Enhancing Underwater Object Detection, Multi-Label Classification, and Out-of-Distribution Detection with Advanced Deep Learning Techniques and Augmentation Methods

Md Sazidur Rahman
Université de Toulon,
Instituto Superior Técnico

David Cabecinhas
Institute for Systems and Robotics
Instituto Superior Técnico

Ricard Marxer
Université de Toulon, Aix Marseille Univ,
CNRS, LIS, Toulon, France

***Abstract*—The thesis explores the advanced deep learning techniques to elaborate on underwater object detection, multi-label classification, and out-of-distribution detection and focuses especially on the Fathomnet competition dataset, a comprehensive and open source dataset of underwater images. We propose a novel augmentation approach, Depth Jitter; a method dedicated to correcting the color distortions introduced by depth-related features. Our method yielded a performance gain in terms of an expected minimum average precision score of 2-3% mAP@20. Despite challenges like the underwater images' data imbalance and environmental variability, the models such as Query2Label and YOLOv9 with our augmentation technique demonstrated robustness and adaptability, and are competently able to obtain results without referring to external datasets. This study uses the advancement of these models as a starting point for exploring their utility in the advancement of key applications in marine studies–codifying species and observing habitat–for the purpose of marine conservation. Future research will focus on increasing dataset diversity, improving techniques for dealing with data imbalance, improving model interpretability, and exploring options for real-time deployment. Integrating hybrid models and improving out-of-distribution detection will help to advance the reliability and applicability of underwater image analysis.**

I. Introduction

Academic studies confirm that marine life substantially predates terrestrial life. According to [1], life in the ocean is documented as having originated approximately 3.7 billion years ago, while terrestrial life is believed to have appeared around 3.1 billion years ago, as suggested by [2]. The fossil record, as detailed by [3], reveals that marine biodiversity has surpassed terrestrial diversity for about 3.6 billion years. Oceans, which encompass 71% of the Earth's surface, support a higher species richness, aligning with bio-geographic theories that correlate habitat extent with biodiversity [4].

Furthermore, the deep sea, which represents about two-thirds of the planet's area and includes 84% of the ocean's surface and 98% of its volumetric expanse below 2,000 meters, remains the least explored region on Earth [4]. This vast and understudied area is likely a reservoir for a multitude of yet-to-be-discovered species. These species are pivotal not only for comprehending adaptive strategies of life under extreme conditions but also play critical roles in global ecological processes such as climate regulation and the carbon cycle, as discussed by [5], [6]. This realization highlights the fundamental importance of marine biodiversity in the global ecological dynamics, and it warrants additional and broad exploration and investigation.

Conducting research in these elusive regions frequently faces technological and logistical challenges, primarily due to high costs and the need for advanced safety features. Recently, the development and deployment of autonomous marine devices, including autonomous underwater vehicles (AUVs) and submarine gliders, have significantly enhanced our ability to map and monitor the marine environment. These devices are equipped with sophisticated acoustic sensors and imaging technologies, allowing for a more detailed exploration of the underwater world [7]–[10].

Recent improvements in computer vision methodologies have greatly enhanced the capabilities for better observing and understanding marine systems. For example, image and video classification methodologies based on Convolutional Neural Networks (CNNs) allow for the automatic identification and labeling of marine species or objects contained in underwater images and videos. For instance, Autonomous Underwater Vehicles (AUVs) using CNN-based models can automatically classify various organisms and debris in situations where real-time identification is useful (e.g., biodiversity monitoring, species health monitoring and invasive species detection), which greatly decreases effort put forth in the detection and classification to any type of manual analysis.

Extending these methodologies, algorithms such as YOLO (You Only Look Once) and DETR (Detection Transformer) use object detection methodologies that identify and not only categorize the objects contained in the images or moving video. The ability to localize objects within multiple images and video sequences is paramount in tracking population dynamics of species, studying patterns such as migration, and notating human impact which can lead to pollution events.

Additionally, segmentation models, such as Segment Anything, extract the object(s) from its background, and they are able to disengage and differentiate objects even though they overlap in the image. This permits scientists to count distinct organisms of species, analyze interactions between organisms in the same group (e.g., schooling applications of fish), and classify between healthy and bleached corals.

An Out-of-Distribution (OOD) detection framework in computer vision datasets has a vast useful science application for identifying observations distinct from input data. OOD detection in practice, can also be useful for flagging unrecognizable fish species or identifying new types of debris, thus avoiding misclassification by the model and highlighting where new science questions might be pursued. Notably, a major limitation of contemporary methods, primarily pertaining to species and marine education and research, is their heavy reliance on existing dataset metrics for model training, of which are likely not comprehensive representations of data. This method of computation and application is especially disadvantageous for unpublished and unstudied deep sea sciences where much of (global) marine biodiversity is considered by definition to be 'unknown.' Both unknown species and unknown or unmodeled morphology, or both, are less likely to exist in trained data, which may prevent the model from classifying or recognizing the species.

The primary aim of this research is to refine the methodologies used for object recognition and the detection of out-of-distribution instances in studies of deep-sea biodiversity. By enhancing these methods, this study seeks to advance the field of ecological informatics and provide a more profound understanding of deep-sea ecosystems. The significance of this research transcends academic confines, offering essential insights for environmental conservation, sustainable management of marine resources, and an enriched comprehension of life in extreme conditions. Consequently, this study is poised to contribute significantly to marine science, technological innovation, and ecological preservation, underscoring its potential to influence a broad spectrum of scientific and practical fields.

A. *Motivation of this Research*

This research was inspired by the vast and mysterious depth of the ocean and the motivation to learn and understand more about these unexplored regions of the ocean. The deep ocean covers around 71% of our planet and remains on e of the unexplored ecosystems on earth with many rare and undiscovered species. The research presented in this dissertation will lead to a greater understanding of some of the obscured aspects of marine biology and oceanography.

A critical motivation for this study is also the pressing need to develop more advanced tools for biodiversity research and underwater exploration. The challenges of studying marine habitats beneath the surface and the limitations imposed by human accessibility often hinder traditional methods of marine research. This research seeks to transcend these barriers by leveraging state-of-the-art deep learning technologies, offering a more efficient, accurate, and comprehensive approach to exploring the deep sea.

Additionally, this effort is driven by an urgent environmental imperative. Understanding the intricate dynamics of these environments is crucial for the protection and sustainable management of marine ecosystems, especially as human activities continue to impact them profoundly. Through this research, we aim to enhance our ability to monitor and safeguard the rich biodiversity found beneath the waves.

Ultimately, This thesis is ultimately driven by scientific curiosity, determination to advance the technical challenges of deep-sea science, and the desire to be a steward of the environment. To answer questions related to the depths of the ocean and preserve it for future generations, we must challenge the limits of what we know and what we can do.

B. *Contribution*

This research investigates the techniques to enhance the out-of-distribution(OOD) detection performance through multilabel classification or object detection.

- We introduce a depth based augmentation technique for underwater datasets which helps to improve the performance of object classification and detection in underwater environments.
- We provide a moderate benchmark of Fathomnet competition dataset on different multilabel classification and object detection models which enables further research opportunities on the same dataset for out of distribution detection.

II. State of the Art

In this chapter, we summarize the state-of-the-art (SOTA) methods for two tasks: **object detection** and **multi-label classification**, which will serve as a foundation for out-of-distribution (OOD) detection. Object detection focuses on identifying and localizing objects in images, outputting bounding boxes and class labels, while multi-label classification predicts several applicable labels for an image. Both tasks are typically evaluated using the mean Average Precision (mAP) metric. We also discuss benchmark datasets like COCO and Pascal VOC, noting the limitations of traditional methods in underwater environments.

A. *Evolution of Object Detection Models*

Object detection can be categorized into two phases: *traditional object detection* before 2014 and *deep learning-based detection* after 2014 [11].

Traditional methods relied heavily on hand-crafted features, limiting their adaptability to complex environments such as underwater. Early examples include the **Integral Image** for real-time face detection by Viola and Jones [12], **HOG** for human detection by Dalal and Triggs [13], and **Deformable Part Models (DPMs)** [14], which modeled objects as sets of deformable parts. These methods struggled with variability in lighting, viewpoint, and occlusion, key challenges in marine environments.

Deep learning revolutionized object detection with the advent of Convolutional Neural Networks (CNNs) in 2012 [15]. This led to two major classes of object detectors:

1. **Two-Stage Detectors**: These include R-CNN [16], Fast R-CNN [17], and Faster R-CNN [18], which first generate object proposals using methods like selective search [19] and then refine them. These methods are highly accurate, but slower than one-stage detectors.

2. **One-Stage Detectors**: These aim for real-time object detection, with YOLO (You Only Look Once) [20] achieving significant speed improvements. SSD [21] further advanced one-stage detection by using multi-resolution feature maps. DETR [22] later introduced a transformer-based end-to-end system, eliminating anchor boxes.

B. Multi-label Classification

Multi-label classification assigns more than one label to an image, which is particularly useful in environments like underwater scenes where multiple species or objects can be present.

Initial works, such as Binary Relevance (BR) [23], treated each label as an independent classification task, which is simple but ignores label correlations. To address this, methods like Classifier Chains (CC) [24] extended BR by considering dependencies between labels. Another notable method, RAkEL [25], generates a pool of classifiers based on random label sets to capture label correlations more effectively.

Recent advances combine deep learning models with label dependency structures. The CNN-RNN framework [26] integrates CNNs to extract features and RNNs to capture label dependencies. Feng Zhu et al. introduced a Spatial Regularization Network [27] to model spatial relationships between labels, improving multi-label classification performance. Furthermore, Renchun You et al. developed a framework using cross-modality attention and semantic graph embedding [28] to handle the challenges of multi-label classification.

C. Underwater Object Detection and Classification

Object detection in underwater environments poses unique challenges due to variable lighting, visibility, and the dynamic nature of aquatic organisms. Traditional methods relying on manually engineered features [29] have been limited by their dependency on specialized knowledge and their inability to adapt to real-world conditions.

The shift towards deep learning has resolved many of these challenges by allowing models to learn feature representations directly from data. Wu et al. [30] investigated the impact of lighting on underwater object detection, revealing that lighting variations significantly affect detection accuracy. Peng et al. [31] reviewed deep learning techniques for preprocessing underwater images, noting that while these methods address many challenges, improvements in architecture design and integration of domain-specific knowledge are needed for further advancements. Jian et al. [32] did a comprehensive survey on the techniques used by the researchers for object detection and tracking in underwater images. The summary is given in a table format below -

TABLE I
SUMMARY OF UNDERWATER OBJECT DETECTION METHODS BASED ON TRADITIONAL ARTIFICIAL FEATURES.

Category	References
Texture features	Han and Choi [33]; Beijbom et al. [34]; Nagaraja et al. [35]; Fatan et al. [36]; Srividhya and Ramya [37]; Shi et al. [38]
Color and motion features	Gordan et al. [39]; Chen and Chen [40]; Singh et al. [41]; Komari Alaie and Farsi [42]; Susanto et al. [43]
Saliency detection	Wang et al. [44]; Zhu et al. [45]; Jian et al. [46]

D. Improving Model Robustness through Data Augmentation

Data augmentation is a key technique to enhance model robustness, especially in challenging environments like underwater imaging. By artificially expanding the dataset, augmentations such as geometric transformations (e.g., rotation, scaling, flipping) and color adjustments (brightness, contrast, hue) help models generalize better to unseen data. Tools like PyTorch's ColorJitter can simulate underwater conditions. Additionally, methods such as adding noise, blurring, and sharpening address challenges like low visibility and image quality variability. Cutout techniques, where parts of an image are randomly obscured, train models to handle occlusion, a common issue with marine organisms or vegetation. Generating synthetic data using GANs also provides diverse training samples when real data is scarce. These augmentation strategies significantly improve the robustness, adaptability, and performance of models in complex underwater environments.

E. Challenges Limitations

1) Technical Challenges:

a) Occlusions in Object Detection: Underwater environments are complex, with marine life and vegetation often occluding objects, reducing detection accuracy. Traditional models struggle with occlusion unless specifically trained on underwater scenes.

b) High False Positives in Anomaly Detection: Lighting variability and swift water flow create challenges in anomaly detection, leading to false positives. Misinterpretations, like harmless objects being identified as anomalies, require models to define 'normality' more accurately in underwater contexts.

c) Complexity in Multi-label Classification: Underwater images often require multi-label classification due to overlapping species and objects. Classical classifiers, which treat labels independently, struggle with the dependencies between different objects, resulting in misclassification.

2) Broader Issues:

a) Dataset Biases: Most datasets are biased toward terrestrial data, leading to poor performance in underwater applications. Developing diverse and representative underwater datasets is essential for improving model accuracy.

b) Model Interpretability: The black-box nature of deep learning models makes it difficult to understand and troubleshoot decisions in critical applications, such as identifying underwater hazards. Explainable AI is needed to build trust in model decisions.

c) Computational Demands: Real-time analysis of high-resolution underwater images is resource-intensive. Models that can run efficiently on low-power devices are necessary, especially for remote underwater monitoring.

d) Lack of Underwater Benchmarks: There are few standardized benchmarks for evaluating models in underwater environments. Developing these benchmarks would help assess model performance under the unique challenges of underwater conditions.

e) Environment Variability: Changing water clarity, light, and depth significantly impact object appearance, making it hard for models to perform consistently. Future models need to be more adaptable to these changes.

f) Unpredictable Elements: Unexpected objects like marine creatures or human debris may confuse models not trained on such data. Developing models that can quickly adapt to novel elements remains a key challenge.

g) Need for Specialized Training Data: To address these challenges, developing comprehensive underwater datasets is crucial. Collaboration between AI, oceanography, and marine biology can facilitate robust training, along with synthetic data generation and augmentation.

III. Methodology

A. Fathomnet Competition Dataset 2023

The Fathomnet 2023 Competition Dataset is divided into depth-specific subsets to explore the distribution of the marine organisms. The training data comprises images from 0 to 800 meters depth while the evaluation data extends up to 1300 meters depth. This setup helps to investigate how species distribution varies with depth. The species distributions in these two regions overlap, but they are not exactly the same and they diverge as the vertical distance increases. Figure 1 shows the depth distribution in both train and test set.

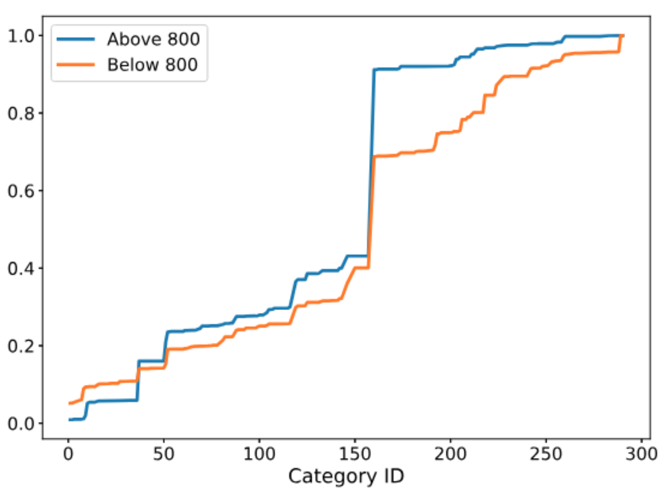

Fig. 1. The overall distribution of categories in the FathomNet 2023 training and evaluation datasets. The datasets differ significantly, with some classes only present in one dataset [47].

Characteristics: The Fathomnet training set contains 5950 images with 23703 localized annotations and the evaluation set contains 10744 images with 49798 localized annotations.Of these, 6,313 images were collected from deeper waters, beyond a certain depth threshold, and are considered as out-of-distribution data [47]. The Fathomnet dataset is relatively long tailed as most of the fine grained image datasets. The most frequently occurred category in both training and evaluation dataset is *S. fragilis*. Beyond *S. fragilis* the order and magnitude of the other categories is quite variable between the sets. The dataset contains 290 categories which belong to 20 semantic supercategories. In the training set 157 categories of the 290 categories do not have any images and almost 80 of the categories have less than 10 samples from which it can be told that the dataset is very imbalanced and long tailed. Figure 2 shows the distribution

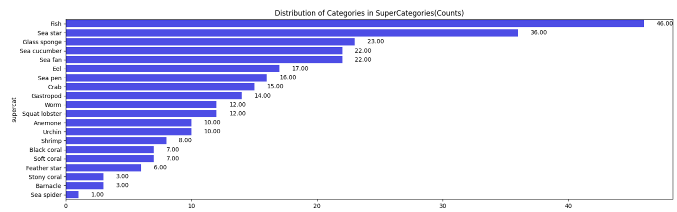

Fig. 2. Categories Count in Supercateogires

of the number of categories in the supercategories. It can be seen that fish is the most represented and sea spider is the least represented supercategory in the Fathomnet 2023 Competition dataset.

B. Depth Jitter

As the Fathomnet dataset comes with object detection and multi-label classification annotations we tried both methods to check which one performs best in terms of out of distribution detection and marine species category classification. Underwater imagery presents challenges due to high variability in appearance caused by light transport in the medium. Factors such as light absorption and scattering significantly affect colors depending on the view and the distance from the object. To improve this variability and improve model robustness we came up with a data augmentation technique to reproduce this variability.

To achieve this, we relied on the existing works on underwater image formation models and color restoration. For our augmentation method, we first estimate depth of the scene in a given image, along with the underwater image formation parameters that are relevant to the particular image. Once these image formation parameters are established, we then apply offsets to the estimated depth and re-render the image with the same parameters. We simulate natural variability of underwater scenes by applying this method, which creates a more diverse training set for species classification, and better prepares our model for actual scenarios. Lastly, in addition to inputs into our model, we performed stratified sampling to mitigate class imbalance in a multi-label classification task with respect to species categories.

C. Underwater Image Formation Model

The underwater image formation model describes how colors are influenced by the water medium, helping restore images affected by underwater conditions [48]. Key variables in this model are summarized in Table II.

TABLE II
VARIABLES IN THE UNDERWATER IMAGE FORMATION MODEL. SOURCE: [48].

Variable	Description	Type
I	Underwater images	$R^{N\times H\times W\times C}$
J	Restored images	$R^{N\times H\times W\times C}$
z	Distance maps	$R^{N\times H\times W}$
B	Veiling light	R^{C}
β	Color absorption coefficient	R^{C}
γ	Backscatter coefficient	R^{C}

Schechner and Karpel's initial model [49] inspired various restoration techniques [50]–[52] addressing backscatter and color absorption. The pixel intensity is modeled by:

$$I_{c,p} = J_{c,p}e^{-\alpha_c z_p} + B_c(1 - e^{-\alpha_c z_p}), \tag{1}$$

where $\alpha \in R^C$ is the wavelength-dependent coefficient. Akkayanak et al. [53] later refined this model to distinguish between backscatter and absorption coefficients more accurately.

D. Using Underwater Image Formation Model for Data Augmentation

The underwater image formation For our multi-label image classification we used the image formation model described in the eq.1 as one of the augmentation techniques. We performed this augmentation on the go while training the classification model. Here we further optimize the calculation of $I_{c,p}$. Suppose $I^{orig}_{c,p}$ is the original image and $I^{mod}_{c,p}$ is the restored modified image. We can denote both of the image by the image formation model as follows:

$$I^{orig}_{c,p} = J_{c,p}e^{-\beta_c z_p} + B_c(1 - e^{-\gamma_c z_p}), \tag{2}$$

and

$$I^{mod}_{c,p} = J_{c_p}e^{-\beta_c z_m} + B_c(1 - e^{-\gamma_c z_m}), \tag{3}$$

we can eliminate $J_{c,p}$ as follows:
First, solve for $J_{c,p}$ from Equation 2:

$$J_{c,p} = \frac{I^{orig}_{c,p} - B_c(1 - e^{-\gamma_c z_p})}{e^{-\beta_c z_p}} \tag{4}$$

Next, substitute this expression for $J_{c,p}$ into Equation 3:

$$I^{mod}_{c,p} = \left(\frac{I^{orig}_{c,p} - B_c(1 - e^{-\gamma_c z_p})}{e^{-\beta_c z_p}}\right) e^{-\beta_c z_m} + B_c(1 - e^{-\gamma_c z_m}) \tag{5}$$

Simplify the equation:

$$I^{mod}_{c,p} = \left(I^{orig}_{c_p} - B_c(1 - e^{-\gamma_c z_p})\right) e^{-\beta_c(z_m - z_p)} + B_c(1 - e^{-\gamma_c z_m}) \tag{6}$$

In the equation 6 we can insert Δz_m as the depth offset added with the original depth map. Thus it will generate synthetic data with depth offsets which we can use as the data augmentation to make out model more robust to different color and depth settings in the underwater environment. Figure 3 illustrates the pixel intensity tracking in different depths. The further we take the depths the pixel intensity reduces and the color of the image becomes more dark. On the other hand, the more less the depth the more pixel intensity and color intensity the image has. The change of intensity can be easily seen in the depth maps of the corresponding images in the figure 3. While applying the augmentation we select a range of depths based on the dataset, in our case the depth range is -5m to 5m which gives the most optimal results. We apply the range randomly so each time the batch is called the depth offset applied to the image is different. That's how this technique makes the dataset more diverse.

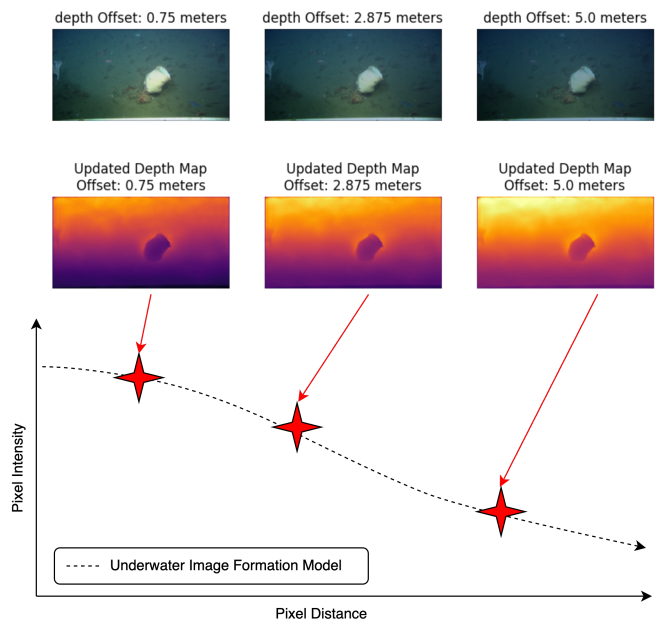

Fig. 3. This figure shows the pixel intensity tracking and change on the image in different depth settings.

It is important to note that the parameter estimation of the Underwater Image Formation Model (UIFM) and DepthAnything inference may be computationally expensive, which can be handled in the pre-training phase. The only computation that will occur in the training iteration is the re-rendering images with the adjusted depth offset (expressed in Equation 6). This makes this augmentation strategy more efficient in the training procedure.

IV. System Overview

A. Multilabel-classification

This section provides a brief overview of the multi-label image classification system based on the Query2Labels (Q2L) model, which is inspired by the DETR architecture. Q2L utilizes a CNN backbone with a transformer for predicting label probabilities. Below, we summarize the key components: feature extraction and query updating.

1) Feature Extraction: Given an input image $x \in R^{H_0 \times W_0 \times 3}$, spatial features are extracted using a backbone network (e.g., ResNet or Vision Transformer). The resulting feature map $F_0 \in R^{H \times W \times d_0}$ is projected via a linear layer to reduce feature dimensions. The reshaped matrix $F \in R^{HW \times d}$ is then processed further [54].

2) Query Updating: Label embeddings $Q_0 \in R^{K \times d}$ act as queries in the transformer. Using self-attention and cross-attention modules, the model pools category-related features from the spatial features. Each transformer decoder layer updates the queries to capture contextual information from the input image, resulting in enriched label embeddings for each category [54].

The queries are treated as learnable parameters to implicitly model label correlations. Each class-specific query is projected to a logit value, and a sigmoid function is applied to predict category probabilities p_k. The asymmetric loss (ASL), a variant of focal loss, handles class imbalance by assigning different γ values to positive and negative samples [55].

B. Object Detection

We evaluated several state-of-the-art object detection models, including DETR, Co-DETR, Deformable DETR, RT-DETR, YOLOv8, YOLOv9, and YOLOv5. After comprehensive analysis, YOLOv9 was selected as the primary model due to its superior accuracy, efficiency, and robustness.

1) YOLOv9 Architecture: YOLOv9 builds on previous YOLO versions with key innovations, including the **Generalized Efficient Layer Aggregation Network (GELAN)** and **Programmable Gradient Information (PGI)**.

a) GELAN:: GELAN efficiently aggregates features across layers, leveraging conventional convolution operators for better parameter utilization. It resolves information bottlenecks by facilitating smoother gradient flow.

b) PGI:: PGI addresses information loss during feedforward processing through an auxiliary reversible branch. This branch generates reliable gradients, improving the accuracy of predictions and parameter updates.

2) Network Components::

1) **Main Branch:** Used for inference and integrates GELAN for feature extraction without additional inference costs.
2) **Auxiliary Reversible Branch:** Maintains information using reversible transformations to generate reliable gradients for the main branch.
3) **Multi-Level Auxiliary Information:** Aggregates gradient information across multiple levels to enhance prediction accuracy.

These components enable YOLOv9 to achieve a high level of efficiency and accuracy, making it suitable for diverse object detection tasks.

C. Out-of-Distribution (OOD) Score Calculation Methods

We implemented two methods to assess the likelihood of a sample being out-of-distribution (OOD):

1) Method 1: Maximum Softmax Probability (MSP): MSP calculates the OOD score based on the maximum softmax probability for a given sample. In-distribution samples will have high confidence (high softmax probability), while lower probabilities indicate the sample is likely OOD. The OOD score is defined as:

$$\text{OOD} = \begin{cases} 1.0 & \text{if } |\hat{C}| = 0 \\ 1 - \max_{c \in C} P(c|x) & \text{otherwise} \end{cases} \tag{7}$$

Where:

- $\hat{C}$ represents the set of predicted classes for the input x.
- C is the set of all possible classes.
- $P(c|x)$ is the softmax probability for class c given input x.

If no classes meet the threshold, the OOD score is set to 1.0, indicating uncertainty.

2) Method 2: Average Confidence Score: This method computes the OOD score by averaging the confidence scores of all predictions. Lower average confidence implies a higher likelihood of being OOD:

$$OOD = 1 - \frac{1}{N} \sum_{n=1}^{N} \text{conf}_n \tag{8}$$

Where conf_n is the confidence score for the n-th prediction, and N is the total number of predictions.

Both methods provide simple and effective ways to measure uncertainty and detect out-of-distribution samples in multi-label classification tasks.

V. Results & Discussions

VI. Quantitative Evaluation

This section evaluates the performance of object detection and multi-label classification models on the complex underwater imagery of FathomNet. The primary focus is on detecting marine species and out-of-distribution (OOD) samples, and the effectiveness of various data augmentation techniques, particularly the proposed Depth Jitter, in improving model robustness.

We conducted experiments using the FathomNet 2023 competition framework hosted by MBARI. Evaluation metrics include mean Average Precision (mAP) for species prediction and Area Under the Receiver Operating Characteristic Curve (AUC) for OOD detection.

A. Evaluation Metrics

To evaluate model performance, we used mAP@20 for multi-label species classification and AUC for OOD detection. These metrics help measure both the model's ability to classify marine species and detect samples outside the training distribution.

B. Out-of-Distribution Detection

We assess OOD detection using the AUC score. The ROC curve, based on True Positive Rate (TPR) and False Positive Rate (FPR), shows how well models differentiate between in-distribution (ID) and OOD samples. The formulae for TPR and FPR are:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}.$$

The AUC score ranges from 0.5 (random guessing) to 1.0 (perfect classifier).

C. Category Predictions

We used mAP@20 to assess multi-label category predictions. The metric evaluates how well the model ranks relevant categories in the top 20 predictions for each image:

$$\mathrm{mAP@20} = \frac{1}{U}\sum_{u=1}^{U} \min\left(\sum_{k=1}^{\min(n,20)} P(k) \times \mathrm{rel}(k), 1\right),$$

where U is the number of images and $P(k)$ is the precision at cutoff k.

D. Final Score

The final score combines the mAP@20 and AUC metrics:

$$\text{Final Score} = \frac{1}{2}(\mathrm{sAUC}+\mathrm{mAP@20}), \quad \mathrm{sAUC} = 2\times\mathrm{AUC}-1.$$

E. Object Detection Performance

We benchmarked various object detection models on the FathomNet dataset. YOLOv9 achieved the highest mAP@20 score of 0.74 on the validation set, outperforming other models. Table III summarizes the performance.

TABLE III
PERFORMANCE OF DIFFERENT OBJECT DETECTION MODELS ON THE FATHOMNET DATASET.

Models	Backbone	Epochs	Train Size	Test Size	mAP@50	mAP@20	Training Time
Deformable DETR	Resnet50	70	640X640	640X640	0.196	0.40	24 hours
DETR	Resnet50	150	640X640	640X640	0.209	0.42	36 hours
Conditional DETR	Resnet50	91	640X640	640X640	0.259	0.50	24 hours
YOLOv8m	cspDarknet	100	640X640	640X640	0.300	0.62	19 hours
lightgray Yolov8m (Baseline)	cspDarknet	50	640X640	640X640	0.330	0.69	-
YOLOv9-e	Gelan	120	640X640	640X640	0.391	**0.74**	22 hours

F. Performance of Query2Label with Augmentations

We evaluated the Query2Label model under different augmentation settings. Depth Jitter, our proposed augmentation, achieved the best validation mAP@20 score of 0.85, outperforming both Clean and ColorJitter augmentations (Figure **??**). Table IV shows the results of different configurations.

TABLE IV
PERFORMANCE OF DIFFERENT MULTI-LABEL CLASSIFICATION MODELS ON THE FATHOMNET DATASET.

Augmentation	Backbone	Train Size	Test Size	Train Loss	Val Loss	Val mAP	Val mAP@20
Clean	ResNest101e	384X384	640X640	0.095	0.234	0.751	0.813
Color Jitter	ResNest101e	384X384	640X640	0.187	0.227	0.762	0.827
Depth Jitter (Ours)	ResNest101e	384X384	640X640	0.165	0.223	**0.803**	**0.855**

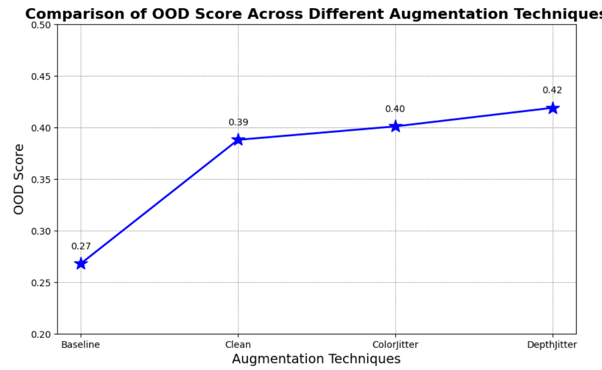

Fig. 4. Comparison of Out-of-Distribution (OOD) Scores Across Different Augmentation Techniques.

G. OOD Score Performance

We also evaluated OOD detection across various augmentations. Depth Jitter achieved the highest OOD score of 0.42, demonstrating its effectiveness in improving model robustness (Figure 4).

H. Kaggle Competition Performance

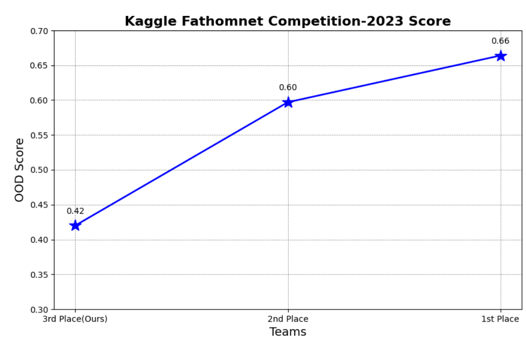

Fig. 5. Comparison of OOD Scores for Top Teams in Kaggle Fathomnet Competition-2023.

In the Kaggle FathomNet Competition 2023, our method would have ranked 3rd. Figure 5 compares OOD scores for the top three teams.

VII. LIMITATIONS

A. Technical Limitations

Training models like Query2Label and YOLOv9 requires high-performance hardware, making them resource-intensive. Query2Label and YOLOv9 each required up to 48 GB of VRAM and lengthy training times.

B. Data-Related Limitations

FathomNet's long-tailed distribution causes poor performance on rare classes. The dataset's limited geographic diversity also affects model generalization to new contexts.

C. Environmental Constraints

Underwater conditions, such as light variability and turbidity, challenge model robustness. Depth Jitter is sensitive to lighting conditions, limiting its effectiveness under extreme scenarios.

D. Interpretability and Usability

Although attention maps offer some interpretability, deep learning models remain black boxes, complicating trust in outputs. Usability for end-users, such as marine biologists, needs further validation.

Addressing these limitations is crucial for improving the system's robustness and applicability in real-world underwater exploration.

VIII. Conclusion

This thesis has demonstrated the effectiveness of advanced deep learning techniques, specifically the Query2Label and YOLOv9 models, for underwater object detection and multi-label classification. The challenging underwater environment, characterized by poor visibility, varying lighting conditions, and complex object appearances, necessitates robust and adaptive methods. Our research introduced the DepthJitter augmentation method, which proved to enhance model performance. The DepthJitter technique addresses the issue of depth-related color distortions, thereby improving the models' ability to generalize across varying underwater conditions. As a result, the models achieved superior mAP@20 scores on the Fathomnet dataset, surpassing other conventional augmentation methods. The application of these models was rigorously tested on the Fathomnet competition dataset, which provided a diverse and realistic set of underwater images. Despite the inherent challenges such as data imbalance and environmental variability, our models performed competitively. One of the noteworthy aspects of this research is the reliance solely on the provided dataset without incorporating external data sources. This constraint underscores the robustness and adaptability of the models and the efficacy of the DepthJitter augmentation. The models' success in handling complex underwater scenes and accurately detecting and classifying multiple marine species highlights their potential for practical applications in marine research and conservation.

A. Future Work

For future research, several key areas need to be addressed to further enhance the performance and applicability of underwater image analysis models. First, enhancing dataset diversity is crucial. Incorporating more varied and extensive datasets can help models learn a wider range of underwater scenarios, thereby improving their generalization capabilities. This can include expanding the dataset with images from different geographic locations, depths, and environmental conditions.

Addressing data imbalance remains a significant challenge. Techniques such as synthetic data generation, oversampling of underrepresented classes, and advanced augmentation methods can be explored to ensure a more balanced representation of different marine species. This can help in improving the detection accuracy for rare and less frequent objects, which are often missed by current models.

Improving model interpretability is another critical area. Understanding how models make decisions and what features they focus on can help in diagnosing errors, improving model architecture, and building trust in automated systems. Techniques such as attention mechanisms and visualization tools can be further developed to provide deeper insights into the models' workings.

Real-time deployment of these models is a promising direction that can significantly impact marine research and conservation efforts. Developing lightweight and efficient models that can operate on low-power devices will enable real-time monitoring and analysis of underwater environments. This can facilitate immediate detection and response to ecological changes, illegal activities, or other significant events.

Exploring hybrid models that integrate multi-modal data, such as combining visual data with sonar or environmental sensors, can provide a more comprehensive understanding of underwater scenes. This multi-modal approach can enhance the accuracy and reliability of detection and classification tasks.

Ensuring robustness to environmental changes is essential for practical deployment. Models need to be resilient to variations in water clarity, lighting conditions, and other environmental factors. Techniques such as domain adaptation and transfer learning can be explored to improve model robustness.

Lastly, improving out-of-distribution detection is vital for enhancing model reliability. Models should be able to identify when they encounter data that is significantly different from their training data and respond appropriately. This capability is crucial for ensuring the safety and effectiveness of autonomous underwater systems.

By pursuing these directions, future work can build upon our findings, contributing to more effective and reliable underwater image analysis techniques. This will not only advance the field of computer vision but also support marine exploration, research, and conservation efforts, helping to protect and preserve our oceans.

Acknowledgment

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Dr. Ricard Marxer, for his continuous support, guidance, and encouragement throughout the course of this research. His insightful feedback and unwavering belief in my capabilities have been invaluable to the completion of this thesis. I am also profoundly grateful to my co-supervisor, Dr. David Cabecinhas, for his expertise, patience, and constructive criticism, which have significantly contributed to the quality and direction of my work.

I would like to extend my sincere thanks to the LIS Lab at Université de Toulon for providing the financial support and resources necessary for this research. The funding and facilities offered by the LIS Lab have been instrumental in facilitating my experiments and enabling me to pursue my research objectives.

Additionally, I would like to acknowledge the faculty and staff of Université de Toulon and Instituto Superior Técnico for their support and assistance during my studies. Special thanks go to my colleagues and friends who have provided a stimulating and supportive environment in which to learn and grow.

Lastly, I would like to thank my family for their unconditional love and support. Their encouragement and belief in me have been a constant source of strength and motivation throughout this journey.

REFERENCES

[1] M. S. Dodd, D. Papineau, T. Grenne, J. F. Slack, M. Rittner, F. Pirajno, J. O'Neil, and C. T. S. Little, "Evidence for early life in Earth's oldest hydrothermal vent precipitates," *Nature*, vol. 543, no. 7643, pp. 60–64, Mar. 2017. [Online]. Available: https://doi.org/10.1038/nature21377

[2] F. U. Battistuzzi and S. B. Hedges, "A Major Clade of Prokaryotes with Ancient Adaptations to Life on Land," *Molecular Biology and Evolution*, vol. 26, no. 2, pp. 335–343, Feb. 2009. [Online]. Available: https://doi.org/10.1093/molbev/msn247

[3] M. J. Benton, "Origins of Biodiversity," *PLOS Biology*, vol. 14, no. 11, pp. 1–7, Nov. 2016, publisher: Public Library of Science. [Online]. Available: https://doi.org/10.1371/journal.pbio.2000724

[4] M. J. Costello, A. Cheung, and N. De Hauwere, "Surface Area and the Seabed Area, Volume, Depth, Slope, and Topographic Variation for the World's Seas, Oceans, and Countries," *Environmental Science & Technology*, vol. 44, no. 23, pp. 8821–8828, Dec. 2010. [Online]. Available: https://pubs.acs.org/doi/10.1021/es1012752

[5] R. Danovaro, M. Canals, C. Gambi, S. Heussner, N. Lampadariou, and A. Vanreusel, "Exploring Benthic Biodiversity Patterns and Hot Spots on European Margin Slopes," *Oceanography*, vol. 22, no. 1, pp. 16–25, Mar. 2009. [Online]. Available: https://tos.org/oceanography/article/exploring-benthic-biodiversity-patterns-and-hotspots-on-european-margin-slo

[6] R. Danovaro, C. Corinaldesi, A. Dell'Anno, and P. V. Snelgrove, "The deep-sea under global change," *Current Biology*, vol. 27, no. 11, pp. R461–R465, Jun. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982217302178

[7] D. R. Yoerger, A. M. Bradley, B. B. Walden, H. Singh, and R. Bachmayer, "Surveying a subsea lava flow using the Autonomous Benthic Explorer (ABE)," *International Journal of Systems Science*, vol. 29, no. 10, pp. 1031–1044, Oct. 1998. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00207729808929596

[8] D. Yoerger, A. Bradley, M. Jakuba, C. German, T. Shank, and M. Tivey, "Autonomous and Remotely Operated Vehicle Technology for Hydrothermal Vent Discovery, Exploration, and Sampling," *Oceanography*, vol. 20, no. 1, pp. 152–161, Mar. 2007. [Online]. Available: https://tos.org/oceanography/article/autonomous-and-remotely-operated-vehicle-technology-for-hydrothermal-vent-d

[9] R. Henthorn, D. Caress, H. Thomas, R. McEwen, W. Kirkwood, C. Paull, and R. Keaten, "High-Resolution Multibeam and Subbottom Surveys of Submarine Canyons, Deep-Sea Fan Channels, and Gas Seeps Using the MBARI Mapping AUV," in *OCEANS 2006*. Boston, MA, USA: IEEE, Sep. 2006, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/4098900/

[10] C. R. German, D. R. Yoerger, M. Jakuba, T. M. Shank, C. H. Langmuir, and K.-i. Nakamura, "Hydrothermal exploration with the Autonomous Benthic Explorer," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 55, no. 2, pp. 203–219, Feb. 2008. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0967063707002580

[11] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," Jan. 2023, arXiv:1905.05055 [cs]. [Online]. Available: http://arxiv.org/abs/1905.05055

[12] I. C. Society, Ed., *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001: 8 - 14 December 2001, Kauai, Hawaii, USA*. Los Alamitos, Calif.: IEEE Computer Society, 2001, meeting Name: Conference on Computer Vision and Pattern Recognition. [Online]. Available: https://www.cs.cmu.edu/ efros/courses/LBMV07/Papers/viola-cvpr-01.pdf

[13] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. San Diego, CA, USA: IEEE, 2005, pp. 886–893. [Online]. Available: http://ieeexplore.ieee.org/document/1467360/

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010. [Online]. Available: http://ieeexplore.ieee.org/document/5255236/

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper$_files/paper/2012/file/c399862d3b9d6b76c84$ $Paper.pdf$

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Oct. 2014, arXiv:1311.2524 [cs]. [Online]. Available: http://arxiv.org/abs/1311.2524

[17] R. Girshick, "Fast R-CNN," Sep. 2015, arXiv:1504.08083 [cs]. [Online]. Available: http://arxiv.org/abs/1504.08083

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Jan. 2016, arXiv:1506.01497 [cs]. [Online]. Available: http://arxiv.org/abs/1506.01497

[19] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective Search for Object Recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, Sep. 2013. [Online]. Available: http://link.springer.com/10.1007/s11263-013-0620-5

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," May 2016, arXiv:1506.02640 [cs]. [Online]. Available: http://arxiv.org/abs/1506.02640

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, *SSD: Single Shot MultiBox Detector*. Springer International Publishing, 2016, p. 21–37. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46448-0_2

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," 2020, publisher: arXiv Version Number: 3. [Online]. Available: https://arxiv.org/pdf/2005.12872.pdf

[23] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0031320304001074

[24] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, pp. 333–359, Dec. 2011. [Online]. Available: http://link.springer.com/10.1007/s10994-011-5256-5

[25] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: An ensemble method for multilabel classification," in *Machine Learning: ECML 2007*, J. N. Kok, J. Koronacki, R. L. d. Mantaras, S. Matwin, D. Mladenič, and A. Skowron, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 406–417.

[26] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," 2016.

[27] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," *CoRR*, vol. abs/1702.05891, 2017. [Online]. Available: http://arxiv.org/abs/1702.05891

[28] R. You, Z. Guo, L. Cui, X. Long, Y. Bao, and S. Wen, "Cross-modality attention with semantic graph embedding for multi-label classification," *CoRR*, vol. abs/1912.07872, 2019. [Online]. Available: http://arxiv.org/abs/1912.07872

[29] D. Yane, L. Daoliang, L. Zhenbo, and F. Zetian, "Review on visual attributes measurement research of aquatic animals based on computer vision," *Transactions of the Chinese Society of Agricultural Engineering (Transactions*

of the CSAE), vol. 31, no. 15, pp. 1–11, 2015. [Online]. Available: http://www.tcsae.org/en/article/doi/10.11975/j.issn.1002-6819.2015.15.001

[30] Y. Wu, Y. Cai, and R. Tang, "Research on the underwater optical imaging processing and identification," *Ship Electron Eng*, vol. 39, no. 5, pp. 93–96, 2019.

[31] X. Peng, Z. Liang, J. Zhang, and R. Chen, "Review of underwater image preprocessing based on deep learning," *Computer Engineering and Applications*, vol. 57, no. 13, pp. 43–54, 2021.

[32] M. Jian, N. Yang, C. Tao, H. Zhi, and H. Luo, "Underwater object detection and datasets: a survey," *Intelligent Marine Technology and Systems*, vol. 2, no. 1, p. 9, Mar. 2024. [Online]. Available: https://link.springer.com/10.1007/s44295-024-00023-6

[33] "Han KM, Choi HT (2011) Shape context based object recognition and tracking in structured underwater environment. In: 2011 IEEE International Geoscience and Remote Sensing Symposium, Vancouver, pp 617–620. https://doi.org/10.1109/IGARSS.2011.6049204."

[34] "Beijbom O, Edmunds PJ, Kline DI, Mitchell BG, Kriegman D (2012) Automated annotation of coral reef survey images. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, pp 1170–1177. https://doi.org/10.1109/CVPR.2012.6247798."

[35] "Nagaraja S, Prabhakar CJ, Kumar PUP (2015) Extraction of texture based features of underwater images using RLBP descriptor. In: Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications, Bhubaneswar, pp 263–272. https://doi.org/10.1007/978-3-319-12012-6_29."

[36] M. Fatan, M. R. Daliri, and A. M. Shahri, "Underwater cable detection in the images using edge classification based on texture information," *Measurement*, vol. 91, 2016. [Online]. Available: https://doi.org/10.1016/j.measurement.2016.05.030

[37] K. Srividhya and M. M. Ramya, "Accurate object recognition in the underwater images using learning algorithms and texture features," *Multimed Tools Appl*, vol. 76, 2017. [Online]. Available: https://doi.org/10.1007/s11042-017-4459-6

[38] "Shi XT, Huang H, Wang B, Pang S, Qin HD (2019) Underwater cage boundary detection based on GLCM features by using SVM classifier. In: 2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, Hong Kong, pp 1169–1174. https://doi.org/10.1109/AIM.2019.8868517."

[39] "Gordan M, Dancea O, Stoian I, Georgakis A, Tsatos O (2006) A new SVM-based architecture for object recognition in color underwater images with classification refinement by shape descriptors. In: 2006 IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca, pp 327–332. https://doi.org/10.1109/AQTR.2006.254654."

[40] "Chen X, Chen HJ (2010) A novel color edge detection algorithm in RGB color space. In: IEEE 10th International Conference on Signal Processing Proceedings, Beijing, pp 793–796. https://doi.org/10.1109/ICOSP.2010.5655926."

[41] "Singh P, Deepak BBVL, Sethi T, Murthy MDP (2015) Real-time object detection and tracking using color feature and motion. In: 2015 International Conference on Communications and Signal Processing, Melmaruvathur, pp 1236–1241. https://doi.org/10.1109/ICCSP.2015.7322705."

[42] H. Komari Alaie and H. Farsi, "Passive sonar target detection using statistical classifier and adaptive threshold," *Appl Sci*, vol. 8, 2018. [Online]. Available: https://doi.org/10.3390/app8010061

[43] "Susanto T, Mardiyanto R, Purwanto D (2018) Development of underwater object detection method base on color feature. In: 2018 International Conference on Computer Engineering, Network and Intelligent Multimedia, Surabaya, pp 254–259. https://doi.org/10.1109/CENIM.2018.8711290."

[44] H. B. Wang, Q. Zhang, X. Wang, and Z. Chen, "Object detection based on regional saliency and underwater optical priors," *Chin J Sci Instrum*, vol. 35, 2014. [Online]. Available: https://doi.org/10.19650/j.cnki.cjsi.2014.02.021

[45] "Zhu YF, Chang L, Dai JL, Zheng HY, Zheng B (2016) Automatic object detection and segmentation from underwater images via saliency-based region merging. In: OCEANS 2016-Shanghai, Shanghai, pp 1–4. https://doi.org/10.1109/OCEANSAP.2016.7485598."

[46] M. W. Jian, W. Y. Zhang, H. Yu, C. R. Cui, X. S. Nie, and H. X. Zhang, "Saliency detection based on directional patches extraction and principal local color contrast," *J vis Commun Image Represent*, vol. 57, 2018. [Online]. Available: https://doi.org/10.1016/j.jvcir.2018.10.008

[47] E. Orenstein, K. Barnard, L. Lundsten, G. Patterson, B. Woodward, and K. Katija, "The fathomnet2023 competition dataset," 2023.

[48] C. Boittiaux, "Visual localization for deep-sea long-term monitoring," Theses, Université de Toulon, Dec. 2023. [Online]. Available: https://theses.hal.science/tel-04482249

[49] Y. Schechner and N. Karpel, "Recovery of underwater visibility and structure by polarization analysis," *IEEE Journal of Oceanic Engineering*, vol. 30, no. 3, pp. 570–587, 2005.

[50] J. Y. Chiang and Ying-Ching Chen, "Underwater Image Enhancement by Wavelength Compensation and Dehazing," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1756–1769, Apr. 2012. [Online]. Available: http://ieeexplore.ieee.org/document/6104148/

[51] T. T. Dana Menaker and S. Avidan, "Color restoration of underwater images," in *Proceedings of the British Machine Vision Conference (BMVC)*, G. B. Tae-Kyun Kim, Stefanos Zafeiriou and K. Mikolajczyk, Eds. BMVA Press, September 2017, pp. 44.1–44.12. [Online]. Available: https://dx.doi.org/10.5244/C.31.44

[52] D. Berman, D. Levy, S. Avidan, and T. Treibitz, "Underwater Single Image Color Restoration Using Haze-Lines and a New Quantitative Dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9020130/

[53] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 04 2019.

[54] S. Liu, L. Zhang, X. Yang, H. Su, and J. Zhu, "Query2label: A simple transformer way to multi-label classification," 2021.

[55] E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," *arXiv preprint arXiv:2009.14119*, 2020.