2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Knowledge Optical To Sonar (KnOTS): Towards the Transfer of Knowledge of Underwater Object Detection from Optical to Forward-Looking Sonar Imagery

Caroline Keenan[1,2] Ella R. Wawrzynek[2] David Whelihan[2]

Ivy Mahncke[2,3] John J. Leonard[4] Madeline D. Miller[2]

***Abstract*— We develop an approach to detect objects in forward-looking sonar (FLS) images using corresponding optical images and without the need for expert manual labeling of sonar images. Sonar sensing is more robust to disadvantageous underwater environmental conditions than optical sensing, but the scarcity of labeled sonar data leads to decreased performance of methods which rely on an abundance of training data. We aim to transfer insights from data-rich applications such as object detection in optical imaging to the data-scarce area of object detection in sonar images. Our approach involves recording of contemporaneous images from commercially available sensors viable for use aboard unmanned underwater vehicles. We collect new optical and sonar data in a shallow, clear-water environment and employ existing object detection techniques for optical images. We leverage the commonality of the sensors' fields of view and our algorithmic processing of the sonar image to transfer knowledge of object bounding boxes to sonar images to create a dataset. Through this transfer, we enable training of a model that detects objects in unseen sonar images and does not require optical images as input at test time.**

I. Introduction

Object detection and localization underwater is a key step in advanced navigation and search algorithms for autonomous underwater vehicles (AUVs), such as simultaneous localization and mapping [9]. Large-scale datasets [12] and advanced algorithms enable high performance when detecting objects with optical cameras [18]. Optical cameras detect objects underwater, but the refraction of light results in objects only being detected at close distances from an optical camera. Additionally, when the water is turbid and has a large amount of sand or other particles, it may be difficult to see an object in imagery from an optical camera. In contrast, multibeam sonar is not affected this way by light or water turbidity. Since sound does not refract as much as light underwater, imaging sonars have longer ranges than optical cameras [4]. It is advantageous to leverage existing advancements for optical images while benefitting from the superior capabilities of sonar sensors underwater.

A. Previous Works

Previous works using sonar data for object detection and localization rely on either the presence of both optical and sonar sensor modalities at deployment, template images for object tracking, manual sonar image labeling, or simulated sonar images.

Raaj et. al [17] localize objects in 3D using an optical camera, a forward-looking sonar (FLS) sensor, and temporal information including vehicle odometry. The Spatial Cross-Attention Transformer Tracker [11] leverages optical and sonar images to track objects underwater, but relies on optical and sonar template images to track a desired object.

Advanced object detection algorithms such as You Only Look Once (YOLO) [18] and Cut-and-LEaRn (CutLER) [23] have been applied to optical images, even those collected underwater. Previous works achieve object detection in underwater optical images with YOLOv8 [25] and underwater optical image segmentation of objects with CutLER [19]. These algorithms do not achieve the same performance on sonar images without training with labeled sonar data. Works in object detection and classification using FLS images use expert-labeled datasets, which remain limited ([20], [8]). Preciado-Grivalva et al. [16] show success with self-supervised methods for sonar image classification, though their datasets include watertank images with only one object per image and wild images with no objects present.

Additional works simulate sonar images for training object detection and tracking methods ([6], [5], [11]), which enable large training datasets without requiring manual labeling. However, simulated training images do not contain the noise which is characteristic of real sonar images, leading to inferior performance when models trained with simulated images are deployed on real sonar images.

[1]Caroline Keenan is with the MIT-WHOI Joint Program in Oceanography/Applied Ocean Science and Engineering, Cambridge and Woods Hole, MA USA caroline.keenan@ll.mit.edu

[2]Caroline Keenan, Ella R. Wawrzynek, David Whelihan, Ivy Mahncke, and Madeline D. Miller are with the Advanced Undersea Systems and Technology Group, MIT Lincoln Laboratory, Lexington, MA 02421 USA ella.wawrzynek@ll.mit.edu, david.whelihan@ll.mit.edu, madeline.miller@ll.mit.edu

[3]Ivy Mahncke is with Olin College of Engineering, Needham, MA 02492 USA imahncke@olin.edu

[4]John J. Leonard is with the Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA jleonard@mit.edu

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Under Secretary of War for Research and Engineering under Air Force Contract No. FA8702-15-D-0001 or FA8702-25-D-B002. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of War for Research and Engineering.

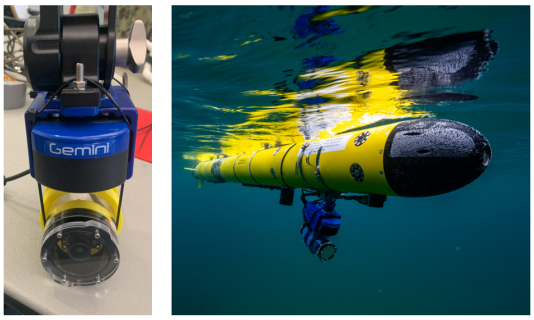

Fig. 1: The optical camera and FLS are mounted to face the same direction and affixed underneath an Iver3 AUV.

B. Our Contribution

We propose a novel algorithm, Knowledge Optical to Sonar (KnOTS), which efficiently produces a model to detect objects in real-world FLS images without requiring manual labeling of sonar images. By eliminating the need for human-annotated sonar datasets, our approach reduces demand for specialized expertise and accelerates model development. We employ self-supervision in the sensor fusion by leveraging an existing object detection framework for optical images and systematically creating labels for sonar images. Our approach is the first we know of to automatically create bounding box labels of objects in FLS images, enabling efficient collection of FLS datasets for future development in a domain with limited training data. A key element of our approach is leveraging the shared azimuthal dimension in the data collected by the carefully positioned optical and sonar sensors to map information across modalities without requiring platform navigation data, identical image capture times between sensors, or temporal association between successive frames from the same sensor.

To support the transfer of knowledge, we constructed a data collection platform with an optical camera and a FLS sensor. We used the dual-sensor system to collect imagery for model training and performance evaluation using a set of known objects. We integrated the system into a commercial AUV for further data collection in Eagle Harbor, Michigan and to demonstrate its suitability for real-world applications.

II. Data Collection

A. Sensor Configuration

We collected images with an Arducam Wide Angle M12 USB Camera [2] and a Tritech Micron Gemini 720s [22] multibeam FLS. The optical camera is affixed in a waterproof housing underneath the imaging sonar and oriented in the same direction, as shown in the left in Fig. 1. The FLS operates at 720 kHz. We set the range to 20 m. We affix the sensors underneath an Iver3 AUV with 3D printed fixtures and band clamps, as shown on the right in Fig. 1.

A schematic of the sensor payload connection is shown in Fig. 2. The battery section of the Iver3 supplies 5V power and 24V power to the extended payload (EP) section. The optical and sonar sensors are controlled by a Raspberry Pi 4 as an adjunct processor in the EP section. The processor is powered with 5V. The optical camera is controlled and powered through a USB connection to the processor. The imaging sonar is controlled via an ethernet connection to the processor and powered with 24V. To pass through the AUV hull, the cabling for both sensors was spliced into a universal wet-mate connector from Teledyne Marine [14].

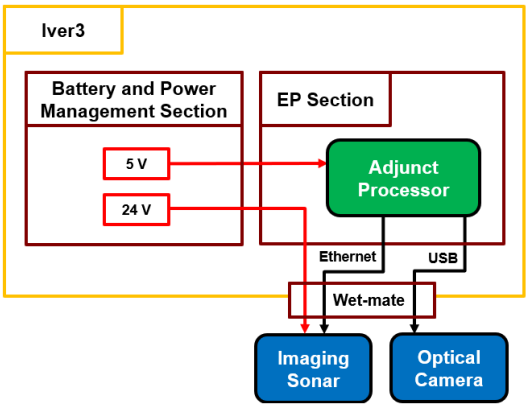

Fig. 2: The sensors are connected to the extended payload section of the Iver3 through a wet-mate connector in the hull. The optical camera is connected via USB to the adjunct processor. The FLS is connected via ethernet to the adjunct processor and powered with 24V from the battery section. The adjunct processor and thus the optical camera are powered with 5V from the battery section.

The Iver3 was deployed in Eagle Harbor, Michigan. The AUV autonomously traverses mission waypoints, recording optical frames at a rate of 20 Hz and sonar frames at a rate of around 10 Hz. The frame rate of sonar recording varies between 7.5-11Hz. Variability of sound speed and absorption underwater affects the time it takes for sound to return to the FLS, causing an inconsistent frame rate. For each sonar frame that is received, the optical frame with the closest timestamp is found. On average, the difference between the timestamps is 0.0133 seconds. The images are treated as a pair for object detection and localization.

B. Objects Imaged

We deploy two moorings. Each mooring consists of a float, an object, and a weight, tied together as shown on the left in Fig. 3. We use a metal stock pot (0.4 m $\times$ 0.4 m $\times$ 0.3 m) and an aluminum pan (0.5 m $\times$ 0.3 m $\times$ 0.1 m) as objects. We chose these objects since they are metal, and thus highly reflective of light and sound and visible in optical and sonar images. Future testing should include a larger number of objects typically found in a marine setting to test practical applicability. During imaging, the moorings are placed 5-10 m apart, as depicted on the right in Fig. 3. The objects hang 2-3 m below the surface float. The moorings are deployed at locations with a water depth of about 3.5 m. The AUV travels at 0-1 m depth both between and around the moorings, capturing diverse views of the objects.

Additionally, our team included multiple divers in the water near the moorings. We included the divers as objects

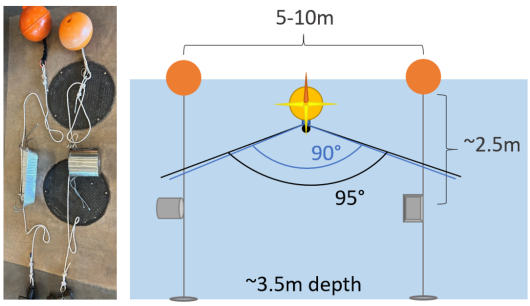

Fig. 3: Each mooring consists of a float, an object, and a weight. The objects were positioned in the view of the sonar (blue 90° field of view) and camera (black 95° field of view) which were angled 45° downward on the traveling AUV.

for detection and localization. Thus the classes of objects in the images are "person", "pot", and "pan".

C. The Need for Manual Image Labeling

We consider the YOLOv11 object detection model pretrained on the Common Objects in Context (COCO) dataset [12] which contains over 200,000 labeled optical images. The COCO-pretrained model achieves a mean average precision at 0.5 intersection over union (mAP50) of 0.195 on the manually labeled optical images. The COCO dataset includes the class "person", but not "pot" or "pan". On the "person" class in our dataset, the COCO-pretrained model achieves a mAP50 of 0.527. Performance is likely affected by the fact that the people are in diving gear and the images are underwater. The COCO training set does not focus on underwater images, in which refraction, absorption, and scattering of light cause degradation [13]. Since the COCO dataset does not include pots and pans, the COCO-pretrained model does not achieve any correct detections of the pot and the pan. Thus, it is necessary to train a YOLOv11 model to recognize our objects underwater.

To do so, we manually labeled optical images. We used Label Studio [21] to draw bounding boxes around the objects in the optical images and assign classification labels. We labeled only 727 of 127,880 recorded optical images. Though not the focus of this work, one can reduce the need for this manual labeling through existing self-supervised object detection algorithms for optical images.

In order to evaluate the accuracy of the automatically generated bounding boxes in sonar images, we manually labeled 1023 sonar images. We used Label Studio to draw bounding boxes around objects in the sonar images while looking at the corresponding optical images for accuracy. Through this work, we aim to eliminate the need for manual labeling to accurately detect objects in sonar images.

D. Details on Field Testing

In the field test, we demonstrate real-time performance while simultaneously collecting a real-world dataset for continued algorithm development and testing, as evaluated below. After a few hours of optical image collection, we train a YOLOv11 model pretrained on the COCO dataset to recognize our objects. We verify that the retrained model detects objects in optical images in real time. The computation of automatic bounding boxes in a sonar image from the corresponding optical image occurs at a rate of 12 image pairs per second on a Raspberry Pi 4 on an Iver3 AUV. The computation can also be done offline on previously collected data. The details of the algorithm are discussed below.

III. Methods

To transfer object detection knowledge from optical to sonar images, processing of optical and sonar images occurs as shown in Fig. 4. Each step is further discussed below.

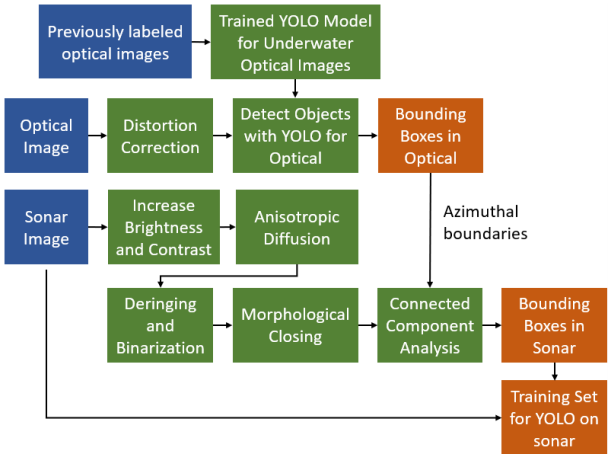

Fig. 4: To transfer knowledge, we first train a YOLO model for optical images, detect objects in optical images and save azimuthal boundaries of the bounding boxes. We process sonar images by increasing brightness and contrast and applying anisotropic diffusion, deringing and binarization, and morphological closing. We use the saved optical azimuthal boundaries in connected component analysis of the processed sonar image. Thus we find bounding boxes in sonar images to create a training set for a YOLO model for sonar images.

A. Optical Image Processing

Bounding boxes are predicted for objects in optical images through a YOLOv11 model [18]. Starting from the YOLOv11 model pretrained on the COCO dataset, we trained the model to recognize our objects in optical images, with 533 training images and 194 validation images. Despite the small training dataset size, the resulting model has 0.985 mAP50 as evaluated on manual labels of all classes. In the confusion matrix in Fig. 5, we see one confusion between "pot" and "pan", and rare false negatives and false positives, as detailed in the "background" row and column.

The pixels across the horizontal axis of the optical image do not correspond to linear spacing throughout the horizontal field of view due to the radial distortion of the camera. With an image of a checkerboard filling the entire camera field of view, we correct the distortion using OpenCV [3] camera calibration. We obtain optical images which better match the true layout of the scene, as shown in Fig. 6.

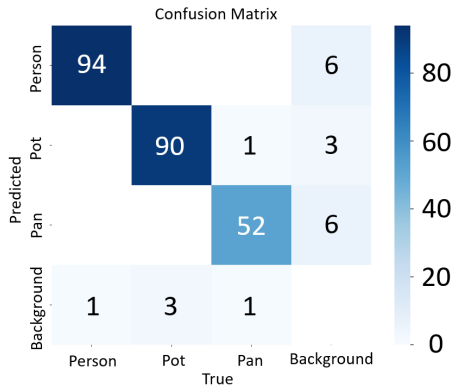

Fig. 5: A confusion matrix showing performance on the manually labeled optical dataset, where counts are the number of predictions.

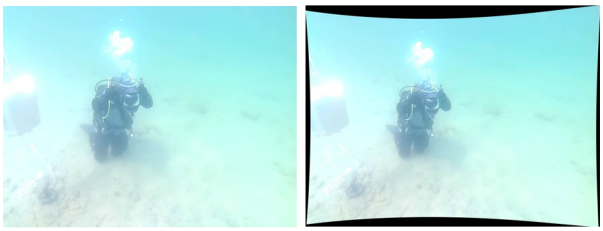

Fig. 6: An optical dataset image before and after correction of radial distortion using OpenCV camera calibration.

B. Sonar Image Preprocessing

The sonar frames are saved in a polar coordinate system, where the rows correspond to ranges from the sensor, and the columns correspond to azimuths across the field of view of the sensor. Each sonar image is preprocessed to determine pixels which correspond to sonar returns from surfaces in the scene. This process involves increasing brightness and contrast of the image, dampening artifacts through anisotropic diffusion and deringing, and identifying pixels corresponding to surfaces through thresholding. To illustrate the sonar image preprocessing steps, we show the sonar image corresponding to the optical image in Fig. 6. To better show the objects, the sonar image brightness is increased, contrast is decreased, and the image is cropped to include ranges 1.9 m to 5.1 m from the sonar sensor.

We increase the brightness and contrast of the image each by 50%, as shown in Fig. 7. Increasing brightness shifts all pixel values closer to the maximum white value of 255, making pixels corresponding to objects more concentrated near 255, as any values above 255 are clipped to 255. This step ensures that weaker sonar returns of objects have similar pixel values as stronger sonar returns of objects. Increasing contrast makes the dark pixel values darker and the light pixel values lighter. This increases the pixel value difference between the areas with no objects and the areas with objects.

We use anisotropic diffusion [15] on the sonar image, as shown in Fig. 8, to preserve object edges and apply a Gaussian blur. Westman and Kaess [24] and Archieri et al. [1] discuss using this method to denoise a sonar image.

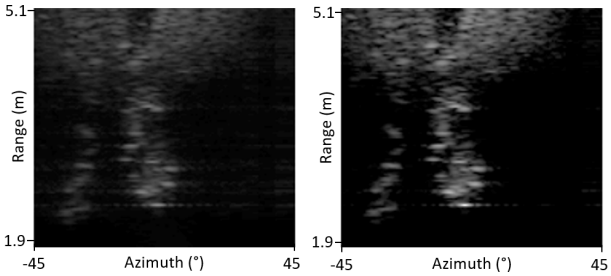

Fig. 7: The sonar image before and after increasing brightness and contrast to increase the difference in pixel value between objects and empty space.

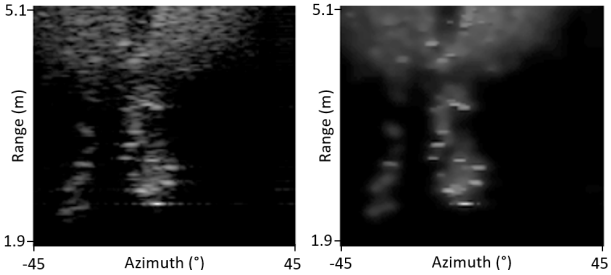

Fig. 8: The brightened sonar image before and after anisotropic diffusion to denoise with a Gaussian blur while preserving object edges.

Sonar images are prone to background noise. As sound propagates underwater, it bounces off small particles, causing speckle noise in the sonar image. Additionally, sound bounces off objects at multiple angles. The sound returns to the sonar sensor through multiple paths which take differing times to traverse, causing noise throughout the sonar image. When there is a sonar return at a certain azimuth and range, there are additional, lesser returns at surrounding azimuths of that same range. This effect can be seen in the sonar image on the right side of Fig. 7 at the bottom of the person near the center of the image. There is a pattern of repeated sonar returns at azimuths on either side of strong sonar return of the person. To reduce the effect of noise on identification of object range, each sonar image is processed to determine which pixels are occupied by an object. This removal of artifacts is often called deringing and is an important step in 3D reconstruction methods which use imaging sonars [7].

To compute the sonar image background noise, we assume the ranges corresponding to less than 0.4 m from the sensor are empty. This is a reasonable assumption for use of sonar sensors on small AUVs of length 1-3 m. Object avoidance and height-from-bottom protocols ensure that the AUV is not within 1 m of an object. With AUV operating speeds of 1-3 m/s and a turning radius around 5 times the length of the AUV, object avoidance is initiated when an object is a few meters away [10].

We compute the mean and standard deviation of these empty ranges to identify the background noise. For each range in the sonar image, if the maximum value in the row is greater than 11 standard deviations above the background mean, we conclude the presence of an object. We experi-

mented by increasing the threshold until most image artifacts were removed. The number of standard deviations is high because the amount of noise is low at these close ranges, so a large threshold is needed to minimize noise at longer ranges. We compute the mean and standard deviation of each row in the image. We set an occupancy threshold to 1.5 times the standard deviation of the row. All azimuths along the range with an intensity value more than the threshold are determined to correspond to an object. This occupancy detection yields a binarized image the same size as the sonar image, with 0 (black) where there is no object, and 255 (white) where there is an object, as shown in Fig. 9.

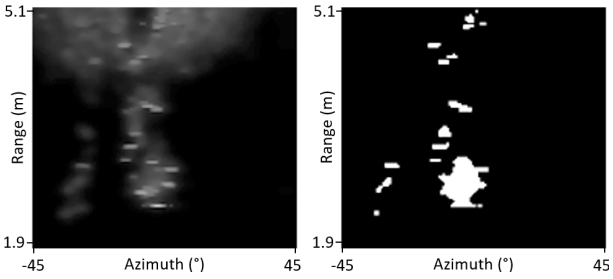

Fig. 9: The diffused sonar image before and after deringing and binarization to remove ringing noise around objects and identify which pixels contain objects.

Once we have a binarized image for the sonar image, we implement morphological closing to connect object occupancies which are close together in the image. We use a square kernel of size 18 by 18 pixels. Morphological closing consists of dilating the occupied regions around each white pixel in the image according to the kernel size, then eroding each white pixel in the dilated image with the same kernel. Thus, small gaps in the binarized image are filled, resulting in a binarized image with continuously connected regions which better encompass the entirety of each object in the scene, as shown in Fig. 10.

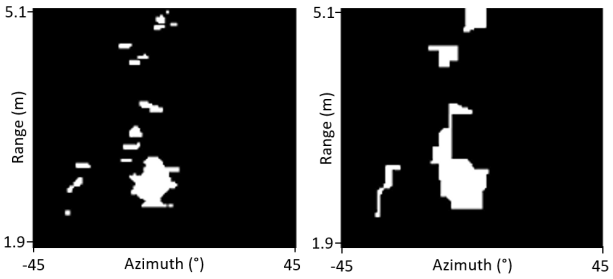

Fig. 10: The binarized sonar image before and after morphological closing to connect nearby pixel occupancies corresponding to the same object.

C. Connected Component Analysis: Identification of Azimuth and Range Bounds for Sonar Bounding Boxes

To calculate normalized minimum and maximum azimuths of the bounding boxes in the optical image, we divide the minimum and maximum azimuths by the optical image width. We convert normalized azimuths to pixels in the width dimension of the corresponding polar sonar image by multiplying by the sonar image width in pixels. To create a bounding box for the object in the sonar image, the object boundaries in the range dimension must be identified.

Ranges in the sonar image between the two azimuths are summed along the azimuthal axis. We identify consecutive nonzero sums as clusters. To avoid determination of artifacts as objects, clusters are only identified as an object if more than 6 consecutive ranges are nonzero. We chose this threshold through observation of artifacts and objects in the sonar images. The optical image bounding boxes signify that the object of interest does not extend significantly beyond the identified azimuths in the sonar image. Thus for each cluster of nonzero ranges as computed above, if the occupancy values are 255 for azimuths farther than 10 pixels outside of the azimuths at those ranges, we conclude that the found object is not the object corresponding to the optical bounding box, and the cluster is disregarded.

Out of the remaining clusters, the cluster closest to the sensor is labeled as the object, since in optical images, the visible object is the one closest to the camera. Through connected component labeling, we identify all connected white pixels in the binarized image which most closely match the computed azimuths and ranges. The boundaries of the component in azimuth and range are defined as the bounding box of the object, as shown in Fig. 11. This process prevents errors from edge misalignment caused by the on-average 0.0133-second gap between optical and sonar image collection and the need to round to the nearest pixel due to image size differences. Results when comparing automatic bounding boxes to manual labels, as shown in Fig. 12, are discussed below.

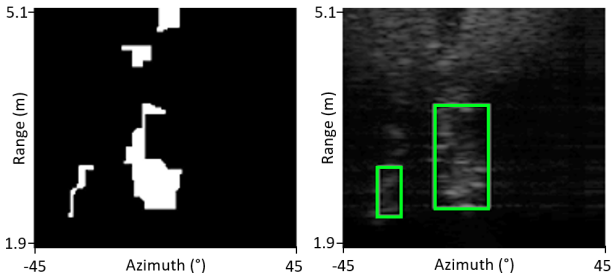

Fig. 11: Borders of connected white components which best match relevant azimuths become the bounding boxes around objects, as shown on the right in the original sonar image.

D. Knowledge Transfer with YOLOv11 to Detect Objects in Sonar Images

We obtain automatic bounding boxes for the optical-sonar image pairs with object classification labels from the YOLOv11 model for optical images. Starting from a YOLOv11 model pretrained on the COCO dataset, we use the boxes and classifications to train a YOLOv11 model to detect objects in polar sonar images. We save the images that were manually labeled to use for testing. There are 6,819 sonar images in the dataset which are not manually labeled. We use a 75/25 split for training and validation. The model is

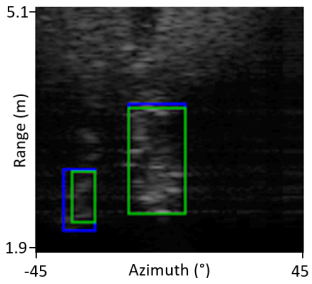

Fig. 12: The automatic bounding boxes are shown in green and the manual bounding boxes are shown in blue.

trained until performance has not improved for 100 epochs. The model trained on images in the polar coordinate system achieves its best epoch at epoch 139.

We convert the images and bounding boxes to cartesian coordinates, as shown in Fig. 13, edited to better show the objects. In the polar coordinates, objects are distorted from their true shape, but there is more resolution at ranges closer to the sonar sensor. In cartesian coordinates, the shapes of objects are true to the real shape, but resolution is lost close to the sonar sensor at the vertex of the image. We train a YOLOv11 model with cartesian images to compare performance with the model trained on polar images. The model is trained until performance has not improved for 100 epochs. The model trained on cartesian images achieves its best epoch at epoch 657.

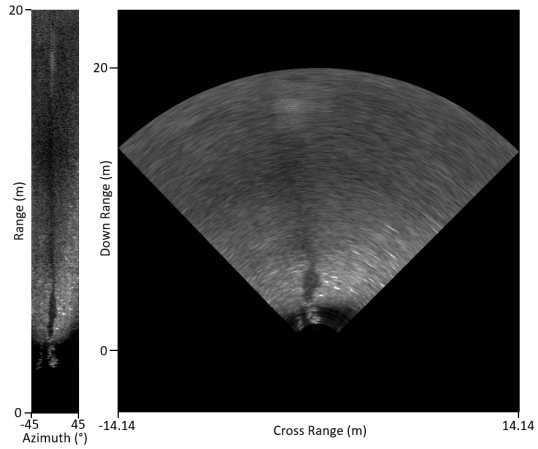

Fig. 13: The polar image is shown on the left, and the cartesian image is shown on the right.

These models detect objects in real time directly in sonar images, without the need for optical images after creating training labels. The models are tested on the manual labels for the sonar images and results are discussed below.

IV. Results

A. Evaluation of Automatic Sonar Bounding Boxes

We evaluate automatic bounding boxes in the sonar images with the manual labels according to the intersection over union (IOU), a fundamental metric for object detection tasks which is used in YOLO error analysis [18]. In the 889 images evaluated, there are 1055 bounding boxes; the average IOU is 0.574. There are 945 out of 1055 automatic boxes which overlap with the manual labels; the average IOU for these boxes is 0.640. Uncertainty is calculated as the standard deviation of the IOUs divided by the square root of the number of labels in the category, or $\Delta\bar{x} = \frac{\sigma}{\sqrt{N}}$.

The IOU for each class is shown in Table I. Note that the performance on the classes "person" and "pot" are better than the performance on the class "pan". There are 918 person objects, 104 pot objects, and 33 pan objects, so the dataset is unbalanced. When the automatic bounding box overlaps with the manual bounding box, the IOU for pan is much higher. Thus, when the identification of the range to the object is a better estimate, the computation of the bounding box is successful. However, the range to the pan is often incorrect, leading to many bounding boxes with an IOU of zero.

TABLE I: IOU by Object Class

Class	**Average IOU Overall**	**Average IOU of Boxes which Overlap**
All	0.574 ± 0.0096	0.640 ± 0.0084
Person	0.587 ± 0.0103	0.643 ± 0.0091
Pot	0.530 ± 0.0284	0.622 ± 0.0227
Pan	0.354 ± 0.0579	0.614 ± 0.0411

There is promising performance on the classes "person" and "pot". There are 918 "person" boxes evaluated, with 840 automatic bounding boxes overlapping with manual labels. There are 628 automatic "person" bounding boxes which have an IOU of over 0.5. There are 104 "pot" boxes evaluated, with 86 automatic bounding boxes overlapping with manual labels and 64 which have an IOU of over 0.5. The performance difference between the object classes may be explained by the difference in signal-to-noise ratio of the objects in the sonar images. The person is much larger than the pot and the pan, and spans more elevations in the sonar field of view than the pot and the pan span. Since FLS sensors collapse the intensities corresponding to all elevations at a certain azimuth and range, there is more signal at the person than at the pot and the pan. Additionally, the pot is thicker than the pan and thus has a larger volume which is a different density than the water around it, resulting in a stronger sonar return than the pan. The higher signal-to-noise ratios of the person and pot improve performance in the calculation of automatic bounding boxes.

B. Evaluation of YOLOv11 Model for Polar Sonar Images

The model trained with automatic bounding boxes achieves a mAP50 of 0.748 as validated on automatic bounding boxes. Further results are shown in Table II. The performance is best on "person", with mAP50 of 0.850.

The model trained with automatic bounding boxes is tested with manual bounding boxes. It achieves a mAP50 of 0.517. Further results are shown in Table III. The performance on the class "pan" is poor, as expected due to low average IOU of the automatic bounding boxes with manual bounding boxes and the low number of "pan" instances in the dataset.

TABLE II: YOLO Summary for Model Trained and Validated with Polar Automatic Bounding Boxes

Class	Images	Instances	Precision	Recall	mAP50
All	1714	1994	0.879	0.679	0.748
Person	1606	1843	0.765	0.838	0.850
Pot	112	112	0.891	0.583	0.678
Pan	39	39	0.980	0.615	0.715

The performance is best on the "person" class, with mAP50 of 0.787. This result is comparable to existing object detection methods for FLS images, such as CCW-YOLOv5 [20] which achieves 0.769 mAP at IOU 0.5 on the "human body" class using manual labels of 7600 polar FLS images for model training. Our method does not rely on manual sonar labels for sonar model training, and only required manually labeling 727 optical images for training.

TABLE III: YOLO Summary for Model Trained with Polar Automatic Bounding Boxes and Tested with Polar Manual Bounding Boxes

Class	Images	Instances	Precision	Recall	mAP50
All	1000	1187	0.787	0.482	0.517
Person	863	990	0.813	0.718	0.787
Pot	139	139	0.699	0.504	0.504
Pan	58	58	0.848	0.224	0.258

C. Evaluation After Conversion to Cartesian Sonar Images

The model as trained and validated on cartesian images achieves increased performance compared to the model on polar images for the class "pot" as validated with automatic bounding boxes, as reported in Table IV.

TABLE IV: YOLO Summary for Model Trained and Validated with Cartesian Automatic Bounding Boxes

Class	Images	Instances	Precision	Recall	mAP50
All	1747	2049	0.881	0.634	0.704
Person	1625	1882	0.827	0.789	0.837
Pot	123	123	0.944	0.681	0.747
Pan	44	44	0.872	0.432	0.528

Performance improves for all classes as evaluated with manual bounding boxes converted to cartesian coordinates, especially for the class "pot", as seen in Table V. This improvement in performance suggests that with the reduction in object distortion due to cartesian images, the YOLO model better recognizes the objects, compensating for inaccuracy in some automatic bounding box labels. The mAP50 of the "person" class is 0.790 and the mAP50 of the "pot" class is 0.644. The pan class achieves a low mAP50, with a high precision value and a low recall value. This result shows that the model has a high true positive rate and a high false negative rate. The model is accurate when it predicts "pan", but misses many instances of "pan" in the dataset. The performance on the class "pan", the smallest class of the dataset, may be improved through further development of automatic bounding box computation and through an increased number of training data images.

TABLE V: YOLO Summary for Model Trained with Cartesian Automatic Bounding Boxes and Tested with Cartesian Manual Bounding Boxes

Class	Images	Instances	Precision	Recall	mAP50
All	1000	1187	0.771	0.527	0.574
Person	863	990	0.774	0.723	0.790
Pot	139	139	0.822	0.600	0.644
Pan	58	58	0.717	0.259	0.288

The performance across the classes varies with the accuracy of the automatic bounding boxes. Further work is needed to collect optical and sonar images of additional classes of objects underwater and improve automatic bounding box calculation of varied objects.

V. Conclusion and Future Work

We present a method to transfer object detection knowledge from underwater optical images to FLS images. We use our integrated sensor platform on a commerical AUV to establish performance benchmarks for automatic labeling and subsequent training of an object detection model using sonar images. This model achieves mean average precision of 0.790 when detecting people in cartesian sonar images, which is comparable to existing object detection of a human body model when trained with manual labels of sonar images. Our method does not require manual labels of sonar images for training. The sonar images in the polar coordinate system are needed for azimuthal association across sensors. Through experimentation, we conclude that converting sonar images to cartesian coordinates for model training leads to increased performance for object detection.

The generalizability of our method should be tested through model deployment on images of many classes of objects typically found in a marine setting. Deploying the model trained for sonar images on a testing dataset taken in turbid water would confirm the knowledge transfer into the sonar domain. Testing on sonar images where objects are far away or obscured allows for evaluation of the model in challenging sonar applications.

Future work includes expanding this method with self-supervised methods for optical object detection to transfer knowledge without relying on manual labels for either modality. To expand upon recent work in self-supervised sonar image classification, we will consider real-world images with multiple objects present per image. Lastly, we may expand the method to automatically label 3D sonar data.

References

[1] S. Archieri et al. "3DSSDF: Underwater 3D Sonar Reconstruction Using Signed Distance Functions". In: *2025 IEEE International Conference on Robotics and Automation (ICRA)*. 2025, pp. 5306–5312. DOI: 10.1109/ICRA55743.2025.11128101.

[2] Arducam. *OV5648 USB2.0 Wide Angle Camera Module B0454*. 2017. URL: https://www.uctronics.com/download/Amazon/B0454_5MP_Wide_Angle_UVC_Camera_Datasheet.pdf.

[3] G. Bradski. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).

[4] Miguel Castillón et al. "State of the Art of Underwater Active Optical 3D Scanners". In: *Sensors* 19.23 (2019). ISSN: 1424-8220. DOI: 10.3390/s19235161.

[5] Hyeonwoo Cho, Jeonghwe Gu, and Son-Cheol Yu. "Robust Sonar-Based Underwater Object Recognition Against Angle-of-View Variation". In: *IEEE Sensors Journal* 16.4 (2016), pp. 1013–1025. DOI: 10.1109/JSEN.2015.2496945.

[6] Louise Rixon Fuchs, Aron Norén, and Philip Johansson. "GAN-enhanced simulated sonar images for deep learning based detection and classification". In: *OCEANS 2022 - Chennai*. 2022, pp. 1–5. DOI: 10.1109/OCEANSChennai45887.2022.9775246.

[7] Thomas Guerneve, Kartic Subr, and Yvan Petillot. "Three-dimensional reconstruction of underwater objects using wide-aperture imaging SONAR". In: *Journal of Field Robotics* 35.6 (2018), pp. 890–905. DOI: https://doi.org/10.1002/rob.21783.

[8] Chengzhou Li et al. *RSOD: Reliability-Guided Sonar Image Object Detection with Extremely Limited Labels*. 2026. arXiv: 2601.12715 [cs.CV]. URL: https://arxiv.org/abs/2601.12715.

[9] Haibin Li et al. "A Review of Underwater SLAM Technologies". In: *2023 5th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI)*. 2023, pp. 215–225. DOI: 10.1109/RICAI60863.2023.10489371.

[10] Juan Li et al. "A Predictive Guidance Obstacle Avoidance Algorithm for AUV in Unknown Environments". In: *Sensors* 19.13 (2019). ISSN: 1424-8220. DOI: 10.3390/s19132862.

[11] Yunfeng Li et al. "RGB-Sonar Tracking Benchmark and Spatial Cross-Attention Transformer Tracker". In: *IEEE Transactions on Circuits and Systems for Video Technology* 35.3 (2025), pp. 2260–2275. DOI: 10.1109/TCSVT.2024.3497214.

[12] Tsung-Yi Lin et al. "Microsoft COCO: Common Objects in Context". In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Cham: Springer International Publishing, 2014, pp. 740–755. ISBN: 978-3-319-10602-1.

[13] Peng Liu et al. "Underwater Image Enhancement With a Deep Residual Framework". In: *IEEE Access* 7 (2019), pp. 94614–94629. DOI: 10.1109/ACCESS.2019.2928976.

[14] Teledyne Marine. *Mini Metal Shell Electrical Connector*. 2015. URL: https://www.teledynemarine.com/brands/impulse/mini-metal-shell-electrical-connector.

[15] P. Perona and J. Malik. "Scale-space and edge detection using anisotropic diffusion". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.7 (1990), pp. 629–639. DOI: 10.1109/34.56205.

[16] Alan Preciado-Grijalva et al. "Self-supervised Learning for Sonar Image Classification". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2022, pp. 1498–1507. DOI: 10.1109/CVPRW56347.2022.00156.

[17] Yaadhav Raaj, Alex John, and Tan Jin. "3D Object Localization using Forward Looking Sonar (FLS) and Optical Camera via particle filter based calibration and fusion". In: *OCEANS 2016 MTS/IEEE Monterey*. 2016, pp. 1–10. DOI: 10.1109/OCEANS.2016.7761077.

[18] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91.

[19] Kurran Singh et al. "Opti-Acoustic Semantic SLAM with Unknown Objects in Underwater Environments". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2024, pp. 1169–1176. DOI: 10.1109/IROS58592.2024.10802819.

[20] Yan Sun and Bo Yin. "CCW-YOLOv5: A forward-looking sonar target method based on coordinate convolution and modified boundary frame loss". In: *PLOS ONE* 19.6 (June 2024), pp. 1–18. DOI: 10.1371/journal.pone.0300976.

[21] Maxim Tkachenko et al. *Label Studio: Data labeling software*. Open source software available from https://github.com/HumanSignal/label-studio. 2020-2025.

[22] Tritech. *MicronGemini 720s*. 2025. URL: https://www.tritech.co.uk/products/microngemini-720s.

[23] Xudong Wang et al. "Cut and Learn for Unsupervised Object Detection and Instance Segmentation". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 3124–3134. DOI: 10.1109/CVPR52729.2023.00305.

[24] E. Westman and M. Kaess. "Wide Aperture Imaging Sonar Reconstruction using Generative Models". In: *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems, IROS*. Macao, Nov. 2019, pp. 8067–8074.

[25] Minghua Zhang et al. "Efficient Small-Object Detection in Underwater Images Using the Enhanced YOLOv8 Network". In: *Applied Sciences* 14.3 (2024). ISSN: 2076-3417. DOI: 10.3390/app14031095.