2024 IEEE 4th International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA),Tripoli, Libya, 19-21 May 2024.

2024 IEEE 4th International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA) | 979-8-3503-7263-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/MI-STA61267.2024.10599670

On Object Detection and Explainability with Sonar Imagery

Tarek A. El-Mihoub
German Research Center for Artificial Intelligence
Oldenburg, Germany
tarek.elmihoub@dfki.de

Abdulhakim El Gadi
CU Coventry
Coventry, England
hakim.elgadi@coventry.ac.uk

Lars Nolle
Jade University of Applied Science
Wilhelmshaven, Germany
lars.nolle@jade-hs.de

Frederic Stahl
German Research Center for Artificial Intelligence
Oldenburg, Germany
frederic_theodor.stahl@dfki.de

***Abstract*— The task of detecting objects in sonar imagery is challenging due to low image resolution, significant noise, and the variable nature of the underwater environment. Artificial Intelligence (AI) methods, such as convolutional neural networks and transformers, can be applied to address these challenges. However, the lack of transparent explanations for the results of deep learning models is a major stumbling block to building trust in these models. This paper evaluates You Only Look Once (YOLO-v8) and DEtection TRansformers (DETR) models, as two prominent tools for object detection, using two sonar datasets as a part of the Hybrid Artificial Intelligence eXplainer (HAI-x) project. The selected datasets include a dataset of sonar raw images and another with pre-processed sonar images. Contrary to the claim that sonar raw images do not require preprocessing for efficient object detection using deep learning techniques, the experiments conducted demonstrate that such preparatory steps can indeed improve the performance of object detectors. Furthermore, they can provide an understandable common ground for explaining the detection process to end users.**

Keywords—Object Detection, SONAR, XAI, YOLOv8, DETR, ViT.

I. Introduction

The HAI-x project aims to equip hybrid AI models with explanation facilities to build user trust in these models [1]. Optimizing the aquatic weed harvester's operation on Lake Maschsee is a use case for this project in a marine environment. To enhance the efficiency and effectiveness of the harvester operations, a hybrid artificial intelligence system is to be developed. This system utilizes a combination of artificial intelligence techniques to optimize the harvester's path, ensuring thorough coverage of the overgrowth weed areas to enable mowing them while minimizing fuel consumption, time, and environmental impact. In defining the optimal path, the system uses sensors to capture and develop a map of the current situations. Among the sensors that the system utilizes are SONAR sensors. The intention is to use a Side-Scan Sonar (SSS), which emits the sound wave to the sides of the sonar device and thus provides a horizontal view of the seafloor and underwater objects. However, the option of using a Forward-Looking Sonar (FLS), which emits sound waves in the direction the sonar is facing to provide a forward view of the underwater features, has not been excluded. Both sonars visualize sonar data as sonar imagery. Sonar imagery can be used to detect areas of avoidance and areas of interest. In line with the project's aim, the system should provide explanations to the users in addition to detecting these areas. One method to achieve this is to provide extra information about such areas to convince the users with the reasons for the classifications of the different areas. Such information can be provided by classifying and locating different entities in the lake and visualizing them on the situation map. In other words, using the sonar imagery to detect the different objects in the lake as a base for defining areas of walk, interest, and avoidance.

In spite that fact that object detection models in general provide to some extent enough information to convince the end users with correctness of their predictions. By displaying the bounding boxes together with classifications and confidence scores, the user can judge and evaluate the object detector's prediction and take an action accordingly. However, the situation may differ in case of using sonar imagery due to their nature. When it comes to implementing cutting-edge AI methodologies in the field of sonar, the research community lags behind by a few years [2]. However, different object detectors have been utilized using sonar imagery [2]. Most researchers apply image enhancement techniques as a preprocessing phase to sonar images to make them recognizable to the bare eye and akin to natural RGB images that are usually presented to object detection models. In contrast, Xie et. al [3] proposed a sonar imagery dataset that consists of raw sonar images captured by FLS claiming that preprocessing sonar images to enhance their visual perception can lead to losing meaningful data and introduce inaccurate information. The scarcity of open-access datasets is notable, and releasing this dataset to the public can be viewed as a positive contribution. The dataset, which is known as underwater acoustic target detection (UATD), was benchmarked on state-of-the-art (SOTA) object detection models considering the efficiency and accuracy measures but ignoring explainability measures. Contrary to Xie et. al [3], this paper emphasizes that visual perception of sonar images is crucial for object detection and explanations of prediction. In addition to enabling the user to visually validate the model predictions, the work presented in this paper facilitates utilizing other XAI methods that rely on samples of training data to rationalize predictions.

In this paper, the performance of YOLOv8 [4] and DETR [5] object detectors are evaluated using UATD dataset. The results are then compared with the performance of both detectors with the marine-debris-fls dataset [6], which contains pre-processed sonar images. Different experimental scenarios have been followed to explore the impact of enhancing the visual perception of sonar data on the performance object detectors considering explainability as a key factor in the evaluation process.

979-8-3503-7263-2/24/$31.00 ©2024 IEEE

This paper is organized into six sections. The second section offers a short introduction to object detection models. The third section briefly reviews their applications in conjunction with sonar imagery. In the fourth section, the explainability of object detectors is discussed. E**X**plainable **A**rtificial **I**ntelligence (XAI) methods are also examined. The fourth section delves into the details of the different experiments conducted and discusses the outcomes of these experiments. The paper concludes by providing a summary and exploring potential directions for future research.

II. Object Detection Using Machine Learning

SOTA AI-based object detectors can be categorized into one-stage and two-stage models [7,8]. The object detection problem can be seen as finding an arbitrary number of objects in an image through extracting its visual features, followed by classifying these objects and estimating their sizes within bounding boxes. Two-stages detectors tackle this problem by separating the previous two tasks into stages. Variants of Regions with Convolutional Neural Network (R-CNN), Spatial Pyramid Pooling deep Network (SPPNet) and Pyramid Networks/FPN follow the two-stages approach. Meanwhile, single-stage detectors combine both tasks into one step. Examples of single-stage detectors are different versions of YOLO, different variants of Single Shot Multibox Detector SSD, RetinaNet and Fully Convolutional One-Stage FCOS. Single-stage detectors are generally characterized by a faster detection speed and greater structural simplicity compared with two-stages detectors. However, their performance trailed that of two-stages detectors, notably for small and dense objects [7].

In addition to single-stage and two-stages detectors, advancements in transformer architecture have shown great promise in achieving competitive results in tackling computer vision problems, including object detection [9]. The Detection Transformer (DETR) [5] employs attention mechanisms to selectively assign importance to various segments of the input data sequence, facilitating object detection. It is usually described as a simple and effective high-level vision framework. In DETR, the object detection problem is defined as a direct set prediction problem to predict a fixed number of bounding boxes and their corresponding class labels in a single pass using a transformer. As other transformers, DETRs are composed of multiple self-attention layers.

A. You Only Look Once (YOLO)

YOLO [10] uses a single-stage architecture to make predictions for bounding boxes and class probabilities. In YOLO, the input image undergoes grid partitioning. For each grid cell, the model predicts multiple bounding boxes along with confidence scores, indicating potential object locations. YOLO also predicts the class probabilities for each bounding box. YOLO uses Non-Maximum Suppression (NMS) to remove redundant or intersecting bounding box predictions and ensure only the most confident and non-overlapping predictions are kept. This single-stage architecture enables efficient and quick predictions, making YOLO suitable for real-time applications. YOLO can tackle reverberation noise in sonar images [11]. Nevertheless, the resolution of the grid of YOLOv1 is not sufficient high, which can cause degrading the prediction accuracy. Various enhancements have been introduced to YOLO producing updated versions. These enhancements include introducing Darknet53, Feature Pyramid Network (FPN), Cross Stage Partial Network (CSPNet), focus layers, Path Aggregation Network (PANet), C3 modules, Extended Efficient Layer Aggregation Network (E-ELAN), Spatial pyramid pooling (SPP) in addition to others.

YOLOv8 is the latest version of YOLO by Ultralytics. It supports classification, segmentation, tracking and pose estimation in addition to object detection. The architecture of YOLOv8 is improved through an anchor-free detection, replacing C3 module with C2f module, using a decoupled head, modifying loss function, and utilizing mosaic data augmentation. Via anchor-free detection, the model needs to anticipate the centre of an object rather than the bounding box coordinates. Such anchor-free detection improves generalization and speed both learning speed and NMS process. In YOLOv8, each bottleneck consists of residual blocks for computation cost reduction during training and C2f module concatenates the output of these bottleneck modules. Through decoupled head by separating classification and regression tasks, the performance is further improved. To avoid the misalignment possibility due to decoupled heads, the loss function of YOLOv8 has been modified by introducing alignment scores. Mosaic augmentation that has introduced in YOLOv4, has been changed in YOLOv8 by stopping it the last 10 training epochs for the sake of performance improvements.

YOLOv8 comes in five variants based on the number of parameters. In this paper, when referring to YOLOv8 in the following sections, it refers to the nano-model, which is the smallest variant.

B. Detection Transformers (DETR)

The Detection Transformer (DETR) [5] introduces a novel object detection approach based on transformers. It employs a transformer encoder-decoder framework, eliminating the need for NMS, and incorporating the Hungarian loss to anticipate sets of objects in a one-to-one mapping, facilitating end-to-end optimization. DETR reframes object detection as a problem of anticipating a set of unordered variables with uncertain relationships. Instead of predicting bounding boxes independently and then applying NMS to remove duplicates, DETR directly outputs a fixed number of bounding boxes for all objects in an image. It utilizes a set of object queries, to interact with image features for predicting the objects and their locations. By restricting the number of object queries to 100 queries, it avoids producing redundant and near-duplicate results, thus eliminating the need for NMS post-processing component. DETR uses the Hungarian algorithm for bipartite graph matching to link forecasted bounding boxes to ground truth objects. This algorithm enables finding the optimal one-to-one mapping between predicted boxes and ground truth boxes based on a similarity metric, often a negative IoU (Intersection over Union) score. The Hungarian matching is incorporated into DETR's loss function, i.e. Hungarian loss, to penalize the model based on the correctness of the associations between forecasted and ground truth bounding boxes. This encourages the model to produce accurate and well-matched predictions.

In addition to DETR, several object detection models based on transformer architectures have been proposed [9]. Deformable-DETR seeks improving the performance of DETR through reducing the density of reference points in the feature space. Sparse R-CNN combines elements of both traditional region-based CNNs and transformers. Pyramid Vision Transformer (PVT) introduces the initial hierarchical architecture for vision transformers and proposes a gradual reduction pyramid and attention mechanism with spatial reduction. Swin Transformers produce hierarchical feature maps, instead of single feature map, through assimilation of image patches in deeper layers. It also restricts self-attention only within each local window, which is partitioned into multiple sub-patches. Focal Transformers introduces focal self-attention to instantaneously grasp both short- and long-range visual relations efficiently and effectively. CrossFormer employs a Cross-scale Embedding Layer (CEL) and Long Short Distance Attention (LSDA) with dynamic position bias to effectively comprehend both local and global visual clues. RegionViT introduces regional-to-local attention to encode hierarchical features. For efficient vision 2D attention, a model structure that enables image encoding at various scales while maintaining a manageable computational cost and adopts the efficient Longformer is incorporated in Multi-Scale Vision Longformer. CrossViT introduces a dual-branch transformer that merges patches with varying sizes of an image to generate more robust features.

Despite various efforts to enhance DETR, the persistent challenge of high computational cost hinders its practical application, limiting its advantages. Although DETR simplifies object detection, its demanding computational requirements make achieving real-time detection challenging. However, in HAI-x project, real-time detection is not an issue and exploring DETR explanation capabilities worth evaluation.

III. Object Detectors for Sonar Imagery

Researchers have investigated utilizing machine learning for underwater object detection using sonar imagery generated from FLS and SSS systems. Key obstacles to achieving precise object detection in sonar imagery include interference from sonar seabed reverberation noise, a limited proportion of pixels depicting foreground object areas, and inadequate imaging resolution [2].

A. Forward-Looking Sonar (FLS)

A CNN-based method is utilized for object detection in FLS images [12]. In this method, objects are separated from the background to estimate their presence while sliding a window over the sonar image. The same concept has been utilized to build object detector for FLS images through using CNN to extract features and then feeding these features into "object classifier" and "object detector" [13]. Pretrained YOLO model has been applied as real-time object detector for localizing an AUV agent using FLS images [14]. The CoordConv has been incorporated in YOLOv5 to introduce the coordinate information of the pixels for improving the accuracy of the object detection in Multi-beam FLS (MFLS) images [15].

The multi-branch shuttle neural network has been incorporated in YOLOv5 to improve its ability to detect small and weak targets using MFLS images [16]. A modified Mask RCNN with less training parameters through replacing the Resnet50/101 with Resnet32 has been proposed for detecting and segmenting objects in MFLS images [17]. Three self-supervised learning methods, which are RotNet, Denoising Autoencoders, and Jigsaw are proposed to classify pre-processed FLS images [18]. RBoxNet as a single model and combined with YOLOv2 are proposed and utilized in a multi-object detection system to locate objects and determine their rotation from sonar imagery for autonomous underwater vehicle (AUV) navigation [19]. The YOLOv3 model has been used to detect obstacles in pre-processed FLS images together with deep reinforcement learning (DRL) for planning a path for AUV [20].

B. Side-Scan Sonar (SSS)

Different ML object detection models have been proposed for dealing with SSS imagery. A CNN based on VGG-16 was proposed for automatic image recognition in SSS images [21]. An improved faster R-CNN detector based on VGG-16 has been developed for automatic wreckage recognition using SSS images [22]. A Self-Cascaded CNN (SC-CNN) detector, which exhibits resilience to speckle noise and variations in intensity, has been proposed to segment the objects, their shadow and seafloor in pre-processed SSS images [23]. YOLOv3 with Darknet-53 network has been used for detecting shipwreck targets in SSS images utilizing multi-scale features fusion with FPN [24]. Inspired by YOLOv3, a Gabor-based detector, which extract features different levels, has been proposed to detect mine like objects in SSS images [25]. A differentiable Architecture Search algorithm with a flexible search space and large inputs (FL-DARTS) has been proposed and used to build self-trained AutoDL object detectors [26]. These detectors have been evaluated on an SSS dataset.

To tackle the target-sparse and feature-scarce attributes of SSS imagery, a transformer model is integrated with YOLOv5s in a real-time TR-YOLOv5s object detector [11]. [21] refers to a lightweight DETR-YOLO model, which combines the global view of the complex marine environment with the lightweight requirements to solve the same problem. To tackle the high false and missed detection rates in the case of multiple densely and overlapping underwater targets in SSS images, a segmentation model that utilizes the blended hybrid dilated convolution and pyramid split attention UNet (BHP-UNet) algorithm is proposed [21]. The multilevel feature fusion network (MLFFNet) [27] has been proposed for accurate underwater object detection. Multiscale convolution (MS-Conv), multilevel feature extraction (ML-FEM), multilevel feature fusion (ML-FFM), multiscale feature pyramid (MS-FPN), and feature association (FA) models together with neighbourhood channel attention mechanism (N-CAM) are combined in the MLFFNet detector to overcome SSS challenges. YOLOv3-DPFIN combined YOLOv3, Dual Path Network (DPN), the fusion transition module, and dense connection method to improve object detection in SSS images [28]. The model has been evaluated on images of a simulation generated dataset and on pre-processed images of a real dataset. Three active-learning-based algorithms have been proposed to reduce the annotation cost through selecting the images, which have the most valuable information, from unlabelled data and continuous retraining to enhance the object detector's performance on pre-processed SSS images [29].

An adaptive global feature enhancement network (GFFNet) has been proposed to detect objects in sonar images. This model combines multi-scale convolution with attention mechanisms and a global receptive field, for extracting multi-scale semantic features from sonar images and improve feature correlation. The detector has been evaluated using SSS and FLS datasets [30]. YOLOv7-based models together with multi-scale information fusion and attention concepts have been utilized for object detection in SSS images. The result of the detection then is passed to an algorithm to determine the target latitude and longitude location [31]. A CNN with bilinear pooling is used to classify objects in SSS images and use a positioning algorithm to determine their locations in an SSS image [32].

IV. Explainability of Object Detectors

In addition to optimize the performance of these detectors taken into account the specific nature of the problem to solve, building trust in these detectors as in any other AI models becomes an essential requirement for efficient deployment [33]. In spite of the fact that the results of object detection models can be viewed as a kind of self-explainable by providing the size of each object within a bounding box in addition including the name of the object. By highlighting the part of the image that contains the object (i.e. emphasizing the visual features of the object), the end user can understand the detector decision. The results of instance segmentation task can be also considered as self-explainable results. They include extra information that enables the end user to understand the reasons for the model decision. Through highlighting the exact boundaries of an instance in addition to the instance name, the instance segmentation model precisely defines the visual reasons for their predictions. A fundamental requirement for self-explainable results of object detection and instance segmentation is that the objects and instances in the input image or frame are recognizable for bare eyes, which is the case for most computer vision problems.

In addition to self-explainable results mentioned above, different AI explanation facilities can be used to build trust in AI models in general. However, the key to offering effective explanations to end users lies in aligning the explanation with both what the end-user can observe and what the model perceives and predicts. Among these explanation approaches, a set of approaches that rely on training examples to reason the model behavior for a given input. Explanations that are based on prototyping and case-based reasoning [34], counterfactual [35] and estimation training data influence [36] are examples of this set. In prototyping and case-based reasoning, a subset of prototypes is chosen from the training dataset. These prototypes are utilized to provide explanations for the model's classification using case-based reasoning. Counterfactual explanation highlights minimum changes in features that lead to a change in the model's outcome. To explain image classifications and object detection algorithms, this approach usually utilizes a subset of the training set to provide such explanations. Estimation training data influence involves identifying a set of training examples that have the greatest influence on the outcome of the model for a specific input. This approach also relies on the training set to identify the proponents and opponents of a given input.

These explanation tools can help to provide efficient explanations for different computer vision tasks. However, this can an obstacle for explaining the prediction models when dealing with sonar data, which is usually presented as an image to ML models for the sake of benefiting from the advances in solving computer vision problems. Typically, sonar data images undergo pre-processing to render them recognizable to the naked eye and to align them to some extent with the images typically processed by computer vision models. Despite the pre-processing and image enhancements phase, the quality of sonar images cannot approach that of camera images due to low resolution, significant noise, and the variable nature of underwater environment. Furthermore, with the continuing roll-out of more powerful deep learning methods, the general thrust has increasingly been to present the model with the raw data, the assumption being that the model is better equipped to recast the raw data into some learned latent space.

V. Experiments and Discussion

Different scenarios were conducted for evaluating the performance YOLOv8 and DETR models on sonar images using two different datasets.

A. Datasets

While sonar image analysis has gained significant research interest, the availability of publicly accessible sonar datasets remains relatively limited. In this paper, two publicly accessible datasets, which are collected in underwater environments, are used in this paper.

The datasets are UATD and marine-debris-fls. Both are collected using FLS. The UATD data set consists of images of 10 objects with larger sizes compared to those of the marine-debris-fls data set, which consists of household objects. Tritech Gemini 1200ik sonar was used to acquire the images of the UATD dataset, whereas marine-debris-fls were acquired using the ARIS Explorer 3000. The UATD collected in real underwater environment from a shallow water lake. On the other hand, marine-debris-fls data are collected in a water tank.

The raw data received from the FLS is presented in polar coordinates, organized into a matrix where rows denote distance, and columns denote direction. However, in this raw data format, object shapes appear distorted. To maintain the accuracy of object shapes, a conversion to Cartesian coordinates is required. This transformation facilitates a more straightforward interpretation of sonar data for individuals, as it represents the shapes of objects without distortion.

The UATD data set consists of raw data of sonar images, while the marine-debris-fls data comprises sonar images that underwent an initial processing phase. The initial processing of phase marine-debris-fls includes image enhancements, such as histogram equalization or logarithmic transformation, in addition to mapping to Cartesian coordinates. This makes them more recognizable for a naked eye than that of UATD images (Fig. 1).

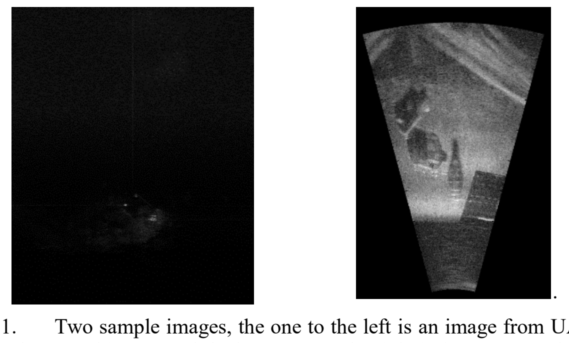

Fig. 1. Two sample images, the one to the left is an image from UATD with BlueRov object [3], while the image on the right is from marine-debris-fls with a standing bottle object [6].

B. Scenarios for Experimentation

A set of experiments has been conducted to evaluateYOLOv8 and DETR on the selected datasets. The aim of these different scenarios is to evaluate the performance of these algorithms in terms of Precision (P), Recall and mean Average precision, as a standard metrics of effectiveness evaluation of object detection models. In the evaluation process, the efficiency was not considered as the object detection process aims to generate a situation map off-line. However, especial attention is paid to the explanation means of the different models using UATD and marine-debris-fls datasets.

1) YOLOv8

Several experiments have been conducted, as shown in Table 1. In experiment E1, A COCO pretrained YOLOv8n model, was fine-tuned trained on the whole dataset of UATD and in a second experiment E2, the same pretrained model was fine-tuned on marine-debris-fls dataset. The experiments show that the model produced a better performance on marine-debris-fls dataset than on UATD dataset as shown in Table 2. Sample of a validation batch of E1 and E2 are shown in Fig 2 and Fig 3. These figures clearly shows that figures produced by E2 can help the user in validating the detection decision, while is quite difficult to use the figures of E1 for such validation.

TABLE I. LIST OF YOLOV8 EXPERIMENTS.

Experiment	Dataset	Dataset size	Pretrained	Pretrained Data
E1	UATD	9200	Yes	COCO
E2	Marine-debris-fls	1870	Yes	COCO
E3	UATD	1840	Yes	COCO
E4	UATD + logarithmic transformation	9200	Yes	COCO
E5	UATD	9200	No	-
E6	Marine-debris-fls	1870	No	-
E7	UATD	1840	Yes	marine-debris-fls
E8	Marine-debris-fls	1870	Yes	UATD

TABLE II. PERFORMANCE OF YOLOV8.

Experiment	P	R	mAP50	mAP50-95
E1	0.859	0.79	0.808	0.36
E2	0.938	0.955	0.971	0.719
E3	0.745	0.724	0.768	0.345
E4	0.803	0.804	0.826	0.371
E5	0.822	0.825	0.827	0.363
E6	0.96	0.982	0.987	0.789
E7	0.721	0.703	0.73	0.326
E8	0.915	0.931	0.962	0.669

To check whether the difference in the performance is related to the differences in the sizes of datasets, another experiment, E3, was conducted. In this experiment, a random selection was made, choosing 20% of the UATD dataset to create a subset with a similar size to that of the marine-debris-fls dataset. The results of this experiment do not show a significant change in the performance compared to that on the complete dataset.

In experiment E4, a simple image enhancement is conducted though applying logarithmic transformation to the images of UATD before fine-tuning a COCO pretrained model on the resulting dataset. This transformation is selected as it can be applied on raw sonar images to make them more recognizable. However, even with this transformation the images still suffer from deformation as they are in the polar coordinates and not in the Cartesian coordinate. Fig 4 shows the impact of applying the logarithmic transformation on a sample image of UATD dataset with an airplane object. The aim of E4 is to evaluate the cost of enhancing the visibility of sonar images in terms of degrading the detection capabilities of YOLOv8n. The results indicate that image enhancement does not incur significant costs in terms of Precision, Recall, and mAPs, as depicted in Table II. In contrast, there are slight improvements in terms of Recall (R) and mean Average precision (mAP50, mAP50-90) compared to E1. In addition to the improvement in appearance of the sonar images that serves explaining the detector behavior as shown in Fig. 4 compared to Fig. 2.

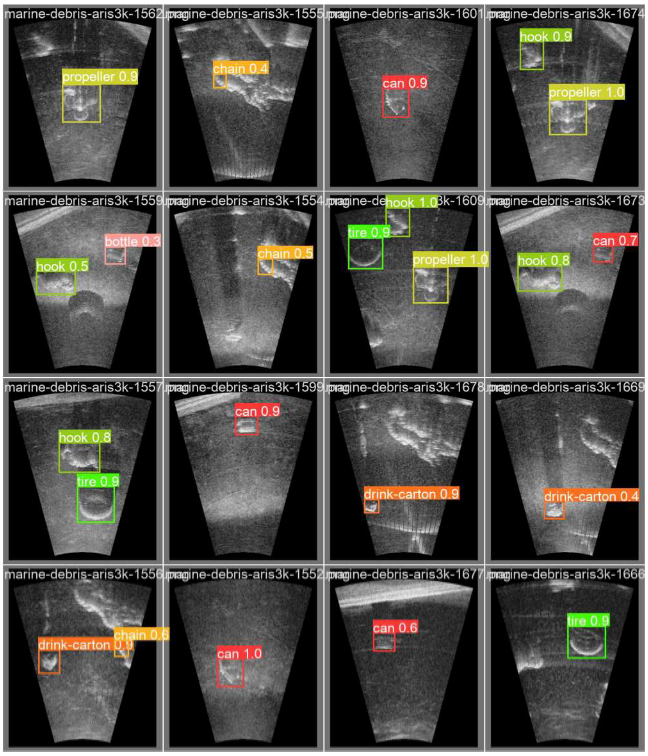

Fig. 2. A sample of a validation batch on marine-debris-fls.

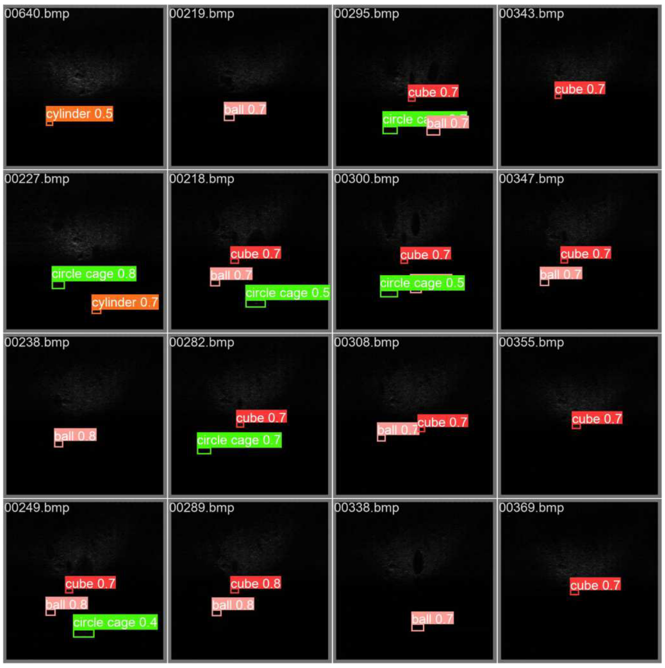

Fig. 3. A sample of a validation batch on UATD.

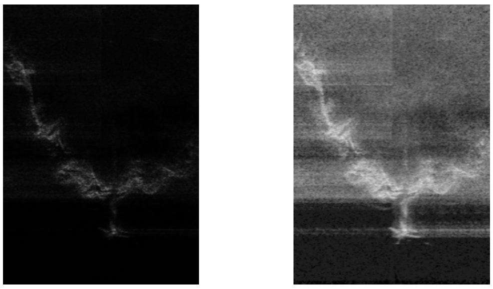

Fig. 4. The image on the left is a raw sonar image with an airplane in polar coordinates, and the image on the right is the same image after applying logarithmic transformation to enhance it appearance.

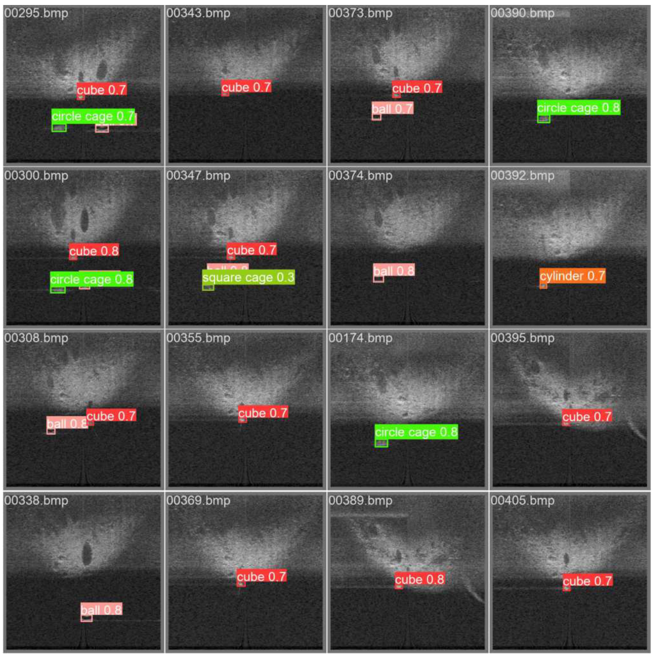

Fig. 5. A sample of a validation batch on UATD with logarithmic transformation.

To fully remove the deformation of the objects in raw sonar images due to polar representation, the images should be represented in Cartesian coordinates. However, since the bounding boxes in UATD are given with reference to these raw images, mapping them to Cartesian can lead to producing images that are outside the boundaries of these. In order words, at least part of the labelling process, which a time-consuming task, needs to be repeated to avoid such possibilities. Hence, despite the considerable potential for enhancing performance and elucidating detector behavior by transforming UATD images into Cartesian coordinates, no additional experiments were conducted.

Two experiments have been carried out, E5 and E6, with the aim of evaluating the impact of the pretrained model on the YOLOv8n performance on both datasets. In E5 and E6, YOLOv8n model was trained from scratch on UATD and marine-debris-fls from scratch. The performance of the model showed similar behavior to that of E1 and E2, with a superiority on marine-debris-fls. Training the YOLOv8n on the sonar datasets produced slightly better detection performances compared with fine tuning a COCO pretrained model, as shown in Table I.

In the last set of experiments, E7 and E8, with YOLOv8n, a marine-debris-fls pretrained model was fine-tuned on UATD and a UATD pretrained model was fine-tuned on marine-debris-fls, respectively. The results of evaluation of the models demonstrate that a UATD pretrained model produced a better performance on marine-debris-fls than that of a marine-debris-fls pretrained on UATD. The superior performance of different YOLOv8n models on marine-debris-fls can be due to the impact of simplifying the detection problem for the model and bare eyes through enhancing the visibility of sonar dataset.

2) DETR

In the first experiment of using DETR, E9, the model from (https://huggingface.co/docs/transformers/model_doc/detr) was fine-tuned on the marine-debris-fls dataset for object detection without any augmentations. The pretrained deter-resnet-50 fine-tuned on marine-debris-fls dataset. Steps defined in (https://github.com/NielsRogge/Transformers-Tutorials/blob/master/DETR/Fine_tuning_DetrForObjectDetection_on_custom_dataset_(balloon).ipynb) were followed to fine-tune DETR.

The same steps followed previously were done to fine-tune a pretrained detr-resnet-50 on UATD in another experiment. The generated models were evaluated using CoCoEvalutor. The results of the evaluation of both experiments are shown in Table III. The model shows a poor performance when fine-tuned on UATD and a good performance when tuned on marine-debris-fls. The poor performance of DETR on UATD can be explained based on the general poor performance of the model in detecting small objects. The objects in UATD images are small compared to that in marine-debris-fls images. In terms of similarities to the images that DETR pretrained on, the image enhancements introduced on marine-debris-fls images make them to some extent similar to the optical images that the model trained on.

TABLE III. PERFORMANCE OF DETR.

Experiment	AR50-90	mAP50	mAP50-95
E1	0.240	0.058	0.017
E2	0.708	0.869	0.616

Fig. 5. depicts the attention weights of the final decoder layer. This entails illustrating, for each identified object, the specific region of the image that the model focused on to predict the corresponding bounding box and class. Fig. 5 shows the queries ids that led to the model prediction together bounding box of prediction on the input image. It also shows the pixels that have a strong impact on detecting each object.

The attention mechanism can also be used to show parts of the image that are highly related, which usually belongs to the same object in the case of object detections. In Fig. 6., the model's self-attention mechanism is visualized. This can provide insight into how the model represents objects. Specifically, by examining the attention response maps, the involvement of the encoder in the process of distinguishing between individual objects can be observed. The visualization provides insight suggesting that the encoder may already be engaged in some form of object separation through the self-attention mechanism. Fig. 6. visualizes relations between different area learned by DETR in an image from marine-debris-fls. In this graph, the areas that highly related to four selected points are visualized. The red points indicate the selected points, and the four images around the marine-debris-fls image visualize the parts that related to each point.

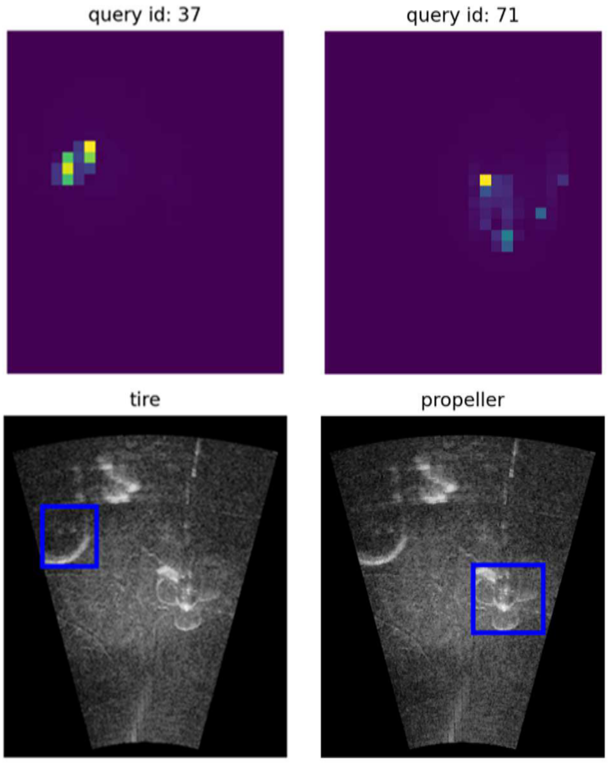

Fig. 6. The upper graphs highlight the areas of the image that lead to the prediction with the query id that is activated in the prediction. The DETR here detects the tire and the propeller as indicated in the lower images but fails to detect a chain which is located above the tire.

The graphs in Fig. 6 show that attention point (220,272), which is on a chain object that DETR failed to detect, is related to the tire object, which DETR has been detected. This can indicate that a static configuration of objects during dataset collection, as the case in marine-debris-fls, can misguide DETR leading to paying attention to correlation between the locations of different objects and such correlation can lead to wrong prediction.

Despite the high computation cost associated with vision transformers, the attention mechanism can be used to explain behaviour of their detection.

VI. Conclusions and Future work

To gain user's trust in object detection models, explainability should be taken into consideration in the various stages of detector development and deployment. Consideration of explanatory aspects should commence from the initial stages of data collection and preparation, extending through to the final deployment stage. As state-of-the-art detectors are data driven models, special attention should be paid to prepare and present these data to the model. In addition to presenting the data in a way that enables efficient and effective utilization of the data in the development process, the presentation should be aligned with the user perception of the model input to facilitate a better perception of the model behavior. In this paper, two prominent object detectors are evaluated with two datasets. One dataset is presented more recognizable for a naked eye than the other. The performance of the two detectors on this dataset whose objects are presented in a form that is aligned with human's perception of these objects outperforms their performance on the other dataset. It is worth mention that in the UATD dataset, the bounding box is defined as a rectangle in the polar coordinates, which is mapped to a disk in the Cartesian coordinate system. Such mapping can result in having objects outside the defined boundaries when mapping is not considered while defining the bounding box. The attention mechanism in vision transformers can be utilized to explain the detector behavior and shed light onto their internal work. This mechanism can be used to visualize relations between different parts of an image, which can be compositional or correlation relations. Such visualization can help the end-user to verify the model predictions and the model developer to take any correction actions in the case of correlation relations. Visualizing the self-attention of DETR on marine-debris-fls raised the authors' concerns regarding the existence of location correlation relations as a result of collecting the sonar images from a single configuration of the objects in a water tank.

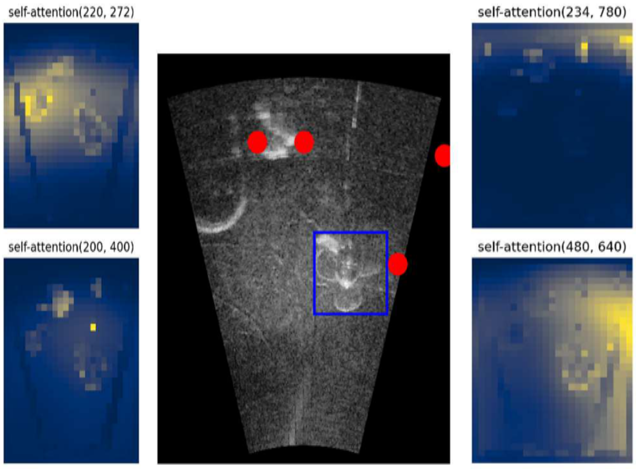

Fig. 7. The self-attention of four points are shown. The self-attention should highlight related areas. A self-attention of a point should show strong relations between points that belongs to the same object.

A possible direction of this research is to further explore the existence of location correlation in marine-debris-fls dataset. Another direction is to explore utilizing different transformer-based detectors for better performance on sonar data in terms of effectiveness, efficiency and explainability.

Acknowledgment

The work presented in this paper is funded by the Federal Ministry of Education and Research, Germany, grant number 01IW23003.

References

[1] L. Nolle, F. Stahl, and T. El-Mihoub, “On Explanations for Hybrid Artificial Intelligence,” in Artificial Intelligence XL, M. Bramer and F. Stahl, Eds., Cham: Springer Nature Switzerland, 2023, pp. 3–15.

[2] Y. Steiniger, D. Kraus, and T. Meisen, “Survey on deep learning based computer vision for sonar imagery,” Engineering Applications of Artificial Intelligence, vol. 114, p. 105157, Sep. 2022

[3] K. Xie, J. Yang, K. Qiu. "A Dataset with Multibeam Forward-Looking Sonar for Underwater Object Detection," in Scientific Data, vol. 9, no. 1, 2022.

[4] Jacob Solawetz and Francesco. What is yolov8? the ultimate guide., 2023. 01-12-2023.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-End Object Detection with Transformers,” in Computer Vision – ECCV 2020, vol. 12346, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., in Lecture Notes in Computer Science,

vol. 12346. , Cham: Springer International Publishing, 2020, pp. 213–229.

[6] M. Valdenegro-Toro, “Deep neural networks for marine debris detection in sonar images,” arXiv preprint arXiv:1905.05241, 2019.

[7] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, “Object Detection in 20 Years: A Survey,” Proc. IEEE, vol. 111, no. 3, pp. 257–276, Mar. 2023

[8] N. Iqbal, C. Manss, C. Scholz, D. Koenig, M. Igelbrink, and A. Ruckelshausen, “AI-based Maize and Weeds detection on the edge with CornWeed Dataset,” presented at the 18th Conference on Computer Science and Intelligence Systems, Sep. 2023, pp. 577–584.

[9] K. Han et al., “A Survey on Vision Transformer,” IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 45, no. 01, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” 2015, doi: 10.48550/ARXIV.1506.02640.

[11] Y. Yu, J. Zhao, Q. Gong, C. Huang, G. Zheng, and J. Ma, “Real-Time Underwater Maritime Object Detection in Side-Scan Sonar Images Based on Transformer-YOLOv5,” Remote Sensing, vol. 13, no. 18, p. 3555, Sep. 2021

[12] M. Valdenegro-Toro, “Objectness Scoring and Detection Proposals in Forward-Looking Sonar Images with Convolutional Neural Networks,” in Artificial Neural Networks in Pattern Recognition, vol. 9896, F. Schwenker, H. M. Abbas, N. El Gayar, and E. Trentin, Eds., in Lecture Notes in Computer Science, vol. 9896, Cham: Springer International Publishing, 2016, pp. 209–219.

[13] 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), Tokyo, Japan: IEEE, Nov. 2016, pp. 144–150.

[14] J. Kim and S.-C. Yu, “Convolutional neural network-based real-time ROV detection using forward-looking sonar image,” in 2016 IEEE/OES Autonomous Underwater Vehicles (AUV), Tokyo, Japan: IEEE, Nov. 2016, pp. 396–400.

[15] 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. 18023–18034, 2022.

[16] J. Wang, C. Feng, L. Wang, G. Li, and B. He, “Detection of Weak and Small Targets in Forward-Looking Sonar Image Using Multi-Branch Shuttle Neural Network,” IEEE Sensors J., vol. 22, no. 7, pp. 6772–6783, Apr. 2022.

[17] Z. Fan, W. Xia, X. Liu, and H. Li, “Detection and segmentation of underwater objects from forward-looking sonar based on a modified Mask RCNN,” SIViP, vol. 15, no. 6, pp. 1135–1143, Sep. 2021.

[18] A. Preciado-Grijalva, B. Wehbe, M. B. Firvida, and M. Valdenegro-Toro, “Self-supervised Learning for Sonar Image Classification,” in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA: IEEE, Jun. 2022, pp. 1498–1507.

[19] G. Neves, M. Ruiz, J. Fontinele, and L. Oliveira, “Rotated object detection with forward-looking sonar in underwater applications,” Expert Systems with Applications, vol. 140, p. 112870, 2020.

[20] X. Cao, L. Ren, and C. Sun, “Research on Obstacle Detection and Avoidance of Autonomous Underwater Vehicle Based on Forward-Looking Sonar,” IEEE Trans. Neural Netw. Learning Syst., vol. 34, no. 11, pp. 9198–9208, Nov. 2023.

[21] Y. Tang, L. Wang, H. Li, and S. Bian, “Side-scan sonar underwater target segmentation using the BHP-UNet,” EURASIP J. Adv. Signal Process., vol. 2023, no. 1, p. 76, Jun. 2023.

[22] Y. Tang, S. Jin, G. Bian, Y. Zhang, F. Li, Wreckage Target Recognition in Side-scan Sonar Images Based on an Improved Faster R-CNN Model. International Conference on Big Data & Artificial Intelligence & Software Engineering (2020), pp. 348–354.

[23] Y. Song, B. He, and P. Liu, “Real-Time Object Detection for AUVs Using Self-Cascaded Convolutional Neural Networks,” IEEE J. Oceanic Eng., vol. 46, no. 1, pp. 56–67, Jan. 2021.

[24] T. Yulin, S. Jin, G. Bian, and Y. Zhang, “Shipwreck target recognition in side-scan sonar images by improved YOLOv3 model based on transfer learning,” IEEE Access, vol. 8, pp. 173450–173460, 2020.

[25] H. Thanh Le, S. L. Phung, P. B. Chapple, A. Bouzerdoum, C. H. Ritz, and L. C. Tran, “Deep Gabor Neural Network for Automatic Detection of Mine-Like Objects in Sonar Imagery,” IEEE Access, vol. 8, pp. 94126–94139, 2020.

[26] Zhang, P.; Tang, J.; Zhong, H.; Ning, M.; Liu, D.; Wu, K. Self-Trained Target Detection of Radar and Sonar Images Using Automatic Deep Learning. IEEE Transactions on Geoscience and Remote Sensing 2022, 60, 1–14.

[27] Z. Wang, J. Guo, L. Zeng, C. Zhang, and B. Wang, “MLFFNet: Multilevel Feature Fusion Network for Object Detection in Sonar Images,” IEEE Trans. Geosci. Remote Sensing, vol. 60, pp. 1–19, 2022.

[28] W. Kong et al., “YOLOv3-DPFIN: A Dual-Path Feature Fusion Neural Network for Robust Real-Time Sonar Target Detection,” IEEE Sensors J., vol. 20, no. 7, pp. 3745–3756, Apr. 2020.

[29] L. Jiang, T. Cai, Q. Ma, F. Xu, and S. Wang, “Active Object Detection in Sonar Images,” IEEE Access, vol. 8, pp. 102540–102553, 2020.

[30] Z. Wang, S. Zhang, W. Huang, J. Guo, and L. Zeng, “Sonar Image Target Detection Based on Adaptive Global Feature Enhancement Network,” IEEE Sensors J., vol. 22, no. 2, pp. 1509–1530, Jan. 2022.

[31] L. Li, Y. Li, C. Yue, G. Xu, H. Wang, and X. Feng, “Real-time underwater target detection for AUV using side scan sonar images based on deep learning,” Applied Ocean Research, vol. 138, p. 103630, 2023.

[32] D. Połap, A. Jaszcz, N. Wawrzyniak, and G. Zaniewicz, “Bilinear Pooling With Poisoning Detection Module for Automatic Side Scan Sonar Data Analysis,” IEEE Access, vol. 11, pp. 72477–72484, 2023,.

[33] T. A. El-Mihoub, L. Nolle, and F. Stahl, “Explainable Boosting Machines for Network Intrusion Detection with Features Reduction,” in Artificial Intelligence XXXIX, vol. 13652, M. Bramer and F. Stahl, Eds., in Lecture Notes in Computer Science, vol. 13652. , Cham: Springer International Publishing, 2022, pp. 280–294. doi: 10.1007/978-3-031-21441-7_20.

[34] O. Li, H. Liu, C. Chen, and C. Rudin, “Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions,” 2017, doi: 10.48550/ARXIV.1710.04806.R. K. Mothilal, A. Sharma and C. Tan, "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations," in FAT* ’20,, Barcelona, Spain, 2020.

[35] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR,” Harvard journal of law & technology, vol. 31, pp. 841–887, Apr. 2018, doi: 10.2139/ssrn.3063289.

[36] G. Pruthi, F. Liu, M. Sundararajan, and S. Kale, “Estimating Training Data Influence by Tracing Gradient Descent,” 2020, doi: 10.48550/ARXIV.2002.08484.