonar	/imagenex831l/range_raw	Acoustic image
Pressure sensor	/bar30/depth	Bathymetric data

Firstly, some comparative experiments on image parameterization are conducted. Five images are selected from the dataset. In the first, second, and fourth images, there are large areas of the seabed and rock regions, while the edges of the rocks are less prominent in the third and fifth images. The first and second images show significant brightness differences, with the distant areas in the field of view appearing as black, invisible regions. The fourth and fifth images have noticeable suspended matter, and in the fourth image, the camera is close to the seabed, resulting in a shorter object distance. The algorithm for image enhancement which employs adaptive histogram equalization (AHE) combined with a dark channel priority (DCP) suspension-removal algorithm is tested against the traditional HE method; the results after processing are shown in Figure 8 with the original and gray-scale images.

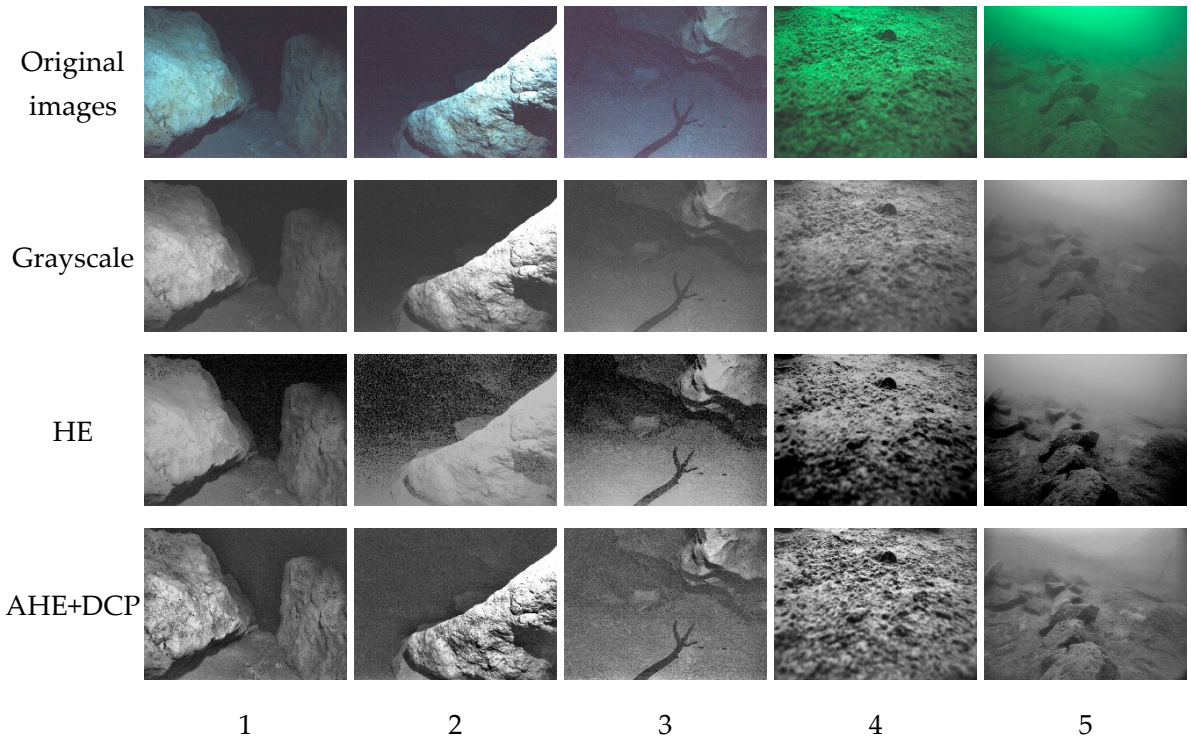


Figure 8. Comparison of the HE and AHE+DCP enhancement effects on five different underwater images.

Next, to evaluate the performance of the image-processing algorithms by quantifying the differences between the original and processed images, a statistical analysis of image-quality parameters is conducted. This includes calculating the mean squared error (MSE) based on image pixels, signal-to-noise ratio (SNR), peak signal-to-noise ratio (PSNR), and processing time; the statistics are shown in Table 3.

Table 3. Comparison of HE and AHE in SNR, MSE, PSNR, and processing time.

Method		1	2	3	4	5
HE	SNR	8.04 dB	6.53 dB	9.64 dB	6.26 dB	6.61 dB
	MSE	2591.091	3618.180	1767.297	3846.195	3552.464
	PSNR	13.9961 dB	12.5461 dB	15.6538 dB	12.2856 dB	12.6368 dB
	Time	0.0963 s	0.0681 s	0.1383 s	0.5485 s	0.5774 s
AHE+DCP	SNR	9.15 dB	7.95 dB	13.31 dB	6.34 dB	6.69 dB
	MSE	1979.318	2606.236	757.346	3778.759	3565.582
	PSNR	14.3877 dB	13.975 dB	19.346 dB	12.369 dB	12.715 dB
	Time	0.0823 s	0.0573 s	0.1454 s	0.5758 s	0.5758 s

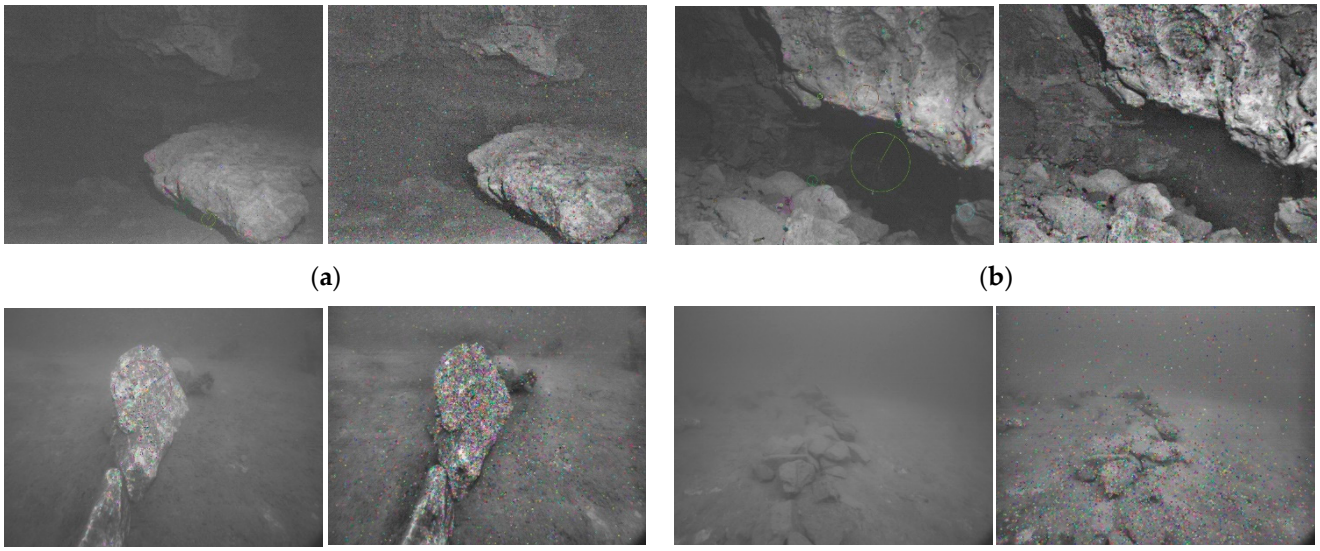
From the data in the table, it is evident that the SNR of the AHE+DCP method has higher values, which is more pronounced in the first three images. The significant differences in SNR across these three images demonstrate that this method has strong adaptability to varying SNR levels. The PSNR values also generally follow this trend. In the comparison of MSE values, the first three images show a clear advantage, while the last two images are roughly equal.

The processing time of the two methods varies according to the complexity of the image scene. In the first three images, the scene complexity is low, visibility is clear, and there are few obstructions, resulting in lower computational demand. In the last two images, the light is absorbed by the water, leading to noticeable color distortion. Additionally, the presence of suspended particles creates fog-like obstructions, resulting in blurred visibility, and the seabed is covered with flocculent substances. Compared to the first three images, the processing time for the last two images is significantly longer. However, there is no noticeable difference in the processing time between the two methods, with AHE+DCP still maintaining a certain advantage.

Feature extraction is then tested using a feature detector to extract SIFT features from the original image and the processed image. The experiment first tested four different scenarios, corresponding to the subfigures in Figure 9 below. From the comparison, it can be seen that the processed images successfully extracted more feature points and distributed them more uniformly.

To further quantify the comparison results directly, the experiment counted the number of feature points extracted from four sets of images. As shown in Table 4, it is clear that the number of feature points increases after using the HE algorithm, which is the first method discussed in this paper, especially for the fourth image, which has a greenish tint and looks blurry. After applying both the AHE and DCP suspension-matter-removal algorithms together, meaning the first and second image-processing methods combined, it is obvious that even more feature points are extracted.

Next, the matching capability of the visual feature points in the processed images was tested, as shown in Figure 10. In the matching process, the red connecting line represents that the pair of matching has higher confidence, while the green connecting line represents that the pair of matching has lower confidence. It can be observed that after processing, in all four scenes, the majority of feature point matches maintained a high level of confidence.



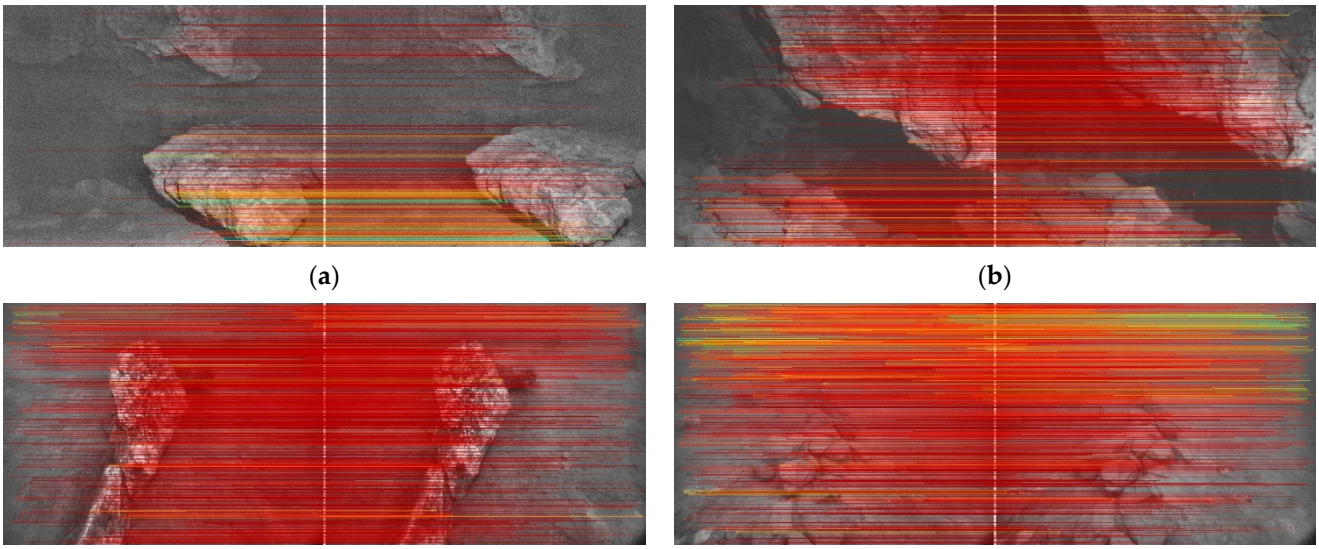
(**a**) (**b**)

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

Figure 9. Visual feature point extraction contrast in different scenes. (**a**) Rocky area; (**b**) Shadowed area; (**c**) Rock surface and seabed; (**d**) Surfaces of the rocks.

Table 4. Comparison of the number of feature points before and after image processing.

	1	2	3	4
Original Image	103	351	806	2
HE	135	460	821	288
AHE+DCP	198	682	1125	523



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

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

Figure 10. Test of visual feature point matching for the processed images in different scenes. (**a**) Rocky area; (**b**) Shadowed area; (**c**) Rock surface and seabed; (**d**) Surfaces of the rocks.

To quantitatively evaluate the performance of the proposed method, this paper conducted statistical comparisons between the number of matched feature points in the original, unprocessed images and the processed ones, as shown in Table 5.

Table 5. Number of successfully matching pairs for the different scenes.

	1	2	3	4
Original Image	85	321	613	0
Enhanced Image	125	456	746	233

In Scene 1, Figure 9a, feature points primarily cluster in the exposed rocky areas above and to the lower right of the image, resembling ground scenes. However, a shadowed area beneath the rock where direct light cannot reach results in lower confidence levels for feature point matching near this region. Consequently, these areas appear as lightly colored matches in the image. The background area at the far end is also constrained by the searchlight's power, appearing black and rendering feature extraction impossible. Nevertheless, the overall quantity and quality of matches meet the system's basic requirements.

In Scene 2, Figure 9b, the primary concentration of feature points lies within the upper-right and lower-left exposed rock areas. The shadowed area within the camera's view is reduced, and appropriate illumination contributes to high confidence levels in the feature points. The only region lacking sufficient brightness for feature extraction and matching is the crevice between the rocks, due to inadequate illumination.

In Scene 3, Figure 9c, feature points are present not only on the rock surface but also in certain areas of the seabed, facilitating feature extraction. The method proposed in this paper can further increase the number of feature point matches based on this foundation.

In Scene 4, Figure 9d, the surfaces of the rocks are covered with flocculent-graded sediment, which may hinder the efficiency of relying on corner points and edges for feature extraction. This is reflected in the lower confidence level of feature matching in the upper half of the image. In the original, unprocessed image, there were no successful matches of any feature point pairs. However, after the image-enhancement process, successful feature matching is achieved.

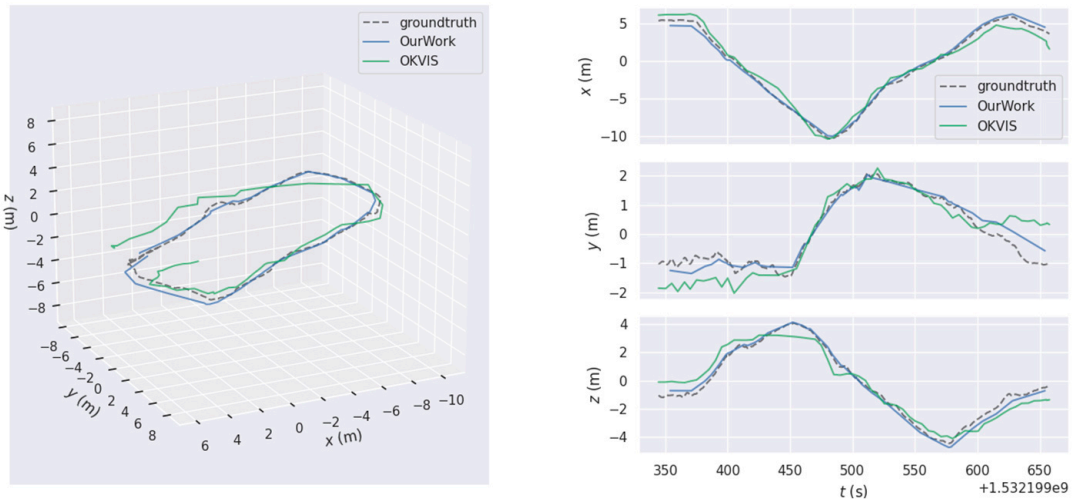
Comparing the number of feature point matches before and after image enhancement reveals that the enhanced images consistently yield more matches than the original ones. In the first, second, and third images, where feature extraction is already feasible in the original image, the enhanced images show varying degrees of improvement in the number of feature matches, with an increase in successfully matching feature pairs.

The primary significance of image processing lies in enhancing the system's robustness in coping with underwater extremes, which is prominently demonstrated in the fourth image. In the fourth image, the contrast is most apparent. The original image has zero feature matches due to the low number of extracted features and the difficulty in finding their counterparts in the other image. However, following the image-enhancement process, the number of extracted features is restored to normal levels, enabling successful feature matching.

The final part of the experiment involves integrating this improved front-end into the SLAM framework and evaluating its trajectory generation results. The dataset provides a set of ground-truth reference trajectories and the results from running OKVIS. Therefore, this experiment also compares the trajectories generated by the improved SLAM framework against these two reference trajectories as shown in Figure 11.

As explained in [23], it is difficult to obtain accurate ground-truth information underwater. This information is derived from manual measurements and subsequent result calibration. In the trajectory plots, the dashed line represents the ground truth, the blue trajectory depicts the path generated by the improved system, and the green trajectory represents OKVIS's path. Observation reveals that the blue trajectory closely aligns with

the reference trajectory, whereas the green trajectory exhibits numerous cumulative errors and considerable fluctuations.



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

Figure 11. Trajectory plot of the dataset: (**a**) comparison of the trajectories for OKVIS, our work, and the ground truth; (**b**) local amplification of the three axes.

On the x-axis, the difference between the two methods is minimal, but the trajectory of the improved system exhibits smoother motion compared to OKVIS, which shows more fluctuations. This trend is also noticeable along the y-axis, particularly in the middle section. In terms of the z-axis, the OKVIS trajectory exhibits significant errors in the middle and front sections, with substantial drift evident toward the tail end. Conversely, the improved system consistently maintains a high level of alignment with the ground truth, demonstrating remarkable consistency throughout.

Next, statistical analysis and presentation of trajectory errors were performed using the EVO plugin. The results are shown in Figure 12 and Table 6. Based on the figures and tables provided, the results obtained from the improved method outperform those of OKVIS comprehensively. In terms of both maximum and average error metrics, the improved method consistently maintains values below 1 when compared to the ground truth. Conversely, the OKVIS results exhibit larger errors, with the maximum error reaching 2.5, indicating significant trajectory shifts in certain instances. Additionally, the sum of squared errors surpasses 50 for OKVIS, suggesting highly unstable trajectories characterized by increased fluctuations and significant drift toward the tail section.

Table 6. Comparison of trajectory statistical results between the two methods.

	The Improved Visual SLAM Results Compared with the Ground Truth	OKVIS Results Compared with the Ground Truth
max	0.829411	2.627724
mean	0.366091	1.148321
median	0.309184	1.020354
min	0.077178	0.295083
rmse	0.420158	1.267575
sse	7.414390	93.191309
std	0.206181	0.536756

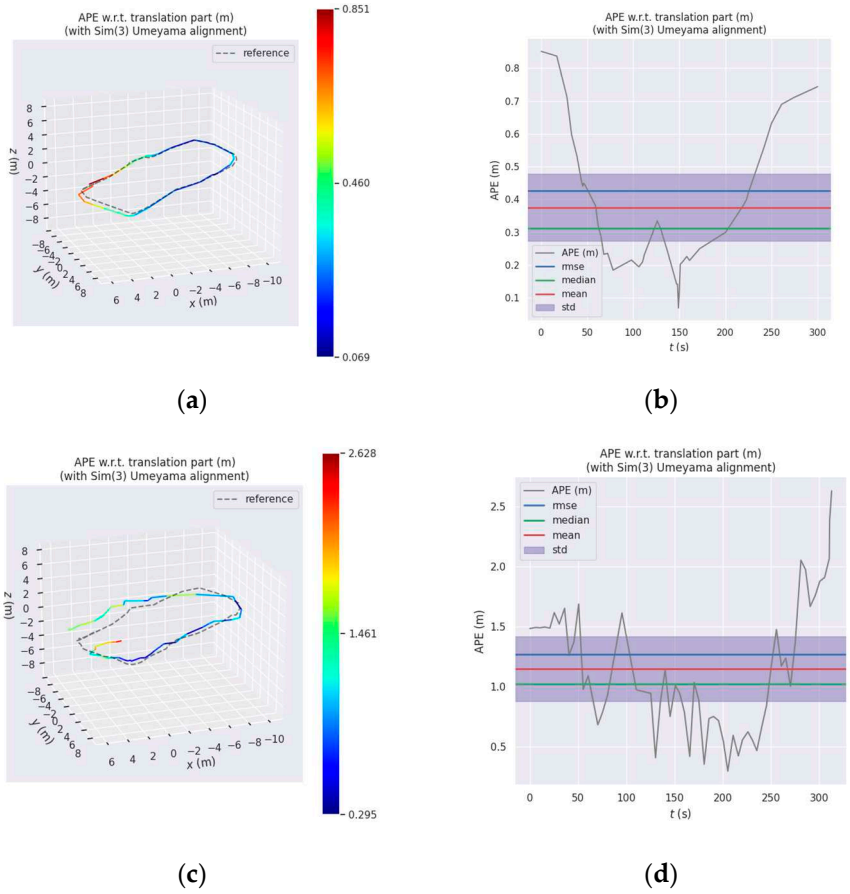


Figure 12. Display of absolute trajectory error using EVO: (**a**) absolute trajectory error (APE) between the improved method and the ground truth; (**b**) display of APE statistics for the improved method compared with the ground truth; (**c**) APE between OKVIS and the ground truth; and (**d**) display of APE statistics for OKVIS compared with the ground truth.

6. Conclusions

OKVIS is a well-established VIO (visual–inertial odometry) visual SLAM framework. However, our data testing indicates that directly applying this framework to underwater environments does not yield effective results. The primary factor affecting accuracy is the insufficient number of visual feature points extracted and successfully matched in underwater images, which significantly degrades estimation accuracy, particularly in the vertical direction. This issue arises because underwater robots may experience jitter or rapid lens switching during ascent and descent, reducing the already limited number of matches and potentially causing misalignment.

Conversely, the enhanced system, after processing image enhancing and feature association between the camera and sonar, demonstrated a commendable performance on the test dataset. It maintained a strong consistency with the ground-truth trajectory and exhibited a robust performance, with no significant fluctuations throughout the trajectory. The number of feature point matches does not have a direct correlation with trajectory-estimation accuracy. However, in extreme cases, such as in this experiment, where a set of images could not extract feature points due to illumination issues, it can lead to failed or erroneous feature point matches because new keyframes may represent new scenes and feature points for matching are insufficient. Additionally, due to the instability of underwater poses, leveraging distance information from sonar also enhances the robustness of feature matching. This is the primary reason why the method in this paper performs well on this dataset compared with a traditional VIO SLAM like OKVIS. Future research will aim to determine whether the image-enhancement method demonstrates good generalization

performance in bright water areas and whether the segmented association method can effectively enhance system performance across different positions of sonar beams.

Author Contributions: This paper is provided by six authors; the authors' contributions are as follows: Conceptualization, H.Q.; data curation, Y.T.; funding acquisition, H.Q. and H.W.; methodology, Y.T. and L.W.; project administration, H.Q.; software, Y.T.; supervision, H.W. and L.W.; writing—original draft, Y.T.; writing—review and editing, D.X. and M.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (52101358, 41906154), the Guangdong Provincial Natural Science Foundation (Youth Enhancement Project) 2024A1515030159, the Jiangsu Provincial Key Research and Development Program Social Development Project (BE2022783), and the Zhenjiang Key Research and Development Plan (Social Development) Project, No. SH2022013.

Data Availability Statement: The data used in this paper is available on the access of the link https://afrl.cse.sc.edu/afrl/resources/datasets/, accessed on 16 May 2024.

Acknowledgments: The authors would like to express their gratitude to all reviewers who have made comments on this article.

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

References

1. Wang, X.; Fan, X.; Shi, P.; Ni, J.; Zhou, Z. An overview of key SLAM technologies for underwater scenes. *Remote Sens.* **2023**, *15*, 2496. [CrossRef]
2. Davison, A.J.; Reid, I.D.; Molton, N.D.; Stasse, O. MonoSLAM: Real-time single camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1052–1067. [CrossRef] [PubMed]
3. Taketomi, T.; Uchiyama, H.; Ikeda, S. Visual SLAM algorithms: A survey from 2010 to 2016. *IPSJ Trans. Comput. Vis. Appl.* **2017**, *9*, 16. [CrossRef]
4. Köser, K.; Frese, U. Challenges in underwater visual navigation and SLAM. *AI Technol. Underw. Robot.* **2020**, *96*, 125–135.
5. Cho, Y.; Kim, A. Visibility enhancement for underwater visual SLAM based on underwater light scattering model. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May 2017–3 June 2017; pp. 710–717.
6. Hidalgo, F.; Bräunl, T. Evaluation of several feature detectors/extractors on underwater images towards vSLAM. *Sensors* **2020**, *20*, 4343. [CrossRef]
7. Zhang, J.; Han, F.; Han, D.; Yang, J.; Zhao, W.; Li, H. Integration of Sonar and Visual Inertial Systems for SLAM in Underwater Environments. *IEEE Sens. J.* **2024**, *24*, 16792–16804. [CrossRef]
8. Leutenegger, S.; Lynen, S.; Bosse, M.; Siegwart, R.; Furgale, P. Keyframe-based visual–inertial odometry using nonlinear optimization. *Int. J. Robot. Res.* **2015**, *34*, 314–334. [CrossRef]
9. Qin, T.; Li, P.; Shen, S. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robot.* **2018**, *34*, 1004–1020. [CrossRef]
10. Geneva, P.; Eckenhoff, K.; Lee, W.; Yang, Y.; Huang, G. Openvins: A research platform for visual-inertial estimation. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 4666–4672.
11. Hidalgo, F.; Kahlefendt, C.; Bräunl, T. Monocular ORB-SLAM application in underwater scenarios. In Proceedings of the 2018 OCEANS-MTS/IEEE Kobe Techno-Oceans (OTO), Kobe, Japan, 28–31 May 2018; pp. 1–4.
12. Ferrera, M.; Moras, J.; Trouvé-Peloux, P.; Creuze, V. Real-time monocular visual odometry for turbid and dynamic underwater environments. *Sensors* **2019**, *19*, 687. [CrossRef]
13. Kim, A.; Eustice, R. Pose-graph Visual SLAM with Geometric Model Selection for Autonomous Underwater Ship Hull Inspection. In Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, USA, 10–15 October 2009.
14. Miao, R.; Qian, J.; Song, Y.; Ying, R.; Liu, P. UniVIO: Unified direct and feature-based underwater stereo visual-inertial odometry. *IEEE Trans. Instrum. Meas.* **2021**, *71*, 8501214. [CrossRef]
15. Palomer, A.; Ridao, P.; Ribas, D. Multibeam 3D underwater SLAM with probabilistic registration. *Sensors* **2016**, *16*, 560. [CrossRef] [PubMed]
16. Suresh, S.; Sodhi, P.; Mangelson, J.G.; Wettergreen, D.; Kaess, M. Active SLAM using 3D submap saliency for underwater volumetric exploration. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 3132–3138.
17. Palomeras, N.; Carreras, M.; Andrade-Cetto, J. Active SLAM for autonomous underwater exploration. *Remote Sens.* **2019**, *11*, 2827. [CrossRef]

18. Huang, C.; Zhao, J.; Zhang, H.; Yu, Y. Seg2Sonar: A Full-Class Sample Synthesis Method Applied to Underwater Sonar Image Target Detection, Recognition, and Segmentation Tasks. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5909319. [CrossRef]
19. Zhou, T.; Si, J.; Wang, L.; Xu, C.; Yu, X. Automatic Detection of Underwater Small Targets Using Forward-Looking Sonar Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4207912. [CrossRef]
20. Abu, A.; Diamant, R. A SLAM Approach to Combine Optical and Sonar Information from an AUV. *IEEE Trans. Mob. Comput.* **2023**, *23*, 7714–7724. [CrossRef]
21. Cheung, M.Y.; Fourie, D.; Rypkema, N.R.; Teixeira, P.V.; Schmidt, H.; Leonard, J. Non-gaussian slam utilizing synthetic aperture sonar. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 3457–3463.
22. Ribas, D.; Ridao, P.; Tardós, J.D.; Neira, J. Underwater SLAM in man-made structured environments. *J. Field Robot.* **2008**, *25*, 898–921. [CrossRef]
23. Joshi, B.; Rahman, S.; Kalaitzakis, M.; Cain, B.; Johnson, J.; Xanthidis, M.; Karapetyan, N.; Hernandez, A.; Li, A.Q.; Vitzilaios, N.; et al. Experimental comparison of open source visual-inertial-based state estimation algorithms in the underwater domain. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 7227–7233.
24. Rahman, S.; Quattrini Li, A.; Rekleitis, I. SVIn2: A multi-sensor fusion-based underwater SLAM system. *Int. J. Robot. Res.* **2022**, *41*, 1022–1042. [CrossRef]
25. Ancuti, C.; Ancuti, C.O.; Haber, T.; Bekaert, P. Enhancing underwater images and videos by fusion. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 81–88.
26. Barros, W.; Nascimento, E.R.; Barbosa, W.V.; Campos, M.F.M. Single-shot underwater image restoration: A visual quality-aware method based on light propagation model. *J. Vis. Commun. Image Represent.* **2018**, *55*, 363–373. [CrossRef]
27. Schettini, R.; Corchs, S. Underwater image processing: State of the art of restoration and image enhancement methods. *EURASIP J. Adv. Signal Process.* **2010**, *2010*, 746052. [CrossRef]
28. Vargas, E.; Scona, R.; Willners, J.S.; Luczynski, T.; Cao, Y.; Wang, S.; Yvan, R. Petillot Robust underwater visual SLAM fusing acoustic sensing. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 2140–2146.
29. Nalli, P.K.; Kadali, K.S.; Bhukya, R.; Palleswari, Y.T.R.; Siva, A.; Pragaspathy, S. Design of exponentially weighted median filter cascaded with adaptive median filter. *J. Phys. Conf. Series. IOP Publ.* **2021**, *2089*, 012020. [CrossRef]
30. Çelebi, A.T.; Ertürk, S. Visual enhancement of underwater images using empirical mode decomposition. *Expert Syst. Appl.* **2012**, *39*, 800–805. [CrossRef]
31. Prabhakar, C.J.; Kumar, P.U.P. Underwater image denoising using adaptive wavelet subband thresholding. In Proceedings of the 2010 International Conference on Signal and Image Processing, Chennai, India, 15–17 December 2010; pp. 322–327.
32. Available online: https://afrl.cse.sc.edu/afrl/resources/datasets/ (accessed on 16 May 2024).