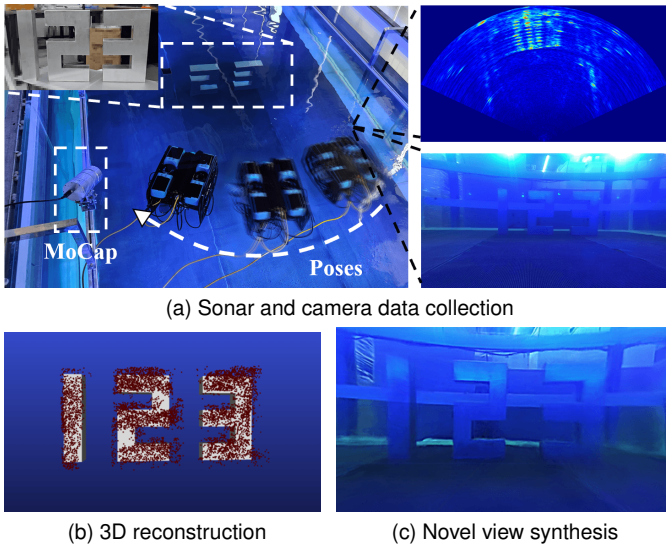
IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Aqua-Splat: Physically-Informed Sonar-Camera Gaussian Splatting for Underwater 3D Reconstruction

Zijie Ling, Yunxuan Feng, Ao Meng, Renxiang Xiao, Shu Pan, Wenjie Lu and Liang Hu*

***Abstract*—Differentiable Gaussian Splatting (GS) has emerged as a powerful paradigm for scene representation, enabling efficient rendering and real-time editing. However, existing GS-based methods, which rely mainly on clear visual images, perform poorly in underwater environments due to camera distortions such as light absorption and backscattering. In contrast, acoustic sensors like Forward Looking Sonar (FLS) offer superior penetration and robustness in such conditions. To leverage the complementary merits of visual and FLS images, we propose a novel GS framework customized for underwater scenarios, termed Aqua-Splat, for robust and accurate underwater perception. It ensures physically consistent reconstruction by incorporating the sonar wave propagation modeling in the image formation process. Moreover, we propose a volume rendering technique for sonar image synthesis, achieving similar speed to visual rendering. Additionally, we introduce a sonar-guided densification strategy to optimize the scene representation. Through extensive experiments on both simulated and datasets from the lab pool, we demonstrate that Aqua-Splat significantly improves image synthesis and 3D scene reconstruction in challenging underwater environments, outperforming existing methods in terms of both geometric accuracy and photometric fidelity. The code of Aqua-Splat will be open-sourced later for the community.**

***Index Terms*—Gaussian Splatting, acoustic-optic vision, sensor fusion.**



(a) Sonar and camera data collection

(b) 3D reconstruction

(c) Novel view synthesis

Fig. 1. **System setup, reconstruction and novel view synthesis by Aqua-Splat.** (a) Data collecting in underwater scenes. We use a ROV to collect visual and sonar data in the lab pool. An underwater motion capture system (MoCap) is used to obtain the ROV's poses. (b) 3D reconstruction of the underwater scene by Aqua-Splat. (c) Novel view synthesis of the underwater scene by Aqua-Splat.

I. Introduction

HIGH-FIDELITY three-dimensional (3D) maps of the underwater environment underpin a wide range of marine applications. Precise geometric models enable infrastructure inspections, facilitate autonomous navigation and manipulation for remotely operated or autonomous underwater vehicles (ROVs/AUVs), and support disciplines such as marine archaeology, environmental monitoring, and habitat mapping [1]–[4]. Optical cameras and acoustic sensors are two common sensors used for underwater 3D reconstruction. The former offers high-resolution imagery and rich textures, but it is susceptible to light attenuation, scattering, and color distortion [5], [6]—limitations that degrade image quality and reconstruction accuracy [6]. Even in low-turbidity waters such as shallow seas or clear lakes, water scattering and color distortion still hinder high-precision, purely visual reconstruction.. Acoustic sensors such as side-scan sonar and FLS, by contrast, provide superior penetration capability and are more robust in such conditions, though sonar images lack elevation information of the recorded echoes. As such, we aim to develop a sonar-visual fusion framework for real-time, high-precision 3D mapping by leveraging the structural penetration capability of sonar and the texture information of optical cameras.

Recovering 3D space from 2D sonar images has long been a challenge in sonar-based reconstruction that classical methods such as space carving has not yet addressed. Until recently, Neural Radiance Fields (NeRF)-based methods, such as Neusis [7] and DSC [8], have adapted implicit neural representations of FLS data to recover 3D scenes effectively. Nonetheless, these approaches rely on dense sampling across the sector-shaped acoustic field, which significantly increases computational cost and reduces rendering speed. Different from NeRF, GS [9] has much faster rendering speed and lower computation costs. UW-GS [10] and RecGS [11] have adapted GS to underwater domains but they still suffer from the inherent limitations of optical sensing as the unimodal vision is considered. A notable attempt Z-Splat [12] introduces z-axis splatting to jointly reconstruct geometry and photometry using camera and sonar data. However, its sonar rendering pipeline does not fully reflect the principles of sonar image acquisition, leading to inaccurate geometry in the reconstructed scene.

To realize accurate and fast underwater reconstruction and novel view synthesis, we propose **Aqua-Splat**, a novel frame-

Manuscript received June 1, 2025; revised July 24, 2025; accepted September 20, 2025.

This paper was recommended for publication by Editor Giuseppe Loianno upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported in part by the National Natural Science Foundation of China under Grant 62573157, and Shenzhen Science and Innovation Committee under Grant JCYJ20241202123714019.

All authors are with the Department of Automation, School of Intelligence Science and Engineering, Harbin Institute of Technology, Shenzhen, China. * For correspondence: l.hu@hit.edu.cn.

Digital Object Identifier (DOI): see top of this page.

©2026 IEEE

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

work that integrates FLS into 3D Gaussian splatting with respect to the physical principles of FLS, as shown in Fig. 1. The main contributions of this paper are summarized as follows:

1) A physically-informed forward model is introduced to render sonar images from 3D Gaussians in the polar domain;
2) A GPU-accelerated volume rendering technique for sonar image generation, along with a sonar-guided densification strategy, is proposed for fast sonar image rendering, achieving a speed of over 120 FPS, comparable to that of visual rendering;
3) The method was evaluated on both simulated and lab pool datasets, demonstrating that fusing GS with sonar imaging leads to superior geometric and photometric reconstruction compared to standard camera-based Gaussian splatting.

The remainder of this letter is structured as follows. Related works about 3D reconstruction using cameras and FLSs are summarized in Section II. The proposed Aqua-Splat method is introduced in detail in Section III, followed by the evaluation experiments in Section IV. Finally, conclusions and future work are presented in Section V.

II. RELATED WORK

A. Visual-GS based 3D reconstruction

Splatting algorithms were introduced more than two decades ago [13] for texture filtering [14] and point cloud rendering [15], [13]. It represented the scene as a sum of anisotropic Gaussian kernels that can be efficiently rendered and without aliasing artifacts. GS uses this scene representation and a fast differentiable rendering pipeline to compute the scene parameters [9]. Gaussian splatting has the merits of decreased computation on empty spaces, analytical derivatives, high-quality reconstructions, and explicit representations, attracting ever-increasing research attention.

UW-GS [10], RecGS [11], SeaSplat [16] and WaterSplatting [17] have proposed GS-based methods tailored for underwater scenes. However, even in low-turbidity settings (e.g., shallow seas or clear lakes), these unimodal vision-based methods struggle: light attenuation, scattering, and color distortion degrade image quality, limiting their reconstruction accuracy for inspection and mapping tasks. In contrast, sonar-based approaches not only remain robust to such visual degradation but also provide inherent scale awareness, further motivating the extension of GS to sonar-integrated imaging.

B. Sonar based 3D reconstruction

3D geometry reconstruction using sonar images with known poses has been investigated for decades. In traditional methods, Westman et al. represent the scene as an undirected albedo field and recover the volumetric albedo through convex linear optimization [18]. Space carving methods do not focus on recovering albedo; instead, they construct the geometry by evaluating occupancy probabilities [19], [20]. However, these methods rely solely on the shape of non-echo regions and assume a known correspondence between sonar returns and elevation angles, limiting applicability in practical scenarios.

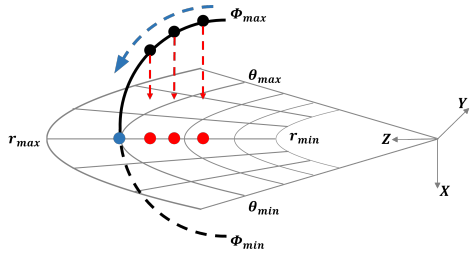


Fig. 2. **Conceptual-level differences between our Aqua-Splat and Z-Splat.** Considering the physical characteristics of sonar imaging, all three points should be projected along the arc direction onto the same point in the zero-elevation plane, as illustrated by the blue points in the figure. However, Z-Splat vertically projects the points onto the horizontal plane in the orthographic camera view, resulting in three different (red) points, which neglects the arc-shaped propagation pattern inherent in sonar imaging.

Neusis [7] firstly integrates the NeRF model into 3D reconstruction using FLS images. DSC [8] has proposed an efficient 3D reconstruction method for FLS images by integrating differentiable rendering and NeRF-based techniques. Unlike [7], DSC introduces echo probability images to replace echo intensity images, reducing computational redundancy.

However, these NeRF-based methods can only reconstruct the geometry of a scene but fail to capture its photometric properties. In contrast to Sonar-NeRF, our work reconstructs both scene geometry and photometry. Besides, [7] and [8] require extensive sampling and computations to simulate sound-wave propagation within the scene, resulting in slower rendering speeds. SonarSplat [21] introduces the first sonar GS framework, but it suffers from lower rendering speed due to a different sonar rendering process from ours.

C. Multimodal reconstruction

Previous studies have explored the integration of complementary information from sonar and camera sensors. Shu et al. [22] proposed the first SLAM system that fuses stereo camera, IMU, and imaging sonar for robust underwater localization under visual degradation. Cardaillac et al. [23] similarly employed a camera and an FLS for 3D reconstruction by matching features between acoustic and optical measurements. However, these methods mainly focus on sparse feature-based reconstruction, which is unsuitable for dense mapping or high-fidelity reconstruction tasks.

Babaee et al. [24] reconstructed 3D objects using RGB images and imaging sonar. However, their method requires 360-degree views of the scene, which limits its applicability to scenarios with small sensor baselines. Qadri et al. [25] recently proposed a method that combines FLS sonar and camera data using implicit neural representations. Unlike our approach, their method focuses solely on reconstructing scene geometry, without modeling photometric properties.

Most related to our work is Z-Splat [12]. Z-Splat utilizes the depth information from sonar to overcome the missing cone present in camera-based 3D reconstruction. However, as illustrated in Fig. 2, the rendering in Z-Splat does not

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

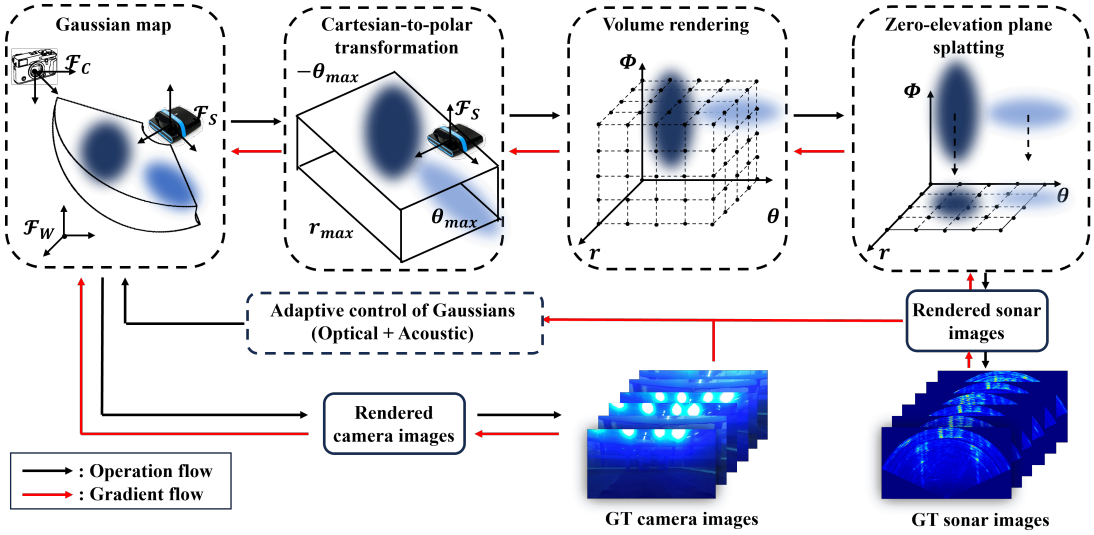


Fig. 3. **System Overview.** The 3D Gaussians are initially defined in world frame. After the Cartesian-to-polar transformation and volume rendering, they are projected onto the zero-elevation plane, generating the rendered sonar image. Both the sonar loss and the camera loss are employed to optimize the 3D Gaussians. Additionally, both losses are used to adaptively densify Gaussians.

follow the principles of sonar imaging, which may lead to inaccuracies in the reconstructed scene geometry.

III. PROPOSED APPROACH

This section describes the details of the Aqua-Splat method, as shown in Fig. 3. It first applies the Cartesian-to-polar transformation to Gaussians, then performs volume rendering in the polar domain, and finally splats the Gaussians to the zero-elevation plane to generate the sonar image.

A. Sonar image formation model

The FLS emits sound pulses and detects echoes at point $\boldsymbol{\mathcal{P}} = [r, \theta, \phi]^T$ (in the polar Sonar Coordinate System (SCS)), provided they fall within its field of view (FoV): $r \in [r_{min}, r_{max}]$, the azimuth angle $\theta \in [-\theta_{max}, \theta_{max}]$ and the elevation angle $\phi \in [-\phi_{max}, \phi_{max}]$. Therefore, the original sonar image is in the polar form. To reduce computational load during sonar image rendering, Gaussians outside the FoV are excluded. Since the FLS cannot measure elevation ϕ, each echo P is projected onto the zero-elevation plane, and only r, θ and echo intensity are recorded. The intensity of a single pixel may result from multiple echoes along the same arc of different elevations. The transformation between the polar and Cartesian domain is given by:

$$\begin{bmatrix} r \\ \theta \\ \phi \end{bmatrix} = \mathbf{m}(\mathbf{P}) = \begin{bmatrix} \sqrt{X^2+Y^2+Z^2} \\ arctan(Y/X) \\ arctan(Z/\sqrt{X^2+Y^2}) \end{bmatrix}, \tag{1}$$

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathbf{m}^{-1}(\boldsymbol{\mathcal{P}}) = \begin{bmatrix} rcos(\phi)cos(\theta) \\ rcos(\phi)sin(\theta) \\ rsin(\phi) \end{bmatrix}, \tag{2}$$

where $\mathbf{P} = [X, Y, Z]^T$ and $\boldsymbol{\mathcal{P}}$ are the Cartesian form and polar form of the same echo point in SCS, respectively.

Our framework optimizes a single Gaussian map shared between camera and sonar images rendering. Similarly to [9], the 3D Gaussians $\mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$ are defined by the point center (mean) $\boldsymbol{\mu}_w = [X_w, Y_w, Z_w]^T$, and a covariance matrix $\boldsymbol{\Sigma}_w$ defined in the Cartesian World Coordinate System (WCS). Following [9], the Gaussians are initialized from SfM point clouds in the Cartesian WCS. Assume that the transformation from the WCS to SCS is $\boldsymbol{T}_{W2S} = \begin{bmatrix} \boldsymbol{R}_{W2S} & \boldsymbol{t}_{W2S} \\ \mathbf{0} & 1 \end{bmatrix}$. The mean of the Gaussian in the Cartesian SCS is obtained as followed:

$$\begin{bmatrix} \boldsymbol{\mu}_s \\ 1 \end{bmatrix} = \boldsymbol{T}_{W2S} \begin{bmatrix} \boldsymbol{\mu}_w \\ 1 \end{bmatrix}, \tag{3}$$

where $\boldsymbol{\mu}_s$ is the mean of the Gaussian in SCS.

Since sonar images are captured in polar form, and sonar image rendering process is performed in the polar domain [7], [8], the Gaussians must be transformed from the Cartesian to the polar form, as shown in Fig. 4. According to Eq. (1), the Gaussian transformed from the Cartesian to polar form does not hold the Gaussian distribution. To simplify the computation, we introduce the local affine approximation $\mathbf{m}_{\boldsymbol{\mu}_s}$ of the projective transformation. It is defined by the first two terms of the Taylor expansion of the function $\mathbf{m}$ at the point $\boldsymbol{\mu}_s$:

$$\mathbf{m}_{\boldsymbol{\mu}_s}(\mathbf{P}) = \boldsymbol{\mu}_P + \mathbf{J}_s(\mathbf{P} - \boldsymbol{\mu}_s), \tag{4}$$

where $\boldsymbol{\mu}_s$ is the center of a Gaussian $\mathcal{N}_s$ in the Cartesian SCS and $\boldsymbol{\mu}_{\boldsymbol{P}} = \mathbf{m}(\boldsymbol{\mu}_s)$ is the center of the Gaussian in the polar SCS. The Jacobian $\mathbf{J}_s$ is given by the partial derivatives of $\mathbf{m}$ at the point $\boldsymbol{\mu}_s$, that is $\boldsymbol{J}_{\boldsymbol{\mu}_s} = \frac{\partial \mathbf{m}}{\partial \mathbf{P}}(\boldsymbol{\mu}_s)$.

Therefore, the covariance of the Gaussian in the polar coordinate is given as follows:

$$\boldsymbol{\Sigma}_{\boldsymbol{P}}^{'} = \boldsymbol{J}_{\boldsymbol{\mu}_s} \boldsymbol{R}_{W2S} \boldsymbol{\Sigma}_w \boldsymbol{R}_{W2S}^T \boldsymbol{J}_{\boldsymbol{\mu}_s}^T. \tag{5}$$

We assume that the acoustic opacity of a Gaussian is equivalent to its optical opacity. The acoustic opacity at a given sampling point $\boldsymbol{\mathcal{P}}_c$ is modeled as the cumulative effect of nearby 3D Gaussians. Consider a Gaussian $\boldsymbol{\mathcal{P}}_i = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i')$ in the polar coordinates that is in the proximity to $\boldsymbol{\mathcal{P}}_c$, with

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

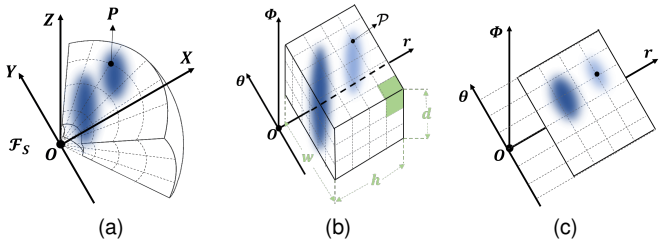


Fig. 4. **Sonar image formation model.** (a) Gaussians inside the FoV contribute to the rendering process. (b) The Cartesian-to-polar transformation, where Gaussians are converted into the polar coordinate and evaluated for occlusion along the radial axis. The green squares in the image represent individual bricks. (c) Rendered sonar image. Gaussians splatted onto the zero-elevation plane generating the rendered sonar image.

an opacity of α_i. Its contribution to the acoustic opacity at $\boldsymbol{\mathcal{P}}_c$ is given by:

$$\alpha_{\mathcal{P}_{c,i}} = \alpha_i \exp\left(-\frac{1}{2}\left(\boldsymbol{\mathcal{P}}_c - \boldsymbol{\mu}_{P,i}\right)^T {\boldsymbol{\Sigma}'_{P,i}}^{-1} \left(\boldsymbol{\mathcal{P}}_c - \boldsymbol{\mu}_{P,i}\right)\right). \quad (6)$$

Given that a total of N Gaussians contribute to the sampling point, the final acoustic opacity at this point is: $\alpha_{\mathcal{P}_c} = \sum_{i=1}^{N} \alpha_{\mathcal{P}_{c,i}}$. After assigning acoustic opacity to all sample points, occlusion along the r-axis must be considered. As shown in Fig. 4(b), the Gaussian $\mathcal{N}_2$ lies behind $\mathcal{N}_1$, thus its energy receiving and reflecting is blocked by $\mathcal{N}_1$. The discrete acoustic transmittance at a given point $\boldsymbol{\mathcal{P}}_c$ is defined as below:

$$T_{\mathcal{P}_c} = \prod_{\mathcal{P}_j \in \mathcal{Q}_{\mathcal{P}_S}} \left(1 - \alpha_{\mathcal{P}_j}\right), \quad (7)$$

where $\mathcal{Q}_{\boldsymbol{\mathcal{P}}_c}$ is the set of sampling points between $\boldsymbol{\mathcal{P}}_c$ and the origin O.

Due to the inability of the FLS to measure the elevation ϕ, the echoes are projected onto the zero-elevation plane. Accordingly, we project the Gaussians onto the zero-elevation plane to construct the sonar image. Finally, the discrete sonar image formation model is:

$$\hat{I}_{son}(r,\theta) = \sum_{\mathcal{P}_S \in \mathcal{A}} \frac{1}{r_{\mathcal{P}_S}} T_{\mathcal{P}_S} \alpha_{\mathcal{P}_S}, \quad (8)$$

where $\mathcal{A}$ is the arc located at (r,θ) and the factor $\frac{1}{r_{\mathcal{P}_S}}$ accounts for spherical spreading on both the transmit and receive paths [7]. So far, the rendering process of transforming a set of Gaussians into the sonar image has been completed.

B. Sonar Volume Rendering and Camera-FLS Fusion

For a sonar image of size $\theta \times r$ in the polar coordinate, we sample along the arc corresponding to each pixel with a fixed number of sampling points per elevation angle. Our design draws on the framework from [7], where the elevation range $[-\phi_{max}, \phi_{max}]$ is discretized into equidistant angles. Unlike their approach, which samples only specific pixels, we discretize the arc corresponding to all pixels into d equidistant angles, ensuring consistent sampling across the entire image.

This implies that within the sonar's FoV, uniform sampling is performed along the r-axis, θ-axis, and ϕ-axis with step sizes Δr, $\Delta\theta$, and $\Delta\phi$, respectively. Here, Δr (range resolution) and $\Delta\theta$ (azimuth resolution) directly correspond to the sonar's hardware-specific resolution. Specifically, the elevation range is discretized into d equidistant angles with step size $\Delta\phi$.

As a result, we obtain a discrete three-dimensional grid of size $w \times h \times d$. In our experiment, w and h of the rendered sonar images are consistent with those of the input sonar images, and d is a hyperparameter which balances reconstruction quality and training efficiency. Roughly, the bigger d is, the higher quality the reconstructions will be, at the cost of a higher computational burden.

Algorithm 1 Volume Rendering of 3D Gaussians.
w, h, d: width, height and depth of the grid
M, S: Gaussian means and covariances in world space
A: Gaussian opacities
F: FoV of sonar

function VOLUMERENDERING(w, h, d, M, S, A, F)
  CullGaussian(p, F)
  $M', S' \leftarrow$ TransformGaussians(M, S, F)
  $B \leftarrow$ CreateBricks(w, h, d)
  $L, K \leftarrow$ DuplicateWithKeys(M', B)
  SortByKeys(L, K)
  $R \leftarrow$ IdentifyBrickRanges(T, K)
  $V \leftarrow \mathbf{0}$
  for all Bricks b **in** V **do**
    for all Voxel v **in** b **do**
      $r \leftarrow$ GetBrickRanges(R, b)
      $V[i] \leftarrow$ BlendInOrder(i, L, r, K, M', S', A)
    end for
  end for
  $V' \leftarrow$ ComputeTransmittance(V)
  $\hat{I}_{son} \leftarrow$ AccumulateOpacity(V')
  return $\hat{I}_{son}$
end function

To render the three-dimensional grid, our method begins by dividing the discrete three-dimensional grid into $16 \times 16 \times 16$ bricks, as shown in Fig. 4(b). Secondly, we instantiate after-transformed Gaussians for every overlapping brick, with keys assigned by combining r and the brick ID, and the instances are sorted using a GPU Radix sort [26]. Thirdly, a thread block is launched per brick and traverses each brick's list from front to back, accumulating acoustic opacity. Finally, acoustic transmittance is computed along the r-axis, and opacity is accumulated along the ϕ-axis. After the opacity accumulation along the ϕ-axis, the resulting intermediate sonar image is further processed via Min-Max normalization. At this point, the forward process from the Gaussian map to the rendered sonar image is complete. A high-level overview of the volume rendering approach is summarized in Algorithm 1.

The visual rendering process follows that proposed in the original 3DGS [9]. After rendering the camera and sonar images, we define the loss function as a weighted combination of the camera and sonar loss functions:

$$\mathcal{L} = \left\| \hat{I}_{\text{cam}} - I_{\text{cam}} \right\|_1 + w \cdot \left\| \hat{I}_{\text{son}} - I_{\text{son}} \right\|_1, \quad (9)$$

where $\hat{I}_{\text{cam}}$ and I_{cam} denote the rendered and the corresponding ground truth (GT) camera image, respectively. I_{son} represents

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

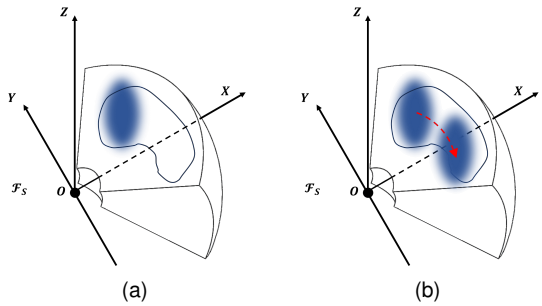


Fig. 5. **Sonar-guided densification strategy.** The operation of copying the Gaussians is performed in the polar domain, here plotted in the Cartesian coordinates just for better visualization.

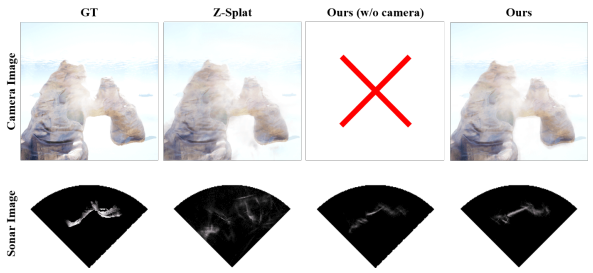


Fig. 6. **Novel view synthesis on the simulated dataset.**

the GT sonar image.

C. Sonar-guided Densification Strategy

Underwater imaging is susceptible to blurring due to light absorption and backscattering, resulting in reduced image sharpness and sparse SfM initialization points. The initially blurry rendered camera images limit the speed and quality of densification under camera supervision alone. To address this limitation, we incorporate a sonar-guided densification strategy to accelerate the process. We observe large position gradients between rendered and GT sonar images, analogous to the view-space gradients observed in camera-based GS [9]. However, in our case, these gradients are defined in the polar coordinate, so we apply a similar densification strategy directly within the polar domain, leveraging sonar-specific spatial cues for more effective scene refinement.

Our introduced sonar-guided densification strategy is illustrated in Fig. 5. For small Gaussians located in insufficiently reconstructed regions, new Gaussians are generated by translating them along the direction of the position gradient. Gaussians with large variances are decomposed into multiple smaller components. Unlike [9], our loss is computed between GT and rendered **sonar** images, and both cloning and pruning operations are performed directly **in the polar coordinates**. This sonar-guided densification effectively compensates for the limitations of camera-only supervision, yielding improved geometric fidelity in underwater environments.

IV. Experiment & Analysis

We conducted experiments in both the simulation environment and lab pool, and compared our proposed Aqua-Splat with existing algorithms vanilla 3DGS [9] and Z-Splat (using FLS and camera input) [12] in two aspects: novel view synthesis and 3D reconstruction. We intentionally used all camera images (both simulated and lab-pool-captured) without preprocessing. Novel view synthesis refers to the generation of camera images and sonar images from camera poses and sonar poses that do not exist in the training data. Ablation studies of our method with and without camera supervision are conducted as well.

We use the metrics PSNR, SSIM, and LPIPS to evaluate the quality of predicted novel views of both camera and sonar images. Specifically: Peak Signal-to-Noise Ratio (PSNR) [27], which measures the ratio between the maximum possible power of a signal and the power of corrupting noise; Structural Similarity Index Measure (SSIM) [28], which assesses the structural similarity between two images; and Learned Perceptual Image Patch Similarity (LPIPS) [29], a learned metric that evaluates image similarity based on perceptual differences.

We use Chamfer distance, Precision, and Recall to evaluate the reconstruction quality. The metrics are calculated by comparing the point cloud of the predicted Gaussian splatting with the GT mesh vertices. The best results in the table are highlighted in bold. All training and testing are performed on an NVIDIA RTX 3060 GPU. We set the weight between the camera and sonar loss terms to $w = 0.2$, and fixed the grid depth d to 256.

A. Simulation experiments

Camera images, sonar images, and poses were collected in the HoloOcean simulation environment [30], [31]. Simulated underwater environments contained objects consistent with the experimental setup in [7], [8]. Sonar images were captured via a simulated FLS with a detection range of $[0, 12\text{m}]$, horizontal aperture of $[-45°, 45°]$, and vertical aperture of $[-10°, 10°]$. The intrinsic parameters of the simulated camera were calibrated by placing a calibration board in the HoloOcean simulation environment.

Table I presents simulated sonar novel view synthesis results: our method outperforms Z-Splat in all scenes (with/without camera supervision), and camera-supervised variants show better sonar quality, verifying dual-modality necessity. Table II (camera novel view synthesis) shows FLS-included methods outperform camera-only ones, and our method outperforms Z-Splat in most cases. Fig. 6 demonstrates our dual (camera+sonar) supervision generates high-fidelity rendered images close to GT. Fig. 7 (GT geometry as mesh, predicted Gaussian kernels in red) shows our dual-modality method has more accurate reconstruction. Table III (3D reconstruction accuracy via multiple metrics) confirms sonar-camera fusion outperforms single-modality, and our dual-input method outperforms Z-Splat in most scenes.

All our experiments were conducted on an NVIDIA RTX 3060 GPU with a training schedule of 30,000 steps, and the total training time ranged from 40 to 50 minutes. Furthermore, we evaluate the rendering speed of our method for data synthesis in frames per second (FPS). The rendering speed of 3DGS for camera images is measured at 122.36 FPS, while our

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

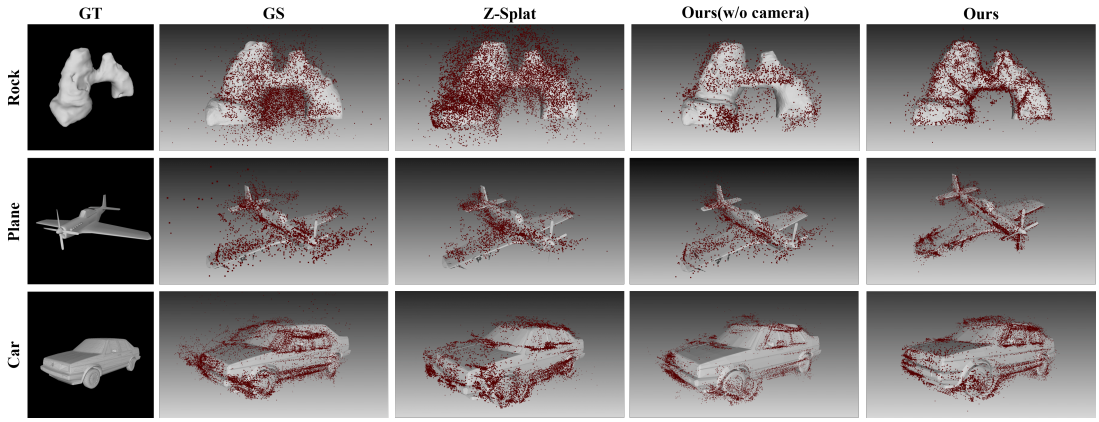


Fig. 7. **Geometry comparison.** We present the GT meshes and overlay the mean positions of the reconstructed Gaussians as point clouds. We observe that our method reconstructs geometric shapes more accurately than both the camera-only approach and Z-Splat.

TABLE I
SIMULATION: SONAR IMAGE NOVEL VIEW SYNTHESIS COMPARISONS

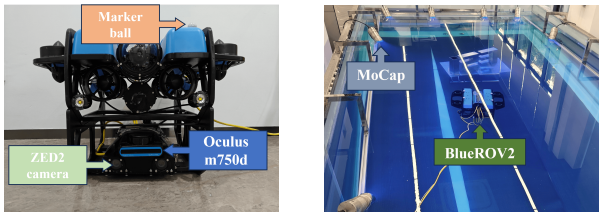
Scene	PSNR↑			SSIM↑			LPIPS↓		
	Z-Splat	Ours (w/o camera)	Ours	Z-Splat	Ours (w/o camera)	Ours	Z-Splat	Ours (w/o camera)	Ours
Rock	17.398	18.852	**19.098**	0.478	0.888	**0.893**	0.335	0.205	**0.170**
Plane	15.881	17.583	**21.835**	0.612	0.752	**0.886**	0.510	0.272	**0.252**
Car	17.644	20.367	**21.597**	0.524	0.682	**0.873**	0.363	0.287	**0.239**

TABLE II
SIMULATION: CAMERA IMAGE NOVEL VIEW SYNTHESIS COMPARISONS

Scene	PSNR↑			SSIM↑			LPIPS↓		
	GS	Z-Splat	Ours	GS	Z-Splat	Ours	GS	Z-Splat	Ours
Rock	30.021	32.251	**34.086**	0.812	**0.837**	0.835	0.382	0.251	**0.187**
Plane	30.963	33.646	**36.185**	0.843	0.928	**0.944**	0.258	0.208	**0.173**
Car	30.149	**34.413**	32.559	0.856	0.895	**0.917**	0.392	0.312	**0.247**

TABLE III
SIMULATION: GEOMETRIC COMPARISONS

Scene	Chamfer↓				Precision↑				Recall↑			
	GS	Z-Splat	Ours (w/o camera)	Ours	GS	Z-Splat	Ours (w/o camera)	Ours	GS	Z-Splat	Ours (w/o camera)	Ours
Rock	0.495	0.324	0.572	**0.168**	0.581	0.738	0.563	**0.869**	0.468	0.698	0.497	**0.752**
Plane	0.673	**0.158**	0.585	0.182	0.482	0.901	0.523	**0.932**	0.423	0.783	0.418	**0.823**
Car	0.723	0.513	0.752	**0.234**	0.425	0.558	0.337	**0.823**	0.337	**0.698**	0.423	0.658



(a) (b)

Fig. 8. **Lab pool dataset collection.** (a) The experimental ROV platform (b) The tank equipped with motion capture system.

method achieves a rendering speed of 121.47 FPS for sonar images. These results indicate that the two approaches exhibit comparable rendering performance, which is much faster than NeRF-based underwater reconstruction methods. As pointed out in [21], the rendering speed of sonar-NeRF-based methods is usually slower than 3DGS by 4 orders of magnitude, and much slower than that of the visual NeRF as well.

B. *Experiments in the Lab Pool*

Apart from the simulator experiments, we also conduct experiments using a ROV in a lab pool. The Bluerov2 robot equipped with a ZED2 stereo camera and an Oculus m750d imaging sonar was used for the experiment, as shown in Fig.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

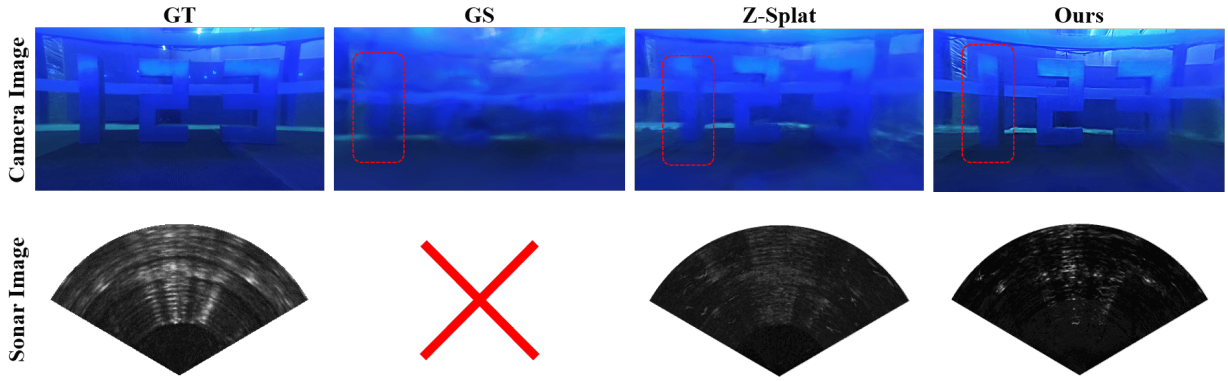


Fig. 9. **Novel view synthesis experiments on the lab pool dataset.** In the highlighted regions of the first row, we can observe that our method reconstructs the metallic model shaped like the number "1" more accurately. The gaps between the models are also clearly visible in the rendered sonar images using our method.

(a) GS (b) Z-Splat (c) Ours

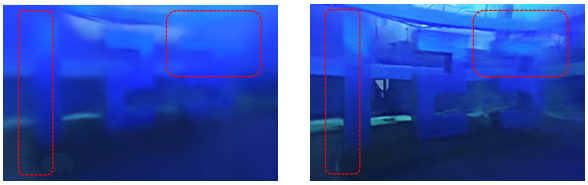
Fig. 10. **Geometry comparison on the lab pool experiment.**

TABLE IV
LABORATORY EXPERIMENTS: QUANTITATIVE GEOMETRY COMPARISONS

	Chamfer↓			Precision↑			Recall↑		
Scene	GS	Z-Splat	Ours	GS	Z-Splat	Ours	GS	Z-Splat	Ours
Numbers Models	1.185	0.526	**0.327**	0.437	0.527	**0.570**	0.523	0.682	**0.723**

8(a). Camera and FLS are well-calibrated underwater before experiments, the details of which follows the method in [22]. We used hand-eye calibration to estimate the 3-DoF relative pose (x, y, yaw) between the sonar and camera, then manually measured the z-axis offset. Given the sonar is closely mounted to the camera, roll and pitch are negligible, resulting in a full 6-DoF relative pose. The tank is about 3.3 m in width and 6.6 m in length. We constructed three metal models in the shapes of numbers "1", "2", and "3" and placed them in the tank. A motion capture system NOKOV equipped with 12 underwater cameras was mounted upside around the tank, which provided the GT pose of the underwater robot, as shown in Fig. 8(b). The sonar image denoising procedure follows the method proposed in [8]: first, a trained Swin-Conv-Unet network (SCUNet) [32] is applied to denoise the original sonar images. Then, a background image (obtained from object-free sonar frames using SCUNet) is subtracted to suppress static noise.

Fig. 9 shows the GT camera image collected in the lab pool and rendered images from different methods. Overall, our method generates higher fidelity images than the others, which can also be observed from Fig. 10 and Table IV.



(a) (b)

Fig. 11. **The ablation experiment.** (a) and (b) illustrate the training outcomes at the 5000th iteration without and with our sonar-guided densification strategy, respectively. As indicated in the highlighted areas, incorporating our strategy leads to the reconstruction of finer structural details.

Furthermore, we evaluated the effectiveness of our proposed sonar-guided densification strategy. Fig. 11(a) and (b) plot the novel view synthesis at the 5000th training epoch without and with the proposed adaptive control of Gaussians, respectively. The results indicate that incorporating the adaptive control mechanism significantly accelerates the densification of the Gaussian map. Specifically, the number of Gaussians increases from 15,204 (without our strategy) to 26,722 (with our strategy), clearly demonstrating the effectiveness of our sonar-guided densification strategy.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

V. CONCLUSION

This work presents Aqua-Splat, a novel sonar-vision underwater reconstruction and novel view synthesis method using 3D Gaussian representation. We introduce an acoustic volume rendering framework along with an adaptive Gaussian control mechanism guided by sonar. Extensive experiments on both simulated and lab pool datasets show that Aqua-Splat yields superior geometric and photometric fidelity compared to traditional camera-only approaches.

However, the current implementation of our method relies on external motion capture systems that provide sonar pose information. To achieve 3D reconstruction in highly turbid natural water, in future work we plan to extend our framework to accommodate sonar-based odometry and sensor fusion techniques to estimate sonar poses directly from sonar imagery and other onboard measurements. As such, no accurate pose prior and motion capture systems will be needed, making it suitable for applications in highly turbid natural water.

Besides, our method still has two key limitations: it may produce reconstruction voids or structural distortions under incomplete view coverage in complex underwater environments, and it fails to handle dynamic objects in water, which can introduce artifacts or lose static structural details. Additionally, the relationship between the optical and acoustic opacity of Gaussians, which is still unclear and assumed to be equivalent in our paper, will be explored in the future.

REFERENCES

[1] W. Wang, B. Joshi, N. Burgdorfer, K. Batsosc, A. Q. Lid, P. Mordohaia, and I. Rekleitisb, "Real-time dense 3d mapping of underwater environments," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5184–5191.

[2] C. Bräuer-Burchardt, C. Munkelt, M. Bleier, M. Heinze, I. Gebhart, P. Kühmstedt, and G. Notni, "Underwater 3d scanning system for cultural heritage documentation," *Remote Sensing*, vol. 15, no. 7, p. 1864, 2023.

[3] A. Abadie, P. Boissery, and C. Viala, "Georeferenced underwater photogrammetry to map marine habitats and submerged artificial structures," *The Photogrammetric Record*, vol. 33, no. 164, pp. 448–469, 2018.

[4] D. Skarlatos and P. Agrafiotis, "Image-based underwater 3d reconstruction for cultural heritage: from image collection to 3d. critical steps and considerations," *Visual Computing for Cultural Heritage*, pp. 141–158, 2020.

[5] H. Lu, Y. Li, S. Serikawa, X. Li, J. Li, and K.-C. Li, "3d underwater scene reconstruction through descattering and colour correction," *International Journal of Computational Science and Engineering*, vol. 12, no. 4, pp. 352–359, 2016.

[6] F. Gu, J. Zhao, P. Xu, S. Huang, G. Zhang, and Z. Song, "Underwater 3d reconstruction based on multi-view stereo," in *Ocean Optics and Information Technology*, vol. 10850. SPIE, 2018, pp. 117–123.

[7] M. Qadri, M. Kaess, and I. Gkioulekas, "Neural implicit surface reconstruction using imaging sonar," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1040–1047.

[8] Y. Feng, W. Lu, H. Gao, B. Nie, K. Lin, and L. Hu, "Differentiable space carving for 3d reconstruction using imaging sonar," *IEEE Robotics and Automation Letters*, 2024.

[9] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[10] H. Wang, N. Anantrasirichai, F. Zhang, and D. Bull, "Uw-gs: Distractor-aware 3d gaussian splatting for enhanced underwater scene reconstruction," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 3280–3289.

[11] T. Zhang, W. Zhi, B. Meyers, N. Durrant, K. Huang, J. Mangelson, C. Barbalata, and M. Johnson-Roberson, "Recgs: Removing water caustic with recurrent gaussian splatting," *IEEE Robotics and Automation Letters*, 2024.

[12] Z. Qu, O. Vengurlekar, M. Qadri, K. Zhang, M. Kaess, C. Metzler, S. Jayasuriya, and A. Pediredla, "Z-splat: Z-axis gaussian splatting for camera-sonar fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[13] M. Zwicker, H. Pfister, J. Van Baar, and M. Gross, "Ewa volume splatting," in *Proceedings Visualization, 2001. VIS'01.* IEEE, 2001, pp. 29–538.

[14] P. S. Heckbert, "Fundamentals of texture mapping and image warping. master's thesis," *University of California, Berkeley*, vol. 2, no. 3, 1989.

[15] M. Zwicker, H. Pfister, J. Van Baar, and M. Gross, "Surface splatting," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 371–378.

[16] D. Yang, J. J. Leonard, and Y. Girdhar, "Seasplat: Representing underwater scenes with 3d gaussian splatting and a physically grounded image formation model," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[17] H. Li, W. Song, T. Xu, A. Elsig, and J. Kulhanek, "WaterSplatting: Fast underwater 3D scene reconstruction using gaussian splatting," *3DV*, 2025.

[18] E. Westman, I. Gkioulekas, and M. Kaess, "A volumetric albedo framework for 3d imaging sonar reconstruction," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9645–9651.

[19] M. D. Aykin and S. Negahdaripour, "Three-dimensional target reconstruction from multiple 2-d forward-scan sonar views by space carving," *IEEE Journal of Oceanic Engineering*, vol. 42, no. 3, pp. 574–589, 2016.

[20] S. Negahdaripour, V. M. Milenkovic, N. Salarieh, and M. Mirzargar, "Refining 3-d object models constructed from multiple fs sonar images by space carving," in *OCEANS 2017-anchorage*. IEEE, 2017, pp. 1–9.

[21] A. V. Sethuraman, M. Rucker, O. Bagoren, P.-C. Kung, N. N. B. Amutha, and K. A. Skinner, "Sonarsplat: Novel view synthesis of imaging sonar via gaussian splatting," 2025. [Online]. Available: https://arxiv.org/abs/2504.00159

[22] S. Pan, Z. Hong, Z. Hu, X. Xu, W. Lu, and L. Hu, "Russo: Robust underwater slam with sonar optimization against visual degradation," *IEEE/ASME Transactions on Mechatronics*, 2025.

[23] A. Cardaillac and M. Ludvigsen, "Camera-sonar combination for improved underwater localization and mapping," *IEEE Access*, vol. 11, pp. 123 070–123 079, 2023.

[24] M. Babaee and S. Negahdaripour, "3-d object modeling from 2-d occluding contour correspondences by opti-acoustic stereo imaging," *Computer Vision and Image Understanding*, vol. 132, pp. 56–74, 2015.

[25] M. Qadri, K. Zhang, A. Hinduja, M. Kaess, A. Pediredla, and C. A. Metzler, "Aoneus: A neural rendering framework for acoustic-optical sensor fusion," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–12.

[26] D. G. Merrill and A. S. Grimshaw, "Revisiting sorting for gpgpu stream architectures," in *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*, 2010, pp. 545–546.

[27] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5470–5479.

[28] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, 2003, pp. 1398–1402 Vol.2.

[29] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[30] E. Potokar, S. Ashford, M. Kaess, and J. G. Mangelson, "Holoocean: An underwater robotics simulator," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 3040–3046.

[31] E. Potokar, K. Lay, K. Norman, D. Benham, T. B. Neilsen, M. Kaess, and J. G. Mangelson, "Holoocean: Realistic sonar simulation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 8450–8456.

[32] K. Zhang, Y. Li, J. Liang, J. Cao, Y. Zhang, H. Tang, D.-P. Fan, R. Timofte, and L. V. Gool, "Practical blind image denoising via swin-conv-unet and data synthesis," *Machine Intelligence Research*, vol. 20, no. 6, pp. 822–836, 2023. [Online]. Available: https://doi.org/10.1007/s11633-023-1466-0