2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

A Fiducial Marker System for ID Recognition in Forward-Looking Sonar Images*

Yixue Zhu[1], Yusheng Wang[2], Hiroshi Tsuchiya[3], Makoto Hiraoka[3], Qi An[1] and Atsushi Yamashita[1]

***Abstract*— We present a fiducial marker system tailored for underwater acoustic imaging, enabling accurate detection and recognition of multiple marker IDs in real-world Forward-Looking Sonar (FLS) images. The marker is physically designed with layered concrete–metal structure to generate strong and distinctive sonar reflections. Our marker detection and recognition pipeline is trained entirely on simulation data, yet it achieves accurate performance on real-world sonar images. By leveraging a custom FLS simulator we generate annotated training samples that closely mimic real sonar characteristics. A YOLO-based detector, trained with these simulated images, localizes markers and regresses corner keypoints. For marker identity recognition, detected regions are rectified and decoded using a grid-based binary recognition scheme. Experiments show that the model achieves a 86.6% true positive detection rate and 100% ID recognition accuracy in the fully visible patch subset of real sonar images, despite being trained solely on synthetic data. This sim-to-real framework offers a scalable solution for underwater localization and inspection in autonomous robotic systems.**

I. Introduction

Underwater sensing presents unique challenges due to strong light attenuation and scattering in turbid and low-light environments. While optical cameras are effective in terrestrial and shallow-water settings, their usability significantly diminishes in deep-sea or murky conditions due to limited resolution and visibility range. In contrast, acoustic sensors—particularly Forward-Looking Sonar (FLS)—enable robust image acquisition based on sound wave reflections, operating reliably even in complete darkness or sediment-rich waters. These sensors have been widely applied in underwater exploration, infrastructure inspection, and autonomous navigation [1]–[4].

In environments where GPS is unavailable and natural features are unreliable, fiducial markers serve as vital anchors for underwater robots by providing discrete, recognizable visual patterns. Marker systems like AprilTag [5] and DeepTag [6] have enabled reliable ID recognition and robot localization in optical domains. Inspired by this success, researchers have explored fiducial marker systems for sonar images. However, marker recognition in sonar remains a challenging problem due to low spatial resolution and speckle noise [7], [8]. To facilitate controlled evaluation, sonar simulators such as ACSim [9] can generate realistic synthetic sonar images.

*This work was partially supported by JSPS KAKENHI under Grant Numbers 23K19993, 24K20867, and 25H01133, as well as JST PRESTO under Grant Number JPMJPR2512.

[1] Graduate School of Frontier Sciences, The University of Tokyo.
[2] Graduate School of Engineering, The University of Tokyo.
[3] Wakachiku Construction Co., Ltd.

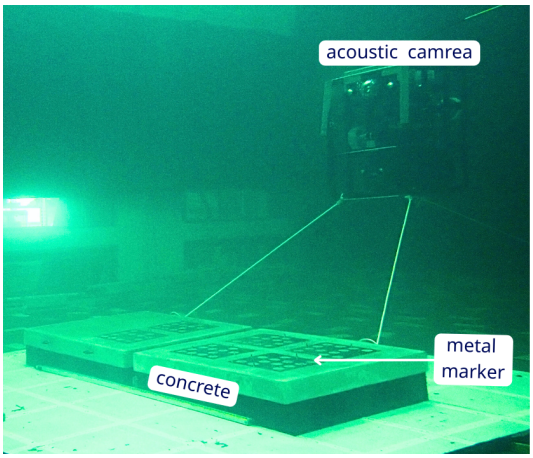

Fig. 1. Overview of the acoustic camera water tank environment setup. An acoustic camera is suspended in water and looks toward the target. Concrete blocks provide a stable base, and metal fiducial markers are mounted on top of concrete blocks.

While some prior systems, such as ACMarker [10] and AcTag [11], have introduced sonar-compatible marker designs, their recognition performance in real-world scenarios often suffers from environmental noise and reflection inconsistencies, frequently yielding true positive rates below 50%. Moreover, existing efforts largely emphasize detection or coarse localization, leaving the ID recognition task underexplored. Yet in many practical applications—such as region tagging, structural inspection, and waypoint verification—reliable marker ID recognition itself is the key to system function, especially when markers are deployed with known spatial semantics (e.g., “ID1 is placed at the entrance”).

To address this gap, we propose a FLS-specific fiducial marker system designed for ID recognition in underwater conditions. Figure 1 illustrates the experimental underwater environment used in this study. An acoustic camera is suspended in water, observing a set of concrete blocks on which our metal fiducial markers are mounted. Instead of focusing solely on marker detection, our system emphasizes accurate classification of marker identities (IDs) especially in the case of fully visible markers, enabling downstream applications such as semantic region identification and task triggering.

Specifically, this work contributes:

- A multi-ID fiducial marker design, using layered metal-concrete structures optimized for sonar reflectivity and distinctiveness.

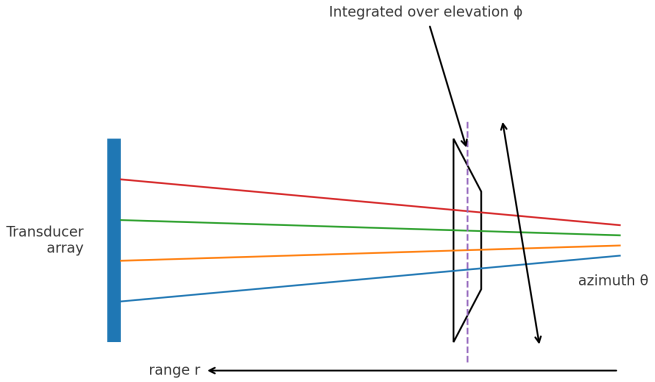

Fig. 2. Principle of Forward-Looking Sonar (FLS) imaging. The transducer array emits acoustic pulses over azimuth θ and elevation ϕ, and the returned backscatter strength is measured at slant range r along each beam. Colored rays illustrate beams at different angles, while the black shape is a reflecting underwater object. The measured ranges form an r–θ 2D sonar image, which can then be transformed into Cartesian coordinates for further analysis.

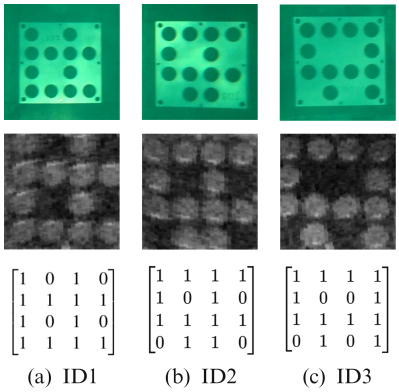

Fig. 3. Visual and acoustic appearances of three kinds of fiducial markers. From top to bottom: Optical camera images, acoustic camera images, and their binary ID code.

- A marker recognition pipeline trained entirely on simulated sonar data, capable of detecting markers, predicting corner keypoints, extracting affine-normalized patches, and recognizing IDs on real FLS images.

Our system achieves high marker ID recognition accuracy in real-world water tank experiments for the case of fully visible markers, laying a foundation for deployable marker-based underwater recognition systems in sonar imaging environments.

II. Methodology

A. Marker Design

As shown in Fig. 2, the FLS forms a 2D polar-coordinate image $I(r,\theta)$ by emitting acoustic pulses over a range of azimuth angles θ and elevation angles ϕ, and then integrating the received backscatter strength $BS(r,\theta,\phi)$ over the elevation direction. Here, r denotes the slant range from the transducer to the target, θ specifies the azimuthal bearing, and ϕ represents the vertical elevation angle. This integration collapses the 3D backscatter field into a 2D slice, producing the polar-form sonar image information that serves as the input for subsequent detection and recognition stages.

As illustrated in Fig. 3, the proposed fiducial marker is designed specifically for underwater sonar imaging. It employs a layered structure composed of a concrete base and a stainless steel top plate perforated with circular holes. The concrete provides a diffuse acoustic scattering background, while the metal layer ensures strong and consistent echo returns from the hole patterns.

The entire marker measures $0.35\,\mathrm{m} \times 0.35\,\mathrm{m} \times 0.2\,\mathrm{mm}$. Each marker encodes a unique 4×4 binary ID through the arrangement of 16 holes, where the presence or absence of a hole corresponds to binary values. The diameter of each hole is $0.06\,\mathrm{m}$, with a center-to-center spacing of $0.022\,\mathrm{m}$.

As shown in the middle panels of Fig. 3, these markers are clearly visible in sonar views. In the middle row of sonar images, it can be observed that the regions immediately below the circular holes of the markers appear brighter compared to the areas above. This phenomenon arises because the FLS transmits acoustic waves toward the marker at an oblique angle, and the finite thickness of the marker plate causes secondary or more reflections from the lower edges of the holes. Such effects can be reproduced in simulation, enabling the generation of synthetic sonar images that exhibit similar bright regions around the circular holes. The bottom panel of Fig. 3 presents the corresponding ID layout of each marker, highlighting the structured nature of the binary code embedded in each marker. These layouts will serve as reference templates for marker recognition in subsequent stages of the processing pipeline.

B. Detection Framework with Corner Points Prediction

We propose a simulation-trained detection framework that simultaneously performs bounding box localization and corner keypoint regression.

As illustrated in Fig. 4, although the FLS software directly outputs a blue-colored sonar image, we instead make use of the raw information from the FLS, which contains (r,θ) coordinates. These values are obtained from the emission of acoustic beams and represent the measured range r and azimuth angle θ. The raw image is inherently in grayscale, as pixel values encode acoustic reflection intensity rather than RGB color as in optical cameras.

To correct the curvature distortion introduced by the (r,θ) sampling geometry, the raw polar image is converted into a Cartesian (x,y) representation. This transformation rectifies the geometric deformation caused by polar sampling in the acoustic camera.

The detection network operates on this corrected (x,y) domain, identifying bounding boxes for each marker and regressing their four corner keypoints. The model is trained using simulated sonar images and evaluated on real-world water tank sonar images.

Training images are generated using a FLS simulator [9], which we further customized to automatically generate four-corner keypoint annotations for each marker based on the FLS projection geometry. The simulation pipeline also

incorporates sonar-specific phenomena such as secondary reflections caused by internal bouncing within the marker structure, as well as Poisson and Rayleigh noise to emulate underwater acoustic interference. Each synthetic image contains multiple markers with varying poses, backgrounds, and degrees of occlusion. These customizations help to bridge the sim-to-real domain gap by encouraging the model to learn robust features and avoid overfitting to clean, idealized data.

The detector is based on a You Only Look Once (YOLO) style backbone [12] and jointly predicts a bounding box and four corner keypoints for each detected marker. Ground truth keypoints are computed from simulator-exported world coordinates. For a 3D marker corner point $\vec{p}_0$ in the global coordinate system, its coordinate in the acoustic camera coordinate system $\vec{p}$ is given by:

$$\vec{p} = \mathbf{R}^\top(\vec{p}_0 - \mathbf{T}), \tag{1}$$

where $\mathbf{R}$ and $\mathbf{T}$ denote the acoustic camera's extrinsic parameters. Figure 5 illustrates this coordinate transformation from the global frame (orange) to the acoustic camera frame (purple), where $\vec{p}_0$ is first translated by $\mathbf{T}$ and then rotated by $\mathbf{R}^\top$ to obtain $\vec{p}$ in the acoustic camera frame. Through the above computation, the ground-truth (r,θ) coordinates of the four marker corners in the simulated images are obtained. These polar coordinates are subsequently transformed into Cartesian (x,y) coordinates, enabling direct integration with the YOLO-style object detection label format. In this representation, each training instance is specified by a class label cls, a bounding box $\text{bbox} = (x_{\text{center}}, y_{\text{center}}, w, h)$, and an associated set of keypoints $\text{pts} = \{(x_i, y_i, v_i)\}_{i=1}^{4}$ describing the four marker corners, where v_i is a binary visibility flag. The coordinates of both bbox and pts are normalized with respect to the full image dimensions, such that all values lie in the interval $[0,1]$.

In the detection stage, we treat all markers as belonging to a single object category labeled as *marker*, regardless of their individual IDs. This enables us to formulate the task as a single-class object detection problem, where the objective is to localize the presence of any valid fiducial marker and regress its associated four corner keypoints.

C. Marker ID Recognition via Patch Rectification

As illustrated in Fig. 4, after detecting the four corner points of multiple markers in the sonar image, each marker is individually rectified using its detected corner coordinates. The objective of this rectification is to transform the marker region, which is generally non-square in the original image, into a square patch, as the markers are originally designed to be square plate. This square normalization facilitates the subsequent ID recognition process.

For the marker patch example shown in Fig. 4, we take the top-right detected marker as a representative case to illustrate the recognition flow. The resulting rectified marker patch is approximately square. The slight residual distortion—appearing, for example, as a mild parallelogram shape with the lower-left circular feature not fully included—is primarily due to small localization errors in the detected corner points in prior detection step. Nevertheless, such rectified patches are sufficiently accurate for reliable ID recognition. The following content details the homography matrix computation used for rectification and the methodology for marker ID recognition.

Once bounding boxes and corner keypoints are predicted, we extract rectified patches using affine homography. These corner points serve as geometric anchors for a subsequent affine homography transformation, which maps the distorted sonar view of the marker into a normalized, front-facing rectangular patch.

Negahdaripour [13] proposed a generalized sonar homography formulation that enables point-wise mapping between acoustic views without restrictive pose assumptions. The homography matrix $\mathbf{H}$ is derived from the sonar projection geometry and rigid body transformation. Unlike optical homographies, this matrix incorporates angular scaling and non-linear projection terms specific to acoustic imaging. Notably, $\mathbf{H}$ is not constant across the image plane—it varies from point to point depending on the incidence angle and local surface normal, reflecting the geometry of slanted sonar beams interacting with 3D underwater structures.

In our application, we focus on fiducial marker ID recognition rather than dense mosaicking or exact 3D reconstruction. As such, we make the following practical approximations:

- The elevation angle ϕ of the acoustic beams is relatively small, and for the ARIS 3000 FLS used in this study it ranges from -7° to $+7^\circ$. Over the limited spatial extent of a single marker, ϕ can be regarded as locally constant.
- The acoustic projection parameters can be regarded as constants across a single marker instance.

Under these assumptions, the full projective homography $\mathbf{H}$ becomes spatially invariant over the marker region and can be well-approximated by a standard 2D affine transformation. Let $\mathbf{p}_i = (x_i, y_i)^\top$ and $\mathbf{p}'_i = (x'_i, y'_i)^\top$ denote the coordinates of the i-th corner point of the marker in the source and target images, respectively. The affine mapping is expressed as:

$$\mathbf{p}'_i = \underbrace{\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}}_{\mathbf{A}\ \text{(linear part)}} \mathbf{p}_i + \underbrace{\begin{bmatrix} t_x \\ t_y \end{bmatrix}}_{\mathbf{t}\ \text{(translation)}}, \tag{2}$$

where $\mathbf{A}$ encodes rotation, uniform or non-uniform scaling, and shear, while $\mathbf{t} = (t_x, t_y)^\top$ represents translation. The parameters $\mathbf{A}$ and $\mathbf{t}$ are estimated from four matched marker corner keypoints between the source and target patches. This affine approximation is sufficient to geometrically normalize the marker patch prior to ID decoding, and is computationally efficient under moderate viewpoint variation.

The resulting rectified patch isolates a single marker instance from the sonar image via affine transformation. To ensure reliable recognition, we first discard any patches containing large pure-black regions, which typically result from partial visibility where parts of the marker fall outside the sonar field of view. This filtering step ensures that only complete and interpretable marker regions proceed to the ID recognition phase.

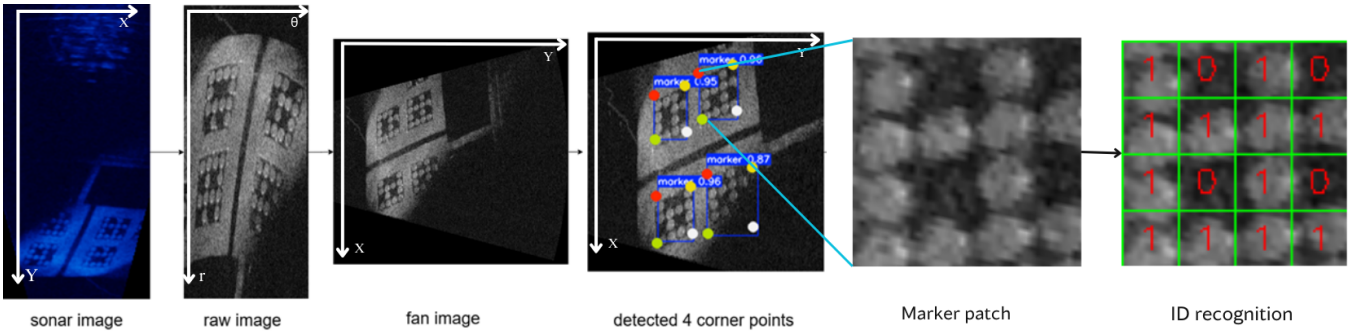

Fig. 4. Overview of the marker detection and ID recognition pipeline for forward-looking sonar images. From left to right: (1) **FLS image** taken in Cartesian (x, y) coordinates, (2) **Raw image** in polar (r, θ) coordinates obtained from FLS raw information, (3) **Fan image** after geometric transformation from raw image back to (x, y), (4) Detection results of four marker corner points with confidence scores, (5) Extraction of the rectified marker patch, and (6) Binary ID recognition by decoding the grid pattern.

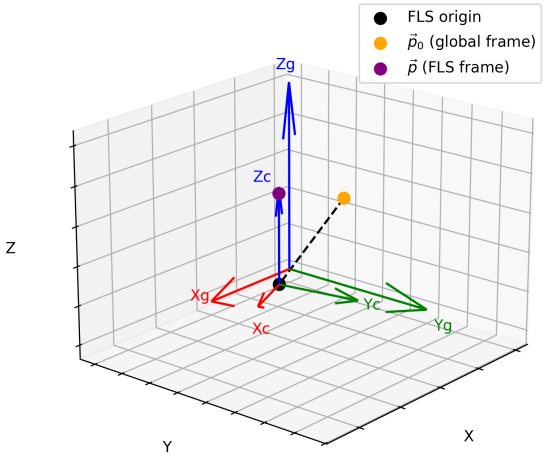

Fig. 5. Coordinate transform from the global frame to the acoustic camera frame. The acoustic camera origin is shown in black; the global frame axes are drawn as Xg, Yg, Zg, and the camera frame axes are drawn as Xc, Yc, Zc). A marker corner point $\vec{p}_0$ (orange) in the global frame is translated by $\mathbf{T}$ and then rotated by $\mathbf{R}^\top$ to obtain its coordinates $\vec{p}$ (purple) in the acoustic camera frame.

The valid patch is then partitioned into a uniform grid that matches the binary structure physically embedded in the marker's hole pattern. Each grid cell is analyzed by applying a fixed intensity threshold to its sonar echo response. Cells with intensities above the threshold are interpreted as binary 1, representing strongly reflective perforated (hole) regions. Conversely, cells with lower intensities are classified as binary 0, corresponding to flat stainless steel regions that reflect less predictably. The resulting binary matrix is then decoded into a digital ID.

To recognize the final marker ID, each valid patch is first decoded into a binary matrix, following the procedure described above. The predicted binary code is then compared against all predefined ID templates illustrated in Figure 3 using the Hamming distance. The ID with the smallest Hamming distance is assigned to the patch. This approach enables reliable recognition by selecting the most plausible match from a fixed ID dictionary. This strategy allows the system to recognize multiple unique marker IDs within a single sonar image.

III. Experiment

We evaluate the proposed fiducial marker system in real-world underwater tank scenarios, focusing on marker detection rate and marker ID recognition accuracy under FLS imaging.

A. Marker Detection

Synthetic sonar images were generated using Blender 3.6.11 with a customized sonar rendering pipeline. The simulated scenes replicate rough concrete environments with variably positioned markers of varying IDs. Sonar-specific phenomena were modeled, including beam divergence, secondary reflections, and Poisson–Rayleigh noise, to reduce the sim-to-real gap. Ground-truth bounding boxes and four-corner keypoints were automatically generated in the (r, θ) domain and transformed to (x, y) coordinates.

A YOLOv11n-based detection model was trained to jointly predict bounding boxes and four corner keypoints for each marker instance. The training dataset consists of 2,000 synthetic sonar images. The model was trained for 100 epochs exclusively on this simulated dataset, without any real sonar images involved during training. In the labeling strategy, all fiducial markers—regardless of their individual ID codes—were annotated as a single object class labeled *marker*.

The input images for marker detection are fan-shaped (x, y) coordinate sonar images, as illustrated in Fig. 4. These images are generated using a forward-looking sonar simulator, which produces both the simulated sonar images and the corresponding ground-truth marker positions. The associated YOLO-format labels are automatically computed from the simulator's ground-truth corner coordinates of each marker, ensuring annotation accuracy without manual labeling.

When tested on a real-world tank dataset of 1,240 sonar frames, the model achieved strong generalization. As summarized in Table I, the system detected 1,074 true marker instances with an average confidence score of 0.93, resulting in a true positive detection rate of 86.6%. Only 166 true

instances were missed, primarily due to extreme occlusion or poor contrast. No false positives were observed, demonstrating high precision.

The experiments were conducted on an NVIDIA GeForce RTX 4090 GPU. The end-to-end pipeline is efficient: per-frame latency was 0.8 ms for preprocessing, 4.5 ms for inference, and 1.2 ms for postprocessing, supporting real-time applications.

B. Marker ID Recognition

The input to the ID–recognition stage consists of approximately square patches of single marker, as shown in Fig. 4. These patches are extracted from real sonar frames that are first converted from raw images into fan–shaped (x,y) coordination images. Marker detection is then performed in this (x,y) domain fan images; for each detected markers, the four corner keypoints are used to estimate an affine transform that rectifies the original, skewed marker region into a normalized, near–square patch. This rectification removes most of the projective distortion and yields a canonical layout of the circular holes, which is essential for accurate template matching.

In our real-world sonar image experiments, recognition is evaluated on patches from three distinct marker IDs. To avoid label ambiguity, we restrict the evaluation set to *fully visible* markers: any patch with occlusions or truncation is discarded beforehand. Operationally, patches exhibiting black borders (indicative of partially captured markers due to the field-of-view limits) are filtered out. In our water-tank experiments, conducted in a $10\,\mathrm{m} \times 10\,\mathrm{m}$ water tank, markers were rarely physically occluded by other objects; however, due to viewpoint constraints, some markers appeared incomplete within the sonar frame. We treat such incomplete captures as partial occlusions, since portions of the marker pattern are missing from the image. Only patches containing fully visible markers, with all pattern elements intact, proceed to the recognition stage.

Each valid patch is converted to a binary code by the following deterministic procedure. The grayscale patch is resized to a fixed resolution (350×350 in our implementation, considering that the actual marker size is $0.35\,\mathrm{m} \times 0.35\,\mathrm{m}$) and binarized.

The binarized image is then partitioned into a 4×4 grid; for each cell (i,j), the mean binary value is computed and mapped to a bit

$$b_{ij} \in \{0,1\} \quad \text{with} \quad b_{ij} = \begin{cases} 1, & \text{if mean(cell)} > \tau, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where τ is a fixed intensity threshold applied to the binarized image. In our experiments, we set $\tau = 50$, selected empirically after evaluating ten candidate values in the range of 0–100 and choosing the one that yielded the most reliable detection results for our dataset. For images captured under different environmental conditions, the threshold can be readjusted accordingly to accommodate variations in illumination and acoustic characteristics.

Recognition is performed by template matching with Hamming distance of a 4×4 binary matrix $\mathbf{B}$ yields by previous steps. As shown in Fig. 3, the Hamming distances between the three marker IDs are $d(\mathrm{ID1},\mathrm{ID2}) = 8$, $d(\mathrm{ID1},\mathrm{ID3}) = 8$, and $d(\mathrm{ID2},\mathrm{ID3}) = 4$. Let $\{\mathbf{T}^{(k)}\}$ denote the predefined ID templates (Fig. 3); for each candidate ID k, we compute

$$d_H\left(\mathbf{B},\mathbf{T}^{(k)}\right) = \sum_{i=1}^{4}\sum_{j=1}^{4}\left[\mathbf{B}_{ij} \neq \mathbf{T}_{ij}^{(k)}\right], \tag{4}$$

and select the ID with the minimum distance, $\hat{k} = \arg\min_k d_H(\mathbf{B},\mathbf{T}^{(k)})$. We also record the distribution of Hamming distances over all evaluated patches to quantify the margin to the nearest competing template.

Table II reports accuracy over all *valid* (fully visible) patches; example visualizations of the grid overlay and predicted bits are provided in Fig. 4. This design produces deterministic, reproducible decisions and provides an interpretable error metric (Hamming distance) that correlates with recognition reliability.

C. Discussion

The experimental results affirm the effectiveness of our simulation-trained sonar marker system, yet also reveal several important insights and future directions.

The quality of ID recognition depends on the accuracy of the detected corner keypoints, since the rectified patch is generated via homography. Small localization errors can distort the binary layout and reduce threshold-based recognition reliability. Furthermore, although the detector achieved a true positive detection rate of 86.6%, some true marker instances were missed (166 out of 1,240 images). These false negatives were primarily due to partial visibility near image borders and background noise. To improve robustness, we also tested a data-driven alternative by training a YOLO classifier on rectified patches. However, preliminary results on ID1 achieved only 62 correct predictions out of 123 samples, indicating that learning-based methods require further tuning due to label imbalance and subtle sonar texture differences. Improving patch alignment and combining learned classifiers with template matching may strengthen overall recognition performance.

One known limitation of our recognition method is that it discards any markers that are partially visible in the sonar image. Real-world sonar images may capture markers that are truncated or occluded, producing incomplete regions with missing returns. If partial occlusion affects only a single bit at one corner, the built-in Hamming distance–based error correction of the ID coding scheme can still recover the correct marker ID. However, as occlusion increases—covering multiple bits or larger areas of the marker—the likelihood of correct classification drops rapidly, and decoding may fail. This explains why our reported 100% recognition rate is a property of the filtered, fully visible subset of patches, rather than a guarantee under all occlusion conditions. Furthermore, the current implementation assumes a fixed grid layout and predefined ID templates, which limits flexibility in dynamic ID generation.

TABLE I

REAL-WORLD MARKER DETECTION RESULTS

Category	Count	Total	Rate
True Positives (TP)	1074	1240	86.61%
False Negatives (FN)	166	1240	13.39%
False Positives (FP)	0	–	0.00%
Avg Confidence		0.93	

TABLE II

REAL-WORLD MARKER ID RECOGNITION ON FULLY VISIBLE PATCHES

Predicted ID	ID1	ID2	ID3
Image Count	496	248	248
Total Accuracy	100% (992/992)		

This filtering strategy directly motivates the design of our processing flow: when a marker is fully visible in the sonar image, the four detected corners allow us to extract a patch that preserves the complete marker geometry. This, in turn, enables reliable ID decoding with minimal risk of misclassification. In other words, the flow is optimized to exploit high-quality detections when they occur, ensuring that the downstream recognition stage operates under conditions where it can achieve maximal accuracy. This design choice allows the system to deliver dependable recognition results in practical deployments, rather than attempting uncertain predictions under degraded visibility.

For long-term deployment, the marker cannot be replaced frequently due to its concrete–metal structure and the cost of underwater installation. Therefore, material durability in seawater is a critical design factor. According to Francis et al. [14], conventional stainless steels such as 316L still exhibit measurable pitting in natural seawater, with reported pit depths on the order of 0.27 mm after exposure, whereas super duplex stainless steels such as Z100 showed no detectable pitting under the same conditions. This demonstrates that super duplex alloys provide substantially higher long-term corrosion resistance and are suitable for multi-year underwater installations. A further practical constraint is that the marker pattern must remain fully visible for reliable identification. In real environments, however, partial occlusion by algae or marine vegetation is often unavoidable. To maintain visibility, the sonar camera can be mounted on a robotic arm, enabling adaptive viewpoint adjustment or simple clearing motions when necessary.

In summary, the proposed framework provides a scalable and practical solution for sonar-based underwater marker recognition, and demonstrates the viability of simulation-only training for real-world deployment.

IV. CONCLUSION

In this work, we proposed a fiducial marker system designed specifically for underwater acoustic imaging. The marker features a layered metal–concrete structure with multiple distinguishable keypoints, offering strong reflectivity and geometric stability under sonar sensing. To enhance generalization across simulation and real-world domains, we trained a detection model using simulation-only data augmented with sonar-specific artifacts such as noise and secondary reflections.

Experimental results show that the method achieves a marker detection rate of 86.6% on real sonar images without false positives, and attains 100% marker ID recognition accuracy across valid patches. The absence of false positives ensures high reliability, especially in safety-critical environments. The proposed method supports real-time inference and offers a practical, scalable solution for underwater localization and inspection tasks in autonomous marine robotics.

Future work will concentrate on the precise recognition of inner keypoints embedded within each marker, aiming to provide a reliable reference structure for accurate 6-DoF pose estimation. This includes designing the keypoint detection network to achieve sub-pixel accuracy in sonar images, as well as improving the geometric consistency of detected keypoints under varying acoustic conditions, enabling robust and accurate localization for underwater robotic applications.

REFERENCES

[1] J. He, J. Chen, H. Xu, and Y. Yu, "Sonarnet: Hybrid cnn-transformer-hog framework and multifeature fusion mechanism for forward-looking sonar image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[2] I. Quidu, L. Jaulin, A. Bertholom, and Y. Dupas, "Robust multitarget tracking in forward-looking sonar image sequences using navigational data," *IEEE Journal of Oceanic Engineering*, vol. 37, no. 3, pp. 417–430, 2012.

[3] Y. Wang, Y. Ji, H. Woo, Y. Tamura, H. Tsuchiya, A. Yamashita, and H. Asama, "Acoustic camera-based pose graph slam for dense 3-d mapping in underwater environments," *IEEE Journal of Oceanic Engineering*, vol. 46, no. 3, pp. 829–847, 7 2021.

[4] Y. Wang, Y. Ji, H. Tsuchiya, J. Ota, H. Asama, and A. Yamashita, "Acoustic-n-point for solving 2d forward looking sonar pose estimation," *IEEE Robotics and Automation Letters*, vol. 9, no. 2, pp. 1652–1659, 2024.

[5] E. Olson, "Apriltag: A robust and flexible visual fiducial system," *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 3400–3407, May 2011.

[6] Z. Zhang, Y. Hu, G. Yu, and J. Dai, "Deeptag: A general framework for fiducial marker design and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2931–2944, 2023.

[7] H. Zhang, M. Tian, G. Shao, J. Cheng, and J. Liu, "Target detection of forward-looking sonar image based on improved yolov5," *IEEE Access*, vol. 10, pp. 18 023–18 034, 2022.

[8] M. Valdenegro-Toro, "End-to-end object detection and recognition in forward-looking sonar images with convolutional neural networks," in *2016 IEEE/OES Autonomous Underwater Vehicles (AUV)*, 2016.

[9] Y. Wang, Y. Ji, H. Tsuchiya, J. Ota, H. Asama, and A. Yamashita, "Acsim: A novel acoustic camera simulator with recursive ray tracing, artifact modeling, and ground truthing," *IEEE Transactions on Robotics*, vol. 41, pp. 2970–2989, 2025.

[10] Y. Wang, Y. Ji, D. Liu, Y. Tamura, H. Tsuchiya, A. Yamashita, and H. Asama, "Acmarker: Acoustic camera-based fiducial marker system in underwater environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5018–5025, 2020.

[11] K. Norman, D. Butterfield, and J. G. Mangelson, "Actag: Opti-acoustic fiducial markers for underwater localization and mapping," in *Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2023.

[12] Ultralytics YOLO, Version 11.0.0, 2024.

[13] S. Negahdaripour, "Visual motion ambiguities of a plane in 2-d fs sonar motion sequences," *Computer Vision and Image Understanding*, vol. 116, no. 6, pp. 754–764, 2012.

[14] R. Francis, G. Byrne, and G. Warburton, "The corrosion of superduplex stainless steel in different types of seawater," in *NACE International Annual Conference*, vol. CORROSION 2011, pp. 1–9, 2011.