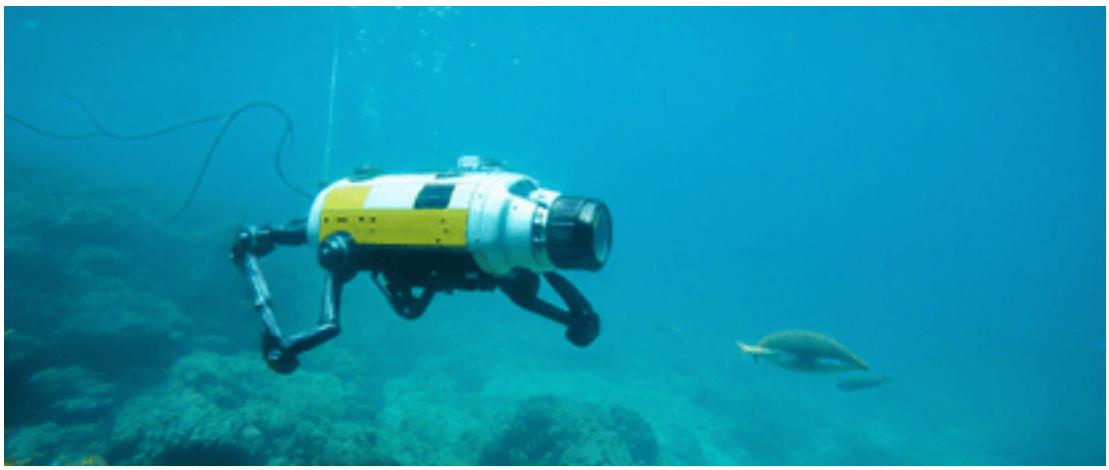
TÉCNICO LISBOA



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

Md Sazidur Rahman

Thesis to obtain the Master of Science Degree in

Electrical & Computer Engineering

Supervisors: Prof. Dr. Ricard Marxer
Dr. David Alexandre Cabecinhas

Examination Committee

Chairperson: Prof. João Manuel de Freitas Xavier
Supervisor: Prof. Dr. Ricard Marxer
Members of the Committee: Prof. Carlos Jorge Andrade Mariz Santiago
Dr. David Alexandre Cabecinhas

October 2024

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

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.

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.

Keywords

Deep learning, Underwater object detection, Multi-label classification, Query2Label, YOLOv9, Fathomnet dataset, DepthJitter augmentation, Marine research, Species identification, Habitat monitoring, Data imbalance, Environmental variability, Model robustness, Out-of-distribution detection, Marine conservation

Resumo

A tese explora as técnicas avançadas de aprendizagem profunda para elaborar a deteção de objectos subaquáticos, a classificação de vários rótulos e a deteção de fora da distribuição e centra-se especialmente no conjunto de dados da competição Fathomnet, um conjunto de dados abrangente e de fonte aberta de imagens subaquáticas. Propomos uma nova abordagem de aumento de dados, denominada Depth Jitter; um método dedicado a corrigir as distorções de cor introduzidas por caraterísticas relacionadas com a profundidade. O nosso método produziu um ganho de desempenho em termos de uma pontuação de precisão média mínima esperada de 2-3% mAP@20. Apesar de desafios como o desequilíbrio de dados das imagens subaquáticas e a variabilidade ambiental, a nossa técnica de aumento de dados permitiu aos modelos como o Query2Label e o YOLOv9 demonstrar robustez e adaptabilidade, e são capazes de obter resultados sem recorrer a conjuntos de dados externos. Este estudo utiliza estes modelos como ponto de partida para explorar a sua utilidade no avanço de aplicações-chave em estudos marinhos - codificação de espécies e observação do habitat - para efeitos de conservação marinha. A investigação futura centrar-se-á no aumento da diversidade do conjunto de dados, na melhoria das técnicas para lidar com o desequilíbrio dos dados, na melhoria da interpretabilidade do modelo e na exploração de opções para a implementação em tempo real. A integração de modelos híbridos e o melhoramento da deteção de elementos fora da distribuição ajudarão a aumentar a fiabilidade e a aplicabilidade da análise de imagens subaquáticas.

Palavras Chave

Aprendizado profundo, Detecção de objetos subaquáticos, Classificação multilabel, Query2Label, YOLOv9, Conjunto de dados Fathomnet, Aumento DepthJitter, Pesquisa marinha, Identificação de espécies, Monitoramento de habitats, Desequilíbrio de dados, Variabilidade ambiental, Robustez do modelo.

Contents

List of Figures

List of Tables

Acronyms

YOLO	You Only Look Once
DETR	Detection Transformer
OOD	Out-of-Distribution
HOG	Histogram Oriented Gradients
DPM	Deformable Part Model
CNN	Convolutional Neural Network
RCNN	Regions with CNN Features
SPPNet	Spatial Pyramid Pooling Networks
RPN	Region Proposal Network
FPN	Feature Pyramid Network
SSD	Single Shot Multibox Detector
BR	Binary Relevance
CC	Classifier Chains
LP	Label Power-set
kNN	k-Nearest Neighbors
RAkEL	Random k-labeLsets
ASGE	Adjacency-based Similarity Graph Embedding
CMT	Cross-Modality Transformer
TDRG	Transformer-based Dual Relation Graph

MBARI Monterey Bay Aquarium Research Institute

ROV Remotely Operated Vehicle

AUV Autonomous Underwater Vehicle

mAP Mean Average Precision

Q2l Query 2 label

1

Introduction

Contents

1.1 Background

Academic studies confirm that marine life substantially predates terrestrial life. According to Dodd et al. [19], 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 Battistuzzi et al. [20]. The fossil record, as detailed by Benton [21], 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 [22].

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 [22]. 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 Danovaro et al. [23, 24]. This realization highlights the fundamental importance of marine biodiversity in the global ecological dynamics, and it warrants additional and broad exploration and investigation.

The ocean's diversity of life provides enormous opportunities for exploration and discovery. Marine organisms often contain unusual biochemical properties that can serve as basis of new explorations and advancements in science and medicine. The urgent need to conserve these systems is aggravated by changes due to activities such as deep-sea mining and climate change, increasing the need to understand deep-sea ecosystems.

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 [25–28].

These above mentioned technological developments are further enriched and supplemented by a number of computer vision techniques that plays a significant role in monitoring and understanding the marine systems. **Image and video classification** techniques which are based on Convolutional Neural Network(CNN) are capable of automatically determining and labelling marine species or objects from images or videos collected under the water. For example, an Autonomous Underwater Vehicle(AUV) may take a video of coral reef that includes a variety of fish, corals and/or marine debris. With the CNN based image/video classification models, researchers can classify marine organisms and objects while at sea, allowing them to monitor local biodiversity, track health of species and identify the presence of invasive

species. This technique reduce the manual labour needed to analyze the substantial amount of image/video data collected from underwater environment making monitoring the marine biodiversity more efficient. **Object detection** algorithms such as YOLO(You only look once) and DETR(Detection transformer) advance this field beyond object identification to include object localization in images/videos. For instance, monitoring of habitats on the ocean floor, these algorithms can detect and visualize individual fish, coral or even pollutants such as plastic waste. In marine applications this is crucial to track the population dynamics of the species, behavioral aspects such as migration and human impact on the ocean with regard to detecting areas of debris. Another important technique is **segmentation**. Segmentation models like Segment Anything can do more than just detect and localize objects. Thus, they can also extract object from background area of image or video. In the case of a test video produced as an example to simulate fish in schools, segmentation identifies and outlines each individual fish within the school even if they overlap making it possible for scientists to count the fish, investigate how they interact with one another and differentiate healthy coral from bleached. In marine monitoring for computer vision, an **out-of-distribution (OOD)** detection framework can be a valuable approach within a marine monitoring workflow by identifying observations that are different from trained data such as unknown marine species, or environmental phenomena that may not have been included in the training dataset. For an illustrative example, in monitoring fish species using image classification, OOD will flag a fish species that has a completely different sign for no recognition. This could be indicative of a rare, or unknown, species. Equally in object detection or segmentation tasks, OOD can also flag observations which do not fit a recognizable category, such as new types of debris or marine life and will prevent the model from falsely recognizing descriptions of a known category. This is especially critical in complex and dynamic engineering environments such as the ocean where unknown species, human-made objects, and recognizing when something is outside the model's parameters, it is an important aspect of the scientific discovery and accurate marine monitoring process. In this research work we will mostly focus on image classification, object detection and out-of-distribution detection as our goal is to identify the category of marine species and to identify if it is from the training set or not.

A major limitation with contemporary object detection methods is their often exclusive dependence on existing datasets for training. This is particularly true in the area of deep-sea science. The deep sea is expansive and essentially uncharted when it comes to biodiversity; thus there is always the potential for occurrences such as depressed morphology or taxonomy that were not in the training datasets. In order to gain a thorough understanding of variations in deep-sea biodiversity, we need to create capabilities to detect and classify the unknown. The ability to detect the unknown represents an exciting new frontier for marine science to open doors for discovery and further, to understand ecological interactions.

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.

1.2 Motivation for 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 need to protect the environment. It is important to understand how these environments work in order to protect and manage marine ecosystems in a sustainable way, 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 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.

1.3 Objective of this Research

The primary objective of this thesis is to develop and evaluate deep learning models tailored for the analysis of marine visual data across varying ocean depths. This initiative aims to tackle the significant challenges posed by disparities in data acquisition conditions in the ocean, especially differences between upper ocean regions, where data is more abundant, and deeper waters, which are less explored

and harder to access. The focus will be on enhancing the model's capabilities in identifying marine species and detecting shifts in species distributions, enabling robust performance not only in familiar but also in novel or less-documented environments.

The thesis will specifically aim to design and train machine learning models that leverage the FathomNet annotated image set. Emphasis will be placed on advanced image processing techniques, including transfer learning, to enable these models to adapt to varying conditions such as changes in lighting, camera specifications, and environmental factors. The robustness of these models will be crucial, as they must perform consistently across a range of oceanic conditions and be capable of managing the transition from data-rich upper ocean scenarios to data-sparse conditions found in deeper waters.

Another key aim is the development of methodologies for out-of-distribution detection. This involves the ability to flag data that represents new or unusual findings, such as unencountered species or significant environmental changes. Such capabilities are vital for expanding our understanding of marine biodiversity and for the ongoing efforts in cataloging marine life, particularly in under-explored regions.

Furthermore, this research intends to test and validate the applicability of the developed models for real-world deployment in marine ecological monitoring and automated data analysis systems. It will provide actionable insights for marine ecologists and recommendations for improving data collection and analysis strategies. Ultimately, by enhancing the adaptability and scalability of machine learning applications, this thesis aims to contribute significantly to marine biology. It seeks to improve our understanding of marine biodiversity, particularly in challenging deep-water ecosystems, and to support the management and conservation of these vital environments.

1.4 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.

1.5 Outline

This thesis is organized as follows:

Chapter 2: State of the Art: This chapter outlines the evolution of object detection models from tradi-

tional methods such as the Viola-Jones and HOG detectors, to models based on CNN, such as RCNN, Fast RCNN, and RTR-CNN, which focus on improving performance using Region Proposal Networks. Further, it has examined the recent developments of innovation with one-stage detectors, for example, YOLO and SSD, and transformer-based models, such as DETR, which have established new records for speed and accuracy. Further, it has discussed work in multi-label classification, from early studies by Boutell et al. to modern deep learning models such as CNN-RNN, Spatial Regularization Networks, and Transformer-based models that explain complex label relationships. Lastly it discussed issues with underwater environments, for example, occlusion of objects, high false positive rate, bias of datasets, and computational limitations.

Chapter 3: Fathomnet Competition Dataset: This chapter provides an overview of the Fathomnet Competition Dataset, which was obtained from MBARI. This section covers how the dataset was prepared, properties of the Fathomnet 2023 dataset and the characteristics of the dataset.

Chapter 4: Methodology: This chapter presents a discussion of the pre-process and application of a known Fathomnet dataset for multi-label classification, through the user of a novel data augmentation method derived from generating underwater images. The application of the Query2Label (Q2L) model for multilabel classification and YOLOv9 for object detection has been presented. Afterwards, methods for OOD detection have been discussed.

Chapter 5: Results & Discussions: This chapter presents a comprehensive evaluation of the proposed deep learning models, Query2Label and YOLOv9, for underwater object detection and multi-label classification using the Fathomnet dataset. Detailed performance metrics, including mAP@20 scores, are provided to illustrate the effectiveness of the proposed augmentation technique. The chapter also includes qualitative evaluations, such as attention map visualizations, which highlight the models' strengths and areas for improvement. The discussion covers the impact of data imbalance, environmental variability, and the robustness of the models in real-world scenarios. Furthermore, the out-of-distribution detection capabilities of the models are examined, showcasing their reliability in handling diverse and unseen data.

Chapter 6: Conclusion: This chapter summarizes the key findings of the research, emphasizing the success of the proposed methods in enhancing underwater image analysis. It reiterates the improvements achieved through the DepthJitter augmentation and the competitive performance of the models without relying on external data sources. The conclusion also identifies several areas for future work, including enhancing dataset diversity, addressing data imbalance, improving model interpretability, exploring real-time deployment, and ensuring robustness to environmental changes. These directions aim to build upon the current findings and further advance the field of underwater image analysis for marine exploration and conservation.

2

State of the Art

Contents

In this chapter, we will summarize the state-of-the-art(SOTA) methods regarding two in-depth tasks: **object detection** and **multilabel classification**, which we will use for out-of-distribution(OOD) detection task later on. Object detection requires identifying and localizing objects in the image, where the inputs are an image, and the outputs are bounding boxes and class labels of the detected objects. The task of multi-label classification requires predicting several applicable labels for an image and aims to predict all valid labels for a given image. Performance of both tasks is typically measured using the mean Average Precision (mAP) metric that calculates how accurate a prediction is across class and threshold. We will then briefly discuss about the benchmark datasets like COCO, Pascal VOC on airborne images and also point out the limitations of these traditional methods in underwater environment.

2.1 Evolution of Object Detection Models

Object detection and localization play a vital role in computer vision, being a critical task. The progress in object detection can be divided into two distinct periods: the era of ”traditional object detection” prior to 2014, and the subsequent era of ”deep learning-based detection” [1].

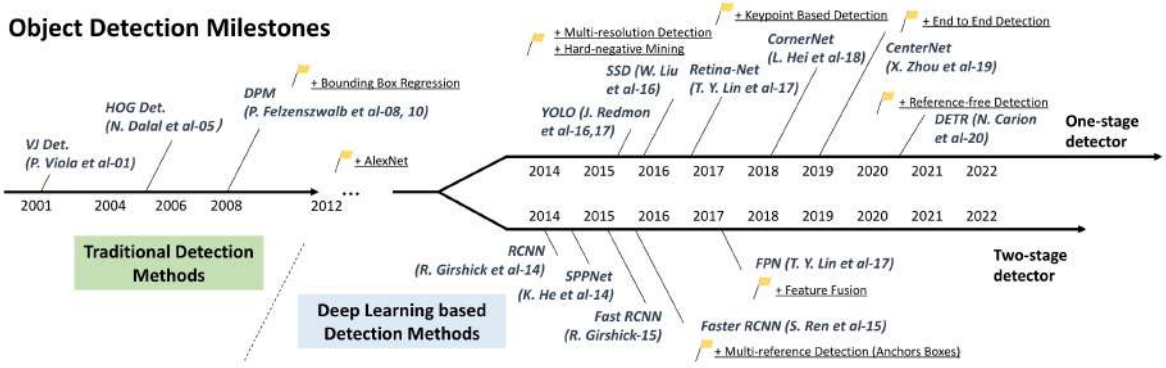


Figure 2.1: The evolution of object detection in the past twenty years.Source: [1]

Figure 2.1 demonstrates a visual summary of the most significant object detectors over the past twenty years, which includes pre-deep learning and post-deep learning development of object detection. The points below summarize the significant advancements in the field of object detection:

2.1.1 Traditional Detectors

Traditional object detection algorithms have been largely depended on hand-crafted features, rather than directly learning from data.Specifically, these methods utilize pre-defined techniques for feature extraction to learn patterns in the images that can be used later for detecting objects. Though they are

foundational algorithms in the field, traditional object detectors are limited in their ability to develop hand-crafted features, thus limiting their ability to generalize complex and diverse data. Rather than learning new features on their own from large datasets, they are less adaptive in various lighting, viewpoints and occluded image data which are very important in marine environment. Below, some of the traditional object detectors discussed -

Viola Jones detector: In the early 2000s, P. Viola and M. Jones were able to achieve real-time human face detection without any limitations [29]. The authors introduced the notion of the **Integral Image**, a concept that allows fast calculations of features used in object detection. The integral image is a data structure that allows computing the sum of pixel values in any rectangular region of the image in constant time. The image is first preprocessed with a limited amount of operations per pixel. This allows computationally fast evaluations of Haar-like features at any location and any scale within the image. The authors also incorporated a learning algorithm based on AdaBoost, which allows for the selection of only a few critical visual features, as opposed to the much larger set of features. The feature selection process is considered to be very important in the development of efficient classifiers. This detection mechanism utilizes **Haar-like features**, which are merely basic rectangular patterns that use special characteristics of images effectively. When using the integral image representation with Haar-like features, the rapid computation and efficiency of the detection system, particularly for face detection, can be facilitated.

HOG Detector: The introduction of the Histograms of Oriented Gradients (HOG)-based approach for human detection by N. Dalal and B. Triggs resulted in significant improvements compared to existing feature sets [30]. The HOG descriptor is calculated based on a dense grid of uniformly spaced cells which capture the particular gradient orientations and thus provide a better representation of the human shape. A linear Support Vector Machine (SVM) classifies each image, balancing efficiency and speed, such that the human detection system can process images very quickly with just a slight cost in accuracy. Additionally, local contrast normalization helps the descriptor with its robustness towards changes in illumination and background clutter, thereby enhancing the reliability of the feature. The method incorporates fine scale gradients and efficiently uses up to 9 orientation bins to further improve the accuracy of detection. Descriptor blocks are used to overlap for segments of the image in order to reduce the miss rate by about 5%. The method also introduced a new dataset that contains more than 1800 annotated images to support analysis, contributing to robustness across a wide variety of poses and backgrounds. Furthermore, the combination of features and techniques yields a very good human detection system, achieving nearly perfect results on the MIT pedestrian database and creating a benchmark in the field.

Deformable Part Model(DPM): Deformable Part Models (DPMs) [31] function as object detection frameworks that conceptualize objects as sets of deformable parts with filters that model local appearance features. The ability of the model to account for variability due to changes in viewpoint, human pose, or occlusion is achieved by modeling the spatial relationships and configurations of parts with the use of

springs.

DPMs can be understood in the context of pictorial structures, representing a star-structured model where the root filter captures the entire object and the part filters capture the local information and object features. The model is trained using a discriminative strategy whereby the decision boundary between positive and negative examples is optimized to improve detection.

A major advance of DPMs is the use of latent variables that allow the model to learn unlabeled part centroid locations through a latent SVM formulation. The optimization proceeds in an iterative manner by alternating between fixing the values of the latent variables to improve the visible part locations or training latent location parameters to separately maximize the fit between the negative and positive examples. The use of latent variables provides a way for the model to accommodate both inter-class variability (e.g. bear versus dog) and non-rigid deformations (e.g. a dog curling up).

DPMs provide state-of-the-art performance on benchmark datasets such as PASCAL VOC by utilizing intelligent optimization and feature representations including Histogram of Oriented Gradients (HOG). By modeling complex categories using mixtures of deformable parts, DPMs established a new foundational technique in object localization and detection.

2.1.2 CNN based Two-stage Detectors

After reaching a peak in the performance of manually designed features, progress in object detection research stagnated after 2010. However, the introduction of convolutional neural networks in 2012 by A. Krizhevsky reinvigorated the field of object detection [32]. These deep convolutional neural networks have the ability to learn robust and high-level features from images. In 2014, R. Girshick et al. published a paper titled ”Regions with CNN Features” which further advanced the field of object detection [2]. Since then, object detectors have continued to rapidly evolve.

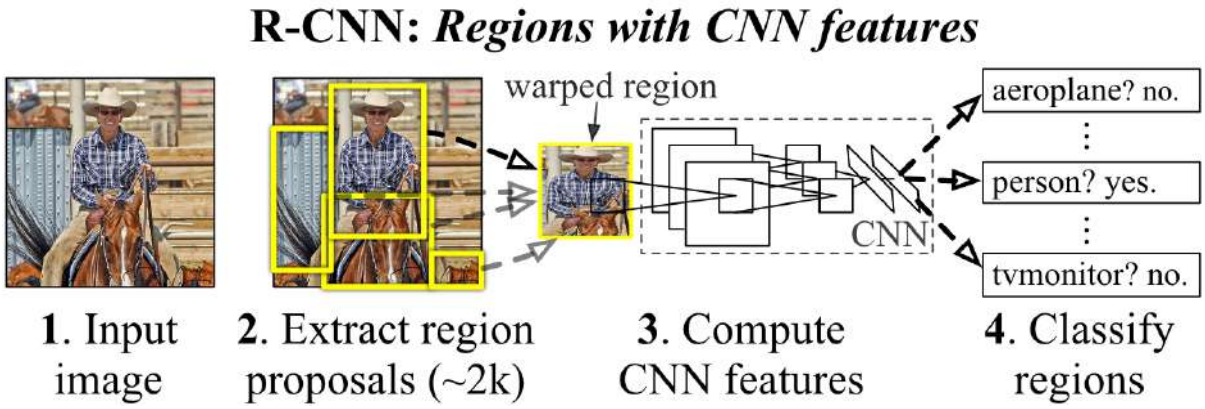


Figure 2.2: Architecture of RCNN [2]

RCNN: The main idea behind RCNN is quite simple, it starts by generating a set of object proposals,

which are candidate bounding boxes, using a method like selective search [33]. These proposed regions are then standardized to a uniform image dimension and passed through a pretrained CNN model, such as AlexNet [32], to extract relevant feature representations. These feature vectors are then used for two main purposes: to determine if an object is present in each proposal and to categorize the object using linear SVM classifiers.

This approach significantly improved the mean Average Precision (mAP) from 33.7% to 58.5% on the VOC07 dataset, outperforming previous methods like DPM-v5 [31]. However, a notable limitation of RCNN is its long detection times, mainly due to repetitive feature computations over a large number of overlapping proposals. This often leads to long processing time which is far from real-time. To overcome this limitation, a subsequent advancement called SPPNet [3] was introduced in the same year, significantly improving the computational efficiency of object detection.

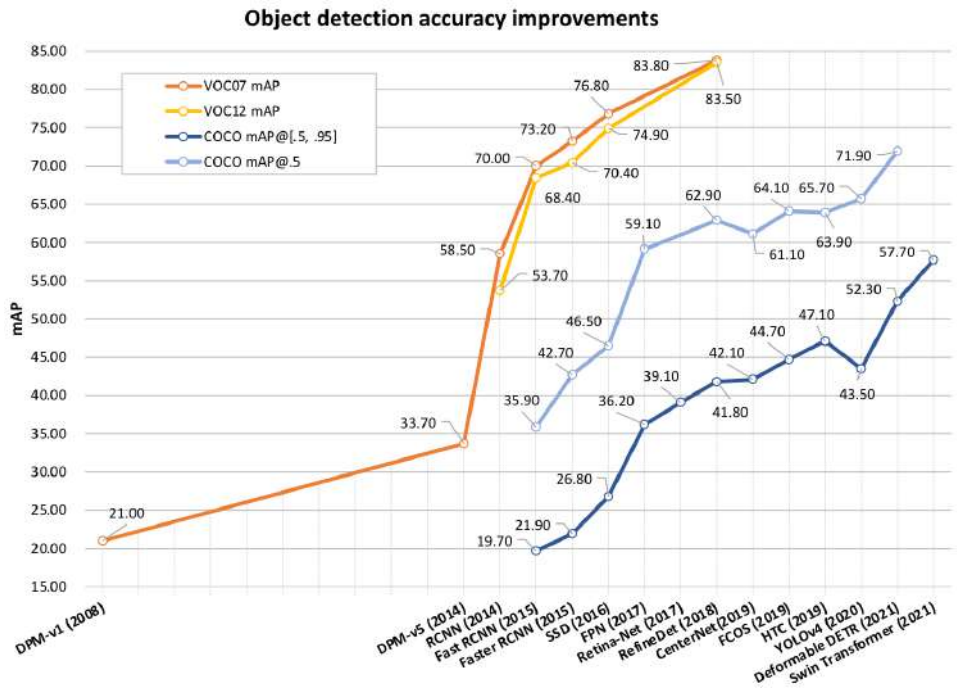


Figure 2.3: Accuracy improvement of the object detectors on VOC and MSCOCO datasets. [1]

SPPNet: In 2014, K. He et al. proposed the Spatial Pyramid Pooling Networks (SPPNet) [3] as a solution to the fixed-size input requirement of previous CNN models like AlexNet [32]. The key innovation of SPPNet is the inclusion of a Spatial Pyramid Pooling (SPP) layer, which enables a CNN to generate a fixed-length representation regardless of the size of the image or region of interest, without the need for rescaling. When SPPNet is used for object recognition, the feature maps are computed once for the entire image and fixed-length representations of random areas are created for training the detectors. This approach eliminates the need to compute the convolutional features again.

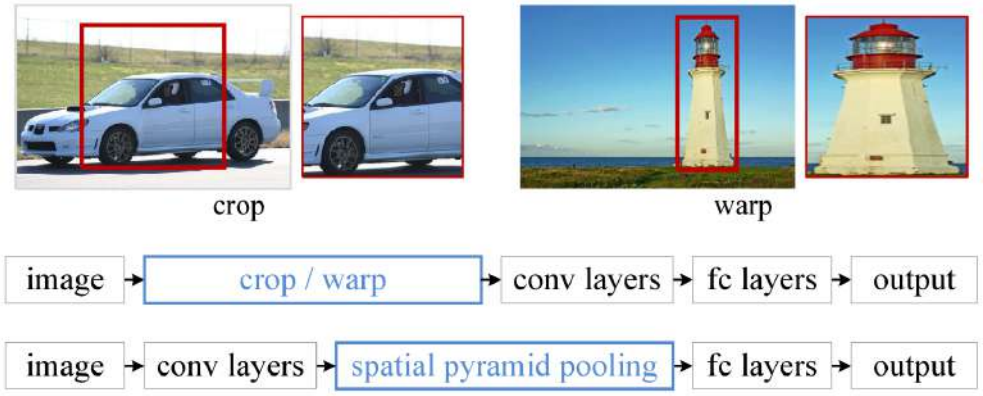


Figure 2.4: Architecture of SPPNet [3]

SPPNet enhances performance compared to R-CNN while preserving detection accuracy (VOC07 mAP = 59.2%). Nevertheless, there are limitations associated with SPPNet: it continues to be training in a multi-stage fashion and continues to only fine-tune the fully connected layers, while the earlier layers remain fixed. To address these concerns, Fast R-CNN [4] was introduced later in the same year.

Fast R-CNN: In 2015, Girshick introduced the Fast RCNN detector [4], which builds on the advances of R-CNN and SPPNet [2] [3]. The Fast RCNN detector enables simultaneous training of a detector and a bounding-box regressor within the same network configuration. By doing so, it significantly improved the mean average precision (mAP) from 58.5% (achieved by RCNN) to 70.0% on the VOC07 dataset, while also being nearly 200 times faster in terms of detection speed compared to R-CNN.

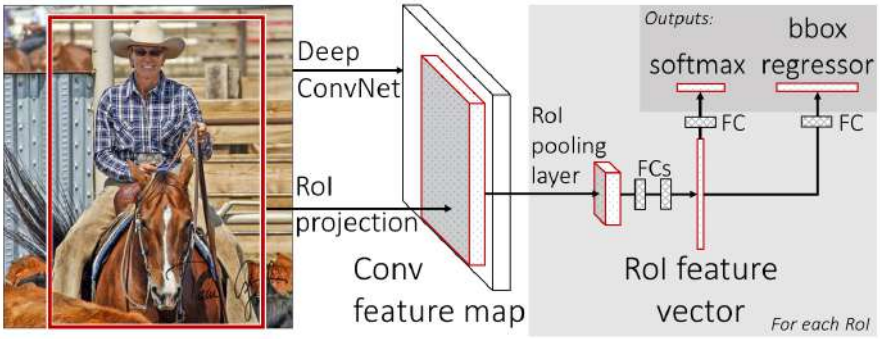


Figure 2.5: Fast RCNN Architecture [4]

However, despite effectively combining the advantages of R-CNN and SPPNet, the detection speed of Fast R-CNN is still constrained by proposal detection. Subsequently, Faster R-CNN [5] was published, offering further enhancements in performance.

Faster R-CNN: In 2015, Ren et al. introduced the Faster RCNN detector [5] as an improvement over the Fast R-CNN. The Faster R-CNN, which operates at near-real-time speeds, achieved a COCO mAP@.5 of 42.7% and a VOC07 mAP of 73.2% when using ZF-Net, with a frame rate of 17fps [34]. The

key contribution of the Faster R-CNN is the introduction of the Region Proposal Network (RPN), which allows for efficient region proposals. Over time, various components of object identification systems, including proposal detection, feature extraction, and bounding-box regression, have been integrated into a single learning framework, starting from R-CNN and progressing to Faster RCNN. Despite the performance improvements of Faster RCNN over Fast RCNN, there still remains some redundancy in the subsequent detection step. Subsequent advancements, such as RFCN [35] and light head RCNN [36], have been proposed to address these issues.

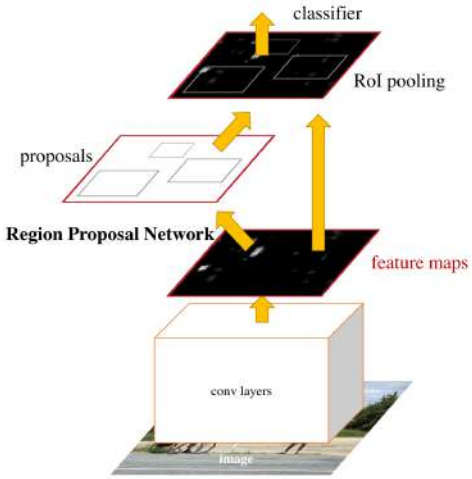


Figure 2.6: Architecture of Faster RCNN [5]

Feature Pyramid Network: In 2017, T.-Y. Lin et al. proposed the concept of FPN (Feature Pyramid Network) [37] as a solution to the limitations of deep learning-based object detectors that only relied on the top layer feature maps of the networks. While the deeper layers of a convolutional neural network (CNN) are valuable for category classification, they are not effective for object localization. In order to address this issue, FPN introduces a top-down architecture with lateral connections, allowing for the development of high-level semantics at all scales. By leveraging the inherent feature pyramid structure created by a CNN during forward propagation, FPN achieves significant advancements in object recognition for objects of different sizes. The application of FPN in a simple Faster R-CNN system yields state-of-the-art results for single model identification on the COCO dataset, with a COCO mAP@.5 of 59.1%.

2.1.3 CNN based One-Stage Detectors

Many object detectors follow a two-stage approach, where the first stage focuses on detecting probable objects to achieve high recall, and the second stage fine-tunes the object location and increases discrimination. While these two-stage detectors can achieve high accuracy without the need for addi-

tional features, their slow processing speed and high computational complexity make them impractical for real-world applications. On the other hand, one-stage detectors aim to detect all objects in a single step, making them popular for mobile devices due to their real-time processing capability and ease of deployment. However, one-stage detectors may struggle to effectively identify densely positioned and small objects.

YOLO: The YOLO (You Only Look Once) framework for object detection was first introduced by R. Joseph et al. [6] in 2015, marking the emergence of one-stage detectors in the deep learning era [6]. YOLO is well-known for its impressive speed, with a faster variant achieving a remarkable 155 frames per second (fps) and a mean Average Precision (mAP) of 52.7% on the VOC07 dataset. An enhanced version maintains a rapid pace at 45 fps while achieving a higher VOC07 mAP of 63.4%.

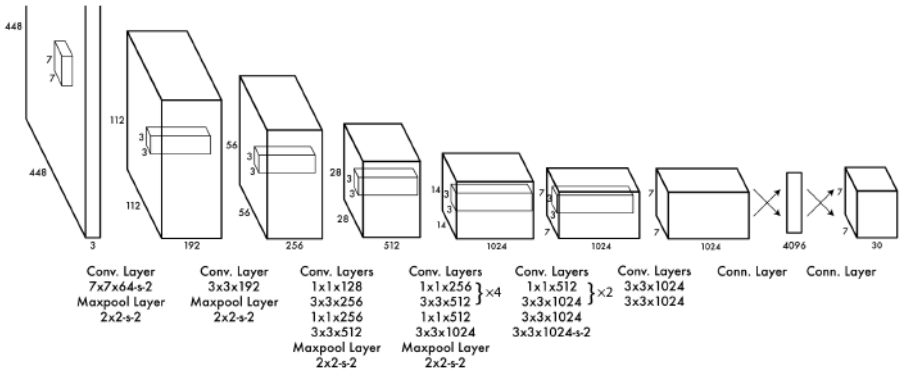


Figure 2.7: YOLO Architecture [6]

Unlike traditional two-stage detectors, YOLO utilizes a single neural network that processes the entire image. This network divides the image into regions and simultaneously predicts bounding boxes and object probabilities for each region. However, compared to two-stage detectors, YOLO sacrifices some localization accuracy, particularly for small objects. To address this issue, subsequent revisions of YOLO, such as YOLOv7 [38] and the addition of Single Shot MultiBox Detector(SSD) [39], have focused on improving localization accuracy. YOLOv7 utilizes two major advancements—dynamic label assignment, which assigns the best-suited labels to anchors during training, and model structure reparameterization, which optimizes the models through changes to physical parameters to provide a more efficient inference time. These developments allow YOLOv7 to outperform all other object detectors currently developed in terms of Speed (5 - 160 FPS) and Model accuracy. [38].

Single Shot Multibox Detector: SSD (Single Shot MultiBox Detector) was introduced by W. Liu et al. in 2015 [39], representing a significant breakthrough in the field of object detection. One of the key contributions of SSD is its utilization of multi-reference and multi-resolution detection techniques, which greatly enhance the accuracy of one-stage detectors, particularly when dealing with smaller objects. This unique approach offers SSD a dual advantage: it achieves impressive detection speed and

accuracy, achieving a COCO mAP@.5 score of 46.5%, with a faster variant capable of running at an impressive 59 frames per second (fps). Unlike previous detectors that mainly focused on detection within their top layers, SSD stands out by its ability to detect objects of different scales across multiple layers of the neural network.

RetinaNet: Despite their advantages in terms of speed and simplicity, one-stage detectors have historically lagged behind two-stage detectors in terms of accuracy. In 2017, T.-Y. Lin et al. sought to understand the reasons behind this performance gap and proposed a solution called RetinaNet [40]. Their investigation revealed that the significant foreground-background class imbalance encountered during the training of dense detectors was the primary cause. To address this issue, RetinaNet introduced a novel loss function called "focal loss." This loss function modifies the conventional cross-entropy loss, giving the detector a stronger focus on challenging and misclassified examples during training. The introduction of focal loss proved to be a game-changer, allowing one-stage detectors to achieve accuracy levels comparable to those of two-stage detectors, while still maintaining a high detection speed. Specifically, RetinaNet achieved a COCO mAP@.5 score of 59.1%.

DETR: In recent years, Transformers have had a significant impact on deep learning, particularly in the field of computer vision. Unlike traditional convolutional operators, Transformers rely solely on attention processes to enable the development of a global-scale receptive field.

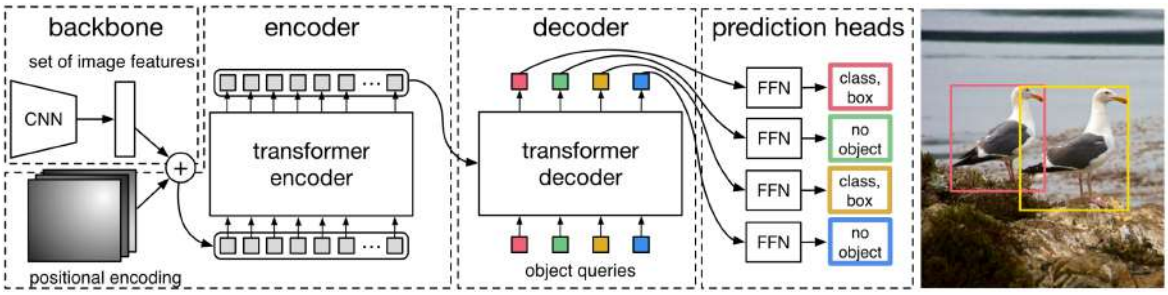


Figure 2.8: DETR architecture [7]

In 2020, N. Carion et al. proposed DETR [7], a groundbreaking approach that treats object detection as a set prediction problems. They put forward an end-to-end detection system, rooted in Transforms, that eliminates the use of anchor boxes or anchor points by directly predicting the location and class of objects. The object detection model, DETR, used bipartite matching to associate predictions to ground truth object during training and enabling the model to learn object detection without any anchor mechanism. This approach transformed object detection due to its engineering simplicity and ability to improve performance. Building on this, X. Zhu et al. introduced Deformable DETR [41] to address DETR's long convergence time and challenges in recognizing small objects. This innovative technique achieved state-of-the-art performance on the MSCOCO dataset, achieving a COCO mAP@.5 score of 71.9%.

2.2 Multi-label Classification

Multi-label image classification must assign more than one relevant label to a single image, in contrast to single-label classification where one label is predicated. This is useful in complicated situations such as the marine environment, where an image may have multiple species of marine habitats, lots of different underwater features, or objects like coral reef and marine debris that must be identified together. This task requires recognizing one or multiple objects, which can also overlap, within one frame, and that in itself is difficult. In this section, we will summarize the ways, challenges, and novel developments in multi-label image classification, while emphasizing underwater image exploration.

2.2.1 Early Research & Foundation

Acknowledging the earlier works for multi-label classification, Boutell et al. (2004) [42] proposed one of the initial frameworks for multi-label classification, taking into consideration situations in which instances could belong to multiple classes at the same time; unlike the traditional single-label classification. The work modified the process of redefining multi-class instances to making a classification assignment for each label. Although applied to scene classification at first, their framework can be applied to many areas; for example in medical diagnosis where a patient can have multiple conditions, or in a document classification where text can belong to multiple categories. Their contribution demonstrated how useful it is to construct models to account for label correlation and dependencies; which was stimulation for many proposed approaches into the multi-label classification space.

Models of Training

In multi-label classification, multiple models are used to deal with the difficulty of instances associated with multiple labels. The main difference among them is how they process multi-label data; to do this, some models will make the problem easier, while others will ignore the label information or utilize the label information to its full extent. In the following list, we summarize several common approaches to training models with multi-label data which was introduced by Boutell et al. [42]:

- **MODEL-s (Single-label):** Assigns the class with the highest dominance to multi-label data.
- **MODEL-i (Ignore):** Ignores multi-label data while training.
- **MODEL-n (New Class):** Creates new classes for multi-label data, though this is done at the expense of data sparsity.
- **MODEL-x (Cross-training):** The use of multi-label data multiple times augments data utilization and enhances accuracy.

Testing Criteria

To effectively evaluate and compare the performance of multi-label classification models, a set of testing criteria is especially useful in scenarios where we have a complex judgment task, such as underwater image exploration, where the criteria would help us judge how well our model predicts a number of labels for each specific instance, and how well it handles overlapping classes. In their article "Learning Multi-Label Scene Classification" published in 2004, Boutell et al. [42] introduced a set of testing criteria as determinants for model performance. The testing criteria that they introduce are as follows:

- **P-Criterion (Positive):** Labels all positive classes.
- **T-Criterion (Top):** Labels with the top scoring class even if scores are negative.
- **C-Criterion (Close):** Multi-class labels if the scores are close for top classes, constrained by a threshold.

Their experiments on 2400 images show that their cross-trained models (MODEL-x) outperform others. The delicate balance between recall and precision shown by the C-Criterion is excellent, and α is versatile for performance evaluation.
Significant results are presented, showing that the introduction of cross-training (MODEL-x) dramatically enhances the potential of multi-label classification. The C-Criterion effectively handles multi-label classifications, with α-evaluation providing a customizable approach to performance evaluation. Both the methods combined result in robust multi-label scene classification, offering one of the promising directions for serving as an example with large datasets and different classifiers.

2.2.1.A Binary Relevance

Multi-label classification is particularly helpful in contexts where one example connects to multiple relevant classes. For example, in underwater images, a single image may display multiple marine species requiring classification and identification. Based on the survey by Tsoumakas and Katakis [43], multi-label methods are broadly categorised into problem transformation methods and algorithm adaptation methods. Problem transformation methods consist of Binary Relevance (BR) and Label Power-set (LP), where multi-label problems are transformed into multiple issues of single-label classification. LP is a straightforward and easily scalable approach but quite often lacks capturing the label correlation. In contrast, LP models the label correlation by treating each unique combination of labels, aiming at increasing computational complexity with possibly low benefits. Advanced transformation methods, such as classifier chains (CC), model label sequences in sequence, thus enhancing predictive power efficiently in BR. Multi-label data handling has been an important direction of research, and many algorithms have been

developed or adapted to make them natively support multi-label data. It does offer better performance by inherently modeling label correlations. Starting from ML-kNN, which extends k-nearest neighbors by considering prior probabilities, AdaBoost extensions like AdaBoost-MH and AdaBoost-MR [44] apply the boosting technique for multi-label problems. Tsoumakas and Katakis [43] proposed label density and label cardinality metrics to quantify how multi-label a dataset is, allowing one to relate method performance across different contexts. They conclude that ”simple” BR methods are sufficiently efficient, while in more sophisticated methods CC, some adapted algorithms include label dependencies and, therefore will increase the importance of balanced methods.

2.2.2 Advances in Multi-label Classification

2.2.2.A Classifier Chains

Binary Relevance (BR) is a very basic technique in multi-label classification, treating each label as an independent binary classification problem. From its simplicity and scalability, although BR has been criticized for not modeling interdependence between labels, these facts can lead to sub-optimal predictive performance. To remedy this drawback, Read et al. (2011) [45] have proposed the Classifier Chains (CC) method by extending the BR to chain-linked binary classifiers to model label dependencies. Each classifier in the chain predicts the relevance of a label using features and predictions made by the preceding classifiers. Still computationally efficient, this approach dramatically improves predictive performance by capturing label correlations.
Authors take the CC method a step further by incorporating it within an ensemble framework, named ECC, in which it trains numerous CC models with random orders of labels, and their predictions are averaged out to offer more robust and accurate predictions at the same time avoiding problems like error propagation along the chain. Extensive empirical studies on several multi-label datasets have shown that CC and ECC have better predictive performance and good computational efficiency than traditional BR and other state-of-the-art methods. The study at this moment indicates the efficiency of BR in considerable improvements due to label correlation incorporation, which is done by means like CC and ECC, making them potent means for large-scale multi-label classification tasks.

2.2.2.B Ensemble Method (Random *k*-Labelsets)

Tsoumakas and Vlahavas (2007) [46] further enhanced the field with the Random k-labeLsets algorithm. To tackle this problem, an RAkEL develops a pool of base classifiers for predicting with the help of a random set of labels generated for each training instance. The difficulty it addresses, due to the label correlations, is solved by learning the single-label classifiers that will predict each element in a powerset of the subset—effectively balancing the gap between the dependencies learned between the labels and

the computational burden of the task.

They have demonstrated that their experiments were validated in quite a few domains, such as protein function classification, document categorization, and semantic scene analysis, in which RAkEL outperformed the traditional measures of BR and Label Powerset (LP). The results showed that both classification accuracy and challenges in multilabel classification are improved by RAkEL.

2.2.3 Deep Learning Approaches

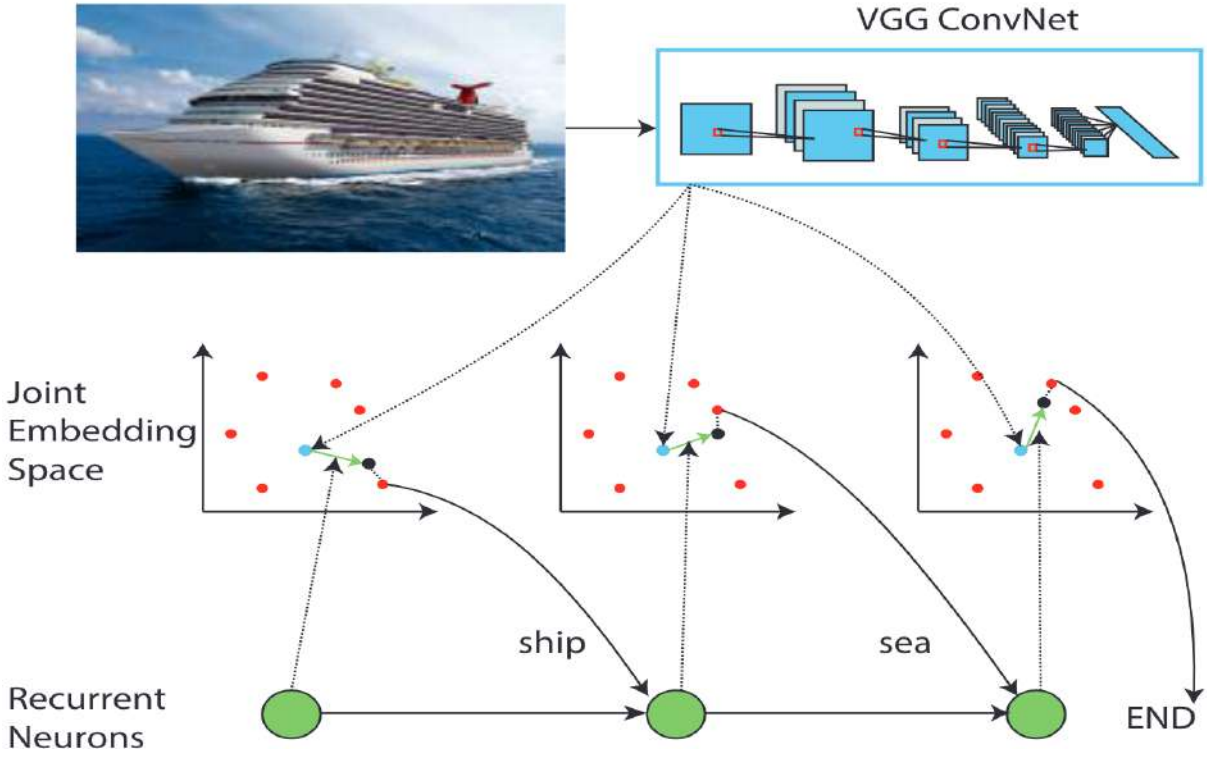


Figure 2.9: An example of the CNN-RNN multilabel classification system for images, where the label dependency and relationship between the picture and label are captured by the framework, which learns a joint embedding space. Here, red and blue points correspond to the label and image embeddings, while the black ones correspond to the sum of the image and recurrent neuron output embeddings. The label embeddings are concatenated in the joint embedding space concerning the co-occurrence dependencies of the labels. Taking the picture embedding and the output of the recurrent neurons, at every time step, an estimation of the likelihood of a label is made [8].

2.2.3.A CNN-RNN: A Unified Framework for Multi-label Image Classification

The unified framework proposed by Wang et al. (2016) [8] extends a combination of Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) for solving multi-label classification challenges. Traditional methods often assume independence among labels, and hence, essential label dependencies are ignored. A CNN-RNN framework incepted from a unified framework learns a joint

image-label embedding in capturing semantic label dependencies and the relevance of the image-label. The CNN extracts the semantic representations from the images, while the RNN models the relationships between labels using its sequential processing capabilities.

The CNN-RNN model [8] can jointly model image features and label embeddings in this shared embedding space. The ability of an RNN to incorporate contextual information of the predicted labels and condition future predictions comes from the use of recurrent neurons. Adaptively using this to focus within an image would be beneficial for the smaller objects, which are missed by using only global image features. End-to-end training is one of the significant advantages of this framework. It can train a model directly without forcing the engineer to manually design the model so that it integrates all image features and besides, all label dependencies. The unified approach also captures the redundancy in label semantics, which reduces computational time and promotes generalization by allowing the use of shared parameters for semantically similar labels. Experiments were done on benchmark datasets like NUS-WIDE, Microsoft COCO, and PASCAL VOC 2007 for the CNN-RNN framework. Among existing methods, it was observed that this method achieved the lead in effective models, especially in its complicated label dependencies and in varying object sizes within images. Deconvolutional networks were employed to visualize the model's capability to focus on different regions of an image in predicting various labels, which was human-like in multi-label classification.

2.2.3.B Spatial Regularization with Image-level Supervisions for Multi-label Image Classification

Feng Zhu et al. (2017) [9] introduced a Spatial Regularization Network for improved multilabel image classification, capturing semantic and spatial relationships between the labels using only image-level supervision. Classic methods fell into the aspect of very often failing in the spatial modeling dependencies since most of them did not have spatial annotations. The spatial regularization network generates attention maps for each label and applies learnable convolutions to capture the underlying relationships among the labels. The classification results of regularization, in turn, become consolidated with those obtained in a ResNet-101 network for enhanced performance.

Additionally, end-to-end training of the SRN framework eliminates extra efforts toward annotation, which is usually difficult. This is further evidenced since the SRN performs remarkably when tested on standard public datasets like NUS-WIDE, MS-COCO, and WIDER-Attribute, proving its strong generalization ability against state-of-the-art methods.

2.2.3.C Cross-Modality Attention with Semantic Graph Embedding for Multi-Label Classification

Renchun You et al. (2020) [10] developed a new framework that considered the shortcomings of previous multilabel classification approaches and integrated cross-modality attention with semantic graph

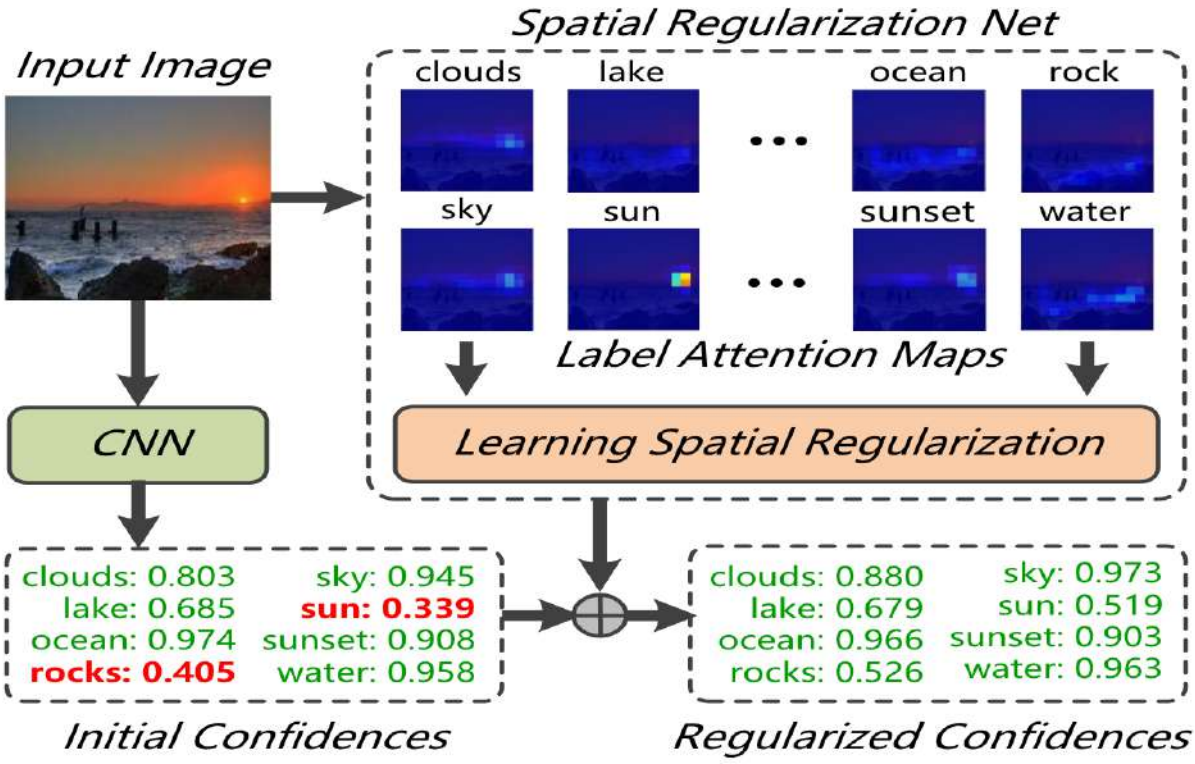


Figure 2.10: Illustration of Spatial Regularization Net(SRN) [9].

embedding. Previously, these methods essentially disregarded the explicit relations between semantic labels and regions of the image, making them underperforming. In light of these shortcomings, they introduced a novel method: Adjacency-based Similarity Graph Embedding (ASGE), a model for learning semantic label embeddings to capture rich label relations. These embeddings are used in guiding the generation of cross-modality attention maps, which is beneficial for improving the model's ability to locate discriminative features and capture spatial dependencies among the labels. The CNN-based backbone is used to extract the visual features, which are projected by the Cross-Modality Transformer (CMT) module into semantic space. The learned label embeddings develop category-specific attention maps for each respective label, thereby bringing out respective relevant regions in the image for each specific label and allowing the model to pay more attention to the relevant areas and less to the not-so-meaningful regions.

The effectiveness of this method has been well explored on benchmark datasets, including MS-COCO, NUS-WIDE, and YouTube-8M Segments. Experiment results showed that the cross-modality attention mechanism achieves state-of-the-art or better performance on the multilabel image classification task. Besides, the model established new performance benchmarks on image and video classification datasets with strong generalization capability.

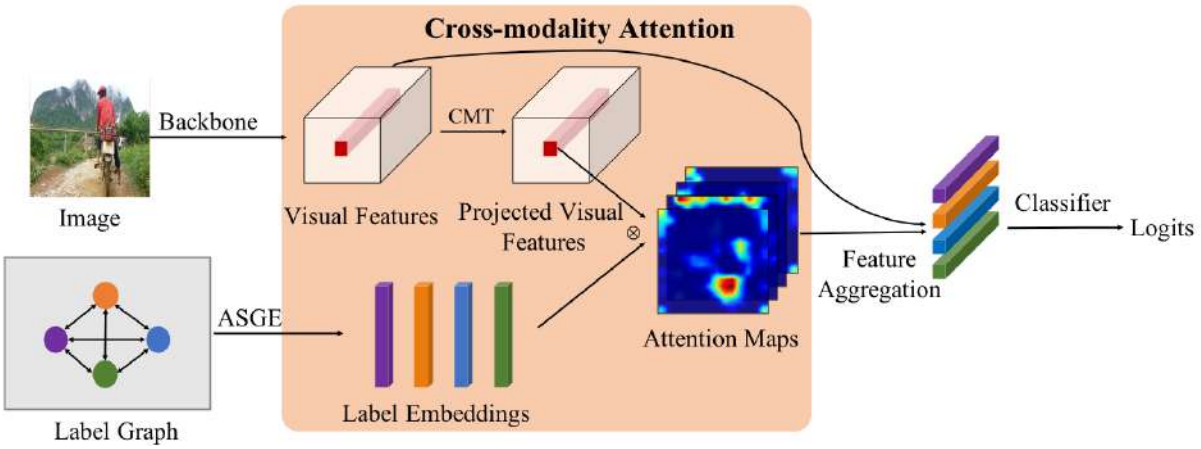


Figure 2.11: General architecture for the MLIC task of the MS-CMA. Label embeddings are given through ASGE. At the early stage, backbone network extraction of the visual data, which are projected in semantic space to get the projected visual features through the CMT module. The projected visual features and learned label embeddings are input into the CMA module to prepare category-wise attention maps. These maps are then used to average the visual features and produce category-wise aggregated features weightedly. The classifier is then utilized to make the last prediction [10].

2.2.3.D Multi-Class Attentional Regions for Multi-Label Image Recognition

The MCAR model presented by Gao and Zhou [11] is a two-stream approach tailored to efficiently and effectively recognize multiple objects in an image. It comprises a global image stream and a local region stream, merging the gap between global image features and the local region.

Global Image Stream: This entire image was processed in this stream to extract global features using a deep convolutional neural network. These global features give a general idea of what the image contains.

Local Region Stream: This stream operates on the areas of the image identified by the global stream. The MCAR module dynamically creates very few diversified attentional regions to maintain a high diversity level, all the while not incurring high computational costs. These areas are then dissected in detail for better accuracy in object recognition.

Key Contributions Region Localization: The MCAR framework designs an extra parameter-free module to suppress irrelevant context, using an attention mechanism on effective regions and relieving heavy object proposals or bounding-box annotations. The underlying idea is that human visual perception relies on global context to steer attention to the particular areas of the visual field.

State-of-the-Art Performance: The new state-of-the-art results by Gao and Zhou's method rely on their extensive benchmarking on MS-COCO and PASCAL VOC datasets in multilabel image classification. It greatly enhances the existing approaches to semantics in images without involving label dependencies.

Robustness and Generalization: The excellent design of the MCAR framework is demonstrated to stand well under various circumstances—worldwide pooling strategies, input sizes, and network archi-

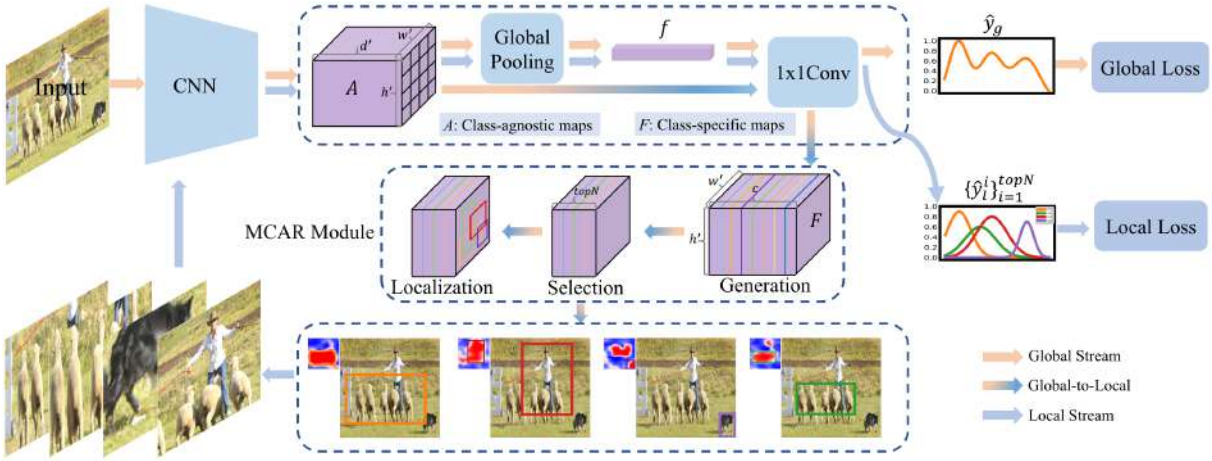


Figure 2.12: The multi-label image recognition pipeline of the MCAR framework commences with extracting the global image stream for feeding an input image into the deep CNN model to obtain its global feature representation. The multi-class attentional region module approximates the localization of regions of potential objects by adding data from the global stream. The MCAR technique is then applied later for inference by aggregating the final prediction through category-wise max-pooling of predictions from both the local and global streams. These localized regions are ultimately input to the shared CNN to acquire the expected class distributions via the local region stream [11].

tectures. High performance can be obtained at much reduced computational costs, which is appropriate for the practical application.

2.2.3.E Transformer-based Dual Relation Graph for Multi-label Image Recognition

After that, Jiawei Zhao et al. (2020) [12] proposed a new Transformer-based Dual Relation Graph (TDRG) framework for multilabel image recognition. Most traditional methods of multi-label classification rely on static label correlations or use simple co-occurrence statistics, which may not model the complex relationship between labels within an image adequately. The TDRG framework thus models structural and semantic information jointly by two mutually complementary relation graphs: a structural relation graph and a semantic relation graph.

Structural Relation Graph The long-range contextual correlation of the regions is learned within the object context by the structural relation graph. This architecture allows position-wise building of spatial relations across scales, which is very important for high-precision recognition of objects with significant variations in size and appearance. The application of transformers in this context widens the receptive capability of conventional CNNs, allowing the model to look at global contextual information effectively. **Semantic Relation Graph** The semantic relations graph manages to model the dynamic semantic meaning of image objects by explicit semantic-aware constraints. Compared to the static approaches, this graph is also adjusted dynamically to the specific content of each image, which, therefore, enhances the model's capability to deal with objects that tend to have less frequent co-occurrences.

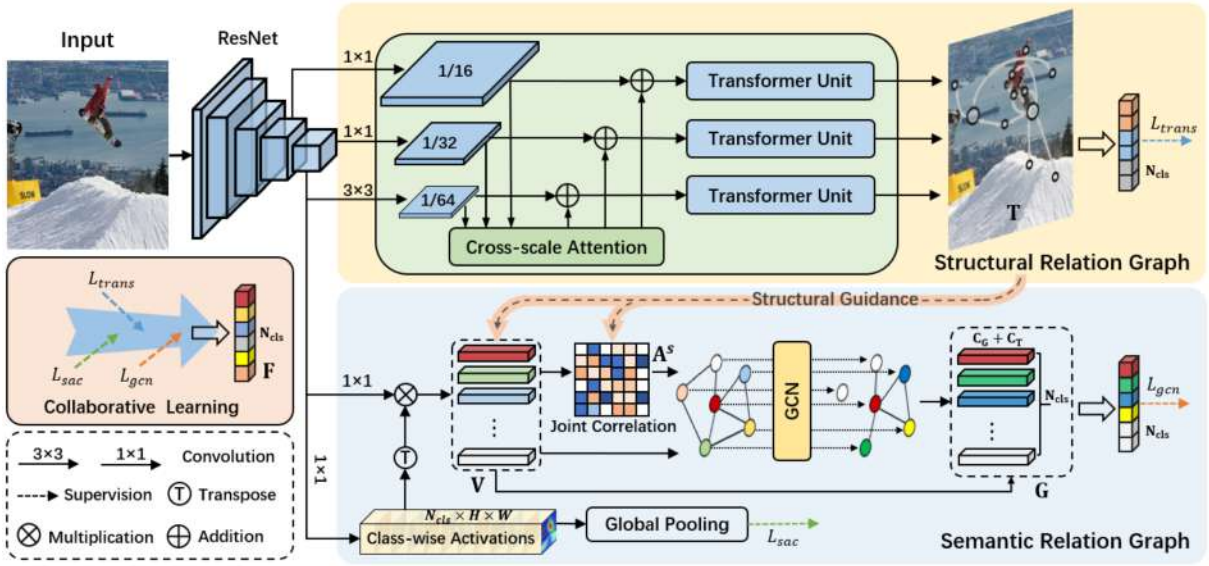


Figure 2.13: The general structure of the Transformer-based Dual Relation Graph (TDRG) network, which is comprised of two fundamental modules: the semantic relation graph module, which models the dynamic class-wise dependencies, and the structural relation graph module, which incorporates long-term contextual information [12].

Moreover, structural information in the semantic graph enhances the robustness of the representations by integrating both the mechanisms of adjacent correlation construction and feature-wise complementary mechanisms.

Collaborative Learning The TDRG framework employs a collaborative learning approach to optimize both the structural and semantic relation graphs jointly. This ensures that, in the final classification model, the spatial relationships captured by the structural graph and the semantic dependencies modeled by the semantic graph can be taken advantage of. The collaborative learning process significantly enhances the model's performance on multi-label classification tasks since it dramatically enriches the depth and breadth of understanding of image content.

2.3 Overview of Underwater Object Detection and Classification Methods

Object detection and classification is an important aspect in underwater computer vision and image processing, which enables computers to understand and analyze underwater scenes. Understanding and interpreting underwater scenes is critical to understanding and utilizing marine environments and shows the possibility of impacts in revealing and utilizing underwater resources. Object detection has been traditionally researched using manually engineered features; although feature engineering has established early research in the field, there are significant limitations in their performance in real-world underwater

situations. Relying on handcrafted features for analysis requires specialized knowledge, with complex algorithmic adjustments and measurements for systems relying primarily on humans to handle variable underwater settings; furthermore, handcrafted features often lack universality and have a lower true positive rate, which limits the contributions to related fields.

Recently, there has been a sea-change towards deep learning methods in the domain of underwater object detection. The deep learning paradigm has the potential for the models to automatically learn hierarchical feature representations from raw data and reduce dependency on manually engineered features. This evolution resolves so many of the challenges of traditional methods, with an increase in flexibility and accuracy around challenging underwater sessions. The use of deep learning does not solely advance an understanding of underwater scenes, but it can significantly move forward exploration capabilities and factually support a sustainable use of marine resources. According to Jian et al. [18], in this field, many methods have been devised over the years, and they can be broadly grouped into two general styles of methods: classical methods using handcrafted features and modern methods based on neural networks or deep learning. Duan et al. [47] thoroughly reviewed the state of research concerned with the visual attributes of marine organisms—including size, shape, and color—seen from a computer vision perspective. Their work analyzed several stages of image processing, from image acquisition to contour extraction to feature calibration to calculations. They also addressed the use of computer vision methods to diagnose, detect, and classify diseases in aquatic organisms. However, such traditional methods are sometimes severely limited in terms of their practical application under real underwater conditions (e.g., variable light, visibility, the dynamic nature of aquatic organisms).

Lighting conditions play a vital role in underwater image quality and object detection & classification accuracy. Wu et al. [48] studied the impact of various lighting environments on the characteristic of underwater images, specifically, modifying various image processing algorithms to extract invariant features that will be less affected by each of the lighting environments during experimentations with underwater detections of red ball objects, to see what was feasible and reliable depending on the underwater illumination conditions. This work led to further understandings of how lighting conditions could effect invariant feature extraction techniques by each of the lighting environments and, overall performance of underwater object detection system. In addition, Yu et al. [49] conducted research on techniques of data collection of marine habitats, comparison of different datasets and some preprocessing techniques of underwater images. They also studied about the application of object detection algorithms in underwater datasets. The study by Peng et al. [50] explored the use of deep learning techniques for preprocessing underwater images. They reviewed the advantages of deep learning approaches to real-world underwater imaging challenges such as varying lighting conditions and water turbidity. They also noted the limitations of existing deep learning models, and suggested improvements such as the formulation of stronger architectures and the inclusion of domain-level knowledge to tackle underwa-

ter application challenges. Jian et al. [18] 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 - Underwater object detection techniques can be classified into three categories based

Table 2.1: Summary of underwater object detection methods based on traditional artificial features. Source: [18]

Category	Reference
Texture features	Han and Choi [51]; Beijbom et al. [52]; Nagaraja et al. [53]; Fatan et al. [54]; Srividhya and Ramya [55]; Shi et al. [56]
Color and motion features	Gordan et al. [57]; Chen and Chen [58]; Singh et al. [59]; Komari Alaie and Farsi [60]; Susanto et al. [61]
Saliency detection	Wang et al. [62]; Zhu et al. [63]; Jian et al. [64]

upon traditional-techniques, as shown in the table2.1 above: object detection using texture features, object detection using color and motion features, and object detection using saliency detection. Generally, most of the initial work in underwater object detection relied heavily on texture-based analysis, as this type of analysis is easier to implement. When considering color and motion features, this technique relies on the dynamic aspects of underwater environments, such as marine life movement and changes in illumination. Commercial and scientific aerial cameras may also exploit color and motion features to facilitate detection. The most recent type of categorization is saliency detection, which attempts to identify the most distinguished visual features, in an image, to aid object detection, especially for complex scenes. Despite the contributions of traditional methods in earlier years, all of the traditional methods are not robust against noise, varying light conditions, and occlusion, perhaps indicating the need for more advanced techniques based upon deep learning that will account for some of these challenges. The table2.2 provides a summary of the various algorithms developed with deep learning for underwater

Table 2.2: Summary of the deep learning methods for underwater object detection. Source: [18]

Category	Reference
Single-stage algorithm	Liu et al. [39]; Bochkovskiy et al. [65]; Hu et al. [66]; Ge et al. [67]; Lei et al. [68]
Two-stage algorithm	Girshick et al. [2]; Girshick [4]; Ren et al. [69]
High noise and low contrast	Chen et al. [70], [71]; Sun et al. [72]
State changes and occlusion	Yang et al. [73]; Lin et al. [74]; Lau and Lai [75]; Zhang et al. [76]
Shadows and uneven illumination	Song et al. [77]; Cao et al. [78]; Li et al. [79]; Ding et al. [80]; Yu et al. [81]; Fan et al. [82]; Wei et al. [83]; Chen et al. [84]
Weak lighting and low quality	Rashwan et al. [85]; Chen et al. [86]; Han et al. [87]; Ge et al. [67], b; Liu et al. [88]
Low data volume	Zurowietz and Nattkemper [89]; Zeng et al. [90]
Saliency detection	Li et al. [79]; Mou et al. [91]; Zhou et al. [92]; Chen et al. [93]

object detection organized by each focus area/technique. Generally, single-stage algorithms are used for real-time detection. Each of the single-stage algorithms listed is able to directly predict the placement

and class of pools in one prediction. Each of the two-stage methods first proposes region proposals for the object or objects and the class regard after which the model is able classify more accurately than can be achieved with a single prediction while sacrificing some speed. Certain methods concentrate on the problem of noise and low contrast of underwater images while others focus on the issues with state change and occlusion that can occur in a dynamic marine environment. Other methods focus on the dealing with shadows and illumination changes to maintain a high accuracy rate for finding objects even with the illumination has changed from undulating object movement. Further, some methods deal with weak lighting and quality of images to mitigate error associated with the generative algorithms of high resolution and reliable computer detection. Further, low data availability is an issue tempoed and dedicated methods have been developed to address all situations in which the data are controversial in the number of training available and generally, the type of model used to implement. Finally, saliency detection consideration considered generated of sensing and situating the most salient or prominent objects from the backgrounds subjects of the underwater environments. This extensive and diverse amount of methods demonstrates how well deep learning can promote and facilitate underwater object detection.

2.3.1 Improving Model Robustness through Data Augmentation

Data augmentation is an important and useful tool for improving the robustness of a model to deep learning failure in computer vision (CV) settings, particularly in challenging environments like underwater imaging. When augmentations are made to a dataset by artificially enlarging the current available dataset, these augmentations improve the ability of the algorithm to generalize to new data and possible future loss of performance in a real-world application.

Leveraging these techniques in computer vision is one of the contributions of this research, to address model robustness. Geometric transformations, which may appear as rotation, scaling, and flipping in addition to color changes (brightness, contrast, hue modifications, etc.) using the PyTorch ColorJitter Library, can simulate the variety of conditions that typically occur with underwater images. Other augmentation methods such as adding noise, blurring and sharpening, to successful computer vision models may also improve the model's ability to adapt to situations regularly encountered underwater associated with low visibility, increased noise and variability in image quality. Methods utilizing cutouts—a methodology where spatial areas of an image are randomly obscured—train models to recognize objects that could be partially occluded at the time of capture, which commonly appears with vegetation or other marine organisms when framing an underwater scene. Generating synthetic data, typically relying on Generative Adversarial Networks (or GANs), provide more examples with desirable properties and variations necessary when real accessible data is limited if available at all.

These augmentation techniques contribute profoundly to the robustness of the underwater object detection system, by improving the performance and adaptability in complex underwater environment.

2.4 Challenges & Limitations

2.4.1 Technical Challenges

Traditionally, most established object detection and classification methods have been evaluated in the context of airborne or terrestrial data, their applicability to underwater imaging an open question. The underwater environment poses challenges that are unique in terms of lighting variability, particulate visibility, and occlusion, and is not similar to airborne environment. The factors require approaches that address the challenges associated with the underwater environment. Here we discuss the challenges with underwater scenes and why these environments need to be separately considered.

A – Occlusions in Object Detection or Classification The underwater world is dynamic and complex, making object detection or classification very challenging. In many cases, marine life, vegetation, and particulates would occlude critical features of the objects, reducing the accuracy of object detection. Traditional object detectors do pretty well in explicit, unobstructed scenes but fail to perform satisfactorily in aquatic scenes unless trained for the same. For instance, a model trained on images of fish in open water may not be able to recognize a fish that is partially covered by algae.

B – High False Positive Rates in Anomaly Detection Identifying anomalies in the unstable underwater environment is challenging. Light behaves differently underwater, affecting object appearance and visibility, and the swift water flow can change the background, causing common objects to be mistakenly identified as anomalies. For instance, unusual shadowing caused by sunlight refraction through water may lead to a harmless object being perceived as a hazardous anomaly. As a result, models might either raise too many false positives or miss actual anomalies necessitating sophisticated definitions and models of 'normality' in underwater contexts.

C – Complexity in Multi-label Classification Underwater views are usually packed with many observed objects and species. This means that multi-label classification is required where a certain image can be assigned to multiple labels. However, this task is complicated by dissimilar appearances of objects, inter-class relations, and object scales vary greatly. These dependencies cannot be captured using the classical multi-label classifiers that typically treat each label as an independent entity resulting into misclassifications.

2.4.2 Broader Issues

A – Dataset Biases The effectiveness of machine learning models heavily relies on the quality and representativeness of training data. Unfortunately, most available datasets are biased towards terres-

trial images. Models trained on such data are likely to perform poorly in marine environments and misidentifying or failing to recognize underwater objects. For example, a common coral structure might be incorrectly classified as a foreign object due to its absence in the training set. Developing diverse datasets that reflect the wide variety of underwater scenes is crucial for creating robust and accurate models.

B – Model Interpretability The black box nature of sophisticated machine learning models presents significant interpretability issues, especially in critical underwater applications. For instance, it would be difficult to correct an autonomous underwater vehicle's mistake if it misidentified a rock as a dangerous underwater mine. This lack of transparency in the model's decision-making process complicates troubleshooting and improvement efforts. There is a strong need for explainable AI where the rationale behind model decisions can be understood and trusted by human experts.

C – Computational Demands Capturing the information needed for precise underwater item and anomaly identification requires high-resolution images. However, real-time analysis of this kind of data necessitates large computer resources, which are frequently unavailable in isolated underwater locales. Creating models that can operate on small, low-power devices without compromising speed or accuracy is a problem. This is especially important for applications where delays might mean missing important occurrences or not capturing fleeting phenomena, such as real-time monitoring of marine habitats.

D – Absence of Benchmark in Underwater Datasets: Setting up benchmarks specifically designed for underwater conditions is essential to efficiently assess and create models for marine applications. In order to create models capable of recognizing underwater items and abnormalities, these benchmarks must take into consideration the particular constraints associated with the underwater environment, such as fluctuating visibility, complicated backdrops, and different objects.

E – Environment Variability Changes in the underwater environment can occur quickly and significantly. Changes in water clarity, depth, and time of day may all have a significant impact on how an item appears in the light. Applications needing constant performance, such tracking marine life across time or traversing underwater terrains for exploration, may find it difficult to use models learned in one set of circumstances to perform effectively in another. The development of flexible models that can react to shifting environmental circumstances must be the main goal of future study.

F – Unpredictable Elements Underwater surroundings present an additional layer of difficulty due to their unpredictable nature. Rare aquatic creatures might suddenly materialize, or human activities could bring inadvertent items like plastic garbage. Models that have not been exposed to these kinds of data

may classify these unanticipated elements incorrectly. One of the main challenges facing researchers is the construction of models that can swiftly adapt to new components on which they have not been explicitly trained.

G – Need for Specialized Training Data Developing specialized underwater datasets that include a wide range of marine settings and species is crucial to addressing these difficulties. Collaborations in AI, oceanography, and marine biology can help achieve this, and robust model training can be supported by the creation of synthetic data and data augmentation.

3

Fathomnet Competition Dataset

Contents

3.1 Dataset Description & Preparation

The FathomNet dataset is an extensive underwater imagery dataset that has been amassed through collaboration with the FathomNet community associated with the Monterey Bay Aquarium Research Institute (MBARI). The dataset provides a variety of marine species images captured at various depths in the ocean, in support of underwater object detection and classification research.
The quality and relevance of the provided imagery data was achieved through a selection process targeting specific depth ranges, along with a specific search area to minimize species variation biases due to location. Additionally, the search was limited to specific geographic regions to minimize the effects of temporal species variation [13]. This dataset was constructed by concentrating on species that live at the bottom, utilizing a list of concepts developed by local taxonomic specialists and all the represented 290 categories are benthic fauna captured by similar camera systems in that area.

Figure 3.1: *S. fragilis.* is the most commonly found concept in the Fathomnet 2023 Dataset both in the training and the evaluation set.

All the camera systems were developed by the MBARI which were mounted on two Remotely Operated Vehicles (ROVs). These data were collected exclusively from the Greater Monterey Bay Area (**35.38N to 37.199N, - 122.8479W to - 121.0046W**) between the surface and 1300 meters depth [13]. The annotation with the dataset is provided in multi-label classification and object detection formats. The multi-label classification annotations provides for each image a set of categories found in it. The object detection annotations are presented in COCO Object detection standard format with each image containing at least one localization. There are 20 semantic supercatogies in addition to the fine-grained labels. Every supercategory is present in both the training and evaluation data, however not all fine-

grained categories are represented in both sets [13].

3.2 Properties of Fathomnet 2023 Dataset

In the field of underwater image analysis and species identification, the properties of Fathomnet Competition 2023 Dataset [13] have important effects for object detection/multilabel classification and out of sample detection. The following are some significant ways that the characteristics of the dataset may affect these areas:

Depth-Based Distribution: 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 3.2 shows the depth distribution in both train and test set.

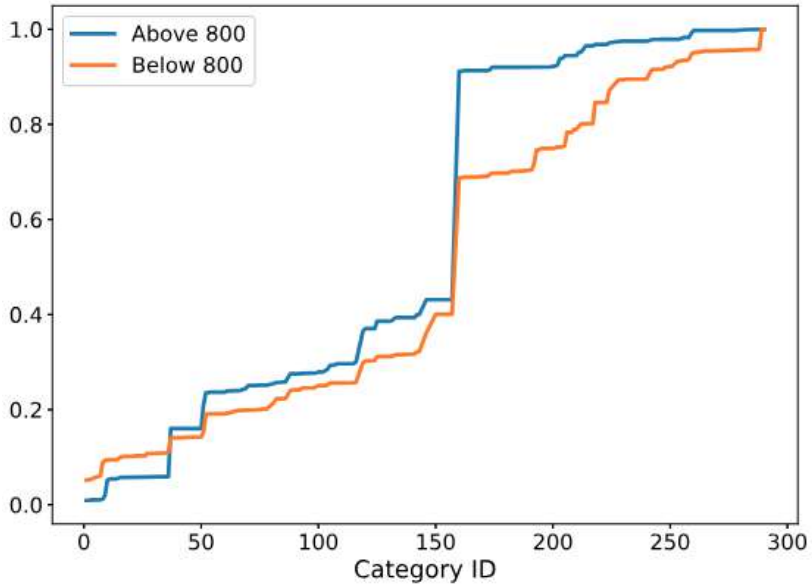


Figure 3.2: The overall distribution of categories in the FathomNet 2023 training and evaluation datasets. They differ greatly from one another, with some classes existing in only one of them [13].

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 [13]. 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.

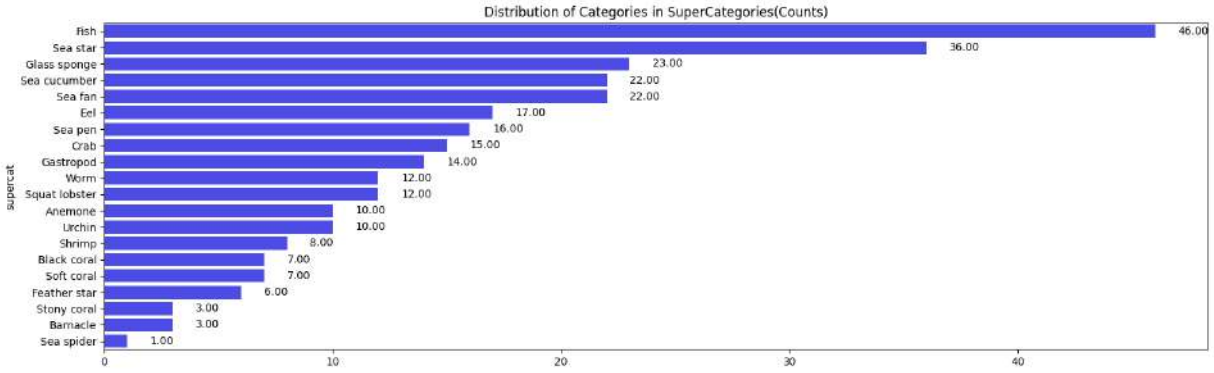


Figure 3.3: Categories Count in Supercateogires

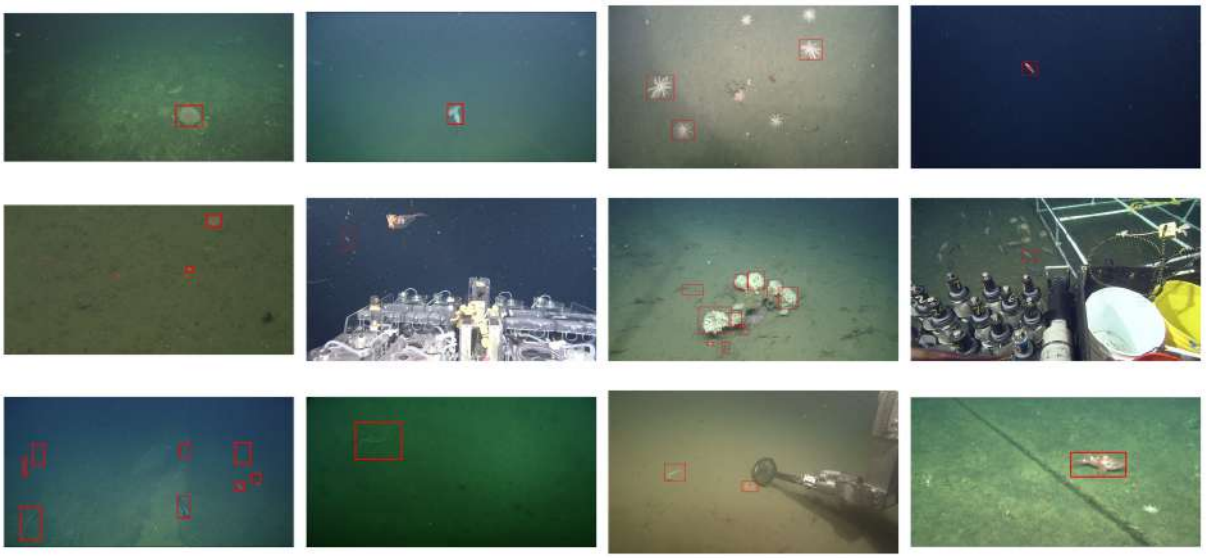


Figure 3.4: Annotation Sample of the Fathomnet Dataset

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 3.3 shows the distribution 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.

Annotation and Localization Quality: In addition to fine-grained labels, the dataset offers annotations in the forms of object recognition and multi-label classification. Every image has at least one localization labeled as a supercategory. To train reliable object detection algorithms, these annotations' quality and detail are essential. Figure 3.4 shows some examples of the annotations which were provided with the Fathomnet dataset. There are insufficient annotations and label noise in certain photos which may result in a decrease in the performance of the model, especially when it comes to differentiating closely

related species or recognizing animals that are partially obscured or fuzzy. **Representation of Super Categories and Fine-Grained Categories:** In the Fathomnet 2023 dataset all the Super Categories are represented in both the training and evaluation dataset but not all fine-grained categories are present in both sets. This imbalance can lead to biased models that perform well on over-represented categories but poorly on underrepresented or unseen categories. In out-of-distribution detection, this could lead to lower accuracy or confidence in detecting less common species.

To sum up, the Fathomnet 2023 dataset's parameters annotation quality, category representation, long-tailed distribution and natural picture variability have a big impact on how successful object detection / classification models are. The capacity of the model to detect a broad range of species under various underwater settings, as well as its generalizability to out-of-sample data, can all be impacted by these aspects. Therefore, while developing and assessing classification or object detection models for underwater imaging and marine biology, it is crucial to take consideration of these factors.

4

Methodology

Contents

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.

4.1 Dataset Pre-Processing

4.1.1 Underwater Light Propagation

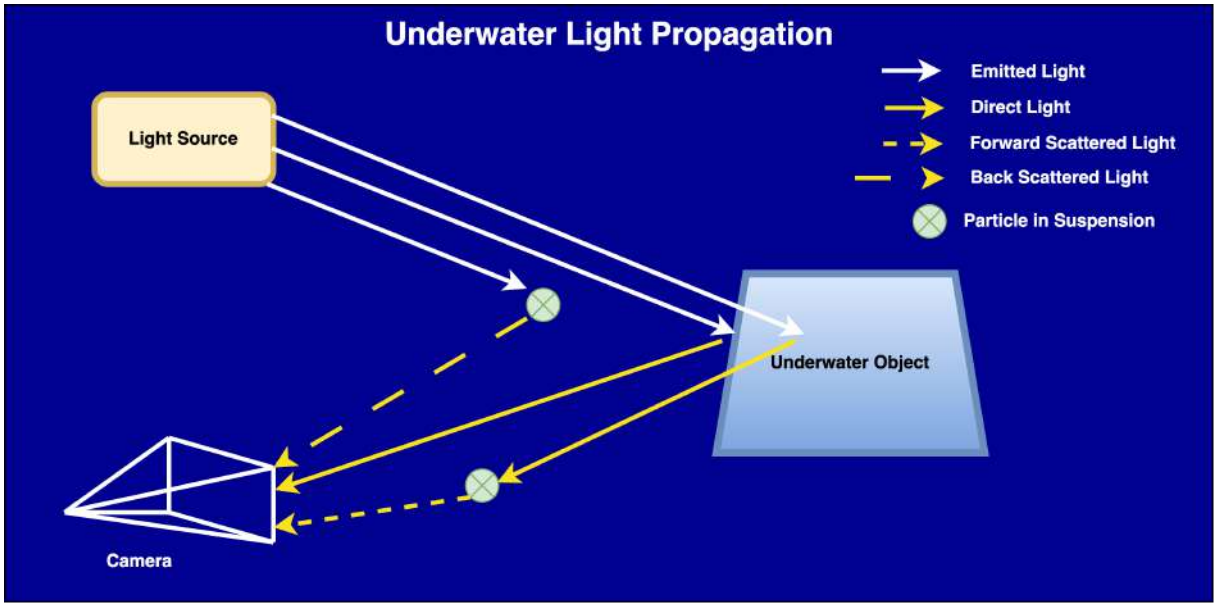


Figure 4.1: As light travels through water, a portion of the emitted light is absorbed and transformed into other forms of energy. Additionally, some photons interact with suspended particles en route to the sensor, causing scattering by acting as secondary light sources [14].

Both absorption and scattering effects have a major impact on the light behaviour in the underwater environment [14, 94–97] as illustrated in figure 4.1. Absorption is converting light energy into other forms of energy in the interaction with water molecules and dissolved substances. Naturally, absorption varies with light wavelengths. Longer wavelengths are absorbed more intensely, for example, red and infrared, compared to blue and green. This results in a decrease in color and contrast while the light travels deeper underwater. On the other hand, scattering occurs when light comes into contact with particles suspended in the water, including tiny solids, phytoplankton, or other contaminants. Underwater scenes appear fuzzy or less defined as a result of these interactions, which also cause light to shift direction and scatter at different angles, reducing visibility and causing image blurring and light diffusion. Similar to absorption, scattering is wavelength-dependent and dependent on light's path length. As shown in the figure 4.1, there are two types of scattering found in practice which are forward scattering and back scattering. Backscattered light, in contrast to forward scatter, does not provide any information about the observed scene [97].

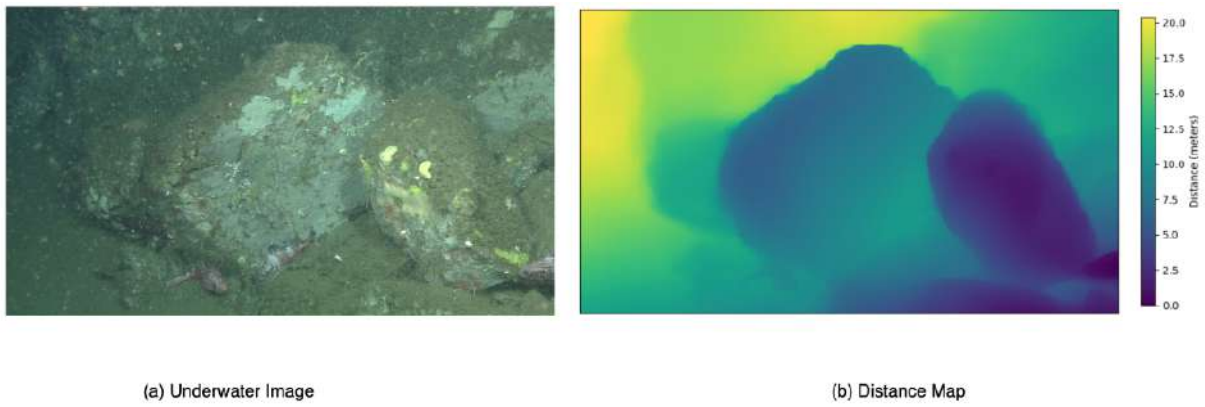


Figure 4.2: This figure presents an underwater image alongside its corresponding depth map, which was generated using Depth Anything [15].

For observing this situation in real life context, we test the relationship between pixel intensity and the corresponding distance of the scene being observed. Figure 4.2 shows an example of underwater image and the distance map of the same image which was obtained from Depth-Anything [15] model. Figure 4.3 illustrates different absorption rates in different wavelength of the red, blue and green channels. The longer wavelength like red is being absorbed quickly than the shorter ones like green and blue.

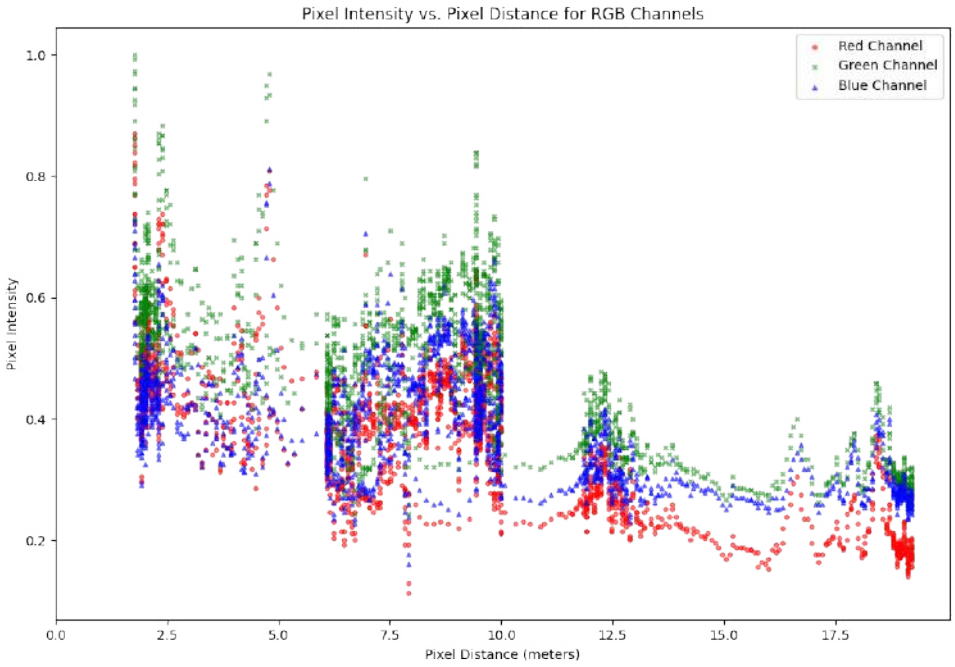


Figure 4.3: Plotting pixel intensities against observation distances of Figure 4.2 shows that underwater absorption causes red light to diminish faster with distance than blue light.

4.1.2 Underwater Image Formation Model

We have seen previously the main effects which occur in the underwater images. Now, we will try to model these effects in the resulting images. The main component we will use for retrieving the images without these disturbing effects is the underwater image formation model which describes how the colors of the underwater scene have impact on the water medium [14].

Variable	Description	Type
I	**underwater images**	$\mathbb{R}^{N\times H\times W\times C}$
J	**restored images**	$\mathbb{R}^{N\times H\times W\times C}$
z	**distance maps of images**	$\mathbb{R}^{N\times H\times W}$
B	**veiling light**	$\mathbb{R}^{C}$
β	**color absorption coefficient**	$\mathbb{R}^{C}$
γ	**backscatter coefficient**	$\mathbb{R}^{C}$
i	**image index**	$[1..N]$
c	**color channel index**	$[1..C]$
p	**pixel index**	$[1..H\times W]$

Table 4.1: Underwater Image Formation model variables. Source: [14].

Based on the initial underwater image formation model proposed by Schechner and Karpel [98], numerous color restoration techniques [99–101] have been proposed which addresses backscatter and

color absorption under natural lightning condition. The underwater image formation model shows that the pixel intensities are driven by the following equation:

$$\boldsymbol{I_{c,p}} = \boldsymbol{J_{c,p}} e^{-\alpha_c z_p} + \boldsymbol{B_c}(1 - e^{-\alpha_c z_p}) \tag{4.1}$$

where $\alpha \in \mathbb{R}^C$ denotes the wavelength-dependent coefficient that influences the distance dependency of color absorption and backscatter. The other variables are described in the table 4.1. However, Akkayanak et al. [96] further refined the model presented in Eq. 4.1 to differentiate between backscatter and absorption coefficients. The revised relationship between pixel intensities and observation distance is expressed as:

$$\boldsymbol{I_{c,p}} = \boldsymbol{J_{c,p}} e^{-\beta_c z_p} + \boldsymbol{B_c}(1 - e^{-\gamma_c z_p}), \tag{4.2}$$

where β and γ are the absorption and backscatter coefficients defined in Table 4.1.
Boittiaux [14] observed that pixel intensities in underwater images tend to follow a normal distribution, with their mean and standard deviation dependent on distance. Leveraging this insight, he proposed an optimization method of Sea-Thru [96] called Gaussian SeaThru, which enhances the Sea-Thru approach by incorporating Gaussian priors over the red, blue, and green channels of the restored image. This technique aims to improve the accuracy of color restoration in underwater images by more effectively modeling the statistical properties of pixel intensities.

$$\boldsymbol{J_{c,p}} \sim \mathcal{N}(\mu_c, \sigma_c^2) \tag{4.3}$$

where the channel-wise mean and standard deviation of the restored image pixel intensities are denoted by μ_c and σ_c, respectively. By adjusting the parameters of Equation 4.3 in accordance with Equation 4.2, we can infer that the acquired image's pixel intensities likewise conform to a normal distribution:

$$\boldsymbol{I_{c,p}} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{m_{c_p}}, \boldsymbol{s_{c,p}^2}), \tag{4.4}$$

where

$$\boldsymbol{m_{c,p}} = \boldsymbol{\mu_c} e^{-\beta_c z_p} + \boldsymbol{B_c}(1 - e^{-\gamma_c z_p}), \tag{4.5}$$

and

$$\boldsymbol{s_{c,p}} = \boldsymbol{\sigma_c} e^{-\beta_c z_p} \tag{4.6}$$

From Equation 4.4, we can express the likelihood of observing I_c:

$$\boldsymbol{L(I_c)} = \prod_{\boldsymbol{p}} \left(\frac{\mathbf{1}}{\boldsymbol{s_{c,p}}\sqrt{\mathbf{2\pi}}} \exp\left(-\frac{\boldsymbol{(I_{c,p} - m_{c,p})^2}}{\mathbf{2}\boldsymbol{s_{c,p}^2}} \right) \right) \tag{4.7}$$

We can then estimate the veiling light ($\boldsymbol{B_c}$), color absorption coefficient ($\boldsymbol{\beta_c}$), and backscatter coefficient (γ_c) by minimizing $\mathbf{log}(\boldsymbol{L}(\boldsymbol{I_c}))$:

$$\arg\min_{B_c,\beta_c,\gamma_c} \sum_{p} \left(\log(s_{c,p}\sqrt{2\pi}) + \frac{(I_{c,p} - m_{c,p})^2}{2s_{c,p}^2} \right) \tag{4.8}$$

Then the restored image $\boldsymbol{J_c}$ is obtained using Equation 4.2:

$$J_{c,p} = \left(I_{c,p} - B_c(1 - e^{-\gamma_c z_p})\right) e^{\beta_c(z_p) z_p} \tag{4.9}$$

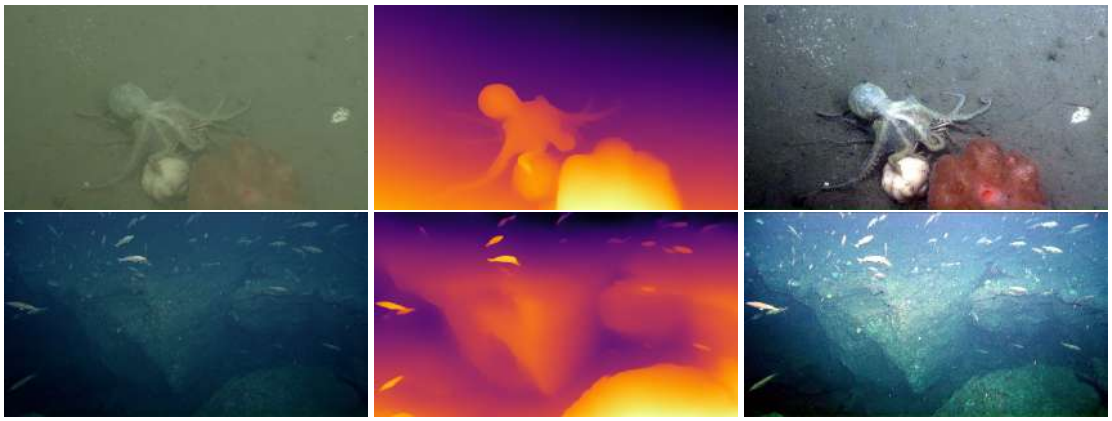


Figure 4.4: Some examples of the visualization of the original image(left), distance map(middle) obtained from Depth-Anything [15] and the restored image(right).

4.1.3 Using Underwater Image Formation Model for Data Augmentation

For our multi-label image classification we used the image formation model described in the eq.4.2 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 $I_{c,p}^{orig}$ is the original image and $I_{c,p}^{mod}$ is the restored modified image. We can denote both of the image by the image formation model as follows:

$$\boldsymbol{I_{c,p}^{orig}} = \boldsymbol{J_{c,p}} e^{-\beta_c z_p} + \boldsymbol{B_c}(1 - e^{-\gamma_c z_p}), \tag{4.10}$$

and

$$\boldsymbol{I_{c,p}^{mod}} = \boldsymbol{J_{c_p}} e^{-\beta_c z_m} + \boldsymbol{B_c}(1 - e^{-\gamma_c z_m}), \tag{4.11}$$

we can eliminate $J_{c,p}$ as follows:

First, solve for $J_{c,p}$ from Equation 4.10:

$$\boldsymbol{J_{c,p}} = \frac{\boldsymbol{I_{c,p}^{orig}} - \boldsymbol{B_c}(1 - \boldsymbol{e}^{-\gamma_c \boldsymbol{z_p}})}{\boldsymbol{e}^{-\beta_c \boldsymbol{z_p}}} \tag{4.12}$$

Next, substitute this expression for $J_{c,p}$ into Equation 4.11:

$$\boldsymbol{I_{c_p}^{mod}} = \left(\frac{\boldsymbol{I_{c,p}^{orig}} - \boldsymbol{B_c}(1 - \boldsymbol{e}^{-\gamma_c \boldsymbol{z_p}})}{\boldsymbol{e}^{-\beta_c \boldsymbol{z_p}}} \right) \boldsymbol{e}^{-\beta_c \boldsymbol{z_m}} + \boldsymbol{B_c}(1 - \boldsymbol{e}^{-\gamma_c \boldsymbol{z_m}}) \tag{4.13}$$

Simplify the equation:

$$\boldsymbol{I_{c_p}^{mod}} = \left(\boldsymbol{I_{c_p}^{orig}} - \boldsymbol{B_c}(1 - \boldsymbol{e}^{-\gamma_c \boldsymbol{z_p}}) \right) \boldsymbol{e}^{-\beta_c (\boldsymbol{\Delta z_m} - \boldsymbol{z_p})} + \boldsymbol{B_c}(1 - \boldsymbol{e}^{-\gamma_c \boldsymbol{\Delta z_m}}) \tag{4.14}$$

In the equation 4.14 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.

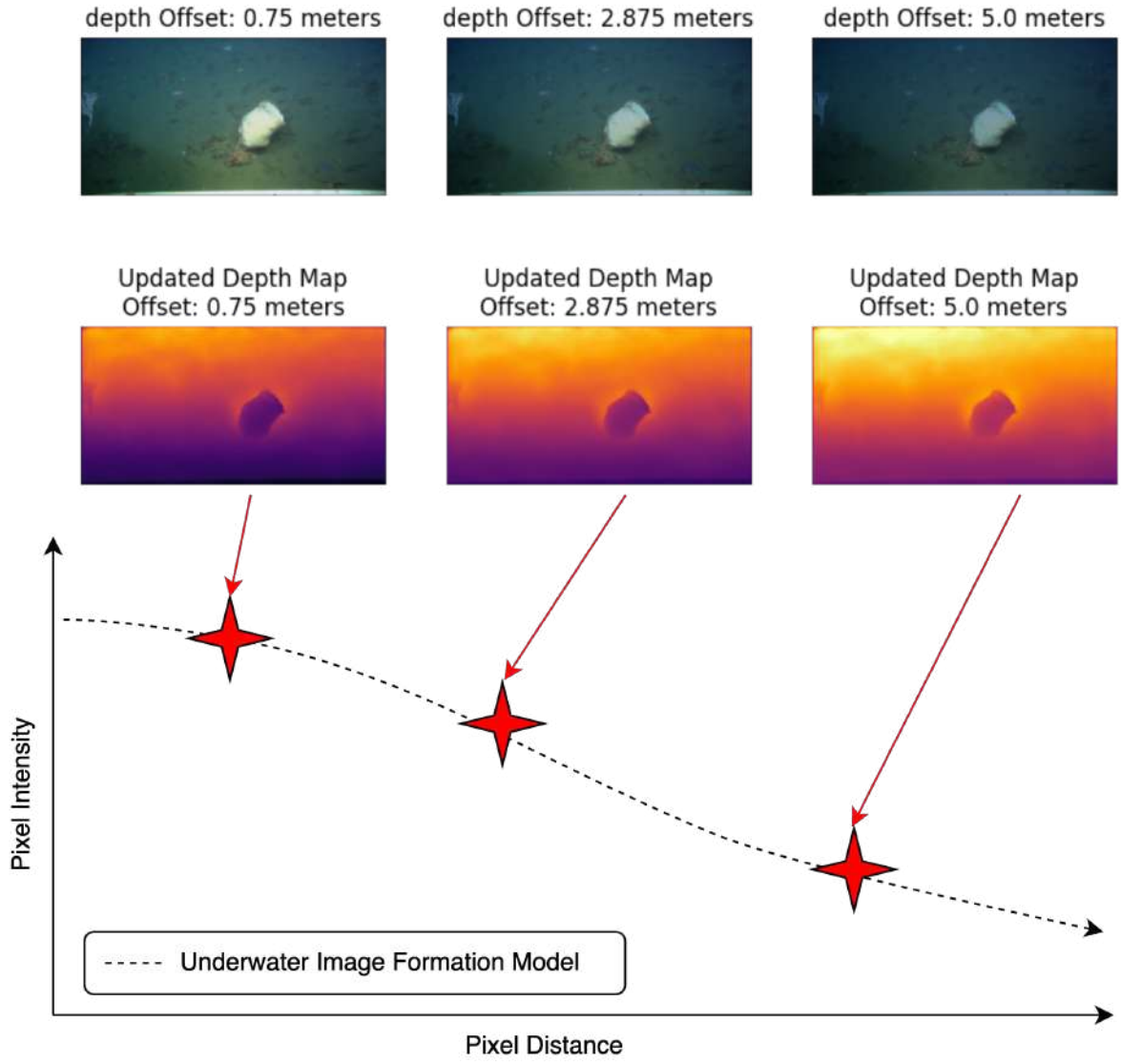


Figure 4.5: This figure shows the pixel intensity tracking and change on the image in different depth settings.

Figure 4.5 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 4.5. 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.

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 4.14). This makes this augmentation strategy more efficient in the training procedure.

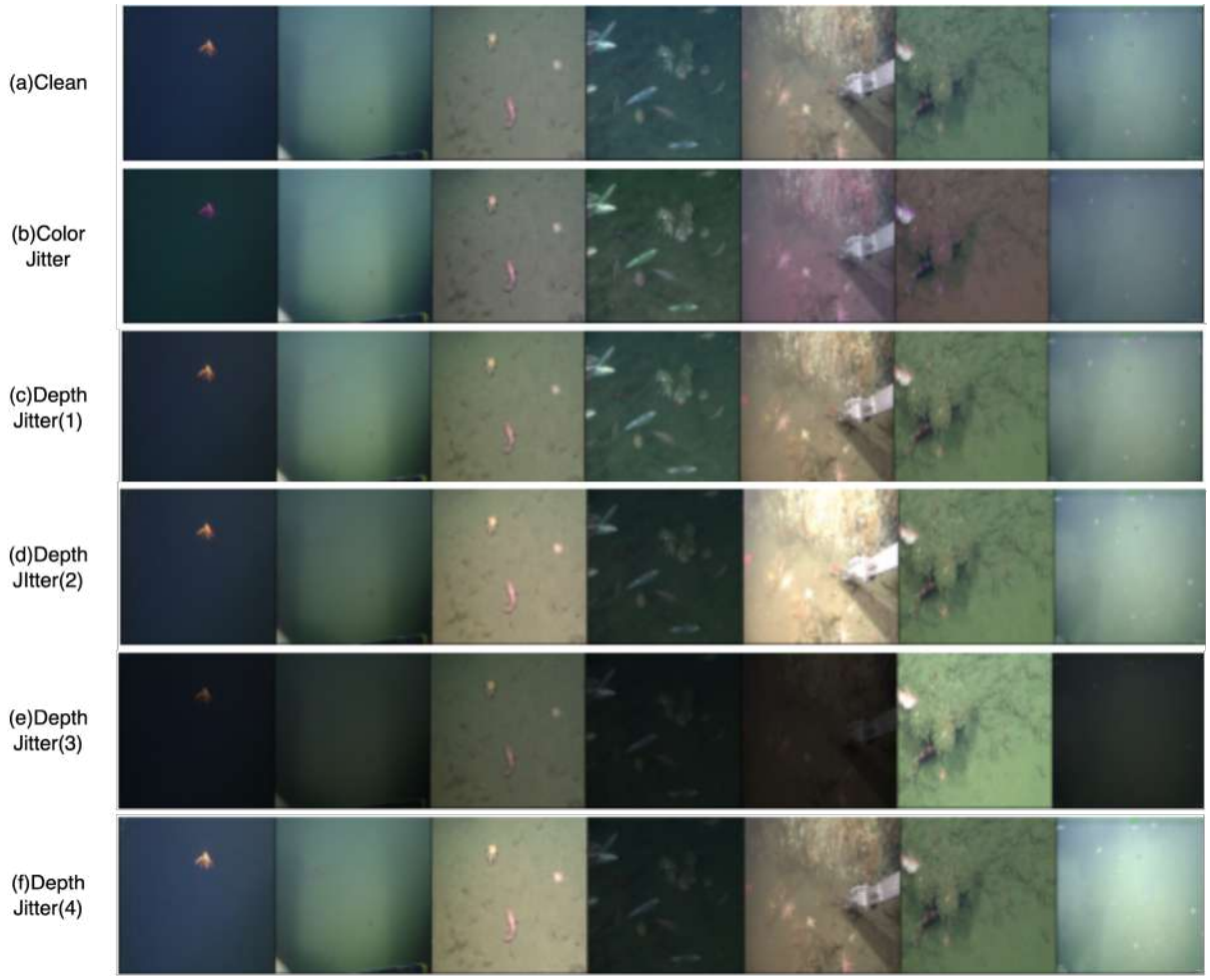


Figure 4.6: Comparison of Different Augmentation techniques.

In figure 4.6 total three(3) types of augmentation techniques were applied. First row represents the images without any augmentation. The second row represents the Pytorch Color Jitter augmentation. This is a fixed augmentation technique which remains same for each epoch for each image. We make our comparison against Color Jitter since it is a simple and commonly used technique to create appearance variability through color change. By serving as a baseline, it helps to highlight the effectiveness of our proposed technique.

Last four(4) rows represent our proposed augmentation technique Depth Jitter. We can see that the depth of each image is different each time the batch is called. This is only possible because we use randomized depth offset to each image in the equation 4.14. Thus we get more balanced and generalized

dataset than to use only the Pytorch Color Jitter to make our neural network model more robust to the real world environment for detecting out of distributed samples relative to the training set.

4.2 System Overview

4.2.1 Multi-Label Image Classification

We utilized the **Query2Label (Q2L)** model for multi-label image classification, which is inspired by Facebook Research's DETR model. Q2L shares a similar architecture with DETR, using a CNN backbone and a transformer, but differs in key aspects. Instead of predicting bounding boxes, it directly associates features with label embeddings, allowing for efficient multi-label classification.

The Q2L architecture is shown below:

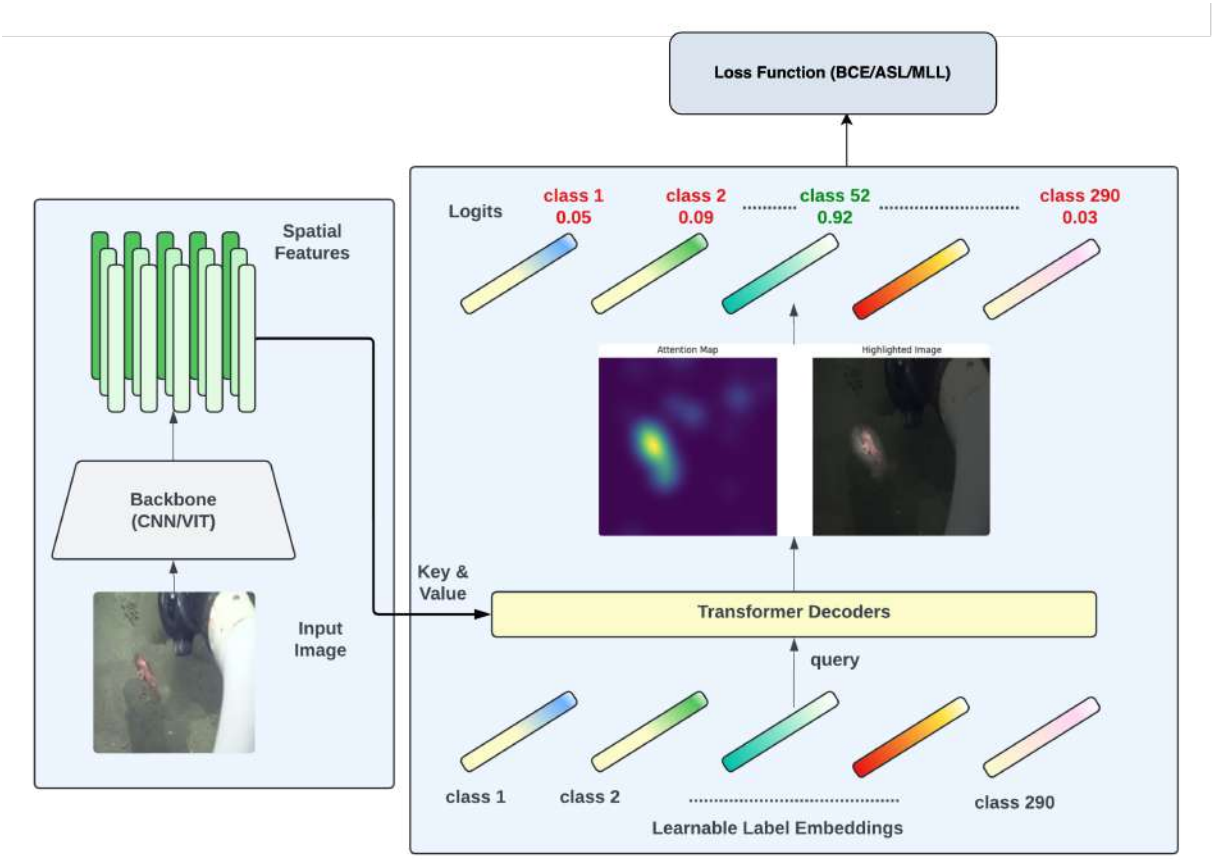


Figure 4.7: Query2Label (Q2L) model framework. The model extracts spatial features from images, processes them through a transformer, and generates attention maps that pool relevant features for label prediction. Source: [16]

Key stages of the Q2L model include:

1. **Backbone Processing**: Images are processed using a pre-trained CNN (e.g., ResNet).
2. **Feature Projection**: Feature maps are reduced via a linear layer.
3. **Position Encoding**: Encodings are added for transformer input compatibility.
4. **Transformer Integration**: A transformer processes the features, using learnable label embeddings to query the feature map.
5. **Classification Head**: The transformer's output is passed through a sigmoid activation to predict label probabilities.

4.2.1.A Feature Extraction

The model starts by extracting spatial features from input images using a CNN backbone. These features are projected to a lower dimension and reshaped for transformer processing. The final feature map is reshaped into a matrix where each spatial location is represented as a feature vector [16].

4.2.1.B Query Updating

Once spatial features are extracted, label embeddings act as queries. These embeddings pool relevant category-specific features using multi-layer transformer decoders. The transformer layers update the embeddings through a series of self-attention and cross-attention operations, enriching the label embeddings with contextual information from the spatial features [16].

- **Self-Attention**: Embeddings interact with themselves to capture dependencies between labels.
- **Cross-Attention**: Embeddings query spatial features to extract relevant information.
- **Feed-Forward Network (FFN)**: Queries are further processed to refine their representations.

Unlike class-agnostic queries in DETR, Q2L's label embeddings are class-specific, allowing for better interpretability and accuracy in multi-label classification.

4.2.1.C Feature Projection

After processing by the transformer, the queried feature vectors are obtained. For each label, a binary classification task is performed. The final feature vectors are projected to logits and passed through a sigmoid function to output the probability for each label [16].

4.2.1.D Loss Function

The model uses an Asymmetric Loss (ASL) function to address sample imbalance. ASL is a variation of focal loss, with separate adjustments for positive and negative samples. This helps in better handling class imbalance, which is common in multi-label classification tasks.

The loss function for each sample is calculated as:

$$L = \frac{1}{K} \sum_{k=1}^{K} \begin{cases} (1 - p_k)^{\gamma^+} \log(p_k), & y_k = 1 \\ p_k^{\gamma^-} \log(1 - p_k), & y_k = 0 \end{cases}$$

Where y_k represents the true label and p_k is the predicted probability. The loss is optimized using stochastic gradient descent, with default parameters $\gamma^+ = 0$ and $\gamma^- = 1$ [16, 102].

4.2.2 Object Detection

In our study, we employed several state-of-the-art object detection models to evaluate their performance and determine the most suitable model for our application. The models tested included DETR, Co-DETR, Deformable DETR, RTDETR, YOLOv8, YOLOv9, and YOLOv5. After extensive experimentation and analysis, we found that YOLOv9 provided the best results in terms of accuracy, efficiency, and robustness. Therefore, we chose YOLOv9 as our primary object detection model. Below, we provide a detailed overview of the YOLOv9 architecture and its components.

4.2.2.A YOLOv9 Architecture

YOLOv9, or "You Only Look Once version 9," builds upon previous iterations of the YOLO series by incorporating novel architectural improvements aimed at enhancing both accuracy and efficiency. The architecture of YOLOv9 is designed to address common issues in deep learning models such as information bottlenecks and the effective utilization of gradient information.

4.2.2.B Generalized Efficient Layer Aggregation Network (GELAN)

The core of YOLOv9's architecture is the Generalized Efficient Layer Aggregation Network (GELAN). GELAN is a lightweight network architecture based on gradient path planning, which facilitates the efficient aggregation of features across multiple layers. This design leverages conventional convolution operators to achieve superior parameter utilization compared to state-of-the-art methods based on depth-wise convolution.

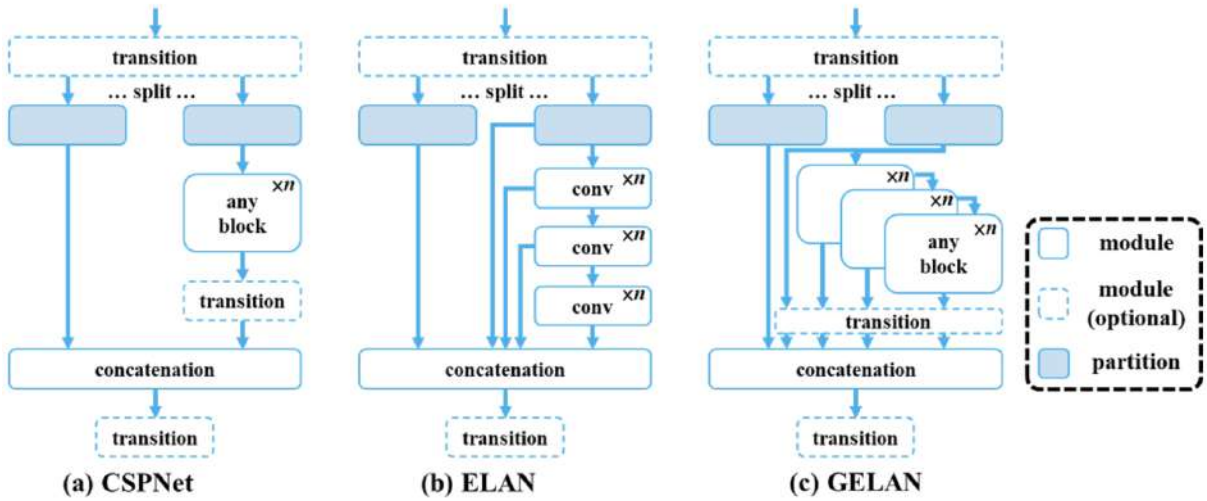


Figure 4.8: The architecture of GELAN: (a) CSPNet, (b) ELAN, and (c) proposed GELAN. GELAN extends ELAN to support any computational blocks. Source: [17]

4.2.2.C Programmable Gradient Information (PGI)

YOLOv9 introduces the concept of Programmable Gradient Information (PGI), which aims to address the problem of information loss during the feedforward process in deep networks. PGI generates reliable gradients through an auxiliary reversible branch, ensuring that deep features maintain key characteristics necessary for the target task. This mechanism allows for better parameter updates and more accurate predictions.

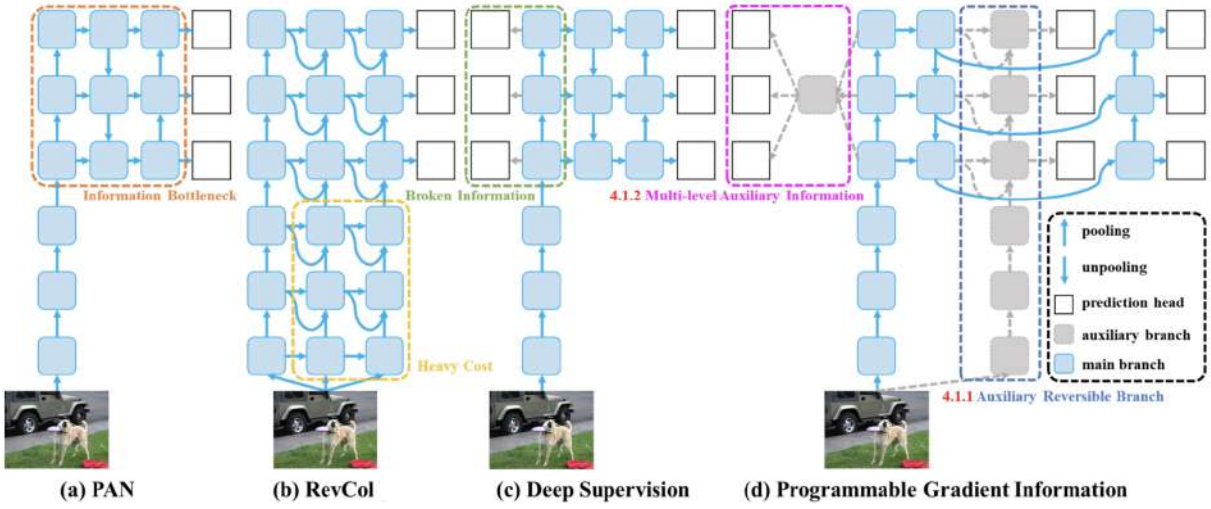


Figure 4.9: PGI and related network architectures: (a) Path Aggregation Network (PAN), (b) Reversible Columns (RevCol), (c) conventional deep supervision, and (d) proposed PGI. PGI comprises three components: main branch, auxiliary reversible branch, and multi-level auxiliary information. Source: [17]

4.2.2.D Network Components

1. **Main Branch:** The main branch is used for inference and incorporates GELAN for feature extraction and aggregation. It does not incur additional inference costs.
2. **Auxiliary Reversible Branch:** This branch generates reliable gradients by maintaining complete information through reversible transformations, which are then used to update the main branch.
3. **Multi-Level Auxiliary Information:** This component aggregates gradient information across multiple levels, ensuring that the main branch retains comprehensive information for accurate prediction.

4.3 Out-of-Distribution (OOD) Score Calculation Methods

4.3.1 Method 1: Maximum Softmax Probability (MSP)

Maximum Softmax Probability (MSP) is a widely used method for out-of-distribution (OOD) detection on classification tasks. For multi-label classification, where a single input may belong to multiple classes, MSP provides a means to assess whether the sample is likely to belong to a known distribution for those classes. The key intuition is that in-distribution samples will yield a model output with high confidence (high softmax probability) for at least one of the classes, whereas low maximum probabilities indicate the sample is more likely to be out-of-distribution.

Mathematically, the OOD score is defined as:

$$\text{OOD} = \begin{cases} 1.0 & \text{if } |\hat{C}| = 0 \\ 1 - \max_{c \in C} P(c|x) & \text{otherwise} \end{cases} \tag{4.15}$$

Where:

- $\hat{C}$ represents the set of predicted classes for the input x, based on a threshold applied to the softmax probabilities.
- C is the set of all possible classes.
- $P(c|x)$ is the softmax probability for class c given input x.

In the event that the predicted class set $\hat{C}$is empty (i.e., no classes meet the threshold of probability), the OOD score is assigned a value of 1.0, indicating a very high level of uncertainty. On the other hand, if the predicted class is not empty, the OOD score is calculated as 1 minus the maximum softmax probability. In other words, the OOD score indicates the model's confidence in its predicted most-likely class. This OOD scoring methodology can be applied to both multi-label classification tasks and out-of-distribution detection tasks, as it provides a simple metric to quantify uncertainty.

4.3.2 Method 2: Average Confidence Score

The Average Confidence Score method evaluates the OOD score by averaging the confidence scores of the predictions. This method provides a measure of how confident the model is about its predictions. The OOD score is computed as:

$$OOD = 1 - \frac{1}{N} \sum_{n=1}^{N} \text{conf}_n \tag{4.16}$$

Here, conf_n represents the confidence score for the n-th prediction, and N is the total number of predictions. The OOD score is one minus the average confidence score, with lower confidence indicating higher likelihood of being out-of-distribution.

5

Results & Discussions

Contents

5.1 Quantitative Evaluation

In this section, we are interested in evaluating the performance of object detection and multi-label classification models on the complex underwater imagery of FathomNet – specifically, we focus on detecting marine species and assessing the models' ability to detect out-of-distribution (OOD) samples. Furthermore, we assess how effective various data augmentation techniques, particularly our proposed technique, Depth Jitter, are in improving model robustness to underwater color and lighting variations with depth. In these experiments, we will evaluate the success of these various approaches on model performance and generalizability to underwater object detection and classification.

To address these questions, we will conduct experiments leveraging the challenge framework from the FathomNet competition 2023 (hosted by the Monterey Bay Aquarium Research Institute (MBARI)) which was held on Kaggle platform from March to May,2023. This framework involves model evaluation on two primary evaluation metrics, category prediction via mean Average Precision (mAP) and OOD via the Area Under the Receiver Operating Characteristic Curve (AUC). We will investigate several types of augmentation techniques and train different object detection and classification models to evaluate performance on these metrics, which will be reported in the subsequent sections.

5.1.1 Evaluation Metrics

In this evaluation, we assess the performance of our object detection and multi-label classification models against the challenging and highly complex underwater data found in the FathomNet dataset. Our aim is two-fold, to not only predict the species of marine life and identify out of distribution (OOD) samples, but also to evaluate how more advanced data augmentation can improve model robustness, specifically the Depth Jitter augmentation method introduced in this study.

In order to complete this evaluation process, we adopt two types of evaluation metrics: mean Average Precision (mAP) for predicting species and Area Under the Receiver Operating Characteristic Curve (AUC) for detecting OOD samples. Both metrics allow us to analyze the ability of deep learning models to classify marine life, and detect OOD samples, given the variability from underwater environmental conditions. The following subsections offer a detailed explanation of these metrics and how they contribute to our analysis of model performance.

5.1.2 Out-of-Distribution Detection

To assess the performance of the models for out-of-distribution detection, we use the Receiver Operating Characteristic ROC curve as it is a common metrics for binary classification tasks. In this topic, it will help us to understand how well or how poorly models ended up separating in-distribution images, denoted as ID images, and out-of-distribution images, denoted as OOD images. As mentioned, the ROC curve has

the True Positive Rate or TPR, which signifies how many OOD samples the model correctly highlighted and the False Positive Rate or FPR, which signifies how many in-distribution ID samples the model mislabeled as being OOD, plotted in relation to multiple classifications thresholds.
As this evaluation metrics is part of the Fathomnet 2023 challenge and there were no labels for the in-distribution and out-of-distribution images for the dataset separately, we could not show the ROC curve explicitly. We had to depend on the final score that we got after submitting the results on the kaggle.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

As the threshold of classification becomes more lenient, both TPR and FPR will increase towards one, meaning that more samples will be classified as OOD, including some that are false positives. The area under the ROC curve (AUC) is an indication of the model's ability to differentiate between in-distribution and OOD samples. AUC of 1.0 is a perfect classifier while an AUC of 0.5 indicates a high level of random guessing.
The AUC is primarily used for OOD detection tasks, where the goal is to identify whether an image belongs to the known training distribution or not.

5.1.3 Category Predictions

For the evaluation of multi-label category predictions, we utilize the mean Average Precision at 20 (mAP@20). This metric assesses how well a model predicts the presence of multiple categories in an image and ranks it accordingly. Mean Average Precision (mAP) is chosen in this case because it takes into account precision (how accurate the predicted categories are) and whether the categories are ranked correctly.

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

where:

- U is the number of images.
- $P(k)$ is the precision at cutoff k.
- n is the number of predictions per image.
- rel(k) is an indicator function that equals 1 if the item at rank k is a relevant (correct) label and 0 otherwise.

This score tells us how well the model ranks the relevant categories among the top 20 predictions for each image.

5.1.4 Final Score

A unified score is computed to rank the models based on their performance with respect to the two main tasks, namely category prediction (mAP@20) and OOD detection (AUC). The motivation behind combining these two tasks is due to the challenge evaluation framework: a unified ranking system is required to conduct evaluations of all of the models. This allows participants to focus on both the accuracy of category predictions and the ability to detect out-of-sample images. The combined score is calculated as follows:

$$\text{Final Score} = \frac{1}{2}(\text{sAUC} + \text{mAP@20})$$

Here, the sAUC is the rescaled AUC score, defined as:

$$\text{sAUC} = 2 \times \text{AUC} - 1$$

To summarize, the AUC is normalized to reach consistency with the second metric (mAP@20), after which we average the rescaled AUC and mAP@20 together to create the final score. This guarantees that both OOD detection as well as category prediction are weighted equally in the final ranking. The overall rank of a model in the challenge is driven by the combined score, which constitutes a single, global metric for performance.

5.1.5 Performance of Object Detection Models

This experiment is aimed to assess the performance of existing object detection (OD) models on a challenging dataset made specifically for underwater imagery, the FathomNet dataset. Specifically, we assess which existing object detection model performs the best in predicting marine species and context of transparent, underwater environments. By comparing and contrasting various models of object detection with distinct parameter configurations, we hope to contribute a benchmark to models of this dataset, where none currently exists. The goal is to determine which model is most suited for accurately detecting objects and categories in the complex underwater environment of FathomNet.

We trained several object detection models on the FathomNet competition dataset. The dataset contains images of different sizes and while we could have run all of the different size images from the original dataset, we decided to resize all images to the same 640x640 pixels size during training using PyTorch while maintaining the aspect ratio of all images. Resizing the images standardized the input size for all models while also allowing for a direct comparison across different architectures.As there is

no predefined benchmark available for this specific task, we established one by testing several object detection models with different configurations. We created a benchmark by testing multiple object detection models, with different structure settings. The Baseline model reported in the FathomNet dataset paper [13] is based on a YOLOv8 model and we have used it for reference when comparing performance. In Table 5.1, we see that many of these models exceed the baseline result. Of the models tested, the YOLOv9 [17] model performs best overall with an mAP@20 score of 0.74 on the validation set, the primary evaluation metric for category predictions.

Because this task was part of a Kaggle competition, the evaluation set does not provide any ground truth to evaluate during development, and as such, it is impossible to consider the individual evaluation metrics for the test set. The only information that will be reported is a final score that combines the results of several evaluation metrics to summarize model performance. The final score will be discussed in more detail later on in this chapter. For every model, we have incorporated early stopping with patience set

Models	Backbone	Epochs	Train Image Size	Test Image Size	mAP@50	mAP@20	Training Time	# of GPUs
Deformable DETR	Resnet50	70	640X640	640X640	0.196	0.40	24 hours	2 X RTX3090
DETR	Resnet50	150	640X640	640X640	0.209	0.42	36 hours	1 X RTX3090
Conditional DETR	Resnet50	91	640X640	640X640	0.259	0.50	24 hours	4 X A40-48
Co-Dino (Photometric Distortion)	Resnet50	50	640X640	640X640	0.327	0.60	58 hours	1 X A100-80
YOLOv8m	cspDarknet	100	640X640	640X640	0.300	0.62	19 hours	1 X A40-48
Yolov8m(BaseLine)	cspDarknet	50	640X640	640X640	0.330	0.69	-	-
RT-DETR	HGnetV2	94	640X640	640X640	0.372	0.69	07 hours	4 X RTX3090
Co-Dino (Photometric Distortion)	SWINL	49	640X640	640X640	0.337	0.70	72 hours	1 X A100-80
RT-DETR (Tuned)	HGnetV2	51	640X640	640X640	0.358	0.72	06 hours	1 X A40-48
YOLOv9-e	Gelan	120	640X640	640X640	0.391	**0.74**	22 hours	1 X A100-80

Table 5.1: Performance of Different Object Detection Models on the FathomNet Dataset.

to 30 epochs, which signifies that if the model performance worsens through 30 epochs, we effectively stop the training early to prevent overfitting to the training data. Each model was trained for 150 epochs maximum, though training could terminate earlier due to early stopping. Throughout training, the best preforming checkpoint was regularly saved.

In order to assess the best checkpoint, we utilized the performance of the model on a validation subset, which is programmed into Ultralytics while training and validating YOLO models. This Validation set was derived from the training set, and the model's performance was monitored within the validation set to determine the best checkpoint. The size and selection of this validation set is conducted internally by the Ultralytics framework to determine the selected model checkpoint based on best validation performance (accuracy) or mAP (mean Average Precision).

For every model we use the default augmentation settings of Ultralytics-YOLO. The learning rate is set to 1×10^{-4} and the SGD optimizer is used as the optimizer. As there is noise and imbalance in the annotations we use label smoothing of 0.1 to regularize. Depending on the availability of the GPU we changed the batch size from 16 to 128. We also use multiple GPU and distributed training for some of the trainings. Those trainings take less time to train as the batch size is higher and parallel computation is higher than one single GPU. Further, we should consider that not all batch sizes produce the same

performance for training a model as the updates to the gradient and convergence can be associated with batch sizes. Further work should address the impact of varying batch size with individual models.

5.1.6 Performance of Query2Label Model in Different Augmentation Settings

The purpose of this experiment is to quantitatively assess various augmentation techniques' utility on the Query2Label model for multi-label classification. In particular, we want to evaluate how augmentation methods -including our method of Depth Jitter- affect the model's generalization capability and accuracy to classify marine species in the depths of the FathomNet dataset.

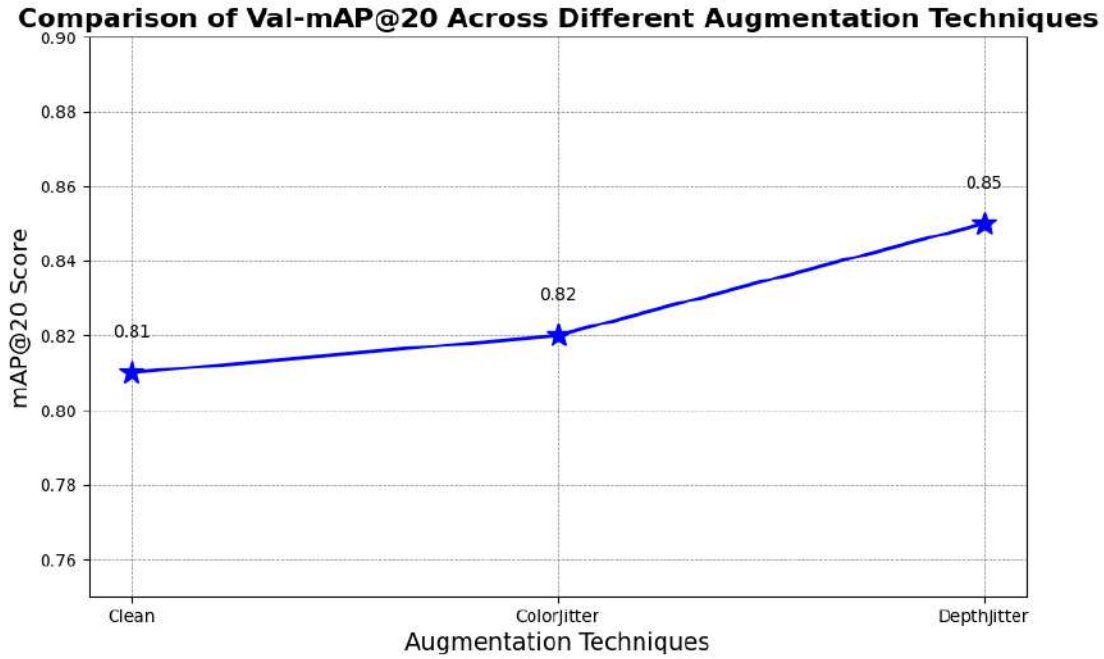


Figure 5.1: Comparison of Val-mAP@20 Scores Across Different Augmentation Techniques: This graph illustrates the validation mAP@20 scores for three augmentation techniques: Clean, ColorJitter, and DepthJitter. The mAP@20 score, which measures the model's average precision at an intersection over union threshold of 20%, is displayed on the y-axis. The x-axis lists the augmentation techniques. The Clean technique shows a baseline score of 0.81, while ColorJitter slightly improves the score to 0.82. DepthJitter achieves the highest score of 0.85, indicating its superior performance in enhancing the model's accuracy on the validation set.

We trained the Query2Label image classification model for category classification alongside object detection models. Although classifying objects and multi-label classification are traditionally viewed as two separate tasks, in this instance the procedures for both of these modes can be employed to predicting the categories of marine species in the datasets that are part of FathomNet competition, which was the reason unified labels were provided for both object detection and multi-label classification. This

creates the opportunity to compare the ability of both image classification and object detection models to predict the correct categories for the images reported.Query2Label is one of the top performing image classification models for multi-label classification and produced improved outcomes than the object detection models that were evaluated as part of this study for dataset completion. Therefore we considered Query2Label to provide categories in images within the multi-label classification task, because this was seems like a more reliable approach than object detection.

Similar to object detection models, we utilize different settings with a variety of backbones such as ResNest and TResnet from the Hugging Face timm library, and attempted several augmentation protocols. The training image was $384 \times 384 pixels$, and the test was $640 \times 640 pixels$. The use of different images during training and testing is simply because the larger images increases model performance through better capturing fine detail, and in particular, the models we evaluated ResNest and TResnet are designed to take advantage of size difference. Consideration of different image sizes is advised in the design of those types of architectures and have been widely observed to produce more favorable outcomes.

To produce meaningful and robust outcomes, we performed the evaluation for each configuration five times with different random seeds and take the mean of the results to reduce random variability during training, thereby considering model performance more consistent and fair.

The table 5.2 illustrates the performance of the Query2Label classification model on the FathomNet competition dataset. Our results demonstrate that the proposed augmentation technique achieves the highest mAP@20 on the validation set, outperforming both the PyTorch ColorJitter and the without augmentation(Clean) techniques. This indicates that our method effectively enhances the model's ability to generalize and accurately classify marine species under diverse underwater conditions.

For training, we initially selected a learning rate of 0.0001 and a weight decay of 0.05. These values were chosen after performing a learning rate range test, which involved gradually increasing the learning rate to observe how the loss evolved during training. The goal was to identify a rate that allowed for rapid convergence without causing instability or overshooting in the training process.

We observed that a learning rate of 0.0001 worked well across most settings, providing a good balance between fast convergence and stable training. However, for some deeper backbones, such as ResNest and TResNet, the training and validation loss showed signs of instability, particularly during later stages of training. This indicated that the learning rate might be too high for these complex architectures, leading to the gradient updates being too aggressive.

In these cases, we reduced the learning rate to 5×10^{-5} to mitigate the issue. This allowed the model to make smaller, more controlled updates to the weights, stabilizing the training process. The adjustment was particularly necessary for models with a large number of parameters, as they tend to be more sensitive to higher learning rates.

For different augmentation techniques, such as ColorJitter and DepthJitter, the learning rate remained consistent across most experiments, as we did not observe significant differences in how these augmentations influenced training stability. The main variations in learning rate adjustments were driven by the complexity of the backbone models rather than the augmentation methods themselves.

Augmentation	backbone_desc	Train Image Size	Test Image Size	loss	train_loss	val_loss	val_mAP	val_mAP@20
Clean	ResNest101e	384X384	640X640	ASL	0.095	0.234	0.751	0.813
Color Jitter	ResNest101e	384X384	640X640	ASL	0.187	0.227	0.762	0.827
Depth Jitter(Ours)	ResNest101e	384X384	640X640	ASL	0.165	0.223	**0.803**	**0.855**

Table 5.2: Performance comparison of different multi-label classification models on the FathomNet dataset using various augmentation techniques. The table evaluates three augmentation strategies: Clean (no augmentation), Color Jitter, and the proposed Depth Jitter method. All models use the ResNest101e backbone for feature extraction. The models are trained on 384x384 image size and evaluated on 640x640 image size. The loss function used is Asymmetric Loss (ASL). The table provides training loss (train loss), validation loss (val loss), mean Average Precision (val mAP), and mAP@20 for each configuration. Depth Jitter outperforms both the Clean and Color Jitter augmentations, achieving the highest validation mAP (0.803) and mAP@20 (0.855), demonstrating its effectiveness in improving model performance on the Fathomnet competition dataset.

For multi-label classification, we used Asymmetric Loss (ASL), which is well-suited for imbalanced data by focusing more on misclassified positive labels and reducing the focus on easily classified negatives. The loss is computed over each batch during training. For each image, ASL calculates the loss by comparing the predicted probabilities for all categories with the ground truth labels. The batch loss is the average of the individual image losses and is used for backpropagation to update the model's weights. Validation loss is similarly computed over the validation set at the end of each epoch to monitor generalization. This approach allows the model to effectively handle multiple labels per image while addressing label imbalance.

Our proposed augmentation technique significantly enhances the performance of the Query2Label model on the FathomNet dataset, improving generalization and classification accuracy. These results also positively impact the out-of-distribution (OOD) score, highlighting the model's robustness in identifying data that deviates from the training distribution—critical for real-world applications. Key factors such as appropriate learning rate adjustments, gradient clipping, and advanced strategies like the one-cycle learning rate scheduler and AdamW optimizer further contribute to the model's effectiveness, boosting both validation mAP@20 and OOD detection.

5.1.7 OOD Score Performance

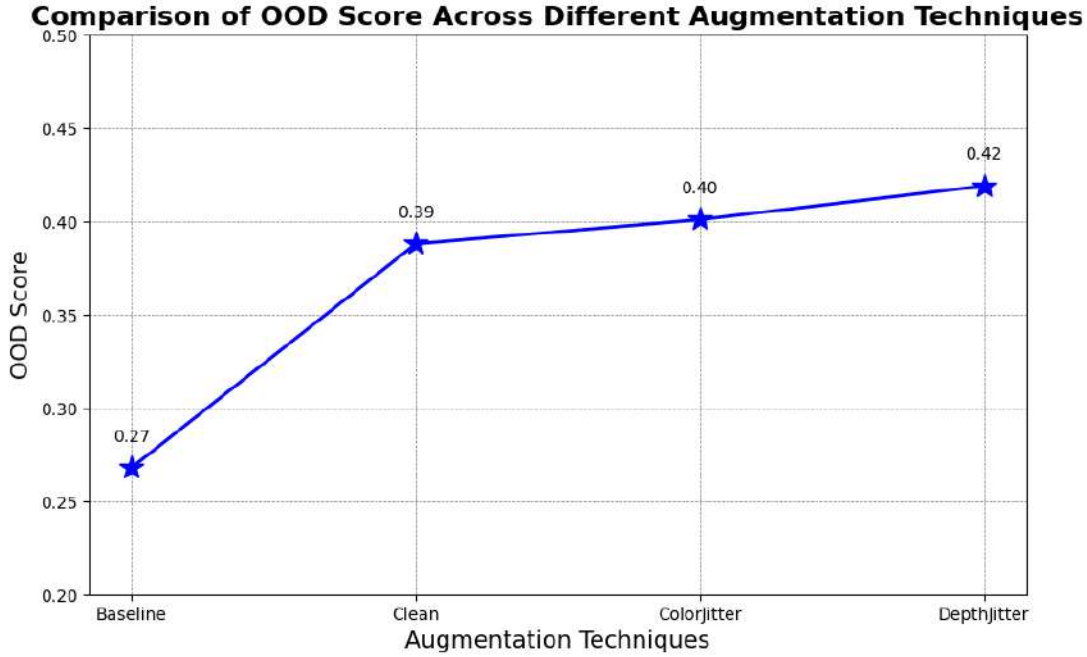


Figure 5.2: Comparison of Out-of-Distribution (OOD) Scores Across Different Augmentation Techniques: This graph depicts the OOD scores for four augmentation techniques: Baseline, Clean, ColorJitter, and DepthJitter. The OOD score, representing the model's ability to handle out-of-distribution data, is plotted on the y-axis. The x-axis lists the augmentation techniques. The Baseline technique shows the lowest OOD score at 0.27, while Clean improves to 0.39. ColorJitter achieves an OOD score of 0.40, and DepthJitter has the highest score at 0.42. These results indicate that DepthJitter is the most effective technique for enhancing the model's robustness to out-of-distribution data.

Figure 5.2 shows the comparison of the OOD score across different augmentation techniques on the evaluation set. The ood score is the average of $sAUC$ and $mAP@20$ which is described in the evaluation metrics section above. The baseline [13] used YOLOv8 for classification and the average confidence score technique for calculating ood score. The other techniques in the above figure use the Query2label [16] for classification and maximum softmax probability technique for ood score calculation. Both of these ood score calculation techniques have been discussed in the Capítulo 4. We see that our proposed technique used for classification scored highest in the ood score calculation than other augmentation techniques. Our proposed technique performs more than 1.5 times than the baseline.

5.1.8 Kaggle Competition Performance

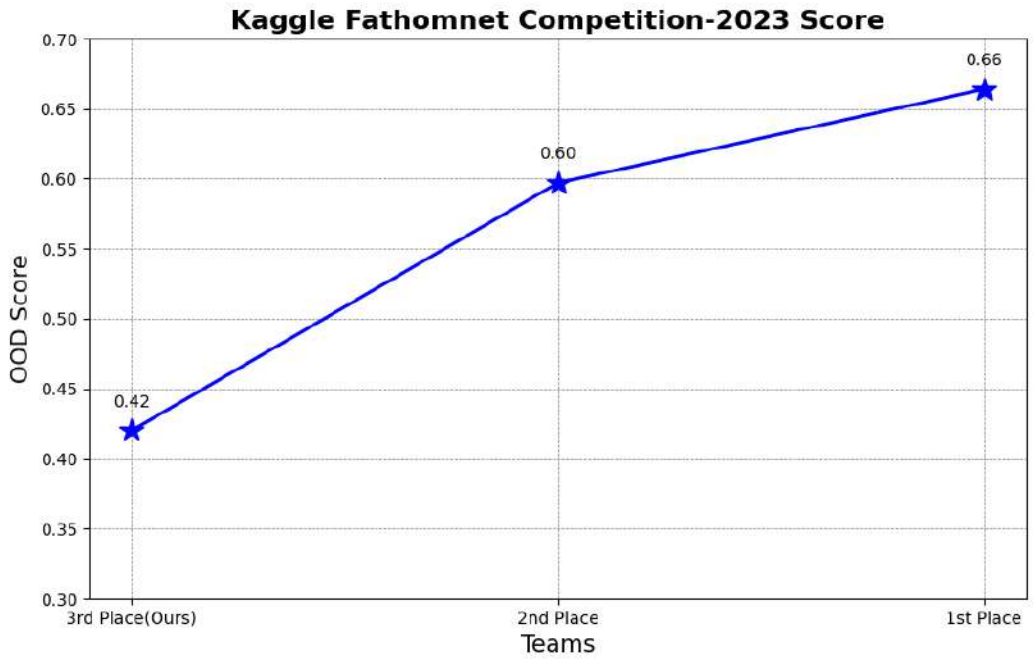


Figure 5.3: Comparison of OOD Scores for Top Teams in Kaggle Fathomnet Competition-2023: This graph presents the out-of-distribution (OOD) scores for the top three teams in the Kaggle Fathomnet Competition-2023. The OOD score, displayed on the y-axis, is a measure of the model's ability to identify out-of-distribution samples. The x-axis lists the teams by their ranking: our team in 3rd place with a score of 0.42, the 2nd place team with a score of 0.60, and the 1st place team with the highest score of 0.66. The plot highlights the progressive improvement in OOD scores from 3rd to 1st place.

This challenge was part of a Kaggle competition that took place in 2023. If we had participated with our proposed method, we would have placed 3rd in the competition. The graph in the figure 5.3 shows a significant gap in scores between the top two teams and our team. One possible reason for this disparity is that the top teams likely used additional images from external sources [103] to make their datasets more balanced, resulting in higher scores. In contrast, we worked exclusively with the provided dataset without incorporating external data.

5.2 Qualitative Evaluation

5.2.1 Object Detection(Visual Inspection of Predictions)

Figure 5.4: (a) The ground labels for object detection. (b) The predicted labels by yolov9 object detection model.

Figure 5.4 presents a comparison between the ground truth labels and the predicted labels of the YOLOv9 object detection model on the Fathomnet competition dataset. The majority of the predictions made by YOLOv9 are accurate, demonstrating the model's strong performance. However, there are some edge cases where the model fails to detect objects. These missed detections may be attributed to the insufficient representation of certain examples in the training set. Overall, the YOLOv9 model performs satisfactorily, but further improvements could be achieved with a more balanced and comprehensive training dataset. The quality of the bounding boxes is also quite good and accurate.

5.2.2 Multilabel Classification(Attention Map Visualization)

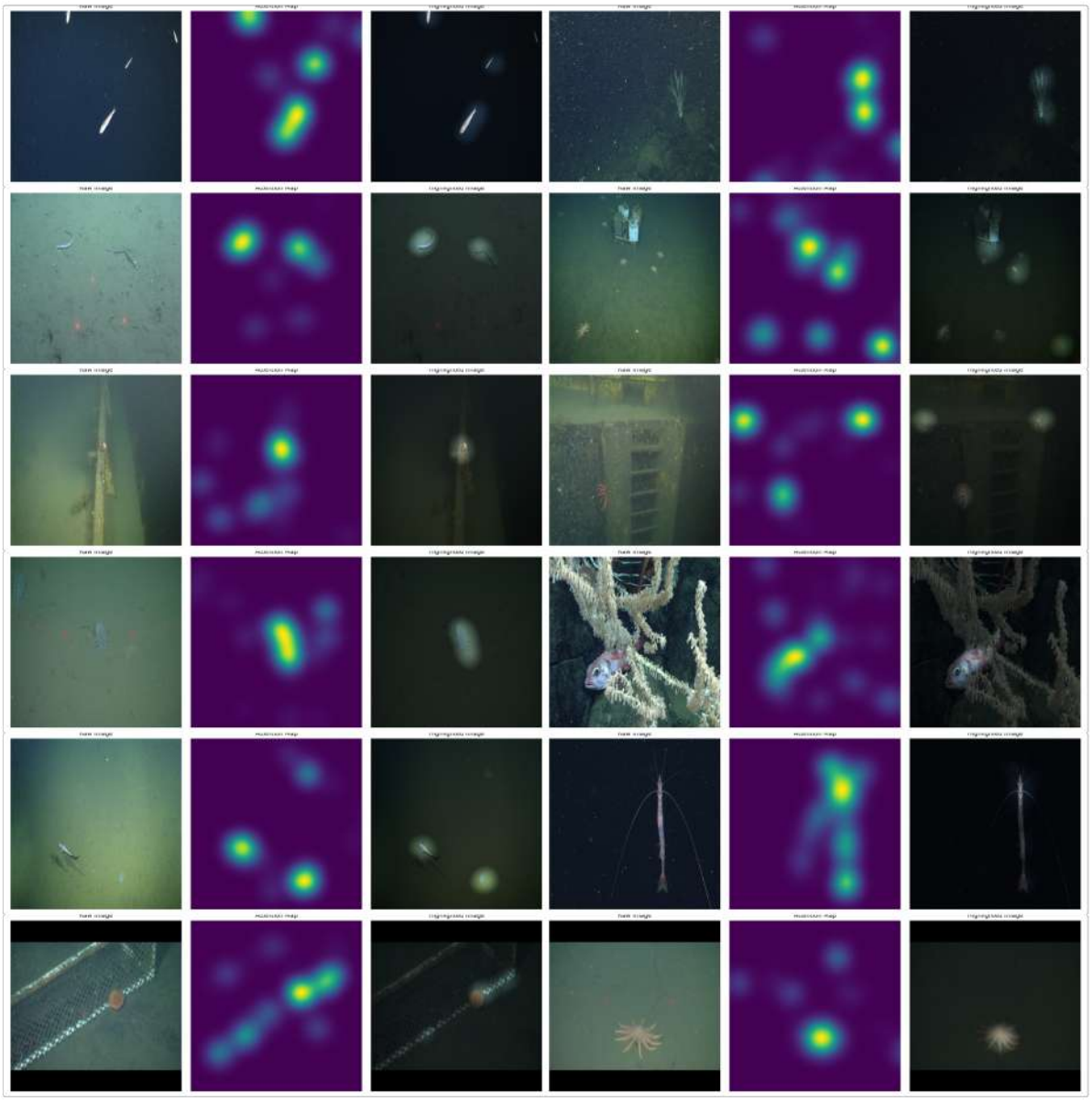


Figure 5.5: Attention maps generated by Query2label [16].

The attention map visualization in Figure 5.5 illustrates how the model focuses on different regions of the images when making predictions. Each set contains 3 images. The left side of each set shows the original image, the middle image depicts the corresponding attention map generated by the model and the right image shows the highlighted areas on the original image. These attention maps highlight the areas that the model considers most relevant for identifying and classifying the objects within the images.

The visualization demonstrates the model's ability to pinpoint key features and regions associated with the objects of interest. In most cases, the highlighted areas correspond well with the actual locations of the objects, indicating that the model is effectively learning to focus on significant parts of the images. However, there are instances where the attention is dispersed or not concentrated on the target objects, which may suggest areas for improvement.
Overall, this attention map visualization provides valuable insights into the model's interpretability and decision-making process, showing that the model generally performs well in identifying relevant regions. Enhancing the model's training data and refining its architecture could further improve its accuracy and focus in challenging scenarios.

5.3 Limitations of the System

Despite the promising performance of our proposed method, there are quite a few limitations that needs to be addresses:

5.3.1 Technical Limitations

The system's computational requirements are high, especially during the training phases. Models like Query2Label and YOLOv9 demand high-performance hardware for training, which may not always be available in all deployment environments. For instance, training Query2Label on a single NVIDIA A100 GPU (80 GB VRAM) took approximately 10 hours, while YOLOv9 required around 24 hours on the same hardware. Both models required significant VRAM, with Query2Label using up to 40 GB and YOLOv9 utilizing up to 48 GB. The fine-tuning process also requires substantial computational resources, which adds to the overall time and hardware demands.

5.3.2 Data-related Limitations

Although extensive, the Fathomnet dataset has a few drawbacks. The dataset has a long-tailed distribution that is imbalanced, with relatively few occurrences in many classes. This imbalance may cause the system to perform poorly on rare classes but well on often occurring classes, resulting in biased model performance. Furthermore, the model's capacity to generalize to new contexts can be constrained by the dataset's lack of diversity in terms of underwater conditions and geographic locations.

5.3.3 Environmental Constraints

The underwater environment presents unique challenges, such as varying light conditions, water turbidity, and occlusions by marine flora and fauna. These factors can significantly affect image quality

and, consequently, the system's performance. The current model may struggle in conditions that are markedly different from those seen during training. Our proposed augmentation technique is highly dependent on the lighting condition. If the lightning condition of the image is not ideal the veiling light and the backscatter coefficient can be wrongly estimated and default to the lower and upper bounds set by the user. This leads to Depth Jitter not affecting the image appearence. One solution can be getting the parameters of the underwater image formation model again by adjusting the lower and upper bounds which can improve the augmentation technique performance.

5.3.4 Interpretability and Usability

Although the attention map visualizations provide some level of interpretability, understanding the model's decision-making process remains challenging. This black-box nature of deep learning models can be a barrier for end-users who need to trust and understand the system's outputs. Additionally, the usability of the system in practical applications needs further validation, particularly in terms of its integration into existing workflows and its user-friendliness for marine biologists and other stakeholders.

By acknowledging these limitations, we aim to provide a balanced view of the system's capabilities and areas for future improvement. Addressing these issues in subsequent research will be crucial for enhancing the robustness, reliability, and applicability of the system in real-world underwater exploration and monitoring tasks.

6

Conclusion

Contents

6.1 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.

6.2 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.

Bibliography

[1] 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

[2] 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

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," 2014, vol. 8691, pp. 346–361, arXiv:1406.4729 [cs]. [Online]. Available: http://arxiv.org/abs/1406.4729

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

[5] 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

[6] 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

[7] 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

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

[9] 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

[10] 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

[11] B. Gao and H. Zhou, “Multi-label image recognition with multi-class attentional regions,” *CoRR*, vol. abs/2007.01755, 2020. [Online]. Available: https://arxiv.org/abs/2007.01755

[12] J. Zhao, K. Yan, Y. Zhao, X. Guo, F. Huang, and J. Li, “Transformer-based dual relation graph for multi-label image recognition,” *CoRR*, vol. abs/2110.04722, 2021. [Online]. Available: https://arxiv.org/abs/2110.04722

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

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

[15] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, “Depth anything: Unleashing the power of large-scale unlabeled data,” 2024.

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

[17] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, “YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information,” Feb. 2024, arXiv:2402.13616 [cs]. [Online]. Available: http://arxiv.org/abs/2402.13616

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

[26] 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

[27] 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/

[28] 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

[29] 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

[30] 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/

[31] 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/

[32] 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/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[33] 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

[34] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” *CoRR*, vol. abs/1311.2901, 2013. [Online]. Available: http://arxiv.org/abs/1311.2901

[35] J. Dai, Y. Li, K. He, and J. Sun, “R-FCN: object detection via region-based fully convolutional networks,” *CoRR*, vol. abs/1605.06409, 2016. [Online]. Available: http://arxiv.org/abs/1605.06409

[36] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, “Light-head R-CNN: in defense of two-stage object detector,” *CoRR*, vol. abs/1711.07264, 2017. [Online]. Available: http://arxiv.org/abs/1711.07264

[37] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, “Feature pyramid networks for object detection,” *CoRR*, vol. abs/1612.03144, 2016. [Online]. Available: http://arxiv.org/abs/1612.03144

[38] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” 2022.

[39] 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

[40] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” 2018.

[41] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” 2021.

[42] 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

[43] G. Tsoumakas and I. Katakis, “Multi-Label Classification: An Overview,” *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, Jul. 2007. [Online]. Available: https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/jdwm.2007070101

[44] R. E. Schapire and Y. Singer, “[No title found],” *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000. [Online]. Available: http://link.springer.com/10.1023/A:1007649029923

[45] 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

[46] 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.

[47] D. Yan′e, 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

[48] 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.

[49] H. Yu, “Research progresson object detection and tracking techniques utilization in aquaculture: a review,” *Journal of Dalian Ocean University*, vol. 35, no. 6, pp. 793–804, 2020.

[50] 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.

[51] “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.”

[52] “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.”

[53] “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.”

[54] 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

[55] 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

[56] “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.”

[57] “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.”

[58] “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.”

[59] “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.”

[60] 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

[61] “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."

[62] 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

[63] "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."

[64] 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

[65] "Bochkovskiy A, Wang CY, Liao HYM (2020) YOLOv4: optimal speed and accuracy of object detection. Preprint at arXiv: 2004.10934."

[66] X. L. Hu, Y. Liu, Z. X. Zhao, J. T. Liu, X. T. Yang, and C. H. Sun, "Real-time detection of uneaten feed pellets in underwater images for aquaculture using an improved YOLO-V4 network," *Comput Electron Agric*, vol. 185, 2021. [Online]. Available: https://doi.org/10.1016/j.compag.2021.106135

[67] H. L. Ge, Y. W. Dai, Z. Y. Zhu, and R. B. Liu, "A deep learning model applied to optical image target detection and recognition for the identification of underwater biostructures," *Machines*, vol. 10, 2022. [Online]. Available: https://doi.org/10.3390/machines10090809

[68] F. Lei, F. F. Tang, and S. H. Li, "Underwater target detection algorithm based on improved YOLOv5," *J Mar Sci Eng*, vol. 10, 2022. [Online]. Available: https://doi.org/10.3390/jmse10030310

[69] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, 2015. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2577031

[70] Z. Chen, Z. Zhang, F. Z. Dai, Y. Bu, and H. B. Wang, "Monocular Vision-Based Underwater Object Detection," *Sensors*, vol. 17, 2017. [Online]. Available: https://doi.org/10.3390/s17081784

[71] "Chen ZY, Zhao TT, Cheng N, Sun XD, Fu XP (2018) Towards underwater object recognition based on supervised learning. In: 2018 OCEANS-MTS/IEEE Kobe Techno-Oceans, Kobe, pp 1–4. https://doi.org/10.1109/OCEANSKOBE.2018.8559050."

[72] X. Sun, J. Y. Shi, L. P. Liu, J. Y. Dong, C. Plant, and X. H. Wang, "Transferring deep knowledge for object recognition in Low-quality underwater videos," *Neurocomputing*, vol. 275, 2018. [Online]. Available: https://doi.org/10.1016/j.neucom.2017.09.044

[73] “Yang HH, Xu GH, Yi SZ, Li YQ (2019) A new cooperative deep learning method for underwater acoustic target recognition. In: OCEANS 2019-Marseille, Marseille, pp 1–4. https://doi.org/10.1109/OCEANSE.2019.8867490.”

[74] “Lin WH, Zhong JX, Liu S, Li T, Li G (2020) ROIMIX: proposal-fusion among multiple images for underwater object detection. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, pp 2588–2592. https://doi.org/10.1109/ICASSP40776.2020.9053829.”

[75] “Lau PY, Lai SC (2021) Localizing fish in highly turbid underwater images. In: International Workshop on Advanced Imaging Technology (IWAIT), pp 294–299. https://doi.org/10.1117/12.2590995.”

[76] M. H. Zhang, S. B. Xu, W. Song, Q. He, and Q. M. Wei, “Lightweight underwater object detection based on YOLO v4 and multi-scale attentional feature fusion,” *Remote Sens*, vol. 13, 2021. [Online]. Available: https://doi.org/10.3390/rs13224706

[77] “Song DL, Sun WC, Ji ZH, Hou GJ, Li XF, Liu L (2014) Color model selection for underwater object recognition. In: 2014 International Conference on Information Science, Electronics and Electrical Engineering, Sapporo, pp 1339–1342. https://doi.org/10.1109/InfoSEEE.2014.6947890.”

[78] “Cao X, Zhang XM, Yu Y, Niu LT (2016) Deep learning-based recognition of underwater target. In: 2016 IEEE International Conference on Digital Signal Processing, Beijing, pp 89–93. https://doi.org/10.1109/ICDSP.2016.7868522.”

[79] C. Y. Li, J. C. Guo, R. M. Cong, Y. W. Pang, and B. Wang, “Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior,” *IEEE Trans Image Proc*, vol. 25, 2016. [Online]. Available: https://doi.org/10.1109/TIP.2016.2612882

[80] “Ding XY, Wang YF, Zhang J, Fu XP (2017) Underwater image dehaze using scene depth estimation with adaptive color correction. In: OCEANS 2017-Aberdeen, Aberdeen, pp 1–5. https://doi.org/10.1109/OCEANSE.2017.8084665.”

[81] “Yu XL, Qu YY, Hong M (2019) Underwater-GAN: underwater image restoration via conditional generative adversarial network. In: 24th International Conference on Pattern Recognition (ICPR), Beijing, pp 66–75. https://doi.org/10.1007/978-3-030-05792-3_7.”

[82] “Fan BJ, Chen W, Cong Y, Tian JD (2020) Dual refinement underwater object detection network. In: 16th European Conference on Computer Vision, Glasgow, pp 275–291. https://doi.org/10.1007/978-3-030-58565-5_17.”

[83] X. Y. Wei, L. Yu, S. W. Tian, P. C. Feng, and X. Ning, “Underwater target detection with an attention mechanism and improved scale,” *Multimed Tools Appl*, vol. 80, 2021. [Online]. Available: https://doi.org/10.1007/s11042-021-11230-2

[84] L. Chen, Y. Y. Yang, Z. H. Wang, J. Zhang, S. W. Zhou, and L. H. Wu, “Underwater target detection lightweight algorithm based on multi-scale feature fusion,” *J Mar Sci Eng*, vol. 11, 2023. [Online]. Available: https://doi.org/10.3390/jmse11020320

[85] “Rashwan A, Kalra A, Poupart P (2019) Matrix Nets: a new deep architecture for object detection. In: 2019 IEEE/CVF International Conference on Computer Vision Workshops, Seoul, pp 2025–2028. https://doi.org/10.1109/ICCVW.2019.00252.”

[86] “Chen L, Liu ZH, Tong L, Jiang ZH, Wang SK, Dong JY et al (2020a) Underwater object detection using Invert Multi-Class Adaboost with deep learning. In: 2020 International Joint Conference on Neural Networks, Glasgow, pp 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207506.”

[87] F. L. Han, J. Z. Yao, H. T. Zhu, and C. H. Wang, “Underwater image processing and object detection based on deep CNN method,” *J Sens*, vol. 2020, 2020. [Online]. Available: https://doi.org/10.1155/2020/6707328

[88] R. S. Liu, Z. Y. Jiang, S. Z. Yang, and X. Fan, “Twin adversarial contrastive learning for underwater image enhancement and beyond,” *IEEE Trans Image Proc*, vol. 31, 2022. [Online]. Available: https://doi.org/10.1109/TIP.2022.3190209

[89] M. Zurowietz and T. W. Nattkemper, “Unsupervised knowledge transfer for object detection in marine environmental monitoring and exploration,” *IEEE Access*, vol. 8, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.3014441

[90] L. C. Zeng, B. Sun, and D. Q. Zhu, “Underwater target detection based on Faster R-CNN and adversarial occlusion network,” *Eng Appl Artif Intell*, vol. 100, 2021. [Online]. Available: https://doi.org/10.1016/j.engappai.2021.104190

[91] “Mou L, Zhang XW, Zhang JJ, Shen XH, Xu XL (2017) Saliency detection of underwater target based on spatial probability. In: 2017 International Conference on Computer Systems, Electronics and Control, Dalian, pp 630–632. https://doi.org/10.1109/ICCSEC.2017.8446733.”

[92] X. Y. Zhou, K. D. Yang, and R. Duan, “Deep learning based on striation images for underwater and surface target classification,” *IEEE Signal Proc Lett*, vol. 26, 2019. [Online]. Available: https://doi.org/10.1109/LSP.2019.2919102

[93] Z. Chen, H. M. Gao, Z. Zhang, H. L. Zhou, X. Wang, and Y. Tian, “Underwater salient object detection by combining 2D and 3D visual features,” *Neurocomputing*, vol. 391, 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2018.10.089

[94] S. Q. Duntley, “Light in the Sea*,” *Journal of the Optical Society of America*, vol. 53, no. 2, p. 214, Feb. 1963. [Online]. Available: https://opg.optica.org/abstract.cfm?URI=josa-53-2-214

[95] B. L. McGlamery, “A computer model for underwater camera systems,” in *Other Conferences*, 1980. [Online]. Available: https://api.semanticscholar.org/CorpusID:122739453

[96] 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.

[97] ——, “A revised underwater image formation model - akkaynak treibitz cvpr 2018,” 06 2018.

[98] 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.

[99] 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/

[100] 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

[101] 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/

[102] 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.

[103] T. T. Huu Nguyen, P. Nguyen, V. P. Nguyen, L. H. G. Tran, M. Van Le, and B. T. Nguyen, “Increased query2label (iq) for small fine-grained multi-label classification,” in *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, 2023, pp. 1–6.

Appendix

More visualization of the Depth Jitter

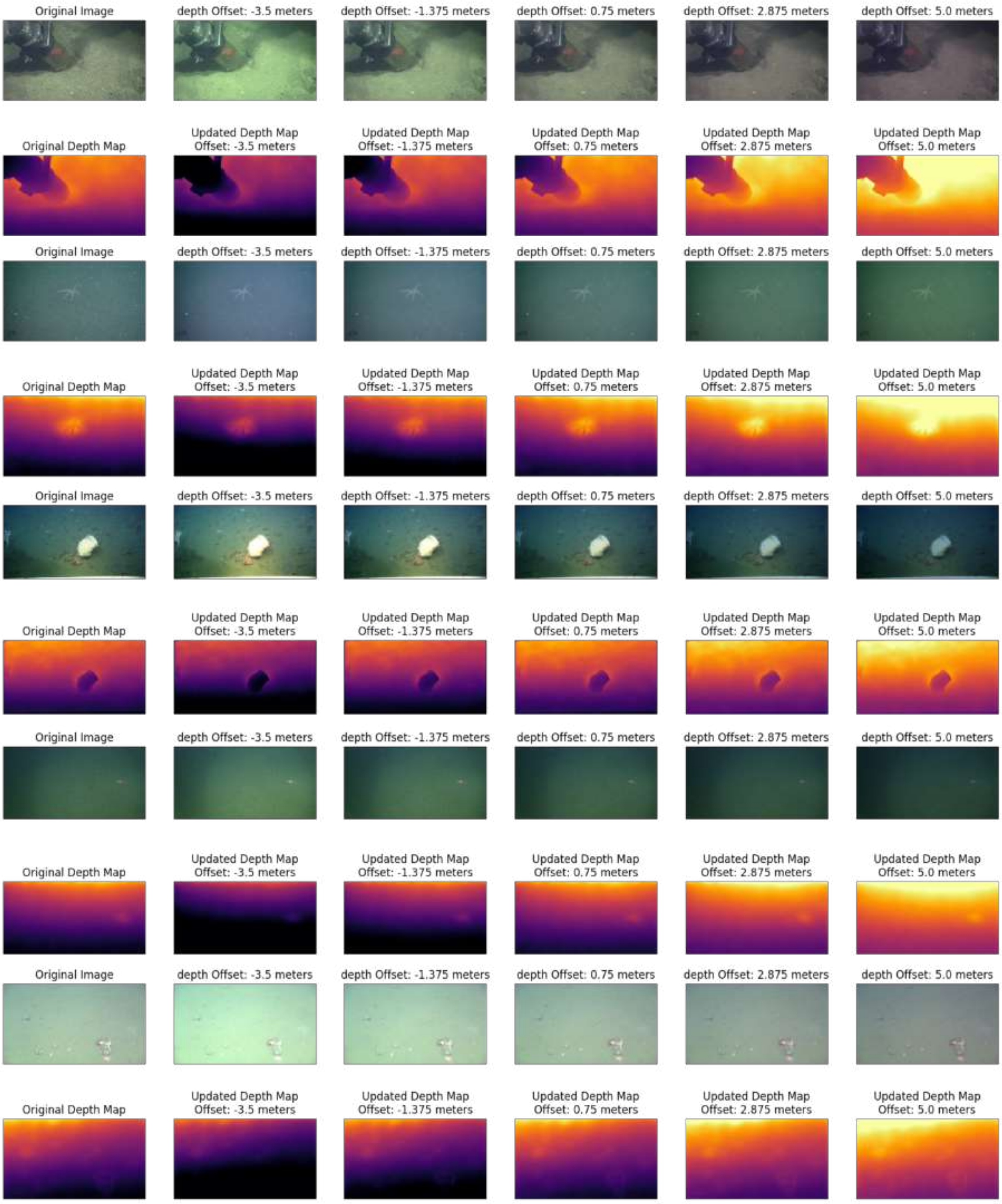


Figure A.1: Visualization of the Depth Jitter Augmentation Technique.