C2ST
C2 Science and Technology

Research Article
https://doi.org/10.26599/C2ST.2026.9700005

Underwater 3D reconstruction via uncertainty-aware binocular-sonar fusion and local transformation refinement

Yang Yang[1,2,3], Yichen Li[1,2,3,*], Chi Zhang[1,2,3], Wenbin Yu[1,2,3], Xinping Guan[1,2,3]

[1] State Key Laboratory of Submarine Geoscience, School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai 200240, China
[2] Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai 200240, China
[3] Shanghai Key Laboratory of Perception and Control in Industrial Network Systems, Shanghai 200240, China

ABSTRACT

Underwater three-dimensional reconstruction is essential for environmental perception, obstacle avoidance, and intervention tasks performed by underwater unmanned vehicles. Acoustic and optical sensing are the two primary means of underwater environmental perception. However, binocular cameras are sensitive to light attenuation and scattering, whereas imaging sonars suffer from low spatial resolution and missing elevation information. Acoustic-optical fusion can provide more accurate and reliable perception information in short-range underwater scenes. However, existing fusion methods remain limited by ambiguous acoustic-optical correspondences, residual sonar-camera calibration errors, heterogeneous measurement uncertainties, and sparse acoustic observations. To address these problems, this paper proposes an uncertainty-aware three-dimensional reconstruction method that fuses a binocular camera with an imaging sonar. The method operates in a single-pose setting using one synchronized binocular image pair and one sonar image. First, candidate 3D positions are generated for each sonar feature by sampling the unobserved elevation angle and projecting them onto the camera image plane. Then, feature-wise local transformation refinement reduces projection errors by dynamically updating acoustic-optical correspondences, and optical and sonar depths are fused according to their uncertainties. Finally, local surface expansion propagates reliable depth corrections to neighboring geometrically consistent pixels to improve reconstruction completeness. Experiments were conducted in three representative underwater scenes, including a ship propeller blade, an underwater vehicle, and an underwater rock wall. The proposed method achieved coverage rates of 58.09%, 46.37%, and 51.21%, respectively, outperforming the comparison methods in all scenes. In a representative refinement case, the mean acoustic-optical depth residual was reduced from 0.2628 m to 0.1233 m. The results demonstrate that the proposed method improves point-cloud coverage and geometric consistency while maintaining competitive local reconstruction accuracy.

KEYWORDS

acoustic-optical fusion; underwater three-dimensional reconstruction; binocular vision; imaging sonar; uncertainty-aware fusion

In recent years, the exploration and development of the ocean have received increasing attention. Consequently, underwater unmanned vehicles (UUVs) have been widely used[1], playing an important role in tasks such as polar scientific research, marine geology, underwater archaeology[2], ship inspection[3], and mineral extraction[4]. During these operations, UUVs rely on robust environmental perception for path planning and obstacle avoidance within unstructured and highly dynamic underwater environments[5]. Conventional underwater perception systems primarily depend on acoustic sensors[6-7] because of their long detection range and robustness to poor visibility. For close-range inspection and fine intervention tasks, both acoustic and optical sensors can provide effective observations[8]. Sonar offers reliable range measurements and robustness to turbidity, whereas optical cameras provide substantially higher spatial resolution, richer texture information, and more detailed object appearance[9]. Underwater three-dimensional (3D) reconstruction is therefore particularly important in close-range operational scenarios because it provides the spatial and geometric information required for accurate localization, manipulation, inspection, and obstacle avoidance[10-11]. Conventional 3D reconstruction methods often require many images acquired from different viewpoints to recover complete target geometry. However, collecting dense multi-view observations by navigating a UUV around a target is difficult in underwater environments because of water currents, limited maneuvering space, and unstable vehicle motion. Therefore, reconstruction from a single sensor pose is highly desirable for close-range inspection and intervention. In this paper, the single-pose setting uses only one synchronized binocular image pair and one sonar image acquired at the same pose, without temporal accumulation or multi-view registration.

Owing to their superior capability to capture fine textures, object contours, and dense geometric details, optical cameras have become indispensable sensing devices for intervention-oriented UUVs operating at close range. Compared with acoustic sensors, optical cameras generally provide richer environmental informa-

Received: June 26, 2026; Revised: July 20, 2026; Accepted: August 18, 2026

*Yichen Li, liyichensjtu@sjtu.edu.cn

Citation: Yang Y, Li Y C, Zhang C, et al. Underwater 3D reconstruction via uncertainty-aware binocular-sonar fusion and local transformation refinement[J]. *C2 Science and Technology*, 2026, 1(1): 49-62. https://doi.org/10.26599/C2ST.2026.9700005.

© Chinese Institute of Command and Control. Publishing services by Tsinghua University Press. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

tion when visibility conditions are sufficient, making them particularly suitable for target identification, surface inspection, and fine manipulation[12]. In standard terrestrial environments, vision-based 3D reconstruction methods have demonstrated remarkable efficacy, relying on distinct feature matching to generate highly accurate spatial models. For instance, an improved RANSAC algorithm for a mobile robot's binocular camera was proposed in Ref. [13], achieving millimeter-level precision in depth calculation. A fast 3D reconstruction framework for warehouse automation was presented in Ref. [14]. However, the translation of these purely visual methods to aquatic environments is severely hindered by the complex physical properties of the water medium. The propagation of light is degraded by rapid absorption and scattering, resulting in image blur, color attenuation, and non-uniform illumination, which compromise reliable feature extraction and drastically reduce depth estimation accuracy. Some researchers have attempted to mitigate these issues using active illumination. In Ref. [15], structured light was projected to enhance binocular stereo matching, while an active light source was utilized for monocular reconstruction of underwater caves in Ref. [16]. More recently, a marker-free underwater reconstruction method combined multi-view dense stereo point clouds with sparse Structure-from-Motion point clouds and employed a two-stage Iterative Closest Point alignment strategy to improve point-cloud registration under underwater imaging distortions[17]. In Ref. [18], line-structured light was integrated with stereo vision, and epipolar constraints were introduced into light-plane calibration to improve the reconstruction accuracy of underwater concrete structures. However, multi-view methods generally require stable acquisition trajectories and sufficiently clear optical observations, whereas structured-light systems require additional active-illumination and calibration equipment. Their performance may therefore deteriorate in turbid water, under nonuniform illumination, or in highly dynamic environments.

To overcome the fundamental limitations of optical sensors, acoustic sensing (e.g., sonar) has emerged as the prevalent alternative. By relying on the intensity and reception time of the echo after emitting sound waves, sonar offers exceptional penetration capabilities that are largely unaffected by underwater illumination and optical turbidity[19]. In particular, the imaging sonars can provide accurate range and azimuth measurements. However, the inherent physical mechanism of sonar imaging inevitably projects 3D space into a two-dimensional (2D) plane, resulting in low spatial resolution and a critical loss of elevation information. This geometric ambiguity makes it exceedingly difficult to restore the true 3D morphology of target objects. Consequently, it is very challenging to perform underwater 3D reconstruction using only sonar. To address the problem of the missing elevation angle in sonar, a method for recovering elevation angles from the perspective differences between consecutive sonar frames was proposed in Ref. [20], although its application was limited to simultaneous localization and mapping (SLAM). In Ref. [21], sonar pairs were utilized to observe points from different perspectives, but their performance was limited in applications requiring large-field-of-view observations. Furthermore, a single-sonar triangulation scheme based on a generative sensor model was introduced in Ref. [22]. Since this method assumes smooth target surfaces, it inevitably loses high-frequency geometric details when dealing with complex targets. Recent studies have further explored learning-based and point-cloud-based sonar reconstruction. A learning-based method estimated a pseudo front view from a single sonar image using a U-Net++-based network, thereby recovering the elevation, azimuth, and depth information required for 3D point-cloud generation[23]. Another study proposed a high-fidelity reconstruction method for noisy 3D underwater sonar point clouds of bridge caisson foundations by combining point-cloud segmentation, two-dimensional boundary extraction, and surface reconstruction[24]. However, learning-based approaches depend on the representativeness and generalization capability of the training data, whereas point-cloud-based methods require specialized 3D sonar measurements or application-specific geometric priors.

Given the respective limitations of single-modality sensing, fusing acoustic and optical sensors has emerged as a compelling paradigm for underwater 3D reconstruction. The synergistic use of these two modalities effectively leverages their complementary strengths: the robust penetration of sonar compensates for the severe depth degradation of cameras in turbid water, while optical sensors provide dense geometric and visual cues that help resolve the elevation ambiguity of imaging sonar. However, executing this multi-modal fusion introduces significant technical challenges. The primary prerequisite is establishing a precise extrinsic transformation matrix between the sonar and camera coordinate systems. Due to their disparate imaging mechanisms, traditional visual calibration protocols cannot be directly applied. In Ref. [25], a custom-designed ArUco board augmented with acoustic reflectors was utilized to manually extract feature correspondences, whereas a calibration method independent of the calibration board was proposed in Ref. [26] by tracking feature points in sonar and camera images. The residual calibration errors remain a non-negligible source of geometric distortion for 3D reconstruction. Beyond calibration, developing robust fusion algorithms presents another major challenge. In Ref. [27], monocular vision and sonar were fused by correcting camera-derived point clouds with sonar depth data within a SLAM framework. A single-beam echo sounder was utilized to constrain the depth of a monocular SLAM system in Ref. [28]. However, these SLAM-dependent schemes impose stringent requirements on real-time performance and stability. In Ref. [29], sonar constraints were integrated directly into the stereo matching algorithm of a binocular system to reduce scale estimation errors. In Ref. [30], the Gaussian splatting algorithm was adopted for camera–sonar fusion to realize 3D reconstruction based on machine learning. More recently, Sonar-MASt3R[31] combined dense optical correspondences extracted by MASt3R with geometric constraints from acoustic reconstruction to achieve real-time opti-acoustic 3D reconstruction. Despite these advances, several challenges remain insufficiently addressed. First, the missing elevation angle of imaging sonar makes direct one-to-one acoustic-optical matching unreliable. Second, residual sonar-camera calibration errors may cause systematic projection deviations. Third, existing fusion strategies often do not explicitly account for the different uncertainty characteristics of optical and acoustic depth measurements. Moreover, reconstruction based only on sparse sonar correspondences may provide limited surface coverage. These issues motivate the development of a reconstruction framework that jointly considers candidate association, local transformation refinement, uncertainty-aware depth fusion, and surface expansion.

To address these challenges, this article investigates the underwater 3D reconstruction problem by fusing a binocular camera and an imaging sonar. The primary idea is to use robust acoustic range measurements to constrain binocular depth estimation while preserving the local geometric details provided by optical observations and reducing the influence of residual sonar-camera transformation errors. The main contributions of this article are summarized as follows:

1) A novel single-pose underwater 3D reconstruction frame-

work based on binocular-sonar fusion is proposed. This framework effectively addresses the critical issues of inaccurate target depth estimation and geometric detail loss by synergistically integrating optical cameras with imaging sonars.

2) An acoustic-optical candidate-matching and feature-wise local transformation refinement strategy is developed. Candidate correspondences are established by sampling the unobserved sonar elevation angle, while residual sonar-camera projection errors are reduced by dynamically updating the correspondences and minimizing the 3D geometric discrepancy between acoustic and optical points.

3) An uncertainty-aware depth fusion and local surface expansion method is introduced. Optical and sonar depths are adaptively fused according to their uncertainties, and reliable depth corrections are propagated to neighboring pixels to improve point-cloud coverage while preserving local surface structure.

1 System modeling and problem formulation

This work considers single-pose underwater 3D reconstruction using a binocular camera and an imaging sonar mounted on a UUV. The binocular camera provides optical observations with relatively high spatial resolution, while the imaging sonar provides robust range and azimuth measurements. Due to the significant differences in imaging mechanisms, coordinate definitions, and measurement characteristics between the two sensors, a unified geometric model is required before acoustic and optical information can be effectively fused.

1.1 Coordinate systems and sensor configuration

The acoustic-optical perception system involves the camera and sonar Cartesian coordinate systems, together with their corresponding image coordinate systems. The binocular camera is composed of a left camera and a right camera. In this paper, the left camera coordinate system is selected as the reference camera coordinate system. A 3D point in the camera coordinate system is denoted as $\boldsymbol{P}_C = [x_C, y_C, z_C]^{\mathrm{T}}$, where x_C, y_C, z_C represent the coordinates along the horizontal, vertical, and optical-axis directions of the camera, respectively. The corresponding pixel coordinate in the optical image is denoted as (u, v).

For the imaging sonar, a 3D point can be represented either in the sonar Cartesian coordinate system or in the sonar spherical coordinate system. In the sonar Cartesian coordinate system, the point is denoted as $\boldsymbol{P}_S = [x_S, y_S, z_S]^{\mathrm{T}}$. In the sonar spherical coordinate system, the same point can be represented by $[r, \theta, \varphi]^{\mathrm{T}}$, where r is the range from the sonar to the target point, θ is the horizontal azimuth angle, and φ is the elevation angle. In practical imaging sonar measurements, r and θ can be directly obtained from the sonar image, whereas φ is not directly observable due to the projection characteristic of the imaging sonar.

1.2 Binocular camera model

For a 3D point $\boldsymbol{P}_C$ in the camera coordinate system, its projection onto the image plane can be described by the pinhole camera model[32]

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \boldsymbol{K}\boldsymbol{P}_C = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_C \\ y_C \\ z_C \end{bmatrix} \tag{1}$$

where λ is the scale factor, $\boldsymbol{K}$ is the intrinsic parameter matrix of the camera, f_x and f_y are the focal lengths along the horizontal and vertical directions, respectively, and (c_x, c_y) is the principal point of the camera. Compared with a monocular camera, a binocular camera can recover depth information through stereo matching. For a pair of rectified stereo images, the disparity $d(u, v)$ between corresponding pixels in the left and right images can be used to calculate the depth value

$$z_c(u, v) = B \cdot f_x / d(u, v) \tag{2}$$

where B is the baseline distance between the two cameras. Once $z_c(u, v)$ is obtained, the 3D coordinate of the corresponding image point can be reconstructed as $[z_c(u, v) \cdot (u - c_x)/f_x, z_c(u, v) \cdot (v - c_y)/f_y, z_c(u, v)]^{\mathrm{T}}$. In this work, the initial optical depth map is obtained through stereo matching. However, underwater optical images are easily affected by light attenuation, scattering, uneven illumination, and refraction[33]. These factors may introduce mismatches and noise into the disparity map, especially in turbid water. Therefore, the optical point cloud obtained directly from stereo matching is usually incomplete and may contain a large number of erroneous points.

1.3 Imaging sonar model

Unlike optical cameras, imaging sonar forms an acoustic image by emitting multiple acoustic beams and receiving echo signals reflected from underwater objects. Each pixel in the sonar image corresponds to a range and an azimuth angle. For a sonar image point (u_s, v_s), the range r and azimuth angle θ can be calculated as

$$r = v_s \eta \tag{3}$$

$$\theta = (u_s - w_s/2)\beta \tag{4}$$

where η is the range resolution of the sonar, w_s is the width of the sonar image, and β is the angular resolution of the sonar. The corresponding 3D point in the sonar Cartesian coordinate system can be written as[34]

$$\boldsymbol{P}_S(\varphi) = \begin{bmatrix} x_S \\ y_S \\ z_S \end{bmatrix} = \begin{bmatrix} r\sin\theta\cos\varphi \\ r\cos\theta\cos\varphi \\ r\sin\varphi \end{bmatrix} = \begin{bmatrix} v_s\eta\sin((u_s - w_s/2)\beta)\cos\varphi \\ v_s\eta\cos((u_s - w_s/2)\beta)\cos\varphi \\ v_s\eta\sin\varphi \end{bmatrix} \tag{5}$$

As shown in Fig. 1, the imaging sonar measurement cannot uniquely determine the 3D position of a target point because the elevation angle φ is missing. Therefore, each sonar feature point corresponds to a set of possible 3D points distributed along an elevation-dependent spatial curve.

1.4 Acoustic-optical projection relationship

To establish the geometric correspondence between acoustic measurements and optical observations, sonar feature points need to be transformed from the sonar coordinate system to the camera coordinate system and then projected onto the optical image plane. The rigid transformation from the sonar coordinate system to the camera coordinate system is defined by a rotation matrix $\boldsymbol{R}_{SC}$ and a translation vector $\boldsymbol{T}_{SC}$. Therefore, a point in the sonar coordinate system can be transformed into the camera coordinate system by

$$\boldsymbol{P}_C = \boldsymbol{R}_{SC}\boldsymbol{P}_S + \boldsymbol{T}_{SC} \tag{6}$$

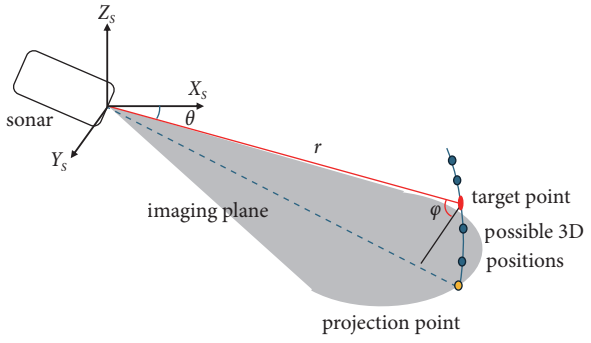


Figure 1 Sonar imaging model.

According to the camera projection model, the corresponding pixel coordinate (u, v) on the optical image plane can be obtained by

$$u = f_x \frac{x_C}{z_C} + c_x \tag{7}$$

$$v = f_y \frac{y_C}{z_C} + c_y \tag{8}$$

Combining the sonar-to-camera transformation and the camera projection model, a sonar measurement can be mapped onto the optical image plane. However, since the elevation angle φ of an imaging sonar measurement is not directly available, a sonar feature point does not correspond to a unique 3D point in space. Instead, varying φ within the vertical field of view of the sonar generates a set of possible 3D positions. After coordinate transformation and camera projection, these possible positions form a line-like or curve-like candidate set on the optical image plane rather than a single deterministic pixel[35].

Accurate estimation of $\boldsymbol{R}_{SC}$ and $\boldsymbol{T}_{SC}$ is essential for reliable acoustic-optical projection. Nevertheless, because of the heterogeneous imaging mechanisms of optical cameras and imaging sonar, residual calibration errors are difficult to completely avoid. These errors may cause the projected sonar features to deviate from their true positions in the optical image, thereby affecting acoustic-optical depth association and the final reconstructed point cloud. Therefore, the fusion process should consider not only sonar range constraints but also the refinement of the sonar-to-camera transformation parameters.

1.5 Problem statement

Based on the above system modeling, this work focuses on the problem of single-pose underwater 3D reconstruction using a binocular camera and an imaging sonar mounted on a UUV. Given a synchronized pair of binocular optical images and the corresponding sonar image, the objective is to reconstruct a 3D point cloud with improved coverage by effectively fusing optical and acoustic measurements.

These two sensing modalities have complementary but heterogeneous characteristics. The binocular camera can provide relatively dense visual observations and preserve local geometric details of the target. However, its depth estimation relies on stereo matching, which is easily affected by underwater light attenuation, scattering, and refraction. As a result, the optical depth map may contain mismatches, noise, and missing regions. In contrast, the imaging sonar provides more robust range and azimuth measurements under degraded underwater visibility, but it cannot directly provide the elevation angle of the target point. Therefore, a sonar feature does not correspond to a unique 3D position, which makes direct acoustic-optical matching difficult. Another challenge comes from the residual calibration error between the sonar and the camera, which may cause sonar features to be inaccurately projected onto the optical image, thereby affecting depth association and final point cloud reconstruction.

Therefore, the key problem addressed in this paper is how to use the robust range information from imaging sonar to constrain and correct the binocular depth estimation while preserving the local structural details provided by optical images. Meanwhile, the fusion process should reduce the influence of residual acoustic-optical calibration errors and improve the coverage and accuracy of the reconstructed underwater point cloud.

2 Acoustic-optical reconstruction framework

2.1 Overview of the reconstruction pipeline

This section presents the basic acoustic-optical reconstruction framework, which processes camera and sonar images and establishes geometric correspondences between the two heterogeneous sensing modalities. The overall process includes optical depth estimation, sonar feature extraction, sonar-to-camera projection, and acoustic-optical candidate matching.

The overall framework of the proposed acoustic-optical 3D reconstruction method is illustrated in Fig. 2. Given a pair of rectified binocular images and a temporally aligned sonar image, the framework combines optical observations with acoustic measurements to generate a fused underwater point cloud.

The framework consists of two parallel sensing branches. In the optical branch, stereo matching is performed on the binocular images to obtain an initial depth map, from which invalid regions and unreliable disparity measurements are removed before generating the optical point cloud. In the acoustic branch, significant echo responses are detected from the sonar image and filtered to obtain reliable sonar features containing range and azimuth measurements.

The two sensing modalities are associated through the sonar-to-camera projection relationship. Because the elevation angle of an imaging-sonar measurement is unobserved, each sonar feature generates multiple possible 3D positions and a corresponding

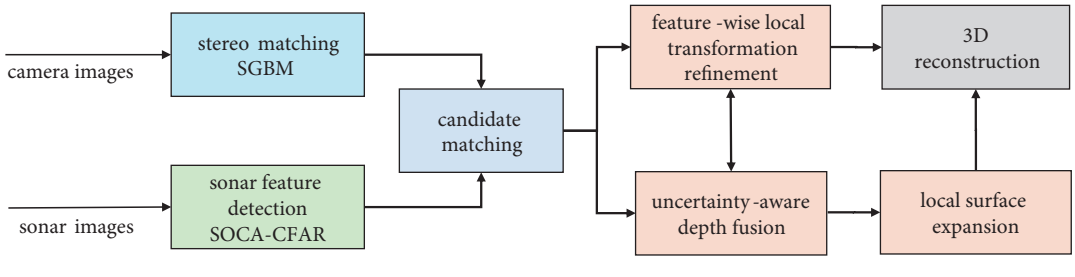


Figure 2 The acoustic-optical 3D reconstruction framework.

candidate region on the camera image plane. Candidate acoustic-optical correspondences are therefore constructed instead of enforcing direct one-to-one matching. Based on these correspondences, feature-wise local transformation refinement is performed to reduce residual projection errors. The optical and sonar depth estimates are subsequently fused according to their uncertainties, and the reliable depth corrections are propagated to neighboring geometrically consistent pixels through local surface expansion. Finally, invalid and duplicated samples are removed, and the retained samples are back-projected to generate the final fused point cloud. The individual modules are detailed in Sections 2.2–2.4 and Section 3.

2.2 Optical point cloud generation based on stereo matching

The binocular camera provides dense optical observations and initial depth information for underwater targets. In this work, stereo matching is used to estimate the disparity between the rectified left and right images. The semi-global block matching (SGBM) algorithm is adopted to obtain the disparity map because it can provide relatively dense matching results while maintaining computational efficiency[36].

According to the binocular camera model introduced in Section 1.2, the disparity map is converted into an initial optical depth map, and each valid pixel can be back-projected into the camera coordinate system to generate the optical point cloud. Let $d(u,v)$ denote the disparity value of pixel (u,v) obtained by SGBM. However, directly using the stereo matching result for underwater 3D reconstruction may introduce many unreliable points. To reduce the influence of invalid optical regions, a valid-region extraction step is applied before generating the optical point cloud. The optical image is first processed by Gaussian filtering, threshold segmentation, and morphological closing operations. A binary valid-region map is then obtained as

$$M(u,v)=\begin{cases}1 & f_I(I(u,v))>\tau\\ 0 & \text{otherwise}\end{cases} \tag{9}$$

where $I(u,v)$ denotes the optical image intensity at pixel (u,v), $f_I(\cdot)$ represents the image preprocessing operation, and τ is the segmentation threshold. Only pixels located in the valid optical region and with valid disparity values are used for point cloud generation. Therefore, the initial optical point cloud is defined as

$$P_o=\{\boldsymbol{P}_C(u,v)\,|\,M(u,v)=1, d(u,v)>0\} \tag{10}$$

where $\boldsymbol{P}_C(u,v)$ is the 3D point reconstructed from pixel (u,v) according to the binocular camera model, and $d(u,v)>0$ indicates that the pixel has a valid disparity value. This filtering process suppresses part of the mismatched points and provides a cleaner optical point cloud for subsequent acoustic-optical association. Nevertheless, the depth accuracy of the optical point cloud is still limited by the quality of stereo matching, and acoustic range information is further required to improve the reliability of reconstruction.

2.3 Sonar feature detection and filtering

The smallest-of cell-averaging constant false-alarm rate (SOCA-CFAR) algorithm is adopted for sonar feature detection[37]. The CFAR-based detector estimates the local background level around each sonar cell and adaptively determines whether the current cell corresponds to a significant acoustic response. Compared with fixed-threshold segmentation, this adaptive detection strategy is more suitable for sonar images with nonuniform intensity distribution.

Let $I_S(u_s,v_s)$ denote the intensity value of a sonar image point. For each candidate cell, the surrounding cells are divided into guard cells and reference cells. The guard cells prevent the target response from being included in the noise estimation, while the reference cells are used to estimate the local background level. A sonar point is regarded as a feature point if its echo intensity is sufficiently larger than the local adaptive threshold.

After the initial detection, false feature points may still exist due to noise and acoustic artifacts. Therefore, an additional filtering step is performed according to the imaging characteristics of the sonar. Since objects closer to the sonar can occlude objects behind them along the same acoustic beam, only the nearest valid feature point is retained for each beam direction. This operation removes part of the false detections and preserves the acoustic feature points that are more likely to correspond to the visible surface of the target. The final detected sonar feature set is denoted as

$$F_S=\{(u_s^i,v_s^i,r_i,\theta_i)\,|\,i=1,\ 2,\ \cdots,N_S\} \tag{11}$$

where N_S is the number of retained sonar feature points, r_i and θ_i are the range and azimuth angle of the i-th sonar feature point calculated from the sonar image model in Section 1.3. These sonar features provide reliable acoustic range cues for subsequent acoustic-optical candidate matching.

2.4 Sonar-to-camera projection and candidate matching

After extracting sonar feature points, the next step is to establish their geometric relationship with the optical image. According to the acoustic-optical projection relationship introduced in Section 1.4, the possible 3D positions of a sonar feature can be transformed into the camera coordinate system and projected onto the optical image plane. By varying φ within the vertical field of view of the imaging sonar, the i-th sonar feature generates a candidate projection set on the optical image plane:

$$C_i=\{(u^i(\varphi),v^i(\varphi))\,|\,\varphi\in[\varphi_{\min},\varphi_{\max}]\} \tag{12}$$

where $[\varphi_{\min},\varphi_{\max}]$ denotes the vertical field-of-view range of the imaging sonar, and $(u^i(\varphi),v^i(\varphi))$ is the projected pixel coordinate corresponding to the elevation angle φ. This set is line-like or curve-like rather than a single deterministic pixel because of the missing elevation information. The acoustic-optical projection geometry and the formation of the candidate projection set are illustrated in Fig. 3.

To construct candidate acoustic-optical correspondences, the projected sonar candidate set is associated with the valid optical depth map obtained from stereo matching. Optical pixels located near the projected candidate set and satisfying basic validity constraints are selected as candidate matching points. Specifically, the pixel-to-projection distance for the i-th sonar feature is defined as

$$\rho_i(u,v)=\min_{(u',v')\in C_i}\|(u,v)-(u',v')\|_2 \tag{13}$$

Then, the candidate matching set corresponding to the i-th sonar feature is defined as

$$M_i=\{(u,v,z_c(u,v))\,|\,\rho_i(u,v)<\delta, M(u,v)=1, d(u,v)>0\} \tag{14}$$

where $z_c(u,v)$ is the optical depth estimated from binocular stereo matching, and δ is the pixel-distance threshold. The construction of M_i avoids enforcing direct one-to-one matching between sonar features and optical pixels. Instead, it preserves multiple possible acoustic-optical correspondences for each sonar feature. These

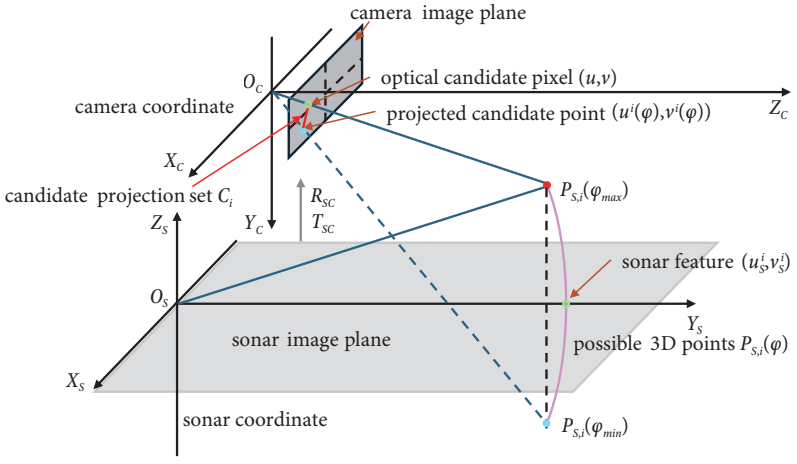


Figure 3 Acoustic-optical projection geometry.

candidate correspondences contain the sonar feature measurements, their projected candidate pixels, and the corresponding optical depth values, which are then used in the following reconstruction stages to obtain the final acoustic-optical 3D point cloud.

3 Uncertainty-aware fusion and point cloud refinement

Based on the acoustic-optical candidate correspondences constructed in Section 2, this section introduces the uncertainty-aware fusion and point cloud refinement method. The basic idea is to use the robust range information from the imaging sonar to constrain the optical depth estimated by the binocular camera, while preserving the local geometric details provided by stereo vision. In addition, the residual extrinsic calibration error between the sonar and the camera is reduced through depth-consistency optimization. Finally, a local surface expansion strategy is introduced to improve point cloud coverage while suppressing invalid and redundant points.

3.1 Feature-wise local transformation refinement

Accurate acoustic-optical projection depends on the transformation parameters between the imaging sonar and the binocular camera. The general sonar-to-camera transformation is represented by $\boldsymbol{R}_{SC}$ and $\boldsymbol{T}_{SC}$, as introduced in Section 1.4. In practical implementation, the local refinement is initialized using the transformation obtained through offline calibration, denoted by $\boldsymbol{R}_{SC}^{0}$ and $\boldsymbol{T}_{SC}^{0}$, where the superscript 0 indicates the initial calibrated values. However, residual calibration errors are difficult to avoid in underwater experiments. Such errors may cause systematic deviations between the projected sonar features and their corresponding optical observations, thereby degrading acoustic-optical association and subsequent depth fusion.

To reduce the influence of residual transformation errors, a feature-wise local transformation refinement strategy is introduced. For each sonar feature, the possible elevation angles are discretely sampled within the vertical field of view of the imaging sonar. Unlike fixing the acoustic-optical correspondences before optimization, the projected pixels and valid correspondences are updated according to the current transformation parameters during local refinement.

Let $\Phi = \{\varphi_1, \varphi_2, \cdots, \varphi_{N_\varphi}\}$ denote the sampled elevation-angle set. For the m-th elevation sample of the i-th sonar feature, its projected pixel coordinate under a candidate transformation $\boldsymbol{R}_{SC}$ and $\boldsymbol{T}_{SC}$ is calculated as

$$(u_i^m, v_i^m) = \mathrm{round}[\pi(\boldsymbol{R}_{SC}\boldsymbol{P}_{S,i}(\varphi_m) + \boldsymbol{T}_{SC})] \tag{15}$$

where $\pi(\cdot)$ denotes the camera projection operation defined by Eqs. (7) and (8), and $\mathrm{round}[\cdot]$ denotes component-wise rounding of the continuous projected coordinates to the nearest integer pixel coordinates. Therefore, each sampled sonar point is directly associated with its projected optical pixel. The corresponding depth component in the camera coordinate system is defined as

$$z_{s,i}^m(\boldsymbol{R}_{SC}, \boldsymbol{T}_{SC}) = [0,0,1](\boldsymbol{R}_{SC}\boldsymbol{P}_{S,i}(\varphi_m) + \boldsymbol{T}_{SC}) \tag{16}$$

Let k_g denote the optical-depth scale used in the depth-consistency gating stage, W and H denote the width and height of the optical image, respectively. The valid correspondence index set under the current transformation is defined as

$$\Gamma_i(\boldsymbol{R}_{SC}, \boldsymbol{T}_{SC}) = \{m | 0 \le u_i^m < W, 0 \le v_i^m < H, d(u_i^m, v_i^m) > 0, e_i(u_i^m, v_i^m, \boldsymbol{R}_{SC}, \boldsymbol{T}_{SC}) < \tau_z\} \tag{17}$$

where

$$e_i(u_i^m, v_i^m, \boldsymbol{R}_{SC}, \boldsymbol{T}_{SC}) = \left| k_g \cdot z_c(u_i^m, v_i^m) - z_{s,i}^m(\boldsymbol{R}_{SC}, \boldsymbol{T}_{SC}) \right|$$

Here, τ_z is the depth-consistency threshold. Since the projected pixels depend on the current transformation parameters, the correspondence set $\Gamma_i(\boldsymbol{R}_{SC}, \boldsymbol{T}_{SC})$ is dynamically updated during optimization.

Let k_{RT} denote the optical-point scale used for local transformation refinement. The refined transformation associated with the i-th sonar feature is obtained by minimizing the three-dimensional geometric discrepancy between the sonar points and their corresponding optical points:

$$\boldsymbol{R}_{SC,i}^*, \boldsymbol{T}_{SC,i}^* = \arg\min_{\boldsymbol{R}_{SC}, \boldsymbol{T}_{SC}} \sum_{\{m \in \Gamma_i(\boldsymbol{R}_{SC}, \boldsymbol{T}_{SC})\}} \| \boldsymbol{R}_{SC}\boldsymbol{P}_{S,i}(\varphi_m) + \boldsymbol{T}_{SC} - k_{RT}\boldsymbol{P}_C(u_i^m, v_i^m) \|_2^2 \tag{18}$$

where $\boldsymbol{P}_C(u_i^m, v_i^m)$ denotes the optical 3D point reconstructed from the binocular depth map. The optimization is initialized using $\boldsymbol{R}_{SC}^0$ and $\boldsymbol{T}_{SC}^0$. The local refinement is performed only when a sufficient number of valid correspondences is available; otherwise, the initial transformation is retained.

When the refinement is performed, the optimized transformation is accepted only if it reduces the mean acoustic-optical depth residual while preserving a sufficient number of valid correspondences relative to the initial transformation. The transformation used for subsequent depth fusion is therefore selected as

$$\left(\boldsymbol{R}_{SC,i}^{sel}, \boldsymbol{T}_{SC,i}^{sel}\right) = \begin{cases} \left(\boldsymbol{R}_{SC,i}^{*}, \boldsymbol{T}_{SC,i}^{*}\right) & \bar{e}_i^{*} < \gamma_e \bar{e}_i^{0} \wedge n_i^{*} \geq \gamma_n n_i^{0} \\ \left(\boldsymbol{R}_{SC}^{0}, \boldsymbol{T}_{SC}^{0}\right) & \text{otherwise} \end{cases} \tag{19}$$

where $\bar{e}_i^0$ and $\bar{e}_i^*$ denote the average depth residuals obtained using the initial and refined transformations, respectively; n_i^0 and n_i^* denote the corresponding numbers of valid correspondences; and γ_e and γ_n are the residual-reduction and correspondence-retention thresholds, respectively.

Finally, the reliable correspondence set used in the subsequent depth-fusion stage is defined as

$$A_i = \{(u_i^m, v_i^m) \mid m \in \Gamma_i\left(\boldsymbol{R}_{SC,i}^{sel}, \boldsymbol{T}_{SC,i}^{sel}\right)\} \tag{20}$$

which is generated using the selected transformation parameters. The dynamic projection and correspondence-update process is illustrated in Fig. 4. For each detected sonar feature, the unknown elevation angle is discretely sampled within the vertical field of view, producing a set of possible 3D candidate points in the sonar coordinate system. Under the current sonar-to-camera transformation, these candidate points are transformed into the camera coordinate system and projected onto the camera image plane. Their projected positions form a line-like or curve-like candidate region, indicated by the dotted curve in the upper part of Fig. 4.

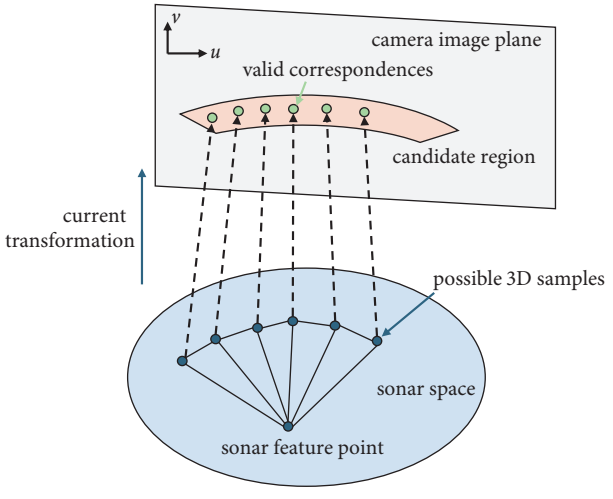


Figure 4 Dynamic projection and correspondence update.

The projected candidate points are then associated with the binocular depth map. Only the projected pixels that fall within the valid optical region, possess valid disparity measurements, and satisfy the acoustic-optical depth-consistency constraint are retained as valid correspondences. During local transformation refinement, the transformation parameters are updated by minimizing the 3D geometric discrepancy between the retained optical and sonar points. Once the transformation parameters change, the sonar candidate points are reprojected, causing the candidate region and the valid correspondence set to be updated accordingly. This projection, correspondence selection, and transformation update process is repeated during optimization until convergence. This strategy performs depth-consistency-constrained three-dimensional geometric refinement while preventing unreliable local optimization results from degrading the subsequent acoustic-optical fusion.

3.2 Uncertainty modeling of acoustic and optical depth

The binocular camera and the imaging sonar provide depth information with different uncertainty characteristics. The binocular camera can generate relatively dense depth information through stereo matching, but its accuracy is highly dependent on the reliability of disparity estimation. In contrast, the imaging sonar provides more robust range measurements, whose depth uncertainty is affected by the range resolution. Therefore, the uncertainty of both optical and acoustic measurements should be modeled before depth fusion.

For binocular vision, the depth value is calculated from the disparity according to the stereo model introduced in Section 1.2. Let σ_d denote the standard deviation of disparity estimation, and let $\sigma_c^2(u,v)$ denote the variance of the optical depth at pixel (u,v). According to first-order error propagation, the camera depth variance can be written as

$$\sigma_c^2(u,v) = \left(\partial z_c(u,v)/\partial d(u,v)\right)^2 \sigma_d^2 \tag{21}$$

Since the binocular depth is calculated by Eq. (2), the camera depth variance can be further expressed as

$$\sigma_c^2(u,v) = \left(Bf_x/d^2(u,v)\right)^2 \sigma_d^2 \tag{22}$$

This equation indicates that the uncertainty of binocular depth increases rapidly when the disparity becomes small. Therefore, optical depth estimation tends to be less reliable in long-range regions or areas with inaccurate stereo matching. For imaging sonar, the range uncertainty is mainly related to the range resolution of the sonar. Let Δr denote the sonar range resolution, and let $\sigma_{r,min}$ denote the minimum allowable range standard deviation. The basic sonar range standard deviation is defined as

$$\sigma_r = \max(\Delta r, \sigma_{r,\min}) \tag{23}$$

The sonar measurement directly provides the target range, whereas the depth used in acoustic-optical fusion is the depth component of the sonar candidate point expressed in the camera coordinate system. Assuming that the uncertainty of the transformed sonar depth is dominated by the sonar range measurement uncertainty, the depth variance associated with the i-th sonar feature is approximated as

$$\sigma_{s,i}^2 = \sigma_r^2 \tag{24}$$

where $\sigma_{s,i}^2$ denotes the variance of the sonar-constrained depth associated with the i-th sonar feature. The minimum uncertainty bound $\sigma_{r,\min}$ prevents the sonar measurement from being assigned an unrealistically small variance. The optical and sonar depth variances obtained above are subsequently used to adaptively determine their contributions during uncertainty-aware depth fusion.

3.3 Uncertainty-aware acoustic-optical depth fusion

After obtaining the candidate acoustic-optical correspondences and the depth uncertainty of each modality, the optical depth and sonar-constrained depth are fused using variance-weighted estimation. For a reliable candidate pixel $(u,v) \in A_i$, the optical depth is first corrected using the depth scale factor k:

$$\hat{z}_c(u,v) = k \cdot z_c(u,v) \tag{25}$$

where $\hat{z}_c(u,v)$ is the scale-corrected optical depth. After feature-wise local transformation refinement, $z_{s,i}^C(u,v)$ denotes the depth component of the sonar candidate point associated with pixel (u,v), expressed in the camera coordinate system using the refined transformation parameters $\boldsymbol{R}_{SC,i}^{sel}$ and $\boldsymbol{T}_{SC,i}^{sel}$. To preserve the

relative depth variation within the reliable candidate region corresponding to the i-th sonar feature, the reference optical depth is defined as the mean optical depth over A_i:

$$\bar{z}_{c,i} = \frac{1}{|A_i|} \sum_{(u,v)\in A_i} z_c(u,v) \tag{26}$$

The sonar-constrained depth is then constructed as:

$$\hat{z}_{s,i}(u,v) = k \cdot z^C_{s,i}(u,v) \cdot z_c(u,v)/\bar{z}_{c,i} \tag{27}$$

where the transformed sonar depth $z^C_{s,i}(u,v)$ provides an absolute acoustic depth reference, whereas the ratio $z_c(u,v)/\bar{z}_{c,i}$ preserves the relative depth variation provided by binocular vision. Given the optical depth estimate $\hat{z}_c(u,v)$ and the sonar-constrained depth estimate $\hat{z}_{s,i}(u,v)$, the fused depth $z_{f,i}(u,v)$ is obtained by variance-weighted fusion:

$$z_{f,i}(u,v) = \left[\frac{\hat{z}_c(u,v)}{\sigma_c^2(u,v)} + \frac{\hat{z}_{s,i}(u,v)}{\sigma_{s,i}^2}\right] / \left[\frac{1}{\sigma_c^2(u,v)} + \frac{1}{\sigma_{s,i}^2}\right] = \frac{\sigma_{s,i}^2}{\sigma_c^2(u,v)+\sigma_{s,i}^2}\hat{z}_c(u,v) + \frac{\sigma_c^2(u,v)}{\sigma_c^2(u,v)+\sigma_{s,i}^2}\hat{z}_{s,i}(u,v) \tag{28}$$

This formulation assigns a larger weight to the modality with smaller depth uncertainty. When the optical depth uncertainty $\sigma_c^2(u,v)$ is large, the fused depth relies more on the sonar-constrained depth. Conversely, when the sonar uncertainty $\sigma_{s,i}^2$ is large, the fused depth relies more on the optical depth. In this way, the proposed fusion strategy adaptively balances the contributions of binocular vision and imaging sonar, improving depth reliability while preserving local structural details.

3.4 Local surface expansion

If reconstruction is performed only at the reliable candidate pixels, the resulting point cloud may remain sparse. Since neighboring optical pixels on the same target surface generally exhibit locally continuous depth variations, a local surface expansion strategy is introduced to propagate the acoustic-optical depth correction from reliable candidate pixels to nearby geometrically consistent pixels.

For convenience, the elements of A_i are indexed by (u_{ij}, v_{ij}), where $j = 1,2,\cdots,n_i$ and $n_i = |A_i|$. For each reliable reference pixel $(u_{ij}, v_{ij}) \in A_i$, a local image neighborhood is defined as

$$N_{ij} = \{(u,v) \mid |u-u_{ij}| \le r_e, |v-v_{ij}| \le r_e\} \tag{29}$$

where r_e is the local expansion radius. To prevent the expansion from crossing target boundaries or propagating to geometrically inconsistent regions, a local depth-continuity constraint is imposed. The local expansion set corresponding to the reference pixel (u_{ij}, v_{ij}) is defined as

$$Q_{ij} = \{(u,v) \in N_{ij} \setminus \{(u_{ij}, v_{ij})\} \mid M(u,v) = 1, d(u,v) > 0, |z_c(u,v) - z_c(u_{ij}, v_{ij})| \le \tau_e\} \tag{30}$$

where τ_e is the local depth-continuity threshold. This constraint ensures that only valid optical pixels whose depths are similar to that of the reliable reference pixel are included in the expansion set. In practical implementation, pixels within N_{ij} are sampled using a stride s_e to reduce redundant reconstruction points and improve computational efficiency. The local depth-scale factor associated with the reference pixel (u_{ij}, v_{ij}) is defined as

$$\alpha_{ij} = z_{f,i}(u_{ij}, v_{ij})/z_c(u_{ij}, v_{ij}) \tag{31}$$

where $z_{f,i}(u_{ij}, v_{ij})$ is the uncertainty-aware fused depth obtained from Eq. (28). The reconstructed depth of an expanded pixel is then calculated as

$$z_{e,ij}(u,v) = \alpha_{ij} \cdot z_c(u,v), \ (u,v) \in Q_{ij} \tag{32}$$

where $z_{e,ij}(u,v)$ denotes the locally propagated depth derived from the reference pixel (u_{ij}, v_{ij}).

Therefore, each reliable candidate pixel retains its uncertainty-aware fused depth $z_{f,i}$, whereas its neighboring pixels use the locally propagated depth $z_{e,ij}$. In this manner, the acoustic-optical depth correction is extended from sparse reliable candidate pixels to nearby geometrically consistent optical pixels while preserving the relative surface variation contained in the binocular depth map. The reliable fused samples and the locally expanded samples jointly provide the reconstruction samples used for subsequent point cloud refinement.

3.5 Point cloud refinement and algorithm summary

The preceding uncertainty-aware fusion and local surface expansion stages generate two types of reconstruction samples: reliable candidate pixels with fused depths and locally expanded pixels with propagated depths. These samples are combined and further refined to generate the final acoustic-optical point cloud. The refinement process includes image-boundary checking, optical-validity checking, disparity-validity checking, depth-range filtering, duplicate removal, and 3D back-projection.

Since the two types of pixels use different depth estimates, each reconstruction sample is represented by a triplet (u,v,z), which contains its image coordinate and corresponding reconstructed depth. The complete reconstruction-sample set is defined as

$$\Omega = \bigcup_{i=1}^{N_S} \left[\{(u,v,z_{f,i}(u,v)) \mid (u,v) \in A_i\} \bigcup \bigcup_{j=1}^{n_i} \{(u,v,z_{e,ij}(u,v)) \mid (u,v) \in Q_{ij}\}\right] \tag{33}$$

The samples generated from A_i retain their uncertainty-aware fused depths, whereas those generated from Q_{ij} retain their locally propagated depths. Representing Ω as a reconstruction-sample set rather than a pure pixel set prevents locally expanded pixels from being incorrectly assigned the fused candidate depth $z_{f,i}$.

Next, invalid pixels are removed according to the image boundary, disparity validity, and optical valid-region constraint. The valid reconstruction-sample set is defined as

$$\Lambda = \{(u,v,z) \in \Omega \mid 0 \le u < W, \ 0 \le v < H, M(u,v) = 1, d(u,v) > 0, z_{\min} < z < z_{\max}\} \tag{34}$$

These constraints remove samples with invalid disparities, invalid region labels, unreliable depth estimates, or out-of-range depth values. For each valid reconstruction sample $(u,v,z) \in \Lambda$, the corresponding 3D point in the camera coordinate system is obtained by

$$\boldsymbol{P}_r(u,v,z) = \left[\frac{z(u-c_x)}{f_x}, \frac{z(v-c_y)}{f_y}, z\right]^{\mathrm{T}} \tag{35}$$

Since the same reconstruction sample may be generated through multiple candidate-matching or local-expansion paths, duplicate samples are removed before insertion into the final point cloud. Accordingly, the final fused point cloud is defined as

$$P_f = \mathrm{Unique}(\{\boldsymbol{P}_r(u,v,z) \mid (u,v,z) \in \Lambda\}) \tag{36}$$

where $\mathrm{Unique}(\cdot)$ denotes the duplicate-removal operation, which

retains only one representative when multiple reconstruction samples produce the same or numerically equivalent 3D point. Together with the validity filtering defined in Eq. (34), this operation removes invalid, out-of-range, and duplicated reconstruction results while preserving the respective depth estimates of reliable fused samples and locally expanded samples.

In summary, the complete procedure of the proposed uncertainty-aware acoustic-optical 3D reconstruction method is summarized in Algorithm 1.

Algorithm 1 Uncertainty-aware acoustic-optical 3D reconstruction

Input: Rectified binocular images, sonar image, camera intrinsic parameters, sonar parameters, and initial sonar-to-camera extrinsic parameters

Output: Fused acoustic-optical 3D point cloud P_f

1: Compute the disparity map $d(u,v)$ using SGBM, and obtain the optical depth $z_c(u,v)$ according to Eq. (2), and extract the valid optical-region map $M(u,v)$ according to Eq. (9).
2: Detect sonar features using SOCA-CFAR and construct the sonar feature set F_S according to Eq. (11).
3: For each sonar feature i, generate its possible 3D positions by varying the elevation angle φ, and project them onto the optical image plane using Eqs. (5)–(8).
4: Construct the candidate projection set C_i according to Eq. (12) and the candidate matching set M_i according to Eq. (14).
5: For each sonar feature, dynamically update the projected acoustic-optical correspondences and refine the local sonar-to-camera transformation by minimizing the three-dimensional geometric discrepancy according to Eqs. (15)–(18). Select the refined or initial transformation according to Eq. (19), and construct the final reliable pixel set A_i according to Eq. (20).
6: Estimate the optical depth variance $\sigma_c^2(u,v)$ using Eqs. (21) and (22), and estimate the sonar depth variance $\sigma_{s,i}^2$ using Eqs. (23) and (24).
7: For each reliable candidate pixel $(u,v) \in A_i$, compute the scale-corrected optical depth $\hat{z}_c(u,v)$ according to Eq. (25), construct the sonar-constrained depth $\hat{z}_{s,i}(u,v)$ according to Eq. (27), and obtain the uncertainty-aware fused depth $z_{f,i}(u,v)$ according to Eq. (28).
8: For each reliable reference pixel $(u_{ij}, v_{ij}) \in A_i$, construct the local neighborhood N_{ij} and expansion set Q_{ij}, estimate the local depth-scale factor α_{ij}, and compute the propagated depth $z_{e,ij}(u,v)$ according to Eqs. (29)–(32).
9: Combine the reliable fused samples and locally expanded samples to construct Ω, and retain the valid reconstruction samples Λ according to Eqs. (33) and (34).
10: Back-project each sample in Λ into the camera coordinate system, remove duplicated or numerically equivalent 3D points, and generate the final point cloud $\boldsymbol{P}_f$ according to Eqs. (35)-(36).

4 Experiments

4.1 Experimental platform

The underwater data acquisition platform used in the experiment is shown in Fig. 5. The platform consists of a binocular camera and a Blueprint Oculus M750d imaging sonar, which provides 512 acoustic beams per ping. The horizontal field of view angle is 130°, and the vertical field of view angle is 20°. During data acquisition, sonar range settings of 1 m and 2 m were mainly used, with a range resolution of 2.5 mm. Each camera of the binocular camera provided images with a resolution of 640×480 pixels. The sonar was mounted below the camera with a downward pitch angle of 15°, while the binocular camera was maintained approximately horizontal.

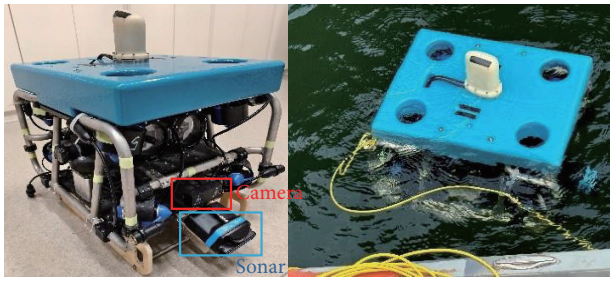


Figure 5 Underwater data acquisition platform equipped with an imaging sonar and a binocular camera.

The experimental data were collected in the Xin'anjiang Reservoir. The binocular camera and imaging sonar recorded data at sampling frequencies of 30 Hz and 15 Hz, respectively. The two sensor streams were temporally aligned before acoustic-optical reconstruction. Three representative scenes were selected for evaluation: a ship propeller blade, an underwater vehicle, and an underwater rock wall. These scenes exhibit different geometric structures, optical textures, acoustic responses, and occlusion conditions. Representative optical images and sonar images of three experimental scenes are shown in Fig. 6. The ship propeller-blade scene contains smooth and partially reflective surfaces, which may result in weak optical texture and unreliable stereo matching. The underwater-vehicle scene contains multiple structural components and discontinuous boundaries, making it sensitive to acoustic-optical projection errors. The rock-wall scene contains a relatively large and irregular surface, and is therefore suitable for evaluating reconstruction coverage and local surface continuity.

The same reconstruction parameters were used for all experi-

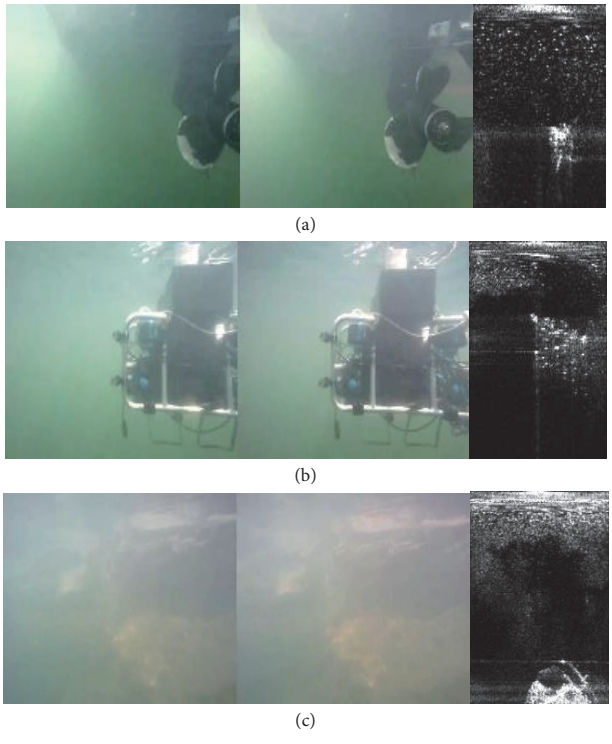


Figure 6 Representative optical and sonar images ((a) Ship propeller blade scene; (b) Underwater vehicle scene; (c) Underwater rock wall scene).

mental scenes unless otherwise specified. The principal parameters included the depth scale factor k, the depth-consistency threshold τ_z, the local expansion radius r_e, the local depth-continuity threshold τ_e, and the valid depth range $[z_{\min}, z_{\max}]$.

4.2 Comparison methods and evaluation metrics

As shown in Table 1, five underwater 3D reconstruction methods were evaluated.

Stereo-only: This method uses SGBM to estimate binocular disparity and directly reconstructs the optical point cloud without using sonar information[38].

Mono-sonar: Sonar features are projected onto the optical image, and sonar-derived depth is assigned to the corresponding image points. This method uses acoustic range information but does not exploit binocular depth for local surface reconstruction.

Stereo-sonar: Binocular depth and sonar depth are fused using the initially calibrated sonar-to-camera transformation. Feature-wise local transformation refinement, uncertainty-aware weighting, and local surface expansion are not applied.

RT-stereo-sonar: This method extends Stereo-sonar by introducing feature-wise local transformation refinement. The refined transformation is retained only when it reduces the acoustic-optical depth residual while preserving a sufficient number of valid correspondences. The same basic depth-fusion strategy as Stereo-sonar is used.

Proposed: The complete method incorporates feature-wise local transformation refinement, optical and sonar uncertainty modeling, variance-weighted depth fusion, local surface expansion, valid-depth filtering, and duplicate-sample removal.

The ground truth in the experiment is based on the manually measured distance and size of the target during the data collection process. The reconstruction results were evaluated using coverage rate (CR), absolute error (AE), mean absolute error within 2 cm (MAE2), and the percentage of reconstructed points with an error below 2 cm (PE2). The coverage rate measures the proportion of the reference target surface represented by the reconstructed point cloud. The reference surface was discretized using a spatial resolution of 1 cm. A reference cell was regarded as covered when it contained, or was sufficiently close to, at least one reconstructed point. The coverage rate is calculated as the ratio of the number of covered reference cells to the total number of reference cells.

For a reconstructed point $\boldsymbol{P}$, the absolute error is defined as the minimum Euclidean distance from $\boldsymbol{P}$ to the reference geometry. MAE2 is the average absolute error of reconstructed points whose errors are below 2 cm. PE2 is the percentage of reconstructed points satisfying the 2 cm error threshold. CR mainly reflects reconstruction completeness, whereas MAE2 and PE2 characterize local reconstruction accuracy. Because local surface expansion naturally increases the number of reconstructed points, point count is reported only as an auxiliary statistic and is not regarded as an independent measure of reconstruction quality.

4.3 Reconstruction results

The qualitative reconstruction results of the five methods for the three scenes are shown in Fig. 7. The five columns correspond to Stereo-only, Mono-sonar, Stereo-sonar, RT-stereo-sonar, and Proposed, respectively. The light-gray points represent the reference geometry, while the colored points represent the reconstructed point clouds. For each scene, all methods are visualized using the same viewpoint and coordinate range.

As shown in Fig. 7, Stereo-only produces a relatively large number of optical points, but the results contain substantial noise and structural distortion, particularly in weakly textured or reflective regions. In the propeller-blade scene, many optical points deviate from the blade surface. In the underwater-vehicle scene, stereo mismatches generate dispersed background points and incomplete structural boundaries. In the rock-wall scene, the optical reconstruction contains severe missing regions.

Mono-sonar provides more stable absolute distance information, but the reconstructed surfaces are sparse and lack fine geometric details. Stereo-sonar improves the overall structural consistency by combining binocular depth with sonar range information. However, residual errors in the initially calibrated transformation still cause projection deviations and limit the quality of acoustic-optical association.

RT-stereo-sonar further improves the alignment between projected sonar samples and optical observations. This reduces part of the local structural distortion, especially in the underwater-vehicle scene, where multiple boundaries and components make the reconstruction sensitive to transformation errors. Nevertheless, because uncertainty-aware weighting and local surface expansion are not used, the reconstructed coverage remains limited.

The proposed method produces the most complete reconstruction in all three scenes. The sonar range constraints suppress unreliable optical depths, while variance-weighted fusion adaptively balances the acoustic and optical observations. Local surface expansion further propagates reliable fused depths to neighboring geometrically consistent pixels, substantially improving surface coverage.

The quantitative reconstruction results for the three scenes are summarized in Table 2. For the propeller-blade scene, the proposed method achieves a coverage rate of 58.09%, compared with 46.69% for Stereo-sonar and 43.75% for RT-stereo-sonar. Its PE2 reaches 61.96%, which is also higher than those of Stereo-sonar and RT-stereo-sonar. The corresponding MAE2 is reduced to 0.65 cm. For the underwater-vehicle scene, the proposed method achieves the highest coverage rate of 46.37%. RT-stereo-sonar provides a slightly higher PE2 of 82.55%, whereas the proposed method obtains 82.08%. This small decrease in PE2 is accompanied by a substantial increase in surface coverage, indicating a favorable balance between reconstruction completeness and local accuracy. For the rock-wall scene, the proposed method increases the coverage rate to 51.21%, whereas Stereo-sonar and RT-stereo-sonar achieve only 20.64% and 23.06%, respectively. The propo-

Table 1 Method component table

Method	Stereo depth	Sonar input	Local refinement	Uncertainty fusion	Surface expansion
Stereo-only	√				
Mono-sonar		√			
Stereo-sonar	√	√			
RT-stereo-sonar	√	√	√		
Proposed	√	√	√	√	√

sed method also obtains an MAE2 of 0.93 cm and a PE2 of 42.63%. These results demonstrate that local surface expansion is particularly effective for large and locally continuous target surfaces.

The coverage-accuracy relationships are further illustrated in Fig. 8. In general, the proposed method is located closer to the upper-right region of the plots, indicating that it achieves a better balance between CR and PE2. Although RT-stereo-sonar attains a

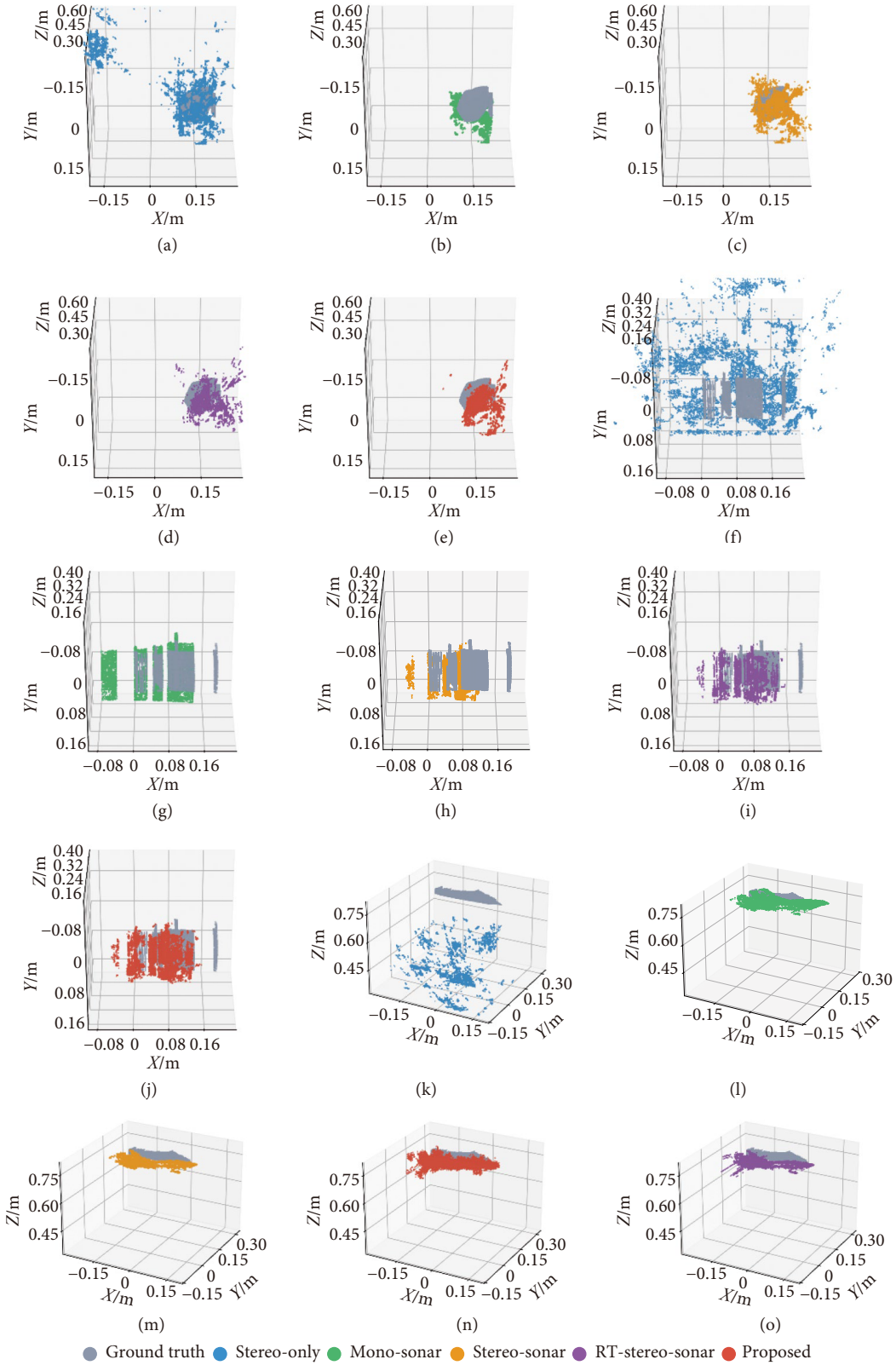


Figure 7 Qualitative comparison of the reconstructed point clouds ((a)-(e) Ship propeller-blade scene reconstructed by Stereo-only, Mono-sonar, Stereo-sonar, RT-stereo-sonar, and Proposed, respectively; (f)-(j) Underwater-vehicle scene reconstructed by the five methods in the same order; (k)-(o) Underwater rock-wall scene reconstructed by the five methods in the same order).

slightly higher PE2 in the underwater-vehicle scene, its coverage remains lower than that of the proposed method.

4.4 Effect of feature-wise local transformation refinement

A representative result of feature-wise local transformation refinement is shown in Fig. 9. The grayscale background represents the optical image, and the cyan contour denotes the extracted target edge. In Fig. 9 (a), the orange points are the initial projection samples obtained by sampling the unknown elevation angle of a detected sonar feature and projecting the resulting 3D candidates onto the optical image plane using the initial transformation. Because these candidates share the same sonar range and azimuth but have different elevation angles, their projections form an approximately line-like distribution. The yellow points denote the initial correspondences that satisfy the optical-validity and acoustic-optical depth-consistency constraints.

Figure 9 (b) compares the initial and refined results. The blue points represent the refined projection samples, and the green points denote the corresponding valid correspondences. After refinement, the projected samples move closer to the optical target edge, and a larger portion of them is retained as valid correspondences, indicating improved geometric consistency between

Table 2 Comparison of 3D reconstruction results

Method	Points			Coverage rate/%			MAE2/cm			PE2/%		
	Scene1	Scene2	Scene3	Scene1	Scene2	Scene3	Scene1	Scene2	Scene3	Scene1	Scene2	Scene3
Stereo-only	42,303	78,123	23,661	12.50	0.00	0.00	1.29	1.66	N/A	7.48	0.36	0.00
Mono-sonar	48,944	136,531	142,652	11.76	7.25	26.27	1.22	**0.61**	1.12	40.94	15.10	33.94
Stereo-sonar	57,998	78,809	68,379	46.69	28.50	20.64	0.87	0.86	0.95	33.29	78.13	42.18
RT-stereo-sonar	32,128	83,301	69,408	43.75	34.20	23.06	0.80	0.96	0.93	49.37	**82.55**	39.82
Proposed	**212,310**	**1,062,641**	**574,233**	**58.09**	**46.37**	**51.21**	**0.65**	0.96	**0.93**	**61.96**	82.08	**42.63**

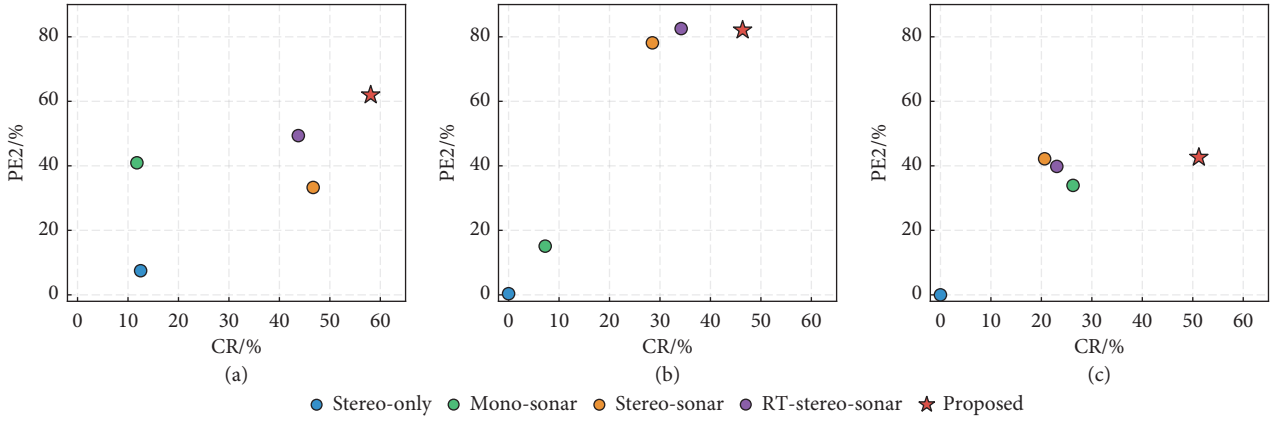


Figure 8 Coverage-accuracy comparison of the five reconstruction methods ((a) Ship propeller blade scene; (b) Underwater vehicle scene; (c) Underwater rock wall scene).

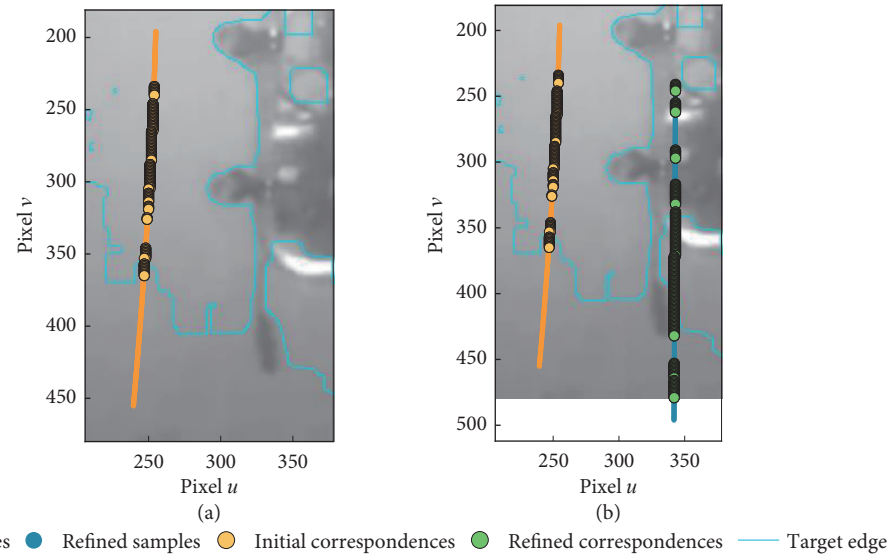


Figure 9 Representative result of feature-wise local transformation refinement ((a) Projection using the initial transformation; (b) Comparison of the initial and refined projection results).

the sonar measurements and the optical observations. For the illustrated sonar feature, the mean acoustic-optical depth residual decreases from 0.2628 m to 0.1233 m. These qualitative and quantitative results demonstrate that the proposed refinement effectively reduces the influence of residual calibration errors.

4.5 Effect of local surface expansion

The effect of local surface expansion is illustrated using the propeller-blade scene in Fig. 10. The gray points represent the reference geometry. In Fig. 10 (a), the red points denote the reliable fused seed points before expansion. Although these points generally follow the target structure, their distribution remains relatively sparse, leaving several parts of the target surface insufficiently reconstructed.

In Fig. 10 (b), the blue points represent the additional points generated through local surface expansion. Compared with the red seed points in Fig. 10 (a), the expanded points are distributed more densely over the target surface, while the overall outer contour of the reconstructed point cloud remains largely unchanged. This indicates that the expansion mainly fills local gaps around reliable seed points rather than extending the reconstruction into geometrically inconsistent regions.

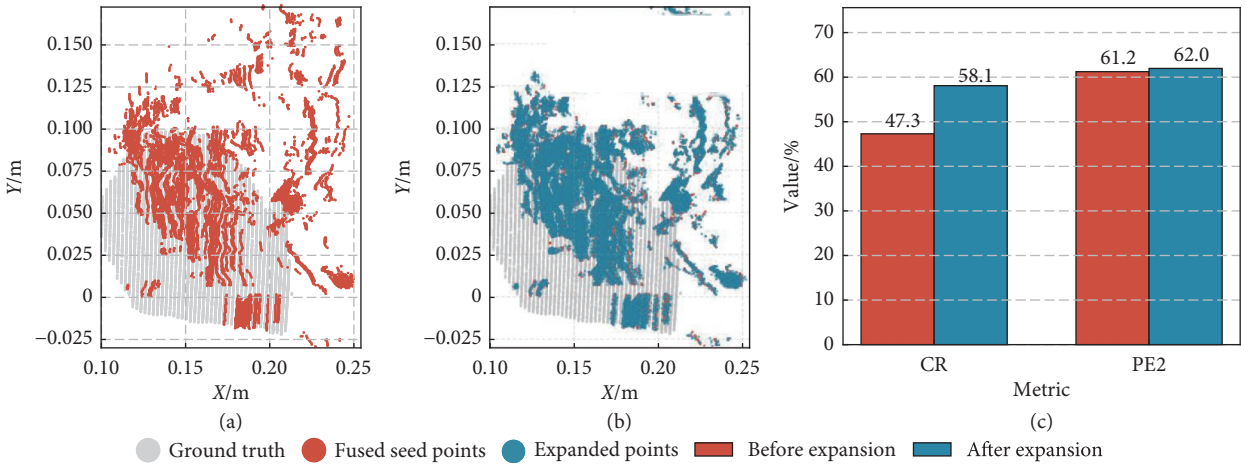


Figure 10 Effect of local surface expansion ((a) Fused seed points before expansion; (b) Point cloud after local surface expansion; (c) Quantitative comparison before and after expansion).

The coverage rate increases from 47.3% before expansion to 58.1% after expansion, while PE2 increases from 61.2% to 62.0%. These results show that local surface expansion improves reconstruction completeness by propagating reliable acoustic-optical depth corrections to neighboring pixels on the same local surface.

Overall, the experiments demonstrate that acoustic range constraints improve the robustness of binocular reconstruction, feature-wise local transformation refinement reduces acoustic-optical projection errors, uncertainty-aware fusion improves depth reliability, and local surface expansion enhances point-cloud coverage.

5 Conclusions

This paper proposes an uncertainty-aware acoustic-optical 3D reconstruction method based on the fusion of a binocular camera and an imaging sonar. The method operates in a single-pose setting, using one synchronized binocular image pair and one sonar image without temporal accumulation or multi-view registration. A feature-wise local transformation refinement strategy is introduced to reduce residual sonar-camera projection errors, while uncertainty-aware depth fusion and local surface expansion are used to improve depth reliability and point-cloud completeness. Experiments on three representative underwater scenes show that the proposed method consistently achieves higher reconstruction coverage than the comparison methods, with coverage rates ranging from 46.37% to 58.09%. In a representative refinement case, the mean acoustic-optical depth residual is reduced from 0.2628 m to 0.1233 m. These results demonstrate that the proposed method can generate more complete and geometrically consistent underwater point clouds while maintaining competitive reconstruction accuracy.

Funding

This work was supported by the National Natural Science Foundation of China (62673331, W2411053, 62373246), the Research Fund of State Key Laboratory of Deep-Sea Manned Vehicles (2024SKLDMV04), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM806), the Oceanic Interdisciplinary Program of Shanghai Jiao Tong University (SL2023MS007), and the Startup Fund for Young Faculty at SJTU (24X010502929).

Data availability

The simulation data and codes supporting the results are available from the corresponding author upon reasonable request.

Declaration of competing interest

None.

Author contribution statement

Yang Yang: Conceptualization, Methodology, Software, Validation, Writing – original draft, Visualization.
Yichen Li: Conceptualization, Writing – review & editing, Supervision, Project administration.
Chi Zhang: Investigation, Methodology, Software, Data curation, Visualization.
Wenbin Yu: Writing – review & editing, Supervision, Project administration, Funding acquisition.
Xinping Guan: Conceptualization, Supervision, Funding acquisition.

Use of AI statement

None.

References

[1] Li Y C, Yang Y, Yu W B, et al. Cooperative highly-maneuvering target tracking using multi-AUV networks: a bearing-only approach[J]. *IEEE Transactions on Mobile Computing*, 2026, 25(2): 2365-2381.

[2] Ferreira F, Veruggio G, Caccia M, et al. A survey on real-time motion estimation techniques for underwater robots[J]. *Journal of Real-Time Image Processing*, 2016, 11(4): 693-711.

[3] Yu M X. Research on 3D structured light vision technology for underwater ship hull fouling detection[D]. Beijing: Beijing University of Chemical Technology, 2024 (in Chinese).

[4] Li R. Development of a stereo vision-based 3D reconstruction system for deep-sea surface mineral resources[D]. Hangzhou: Hangzhou Dianzi University, 2025. (in Chinese).

[5] Li J Q, Zhang G Q, Jiang C Y, et al. A survey of maritime unmanned search system: theory, applications and future directions[J]. *Ocean Engineering*, 2023, 285: 115359.

[6] Li Y C, Yu W B, Xu H T, et al. Cooperative trajectory planning for simultaneous target approaching in multi-AUV systems[J]. *IEEE Transactions on Industrial Informatics*, 2025, 21(11): 8870-8881.

[7] Li Y C, Yu W B, Guan X P. Trajectory planning-aided cooperative localization for multi-AUV networks under harsh communication conditions: a co-designed approach[J]. *IEEE Transactions on Networking*, 2025, 33(6): 3088-3103.

[8] Yang Y, Li Y C, Yu W B, et al. Underwater target tracking based on acoustic-optical fusion for multi-AUV systems[J]. *IEEE Transactions on Network Science and Engineering*, 2026, 13: 7226-7243.

[9] Zhou J C, Liu Q, Jiang Q P, et al. Underwater camera: improving visual perception via adaptive dark pixel prior and color correction[J]. *International Journal of Computer Vision*, 2025, 133(11): 8215-8233.

[10] Zheng T X, Huang S, Li Y F, et al. Key techniques for vision based 3D reconstruction: a review[J]. *Acta Automatica Sinica*, 2020, 46(4): 631-652 (in Chinese).

[11] Hu K, Wang T Y, Shen C W, et al. Overview of underwater 3D reconstruction technology based on optical images[J]. *Journal of Marine Science and Engineering*, 2023, 11(5): 949.

[12] Xu H, He J, Pang C S, et al. Cross-modality feature fusion for forward-looking sonar image segmentation in complex underwater environments[J]. *IEEE Transactions on Multimedia*, 2026, 28: 7488-7501.

[13] Li Y. Research on visual fusion and environment 3D reconstruction for mobile robots[D]. Handan: Hebei University of Engineering, 2024 (in Chinese).

[14] Deng J. Research and implement of warehouse three-dimensional reconstruction method based on mobile robot vision[D]. Mianyang: Southwest University of Science and Technology, 2024 (in Chinese).

[15] Fan J F, Wang X, Zhou C, et al. Structured light vision 3-D reconstruction system for different media considering refraction: design, modeling, and calibration[J]. *IEEE/ASMe Transactions on Mechatronics*, 2024, 29(3): 1997-2008.

[16] Roznere M, Mordohai P, Rekleitis I, et al. 3-D reconstruction using monocular camera and lights: multi-view photometric stereo for non-stationary robots[C]//Proceedings of 2023 IEEE International Conference on Robotics and Automation. London: IEEE, 2023.

[17] Hou S T, Wang Y X, Wu T, et al. Robust multi-view underwater 3D point cloud stitching without feature markers for structural inspection[J]. *Measurement*, 2026, 264: 120278.

[18] Lin H T, Zhang H, Huo J W, et al. High-precision 3D reconstruction of underwater concrete using integrated line structured light and stereo vision[J]. *Automation in Construction*, 2025, 169: 105883.

[19] Kleeman L, Kuc R. Sonar sensing[M]//Siciliano B, Khatib O. Springer Handbook of Robotics. 2nd ed. Cham: Springer, 2016.

[20] Zhang J W, Han F L, Han D F, et al. Integration of sonar and visual–inertial systems for SLAM in underwater environments[J]. *IEEE Sensors Journal*, 2024, 24(10): 16792-16804.

[21] McConnell J, Martin J D, Englot B. Fusing concurrent orthogonal wide-aperture sonar images for dense underwater 3D reconstruction[C]//Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas: IEEE, 2020.

[22] Westman E, Kaess M. Wide aperture imaging sonar reconstruction using generative models[C]//Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau: IEEE, 2019.

[23] Ma S X, Wang J, Huang Y P, et al. Underwater 3-D point cloud reconstruction of a single sonar image based on pseudo front view[J]. *IEEE Transactions on Instrumentation and Measurement*, 2025, 74: 4513113.

[24] Yan J, Zhu Y J, Xiong W, et al. High-fidelity 3D sonar point cloud data reconstruction of underwater bridge caisson foundation for scour inspection[J]. *Measurement*, 2025, 244: 116480.

[25] Lindzey L, Marburg A. Extrinsic calibration between an optical camera and an imaging sonar[C]//Proceedings of the OCEANS 2021: San Diego-Porto. San Diego: IEEE, 2021.

[26] Pecheux N, Creuze V, Comby F, et al. Self calibration of a sonar-vision system for underwater vehicles: a new method and a dataset[J]. *Sensors*, 2023, 23(3): 1700.

[27] Kim J, Lee M, Song S, et al. 3-D reconstruction of underwater objects using image sequences from optical camera and imaging sonar[C]//Proceedings of the OCEANS 2019 MTS/IEEE SEATTLE. Seattle: IEEE, 2019.

[28] Roznere M, Li A Q. Underwater monocular image depth estimation using single-beam echosounder[C]//Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas: IEEE, 2020.

[29] Zhang J W, Han F L, Han D F, et al. Advanced underwater measurement system for ROVs: integrating sonar and stereo vision for enhanced subsea infrastructure maintenance[J]. *Journal of Marine Science and Engineering*, 2024, 12(2): 306.

[30] Qu Z Y, Vengurlekar O, Qadri M, et al. Z-splat: Z-axis Gaussian splatting for camera-sonar fusion[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, 47(9): 7255-7267.

[31] Phung A, Camilli R. Sonar-MASt3R: real-time opti-acoustic fusion in turbid, unstructured environments[PP/OL]. arXiv(2026-03-13) [2026-04-11]. https://arxiv.org/abs/2603.13585.

[32] Mezenceva E, Malakhov S. The study of the semi-global block matching algorithm implementing parallel calculation with GPU[C]//Proceedings of 2021 IEEE 8th International Conference on Problems of Infocommunications, Science and Technology. Kharkiv: IEEE, 2021.

[33] Zhu Z K, Li Y C, Yu W B, et al. Underwater vision-based pose estimation for AUV landing recovery: a monocular-binocular fusion approach[J]. *IEEE Transactions on Automation Science and Engineering*, 2026, 23: 11881-11901.

[34] Cardaillac A, Ludvigsen M. Camera-sonar combination for improved underwater localization and mapping[J]. *IEEE Access*, 2023, 11: 123070-123079.

[35] Phung A, Camilli R. OASIS: real-time opti-acoustic sensing for intervention systems in unstructured environments[C]//Proceedings of 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems. Hangzhou: IEEE, 2025.

[36] Hartley R, Zisserman A. Multiple view geometry in computer vision[M]. Cambridge: Cambridge University Press, 2003.

[37] Wang J K, Chen F F, Huang Y W, et al. Virtual maps for autonomous exploration of cluttered underwater environments[J]. *IEEE Journal of Oceanic Engineering*, 2022, 47(4): 916-935.

[38] Li X X, Chen S Q, Xiao H J, et al. 3D reconstruction analysis based on SGBM algorithm and BM algorithm[J]. *Automation & Information Engineering*, 2019, 40(5): 6-12.