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 Cabecinhas

Examination Committee

Chairperson: Prof. Name of the Chairperson
Supervisor: Prof. Dr. Ricard Marxer
Members of the Committee: Prof. Name of First Committee Member
Dr. Name of Second Committee Member
Eng. Name of Third Committee Member

July 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 her 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

This thesis explores the utilization of cutting-edge deep learning techniques, namely Query2Label and YOLOv9 models, for the tasks of underwater object detection and multi-label classification, with a primary focus on the Fathomnet dataset. We propose the DepthJitter augmentation method, which specifically targets the color distortions caused by depth-related factors, leading to a notable enhancement in model performance, evidenced by superior mAP@20 scores. Despite the inherent challenges of data imbalance and environmental variability in underwater imagery, our models exhibited robustness and adaptability, achieving competitive results without the need for external data sources. This research underscores the potential of these advanced models for practical applications in marine research, such as species identification and habitat monitoring, thereby contributing to marine conservation efforts. Looking ahead, future research should aim to increase dataset diversity, address data imbalance more effectively, enhance model interpretability, and explore real-time deployment possibilities. Additionally, investigating hybrid models and improving out-of-distribution detection will be crucial to further advancing the reliability and applicability of underwater image analysis techniques.

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

Este trabalho demonstra a eficácia das técnicas avançadas de aprendizado profundo, especificamente os modelos Query2Label e YOLOv9, para detecção de objetos subaquáticos e classificação multilabel. Nosso método de aumento proposto, DepthJitter, melhorou o desempenho do modelo, alcançando pontuações superiores de mAP@20 no conjunto de dados Fathomnet. Apesar dos desafios de desequilíbrio de dados e variabilidade ambiental, nossos modelos apresentaram desempenho competitivo sem depender de fontes de dados externas, destacando sua robustez. Para pesquisas futuras, é crucial aumentar a diversidade do conjunto de dados, abordar o desequilíbrio de dados e melhorar a interpretabilidade do modelo. A implantação em tempo real, a exploração de modelos híbridos e a garantia de robustez às mudanças ambientais avançarão ainda mais as aplicações práticas desses modelos. Além disso, a melhoria na detecção de fora de distribuição permanece uma área vital para aumentar a confiabilidade do modelo. Ao seguir essas direções, trabalhos futuros podem se basear em nossas descobertas, contribuindo para técnicas de análise de imagem subaquática mais eficazes e confiáveis para a exploração e conservação marinha.

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, Detecção fora de distribuição, Conservação marinha

Contents

List of Figures

List of Tables

Acronyms

1

Introduction

Contents

1.1 Background

Academic studies confirm that marine life substantially predates terrestrial life. According to [16], 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 [17]. The fossil record, as detailed by [18], 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 biogeographic theories that correlate habitat extent with biodiversity [19].

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 [19]. 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 [20, 21]. This understanding underscores the essential nature of marine biodiversity in global ecological dynamics and warrants further extensive exploration and study.

The ocean's rich biodiversity offers vast opportunities for research and discovery. Marine organisms often harbor unique biochemical properties that hold the potential to propel advancements in both science and medicine. Additionally, the urgent need to conserve these environments is underscored by the rapid impacts of activities such as deep-sea mining and climate change, necessitating a deeper understanding of 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 [22], [23], [24], [25].

The prevalent reliance of current object detection methodologies on existing datasets for training represents a significant limitation, particularly in the context of deep-sea research. The deep sea, with its vast and largely unexplored biodiversity, often presents instances—such as unknown species or previously unobserved events—that are not represented in the training data. To achieve a thorough understanding of deep-sea biodiversity, it is imperative to develop capabilities that can identify and categorize these novel elements. This challenge opens an exciting frontier in marine science, providing opportunities to discover new species and deepen our knowledge of ecological interactions.

The primary aim of this research is to refine the methodologies used for object recognition and the detection of out-of-sample 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 Challenges in Deep Sea Exploration

Deep-sea exploration presents a myriad of formidable challenges, primarily arising from the extreme conditions of these remote environments. One of the most significant obstacles is the profound water pressure encountered at great depths. As depth increases, pressure escalates rapidly, posing a severe risk to traditional underwater sensors and equipment, which may succumb to crushing under such intense forces. Developing technology robust enough to withstand these pressures remains a considerable technological hurdle.

Moreover, deep-sea exploration is further complicated by the absence of natural light. Beyond certain depths, sunlight cannot penetrate, rendering the environment pitch black, which complicates navigation and visual inspection. This lack of visibility, combined with temperatures that are often near or below freezing, significantly limits the use of conventional exploration techniques, such as diving and manned submarines. The cold temperatures pose additional challenges for electronic devices and equipment, which may struggle to operate reliably over prolonged periods under such conditions.

The topology of the deep-sea floor is another major challenge. It encompasses a vast and varied landscape, including undersea mountains, valleys, and tunnels. Mapping and exploring this extensive and diverse terrain is not only costly but also time-consuming. Furthermore, the immense depths and properties of water severely affect signal transmission between deep-sea equipment and surface controllers, leading to potential delays or disruptions. This limitation complicates real-time data collection and decision-making, adding an additional layer of difficulty to deep-sea research and exploration efforts.

These combined challenges underscore the need for innovative solutions and advancements in technology to effectively study and understand the mysterious and inhospitable depths of the deep sea.

1.3 Role of Technology in Overcoming These Challenges

Advancements in technology, particularly sonar and satellite-based instruments, have been instrumental in enabling the exploration and study of the deep-sea environment. Remote sensing technologies

like sonar utilize sound waves to map and identify underwater objects and terrain, providing crucial data even in the absence of visible light.

Autonomous underwater vehicles (AUVs) have revolutionized deepwater exploration. These vehicles can navigate independently through the deep sea and are equipped with research instruments, cameras, and sensors. AUVs are capable of collecting data and capturing images over vast areas, even in complete darkness and under extreme pressure conditions.

Remotely operated vehicles (ROVs), also known as unmanned underwater vehicles, are operated from ships and do not require human divers. Equipped with cameras, lights, and manipulation tools, ROVs can explore and collect samples from areas that are otherwise inaccessible to humans.

Engineering materials and solutions that can withstand the challenging conditions of the deep sea are of paramount importance. This includes the development of reliable communication networks and camera housings that can resist high pressures. Efforts are also being made to create deep-water bases and ecosystems that can support long-term human habitation and scientific exploration. These habitats aim to provide researchers with a stable environment for conducting extensive research.

In summary, while the exploration of the deep sea presents numerous challenges, continuous advancements in engineering, autonomous vehicles, and remote sensing technologies are progressively overcoming these barriers. These innovations allow for a deeper understanding of one of the most unexplored areas of the globe, expanding our knowledge and capabilities in marine science.

1.4 Motivation for this Research

The profound allure of the ocean's mysteries and a fervent desire to expand our understanding of its vast, largely unexplored depths are the primary motivators for this research. The deep sea constitutes the majority of our planet's surface and remains home to numerous enigmatic phenomena and undiscovered species, making it one of Earth's least understood ecosystems. This research aims to unearth these obscure aspects of marine life, endeavoring to make significant contributions to the fields of oceanography and marine biology.

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

Additionally, this effort is driven by an urgent environmental imperative. Understanding the intricate dynamics of these environments is crucial for the protection and sustainable management of marine

ecosystems, especially as human activities continue to impact them profoundly. Through this research, we aim to enhance our ability to monitor and safeguard the rich biodiversity found beneath the waves. Ultimately, this thesis is propelled by a blend of scientific curiosity, a resolve to tackle the technological challenges associated with deep-sea research, and a commitment to environmental stewardship. To unravel the mysteries of the deep sea and preserve it for future generations, it is imperative to push the boundaries of our current knowledge and capabilities.

1.5 Objective of this Research

The primary objective of this thesis is to develop and evaluate machine 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-sample 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.

1.6 Contribution

This research investigates the techniques to enhance the out of distribution 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.

1.7 Outline

This thesis is organized as follows:

Capítulo 2 **State of the Art:** This chapter describes the evolution of object detection models, highlighting the transition from traditional methods like the Viola-Jones and HOG detectors to advanced CNN-based approaches such as RCNN, Fast RCNN, and Faster R-CNN, which introduced Region Proposal Networks for improved performance. It also covers the recent advancements in one-stage detectors like YOLO and SSD, and transformer-based methods like DETR, which have achieved remarkable speed and accuracy. Furthermore, the chapter explores multi-label classification advancements, from early frameworks by Boutell et al. to deep learning approaches like CNN-RNN, Spatial Regularization Networks, and Transformer-based models that capture complex label dependencies. Key challenges specific to underwater environments, such as occlusions, high false positive rates, dataset biases, and computational demands, are discussed. Addressing these challenges requires developing specialized training data and models capable of adapting to dynamic underwater conditions.

Capítulo 3 **Fathomnet Competition Dataset:** This chapter provides an overview of the Fathomnet Competition Dataset, which was obtained from MBARI. The dataset pertains to benthic fauna and comprises 290 categories that were recorded in the Greater Monterey Bay Area utilizing ROV-mounted cameras. The dataset has depth-specific subsets to examine differences in species distribution, and it is annotated for both object identification and multi-label classification. There are 10744 photos in the assessment set and 5950 images in the training set. The images have a large category variability and a long-tailed distribution. Reliable object detection depends on the quality of the annotations, yet label noise and inadequate annotations can be present in some photos. These characteristics affect how well the model performs, emphasizing the importance of giving them serious thought while creating and assessing models for marine biology and underwater imaging.

Capítulo 4 **Methodology:** This chapter describes how the Fathomnet dataset was evaluated, balancing

the dataset for multi-label classification through the use of our proposed data augmentation technique based on the underwater image generation model. It covers the use of the Query2Label (Q2L) model, which integrates a CNN backbone and transformer for feature extraction and label embedding, utilizing Asymmetric Loss (ASL) to handle sample imbalance effectively. YOLOv9 was chosen for object detection due to its exceptional performance, which includes Programmable Gradient Information (PGI) and Generalized Efficient Layer Aggregation Network (GELAN) for improved accuracy. To increase the model's resilience in managing a variety of unusual and diverse data sets, three different techniques for computing out-of-distribution (OOD) scores—Maximum Softmax Probability (MSP), Average Confidence Score, and Gaussian Mixture Model (GMM)—were applied.

Capítulo 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 DepthJitter augmentation method. 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.

Capítulo 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 significant 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

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]

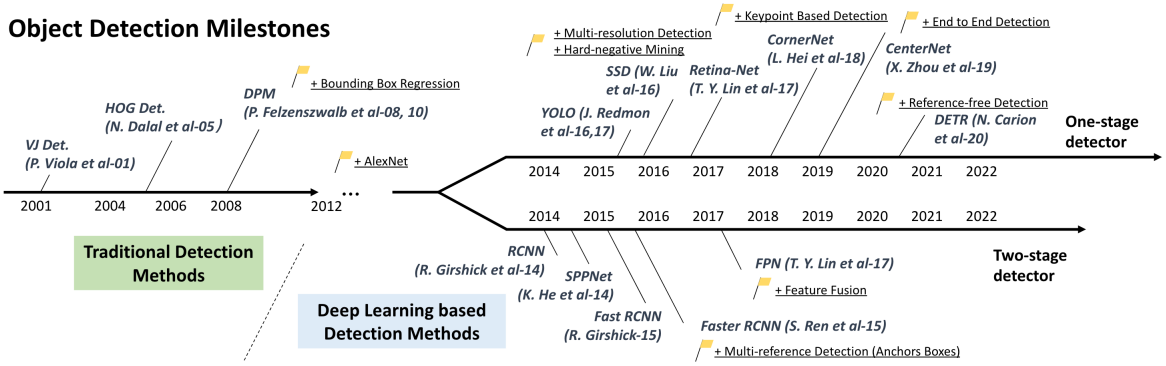


Figure 2.1: The evolution of object detection in the past twenty years. [1]

Figure 2.1 illustrates a graphical depiction of the major object detectors in the last two decades, covering both the periods of pre-deep learning and deep learning in object detection. The subsequent points highlight the noteworthy advancements in object detection:

2.1.1 Traditional Detectors

Viola Jones detector: In the early 2000s, P. Viola and M. Jones were able to achieve real-time human face detection without any limitations [26]. They accomplished this by achieving a frame rate of 15 fps on a 700 MHz Pentium III CPU. The method they proposed in their paper, ”Rapid Object Detection using a Boosted Cascade of Simple Features,” uses the Integral Image representation to efficiently evaluate visual features. It employs a learning algorithm based on AdaBoost to select important features and utilizes a cascade of classifiers to focus computation on regions that are likely to contain objects while quickly discarding background regions. This approach allows for high detection rates and fast processing, making it suitable for real-time applications. The Viola Jones detector has proven to be effective in achieving high frame rates in face detection using only single grayscale images, without the need for additional information.

HOG Detector: The introduction of the HOG-based approach for human detection by N. Dalal and B. Triggs resulted in significant improvements compared to existing feature sets [27]. The authors demonstrated that their approach surpassed wavelet-based methods and other feature sets in terms of accuracy and ability to handle variations in pose and backgrounds. The HOG descriptors effectively captured the edge or gradient structure that is a distinct characteristic of local shape. Furthermore, this

was achieved through a local representation that can be easily adjusted to achieve invariance to local geometric and photometric transformations. Additionally, the authors introduced a new and more challenging pedestrian database, which is publicly accessible. This contribution has played a crucial role in advancing the field of human detection in images.

DPM: This paper [28] presents a system for object detection that utilizes multiscale deformable part models to represent objects with high variability. The system achieves impressive results on challenging datasets by employing a discriminative training procedure and relying solely on bounding boxes for the objects in a given set of images. The system builds upon the pictorial structures framework, which represents objects as a collection of parts arranged in a deformable configuration. These models effectively capture significant variations in appearance and employ mixture models to handle structural and appearance variations within object categories. The system enhances the model through the utilization of a star-structured part-based model, which consists of a ”root” filter and a set of parts filters with associated deformation models. The models are based on linear filters applied to dense feature maps, and the matching process involves dynamic programming and generalized distance transforms to efficiently search for all possible object configurations within an image. Additionally, the system introduces novel local and semi-local features, such as histogram of gradient (HOG) features, and incorporates context for object detection and recognition. The approach is both efficient and accurate, achieving the highest average precision (AP) score across multiple object categories. The system's performance is evaluated on various datasets, demonstrating its ability to effectively represent object classes with high variability and achieve state-of-the-art results.

2.1.2 CNN based Two-stage Detectors:

After reaching a plateau 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 [29]. 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. Currently, there are two main types of object detectors: ”Two-Stage detectors” and ”One-Stage detectors”.

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 [30]. These proposed regions are then standardized to a uniform image dimension and passed through a pretrained CNN model, such as AlexNet [29], 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.

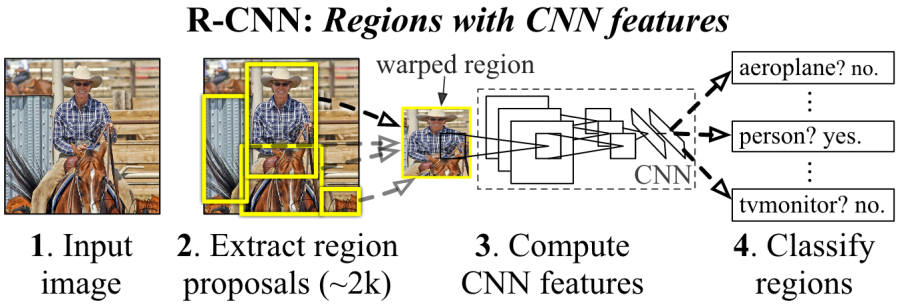


Figure 2.2: Architecture of RCNN [2]

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 [28]. 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 an approximate processing time of 14 seconds per image when using GPU acceleration. To overcome this limitation, a subsequent advancement called SPP-Net [3] was introduced in the same year, significantly improving the computational efficiency of object detection.

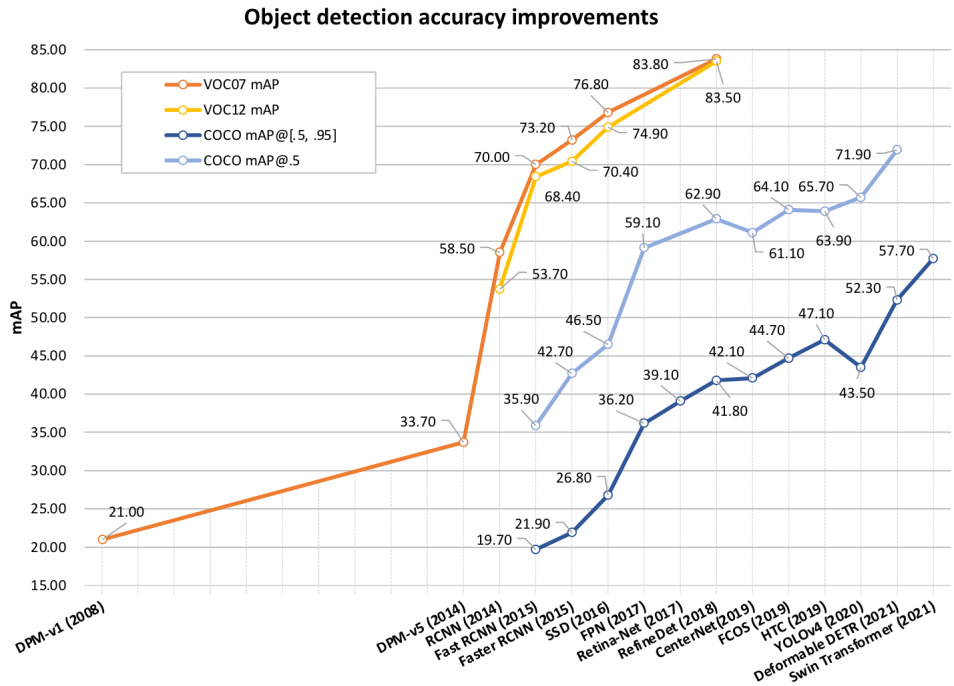


Figure 2.3: Accuracy improvement of the object detectors on VOC and MSCOCO datasets. [1]

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 [29]. 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.

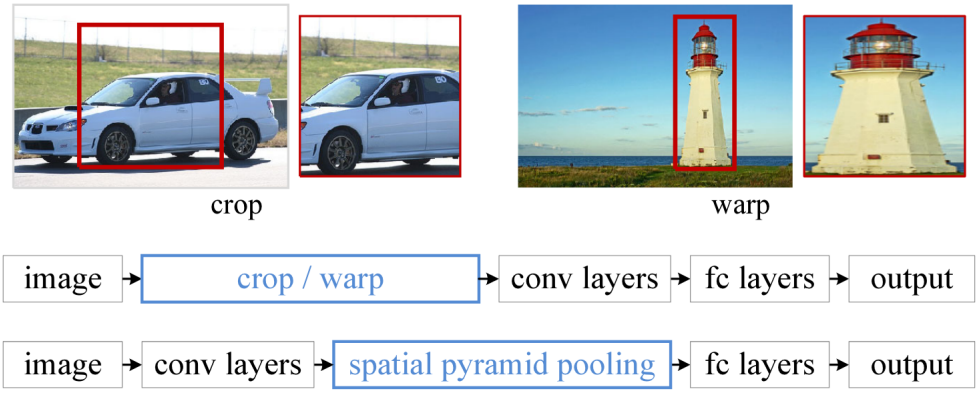


Figure 2.4: Architecture of SPPNet [3]

SPPNet achieves better performance compared to R-CNN, with an improvement of more than 20 times without compromising detection accuracy (VOC07 mAP=59.2%). However, SPPNet still has some limitations. Firstly, its training process remains multistage. Second, SPPNet only fine-tunes its fully connected layers, disregarding all preceding layers. To address these concerns, Fast R-CNN [4] was introduced later in the same year.

Fast RCNN: 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.

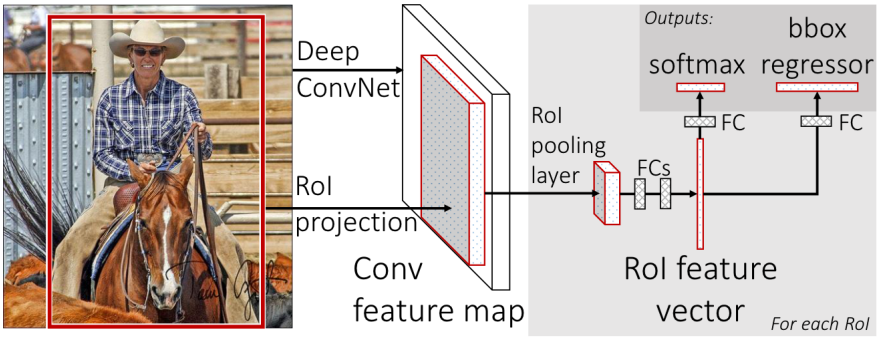


Figure 2.5: Fast RCNN Architecture [4]

However, despite effectively combining the advantages of R-CNN and SPPNet, the detection speed of Fast RCNN 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 RCNN. The Faster RCNN, 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 [31]. The key contribution of the Faster RCNN 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 [32] and light head RCNN [33], have been proposed to address these issues.

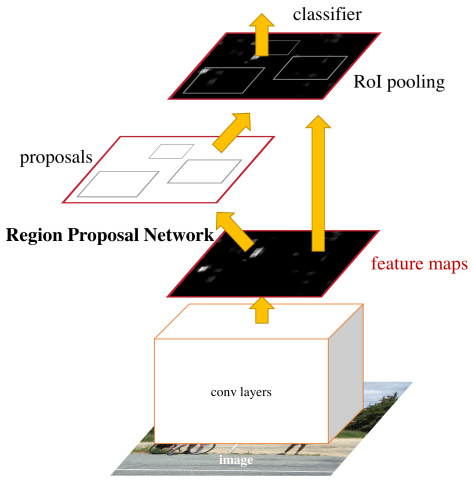


Figure 2.6: Architecture of Faster RCNN [5]

Feature Pyramid Network: In 2017, T.-Y. Lin et al. proposed the concept of FPN (Feature Pyramid Network) [34] 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 additional 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 dense and small objects, which can negatively impact their overall performance in such scenarios.

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%.

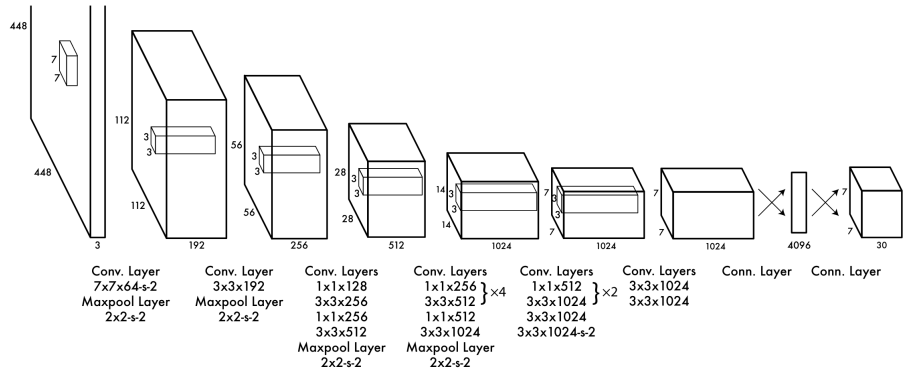


Figure 2.7: YOLO Architecture [6]

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 [35] and the addition of SSD [36], have focused on improving localization accuracy. YOLOv7, in particular, surpasses most existing object detectors in terms of both speed (ranging from 5 to 160 FPS) and accuracy, thanks to enhancements such as dynamic label assignment and model structure reparameterization [35].

Single Shot Multibox Detector: SSD (Single Shot MultiBox Detector) was introduced by W. Liu et al. in 2015 [36], 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 [37]. 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 overcome the limitations of Convolutional Neural Networks (CNNs) and enable the development of a global-scale receptive field.

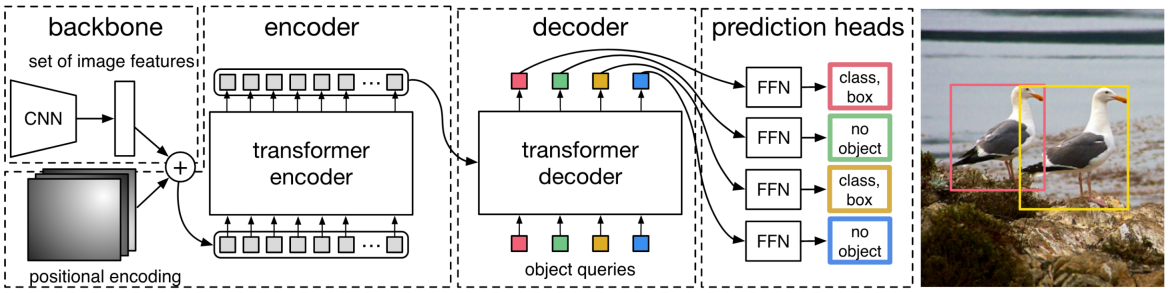


Figure 2.8: DETR architecture [7]

In 2020, N. Carion et al. proposed DETR [7], a groundbreaking approach that treats object detection as a set prediction problem. They introduced an end-to-end detection network powered by Transformers, revolutionizing object detection by eliminating the need for anchor boxes or anchor points. Building on this, X. Zhu et al. introduced Deformable DETR [38] 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 has recently grabbed one of the focal points in the computer vision community: it deals with assigning multiple labels to a single image. This is important for complex scenarios, especially when an image has several objects or features that must be identified simultaneously. Use cases of multi-label image classification in health care, autonomous driving, and underwater exploration come into play. This section provides an in-depth summary of the methods, difficulties, and enhancements of multi-labelled image classification.

2.2.1 Early Research & Foundation

Boutell et al. (2004) [39] address multi-label classification, an aspect of multi-class classification where the instances can be associated with several classes simultaneously, different from single-label classification. That is an essential task in many problems, such as semantic scene classification, medical diagnosis, and document classification. They proposed a framework for multi-label classification and applied it specifically to scene classification.

Models of Training:

- **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:

- **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 can be helpful when texts need to be categorized, medical diagnosis needs to be performed, and protein function needs to be classified because more than one class is desired for a single example. Based on the survey by Tsoumakas and Katakis [40], 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 apply the boosting technique for multi-label problems. Tsoumakas and Katakis 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) [41] 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) [42] 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

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.

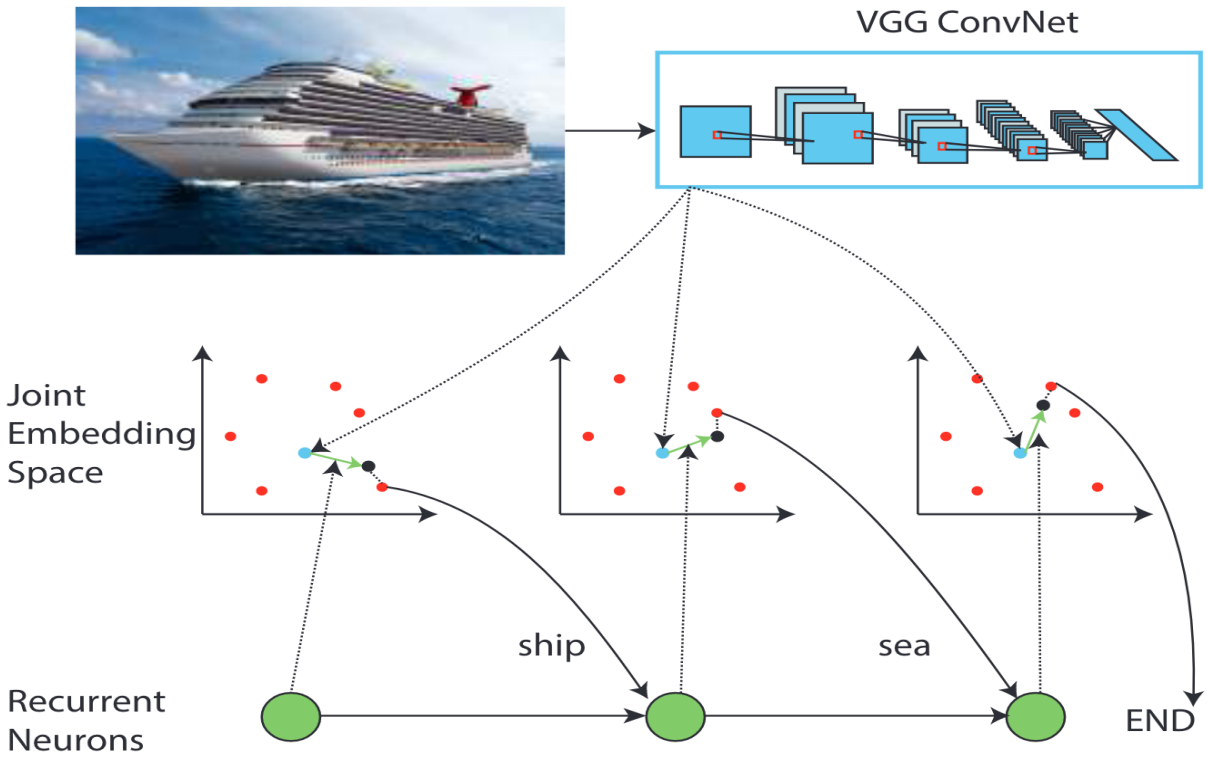


Figure 2.9: An example of the CNN-RNN multilabel classification system for images, where the label dependency and relationship between the picture and label are captured by the framework, which learns a joint embedding space. Here, red and blue points correspond to the label and image embeddings, while 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].

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.

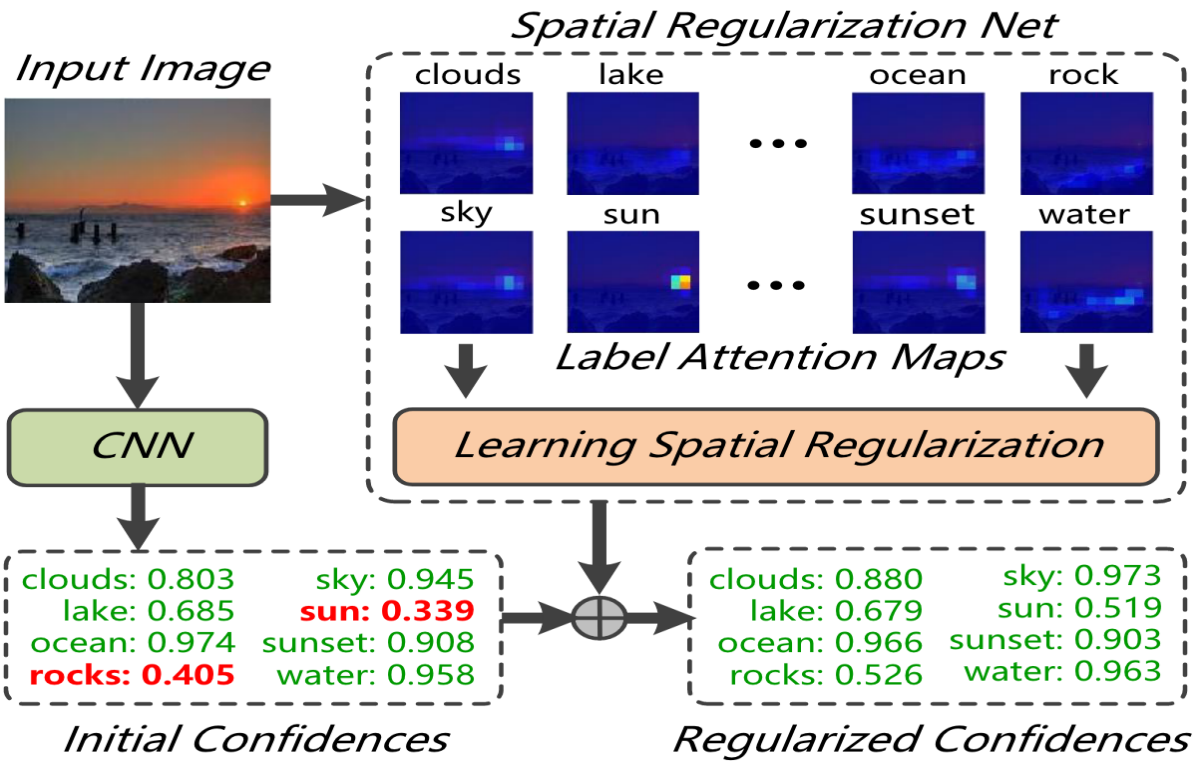


Figure 2.10: Illustration of Spatial Regularization Net(SRN) [9].

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 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.

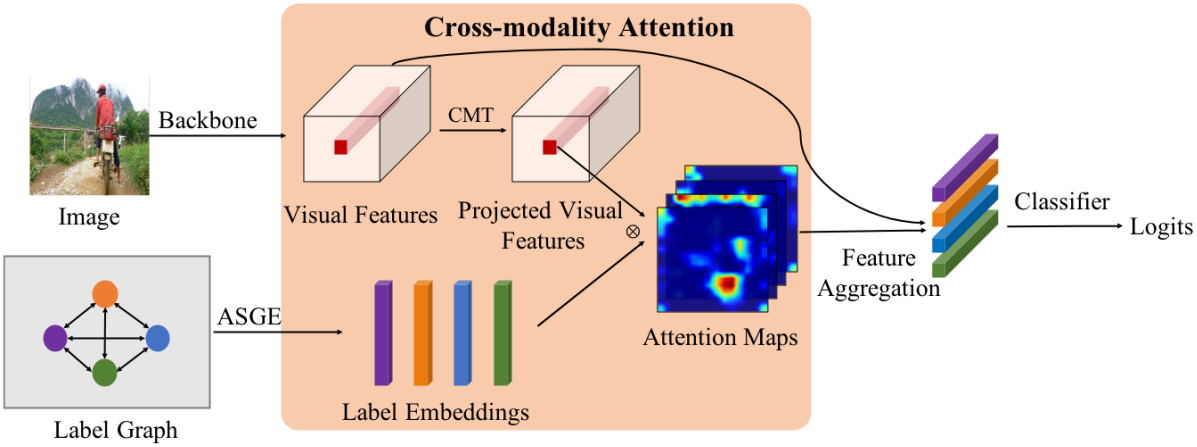


Figure 2.11: General architecture for the MLIC task of the MS-CMA. Label embeddings are given through ASGE. At the early stage, backbone network extraction of the visual data, which are projected in semantic space to get the projected visual features through the CMT module. The projected visual features 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].

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.

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.

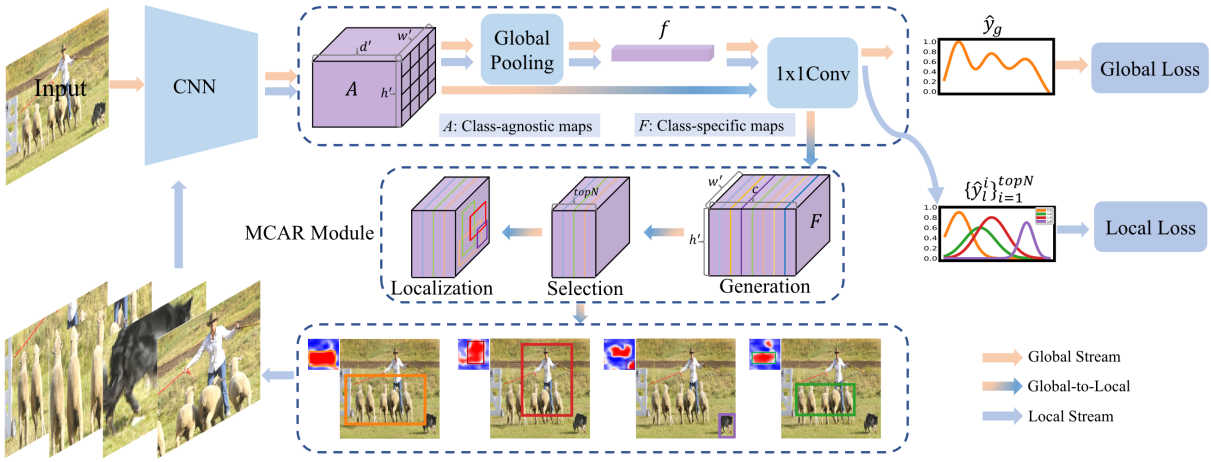


Figure 2.12: The multi-label image recognition pipeline of the MCAR framework commences with extracting the global image stream for feeding an input image into the deep CNN model to obtain its global feature representation. The multi-class attentional region module approximates the localization of regions 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].

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 architectures. 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.

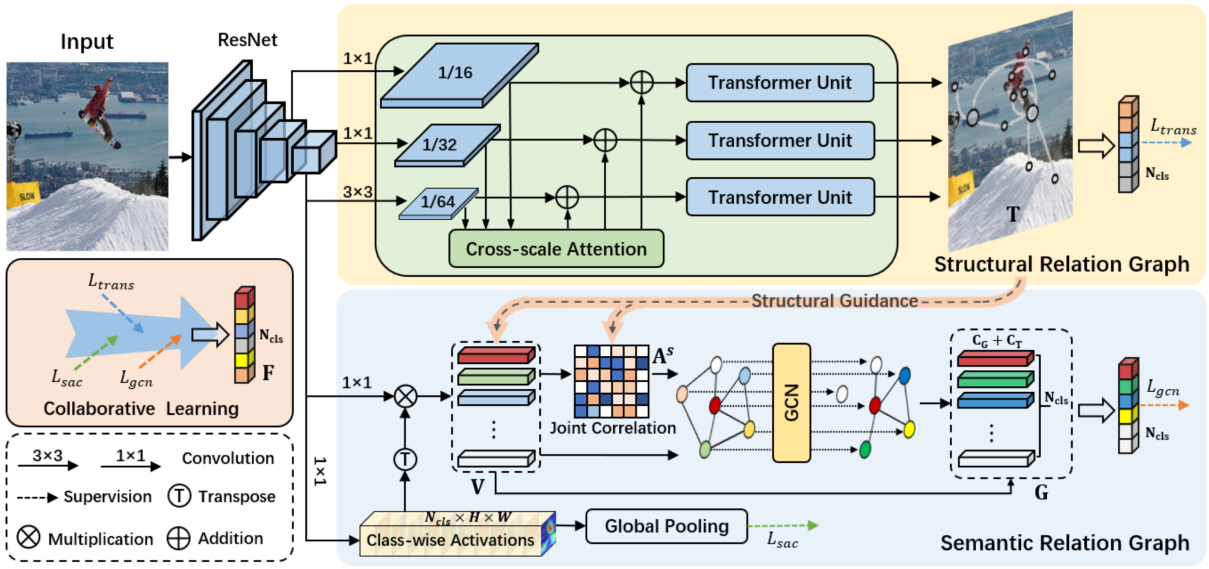


Figure 2.13: The general structure of the Transformer-based Dual Relation Graph (TDRG) network, which is comprised of two fundamental modules: the semantic relation graph module, which models the dynamic class-wise dependencies, and the structural relation graph module, which incorporates long-term contextual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

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. 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 Challenges & Limitations

2.3.1 Technical Challenges

2.3.1.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.

2.3.1.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.

2.3.1.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.

2.3.2 Broader Issues

2.3.2.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 terrestrial 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.

2.3.2.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.

2.3.2.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.

2.3.2.D Absence of Benchmark in Underwater Datasets:

Setting up benchmarks specifically designed for submerged 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.

2.3.2.E Environment Variability

Changes in the undersea 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.

2.3.2.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.

2.3.2.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 dataset was taken from the Fathomnet's large database by Fathomnet Community which is a part of the Monterey Bay Aquarium Research Institute (MBARI) to isolate the depth shift by restricting the search spatially lessen the effect of longitudinal changes of species over time, institutionally to make every effort to guarantee uniformity in personnel and equipment and temporally to guarantee that the data collection cameras are identical [43]. 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.

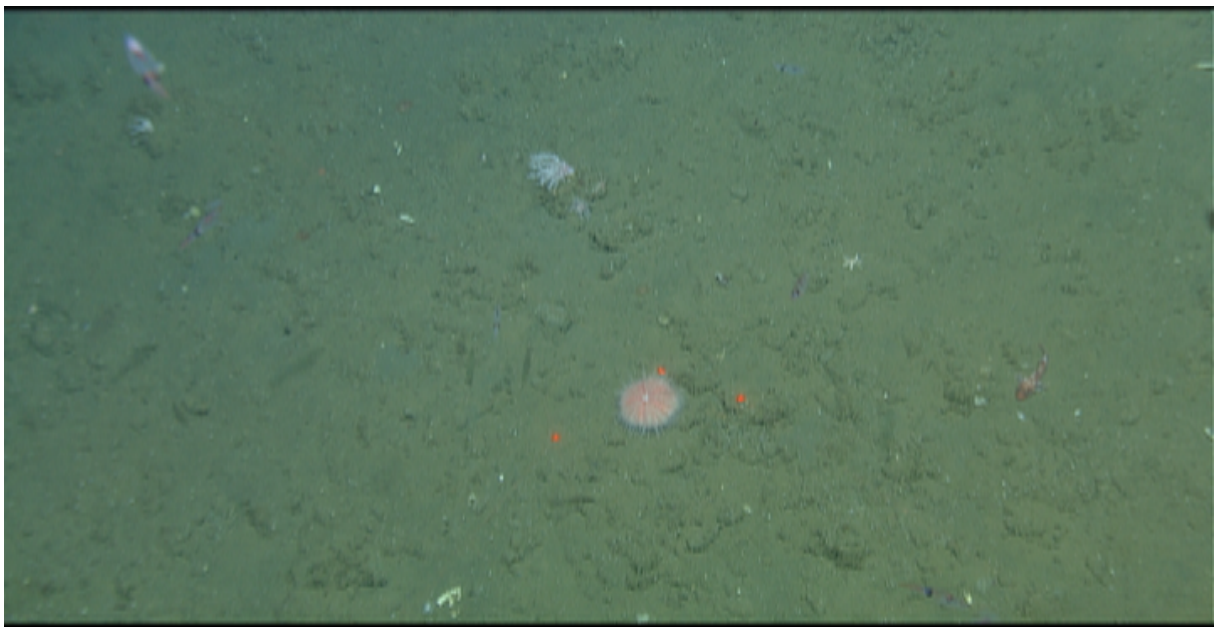


Figure 3.1: *S. fragilis.* is the most commonly found concept in the Fathomnet 2023 Dataset both in the training and the evaluation set.

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 [43]. The annotation with the dataset is provided in multi-label classification and object detection formats. The multi-label classification annotations are presented in comma separated list of image ids and unique set of categories found in each image. 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, even though not all fine-grained categories are represented in both sets [43].

3.2 Properties of Fathomnet 2023 Dataset

In the field of underwater image analysis and species identification, the properties of Fathomnet Competition 2023 Dataset [43] 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 from0-800 meters depth while the evaluation data extends upto 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.

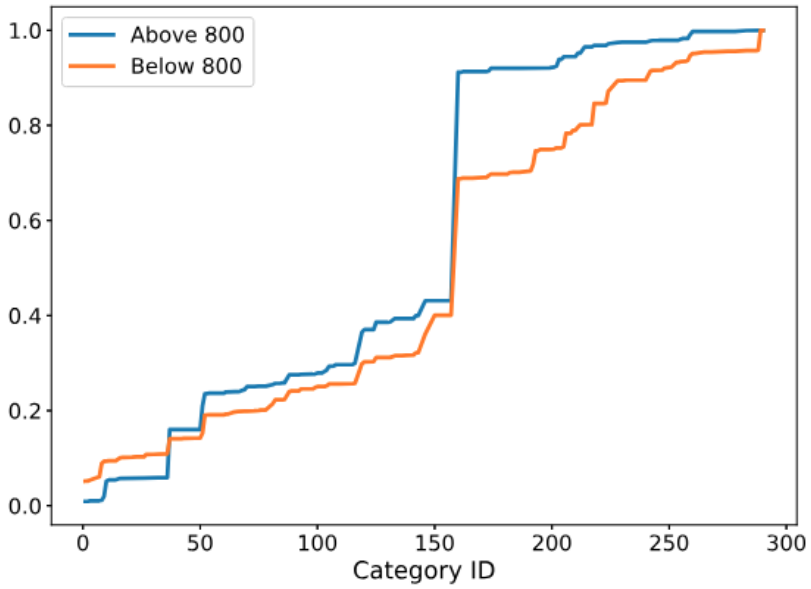


Figure 3.2: The overall distribution of categories in the FathomNet 2023 training and evaluation datasets. They differ greatly from one another, with some classes existing in only one of them [**?**].

Characteristics: The Fathomnet training dataset contains 5950 images with 23703 localized annotations and the evaluation set contains 10744 images with 49798 localized annotations where 6313 images were collected from below the 800 meters threshold which are considered as out of sample [43]. The Fathomnet dataset is relatively long tailed as most of the fine grained image datasets. The most frequently occurred concept in both training and evaluation dataset is *S. fragilis*. Beyond *S. fragilis* the order and magnitude of the other concepts is quite variable between the sets. The dataset contains 290 categories which belong to 20 semantic supercategories. In the training set 157 categories of the 290 categories do not have any images and almost 80 of the categories have less than 10 samples from

which it can be told that the dataset is very imbalanced and long tailed.

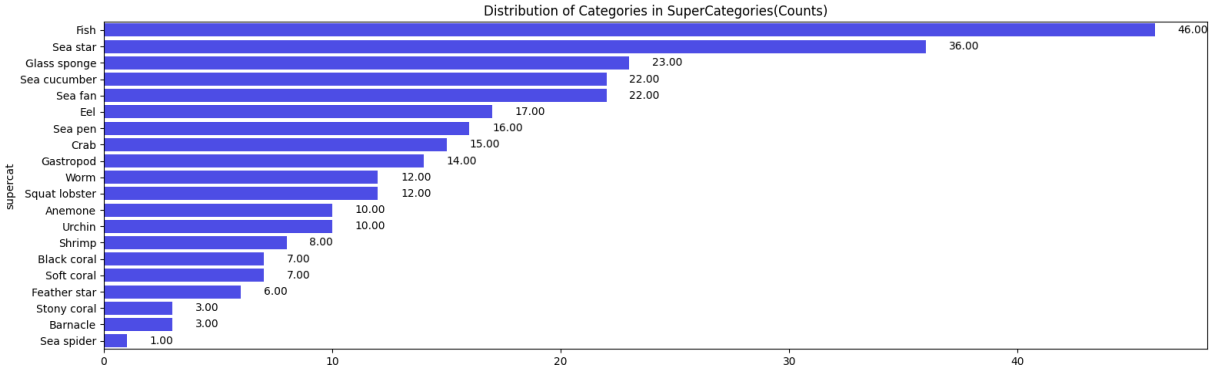


Figure 3.3: Categories Count in Supercateogires

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 example 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.

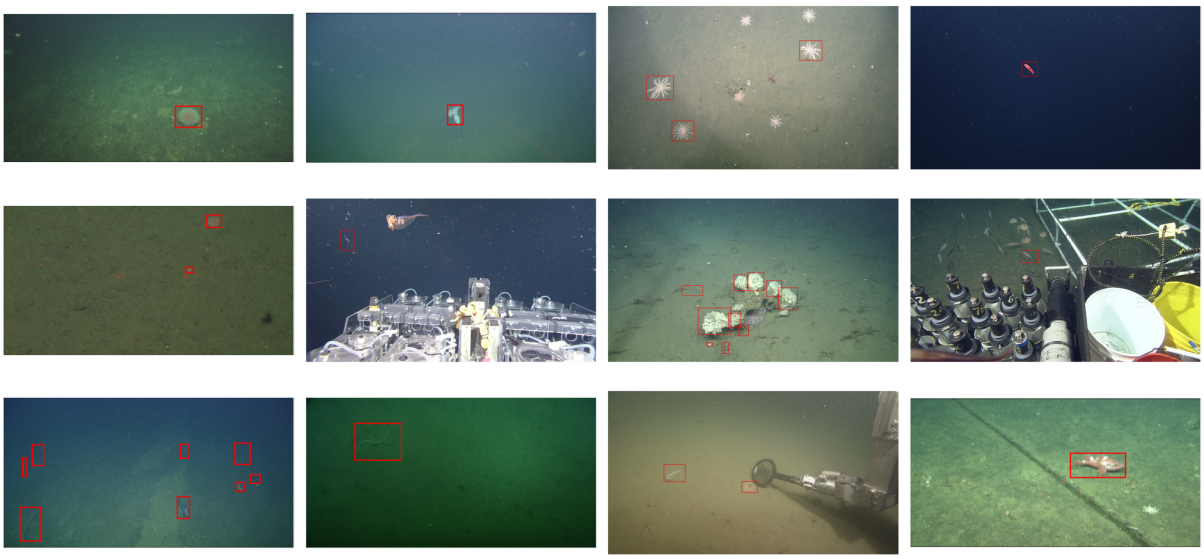


Figure 3.4: Annotation Sample of the Fathomnet Dataset

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 the both dataset. This imbalance can lead to biased models that perform well on over-represented categories but poorly on underrepresented or unseen categories. In out-of-sample 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, care consideration of these factors are crucial.

4

Methodology

Contents

As the Fathomnet dataset comes with object detection and multi-label classification annotations we tried both methods to check which method performs the best in terms of out of distribution detection and marine species category classification. But before that to tackle this challenging underwater dataset we applied some data augmentation techniques to make the dataset more balanced for multi-label classification method.

4.1 Dataset Pre-Processing

4.1.1 Underwater Light Propagation

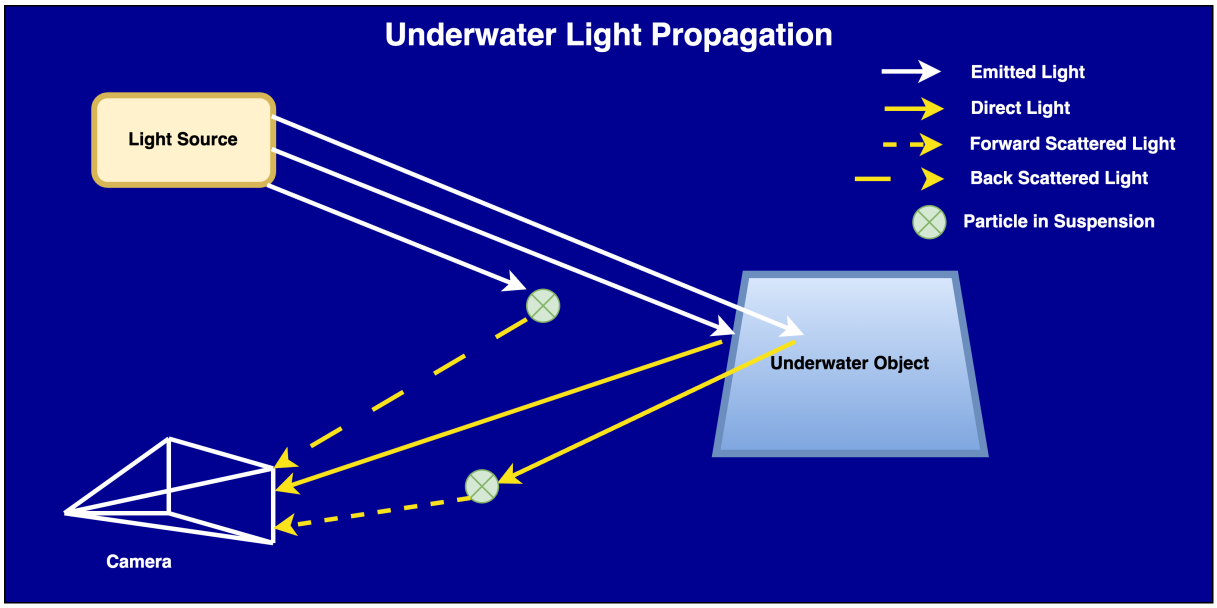
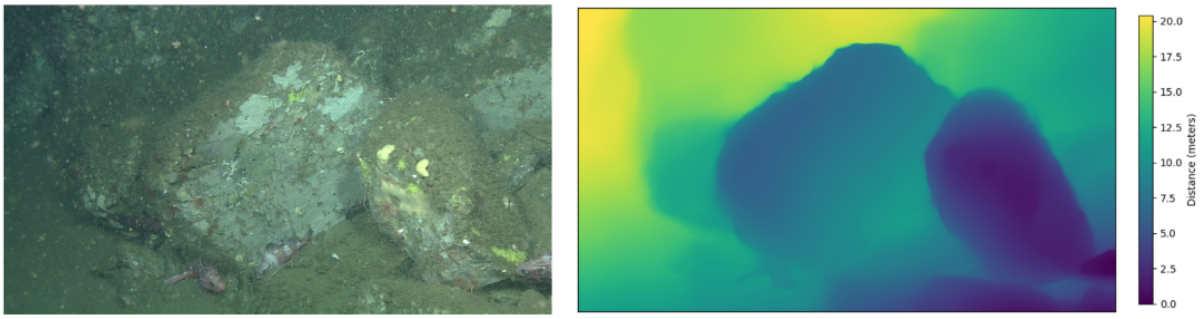


Figure 4.1: As light travels through water, a portion of the emitted light is absorbed and transformed into other forms of energy. Additionally, some photons interact with suspended particles en route to the sensor, causing scattering by acting as secondary light sources [13].

Both absorption and scattering effects have a major impact on the light behaviour in the underwater environment [13, 44–47] 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 figure4.1, there are two types of scattering found in practice which are forward scattering and back scattering. Backscattered light, as contrast to forward scatter, does not provide any information about the observed scene [47].



(a) Underwater Image (b) Distance Map

Figure 4.2: This figure presents an underwater image alongside its corresponding depth map, which was generated using Depth Anything [14].

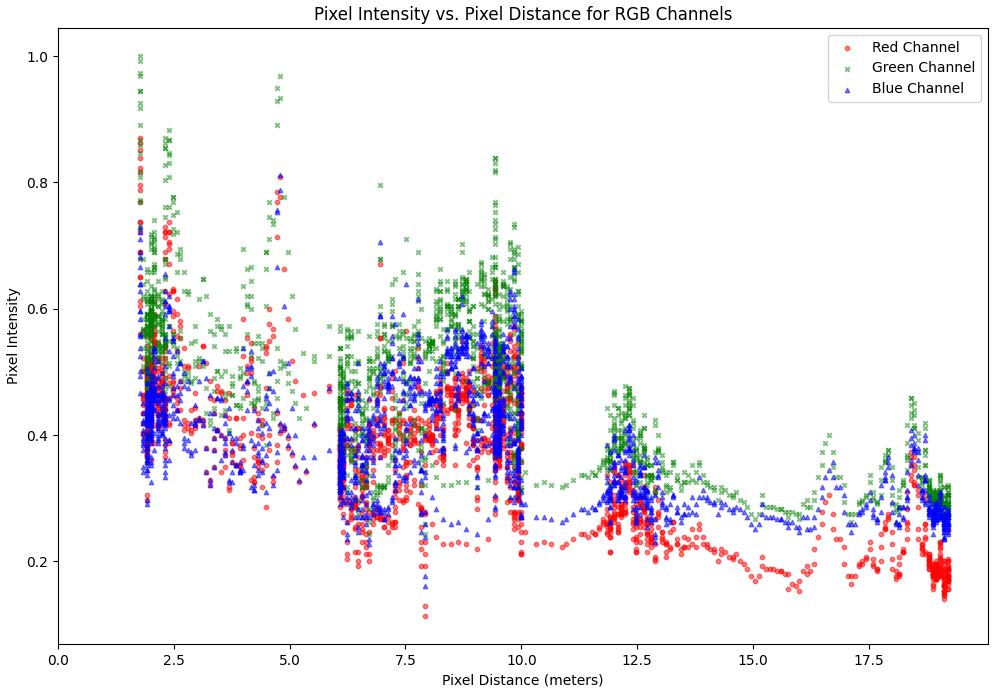


Figure 4.3: Plotting of the pixel intensities against their respective observation distances of the image presented in figure 4.2, the absorption and scattering effects typical of underwater environments are illustrated. The comparison between different color channels demonstrates how longer wavelengths, such as red, are absorbed more quickly than shorter wavelengths, like blue.

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 [14] 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.

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 [13].

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 [13].

Based on the initial underwater image formation model proposed by Schechner and Karpel [48], numerous color restoration techniques [49–51] 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. [46] 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 [13] 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 [46] 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:

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

where

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

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 ($\boldsymbol{\gamma_c}$) by minimizing $\log(\boldsymbol{L(I_c)})$:

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

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

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

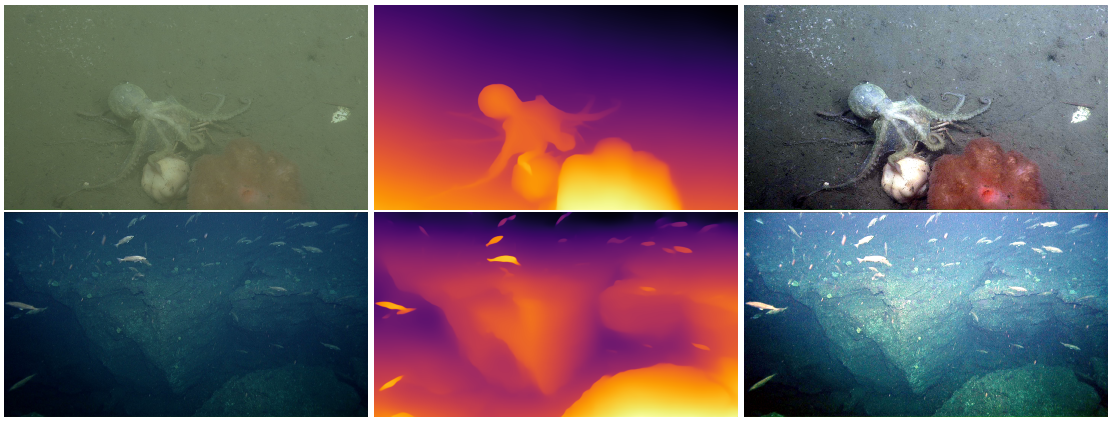


Figure 4.4: Some examples of the visualization of the original image(left), distance map(middle) obtained from Depth-Anything [14] and the restored image(right).

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:

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

and

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

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

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

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

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

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

Simplify the equation:

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

In the equation 4.14 we can insert z_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.

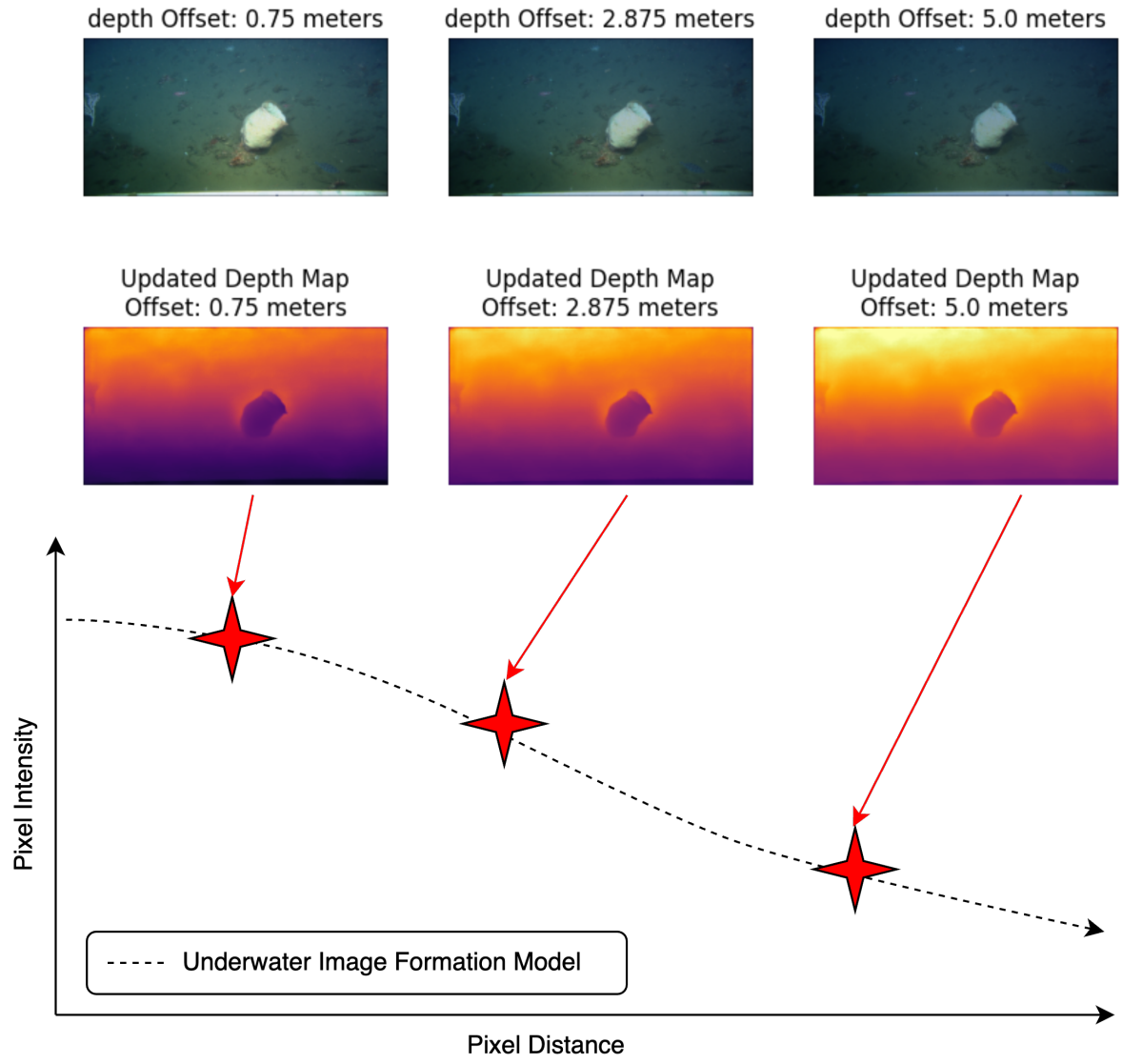


Figure 4.5: This figure shows the pixel intensity tracking and change on the image in different depth settings.

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 dataloader is

called the depth offset applied to the image is different. That's how this technique makes the dataset more diverse.

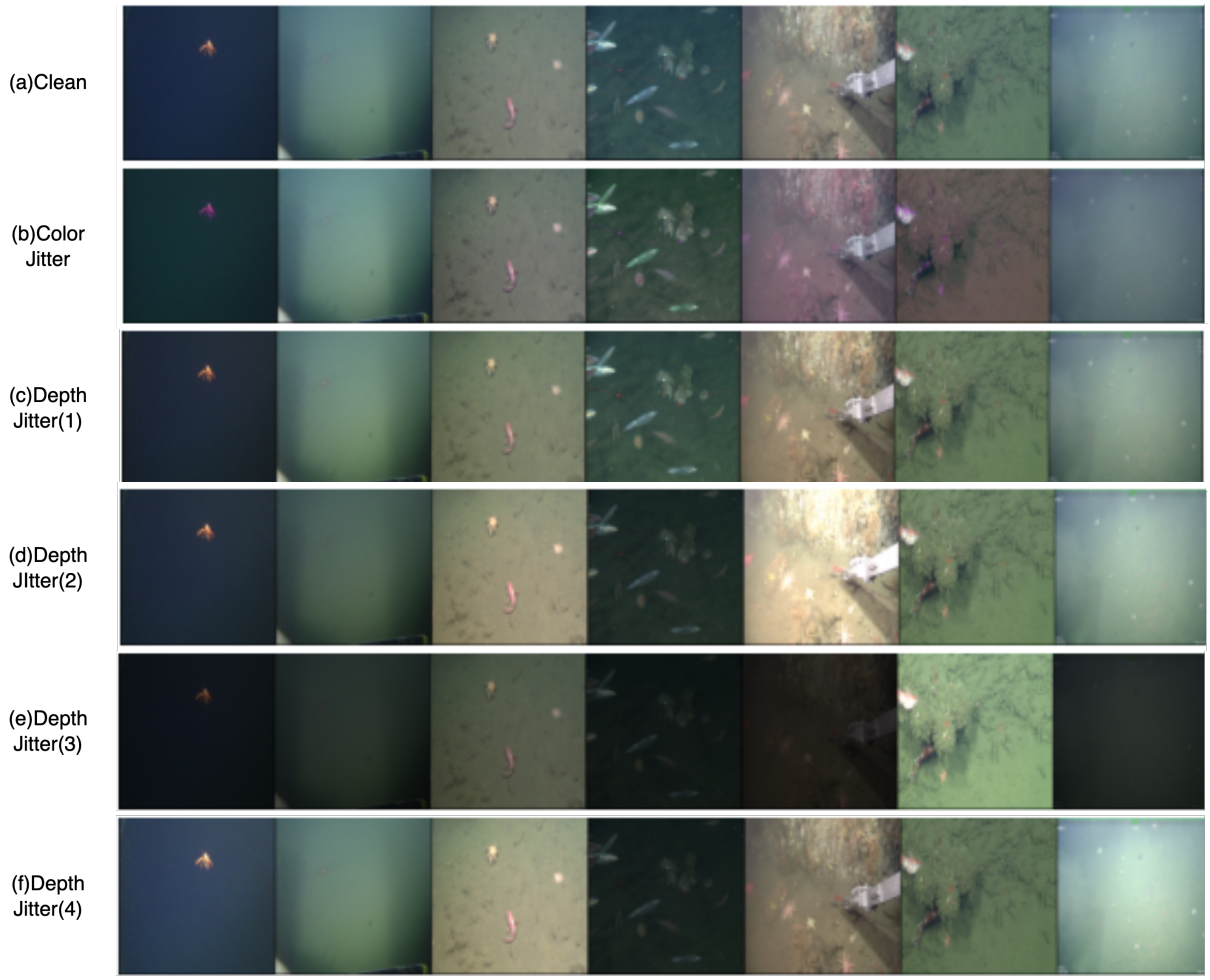


Figure 4.6: Comparison of Different Augmentation techniques.

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. 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 dataloader 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 Multi-Label Image Classification

For multi-label image classification we used Query2Labels [15] model which is one of the best for performing multi-label classification task. This approach draws significant inspiration from the prominent DETR model developed by Facebook Research, which simultaneously performs object detection (with bounding boxes) and labeling. Similar to DETR, the Query2Label (Q2L) model employs a CNN backbone integrated with a transformer. However, Q2L diverges from DETR in several critical aspects, as outlined below.

The architecture of the Q2L model is visually and conceptually represented as follows:

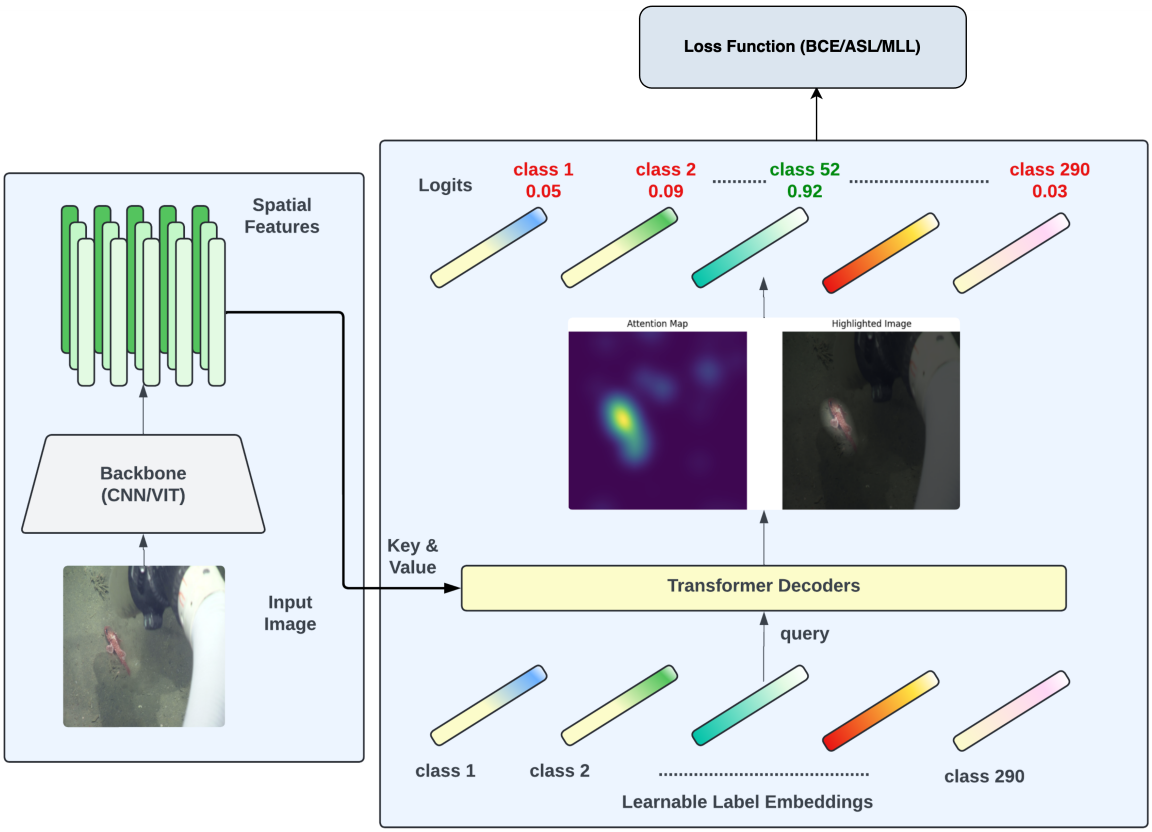


Figure 4.7: Framework of the Proposed Query2Label (Q2L) Model. The Q2L framework begins with the extraction of spatial features from an input image. Each label embedding is then processed by the Transformer, which query the features by comparing the label embedding with the spatial features to generate attention maps. These attention maps are used to adaptively pool the relevant features by linearly combining the spatial features. The pooled feature is subsequently utilized to predict the presence of the corresponding label.

1. **Backbone Processing**: Input images are fed through a pre-trained backbone, typically a variant of ResNet or a Vision Transformer.

2. **Feature Projection**: The output from the backbone is passed through a linear layer to reduce the number of feature planes. For instance, the larger ResNet models produce 2048 feature planes, which are reduced by a specified factor.

3. **Position Encoding**: Position encodings are added to the feature planes. The resulting tensor is reshaped to be compatible with the transformer input requirements.

4. **Transformer Integration**: The reshaped tensor is then fed into a transformer. The default configuration, as used in the original paper, consists of one encoder layer and two decoder layers. Additionally, learnable label embeddings are introduced at this stage. The backbone output is fed into the transformer in place of source word embeddings (as used in language models), and the label embeddings replace target word sequences. Masking is omitted since the model does not output a sequence.

5. **Classification Head**: The transformer's output is directed to a classification head, followed by a softmax layer, resulting in a tensor that represents the probabilities for each potential label.

This model architecture effectively combines CNN and transformer mechanisms to enhance the accuracy and efficiency of multi-label classification tasks.

4.2.1 Feature Extraction

Given an input image $x \in \mathbb{R}^{H_0 \times W_0 \times 3}$, we begin by extracting its spatial features using a backbone network. The resulting feature map $F_0 \in \mathbb{R}^{H \times W \times d_0}$ has dimensions $H \times W \times d_0$, where H_0 and W_0 are the height and width of the original image, H and W are the height and width of the feature map, and d_0 represents the depth of the feature map [15].
To ensure these features are compatible with the desired query dimension for the subsequent processing stage, we apply a linear projection layer. This layer adjusts the feature dimensions from d_0 to d. Finally, we reshape the projected features into a matrix $F \in \mathbb{R}^{HW \times d}$, where HW is the total number of spatial locations in the feature map [15].

4.2.2 Query Updating

After extracting the spatial features of the input image in the first stage, we utilize label embeddings as queries $Q_0 \in \mathbb{R}^{K \times d}$ and perform cross-attention to pool category-related features from the spatial features using multi-layer Transformer decoders, where K is the number of categories [15]. The standard Transformer architecture, comprising a self-attention module, a cross-attention module, and a position-wise feed-forward network (FFN), is employed.

Each Transformer decoder layer i updates the queries Q_{i-1} from the output of its previous layer as follows:

- **Self-Attention:**

$$Q_i^{(1)} = MultiHead(Q_{e,i-1}, Q_{e,i-1}, Q_{i-1}),$$

  where the tilde indicates the original vectors modified by adding position encodings.

- **Cross-Attention:**

$$Q_i^{(2)} = MultiHead(Q_i^{(1)}, F_e, F),$$

  where F_e represents the spatial features.

- **Feed-Forward Network (FFN):**

$$Q_i = FFN(Q_i^{(2)}).$$

In these equations, $Q_i^{(1)}$ and $Q_i^{(2)}$ are intermediate variables. Both the MultiHead(query, key, value) and FFN(x) functions follow the standard definitions in the Transformer decoder [15]. We omit their parameters for simplicity. Since autoregressive prediction is unnecessary, we do not use attention masks, allowing the M categories to be decoded in parallel in each layer.

The self-attention and cross-attention modules use the same MultiHead function, differing only in the sources of the key and value. In the self-attention module, the query, key, and value all come from label embeddings. In contrast, in the cross-attention module, the key and value are derived from the spatial features. This cross-attention process can be described intuitively: each label embedding $Q_{i-1,k} \in \mathbb{R}^d$ for $k = 1, ..., K$ examines the spatial features F_e to determine where to focus and selects the relevant features to combine. Consequently, each label embedding acquires enhanced category-related features and updates itself. Through this iterative process, the label embeddings Q_0 are progressively enriched with contextual information from the input image via cross-attention [15].

Inspired by DETR, we treat the label embeddings Q_0 as learnable parameters. This end-to-end learning approach allows the embeddings to implicitly model label correlations from data. Unlike DETR, where queries are class-agnostic and predicting which query will detect which category is challenging, our queries are class-specific and have clear semantic meanings [15].

4.2.3 Feature Projection

Assuming a total of L layers, the queried feature vectors $Q_L \in \mathbb{R}^{K \times d}$ for K categories are obtained at the last layer [15]. For multi-label classification, each label prediction is treated as a binary classification task. The feature of each class $Q_{L,k} \in \mathbb{R}^d$ is projected to a logit value using a linear projection layer

followed by a sigmoid function:

$$p_k = Sigmoid(W_k^T Q_{L,k} + b_k),$$

where $W_k \in \mathbb{R}^d$, $W = [W_1, ..., W_K]^T \in \mathbb{R}^{K \times d}$, and $b_k \in \mathbb{R}$, $b = [b_1, ..., b_K]^T \in \mathbb{R}^K$ are parameters in the linear layer, and $p = [p_1, ..., p_K]^T \in \mathbb{R}^K$ are the predicted probabilities for each category. Note that p is a function which maps an input image x to category prediction probabilities, where x is omitted for notation simplicity [15].

4.2.4 Loss Function

Thanks to the built-in cross-attention mechanism in Transformer decoders, the framework does not require a new loss function. Three different loss functions were experimented with: binary cross-entropy loss (BCE), focal loss, and two-way multilabel loss. To more effectively address the sample imbalance problem, a simplified asymmetric loss (ASL), which is a variant of focal loss with different γ values for positive and negative samples, is adopted. Experimental results indicate that ASL provides the best performance.

Given an input image x, the framework predicts its category probabilities $p = [p_1, ..., p_K]^T \in \mathbb{R}^K$. The asymmetric focal loss is used to calculate the loss for each training sample x as follows:

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

where y_k is a binary label indicating if image x has label k. The total loss is computed by averaging this loss over all samples in the training dataset D, and optimization is performed using stochastic gradient descent. By default, the parameters are set to $\gamma^+ = 0$ and $\gamma^- = 1$ in the experiments [15, 52].

4.3 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.

4.3.1 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.

4.3.1.A 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.

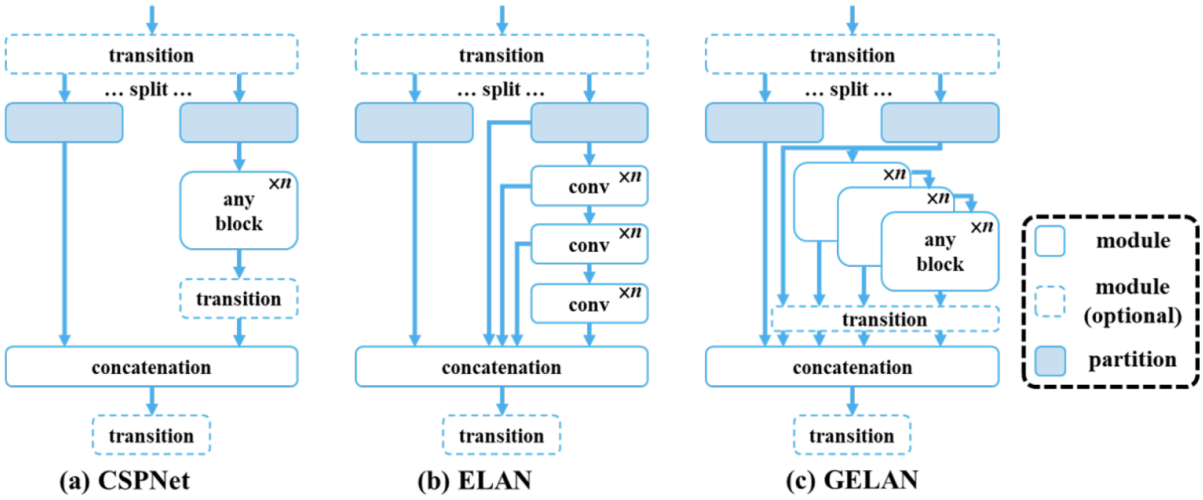


Figure 4.8: The architecture of GELAN: (a) CSPNet, (b) ELAN, and (c) proposed GELAN. GELAN extends ELAN to support any computational blocks.

4.3.1.B 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.

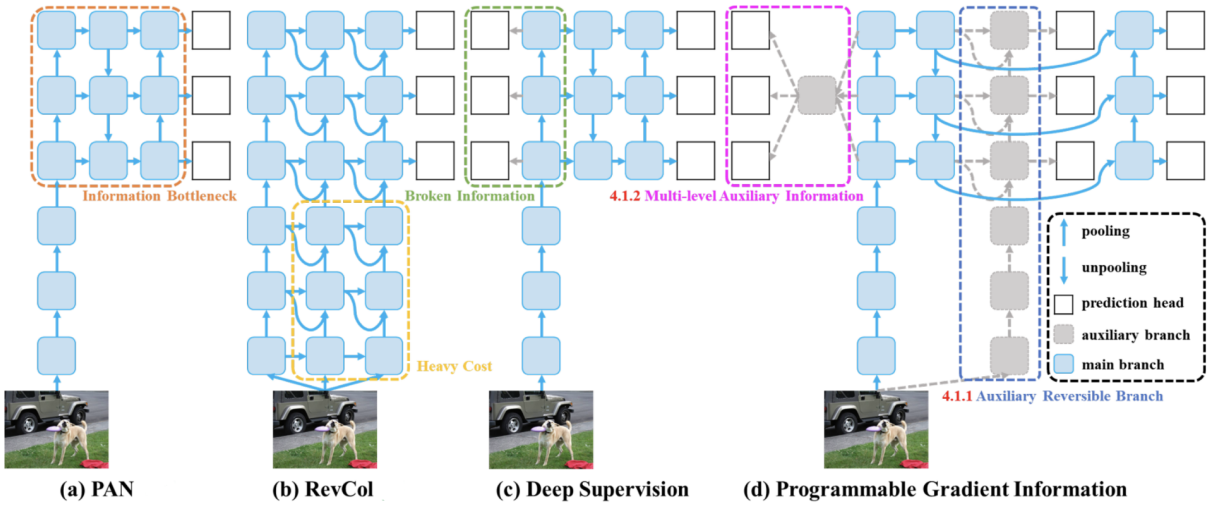


Figure 4.9: PGI and related network architectures: (a) Path Aggregation Network (PAN), (b) Reversible Columns (RevCol), (c) conventional deep supervision, and (d) proposed PGI. PGI comprises three components: main branch, auxiliary reversible branch, and multi-level auxiliary information.

4.3.1.C 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.

4.3.1.D Training and Performance

YOLOv9 is trained using the MS COCO dataset, employing a train-from-scratch strategy over 500 epochs. The training process includes linear warm-up for the initial epochs followed by learning rate decay. Data augmentation techniques, such as mosaic augmentation, are used during training to enhance model robustness.Experimental results demonstrate that YOLOv9 achieves top performance across various benchmarks. The model significantly surpasses previous YOLO versions and other state-of-the-art object detectors in terms of accuracy, parameter efficiency, and computational complexity. Detailed comparisons show that YOLOv9 requires fewer parameters and computations while maintaining or improving detection accuracy.The combination of GELAN and PGI in YOLOv9 results in a highly efficient and accurate object detection model, suitable for a wide range of applications.

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

4.4.1 Method 1: Maximum Softmax Probability (MSP)

The Maximum Softmax Probability (MSP) method calculates the out-of-distribution (OOD) score based on the highest probability predicted by the softmax function. The rationale behind this method is that if the highest softmax probability is low, the sample is likely to be out-of-distribution. The OOD score is computed as follows:

$$OOD = \begin{cases} 1.0 & \text{if len(predicted\_classes\_list)} = 0 \\ 1 - \max(\text{probabilities}) & \text{otherwise} \end{cases} \tag{4.15}$$

If no classes are predicted (i.e., the length of the predicted classes list is zero), the OOD score is set to 1.0, indicating high uncertainty. Otherwise, the OOD score is calculated as one minus the maximum probability from the softmax output.

4.4.2 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

5.1.1 Evaluation Metrics

For each image in the test set, participants are required to predict an ordered list of categories present in the image and determine whether the image is out-of-sample (OOD). The performance is evaluated using mean Average Precision (mAP) for category predictions and Area Under the Receiver Operating Characteristic Curve (AUC) for out-of-sample classification. The combined results determine the ranking on the Kaggle leaderboard.

5.1.2 Out-of-Sample Detection

The Receiver Operating Characteristic (ROC) curve is used to visualize the performance of the classifier across different classification thresholds. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various thresholds, defined as:

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

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

As the classification threshold becomes more permissive, both TPR and FPR increase towards 1. The Area Under the ROC Curve (AUC) is calculated by integrating the area under the ROC curve, with values ranging from 0 to 1, where an AUC of 1 indicates a perfect classifier.

Out-of-sample predictions are evaluated using AUC. To combine this score with the category evaluation, AUC is rescaled as follows:

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

5.1.3 Category Predictions

Category predictions are evaluated using mean Average Precision at 20 (mAP@20), defined as:

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

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.

5.1.4 Final Score

The final score is computed as the simple average of sAUC and mAP@20:

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

5.1.5 Performance of Object Detection Models

We train several object detection models on fathomnet competition dataset. All the images of the dataset are trained and tested in 640X640 size. All the tests in table 5.1 are performed by us as there is no benchmark available for this particular task. Therefore we generate the benchmark with several different object detectors with different settings. In the table 5.1, we can see that several models outperform the baseline [43] result. Among thos we get the best performance from YOLOv9 [53] object detection model which achieved mAP@20 0.74 on the validation set which is the main evaluation metric for categories prediction. As this task was a kaggle competition and there is no ground truth for the evaluation set it is impossible to know the evaluation metrics individually. We are able to know only the final score which will be discussed later in this chapter. For every model we used patience 30 which means if the

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.

performance of the model does not improve for 30 epochs it will keep the best from the last 30 epochs. That is why there are difference in the epochs as we can see in the table 5.1. 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 in the annotation we use label smooting 0.1 to make the labels in the annotation more smoother and balanced. Also, among 290 classes in the dataset there are some classes which don't have any examples in the train or test set. So, we filter them while training

the object detection model to make it more robust to our dataset. 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.

5.1.6 Performance of Query2Label Model in Different Augmentation Settings

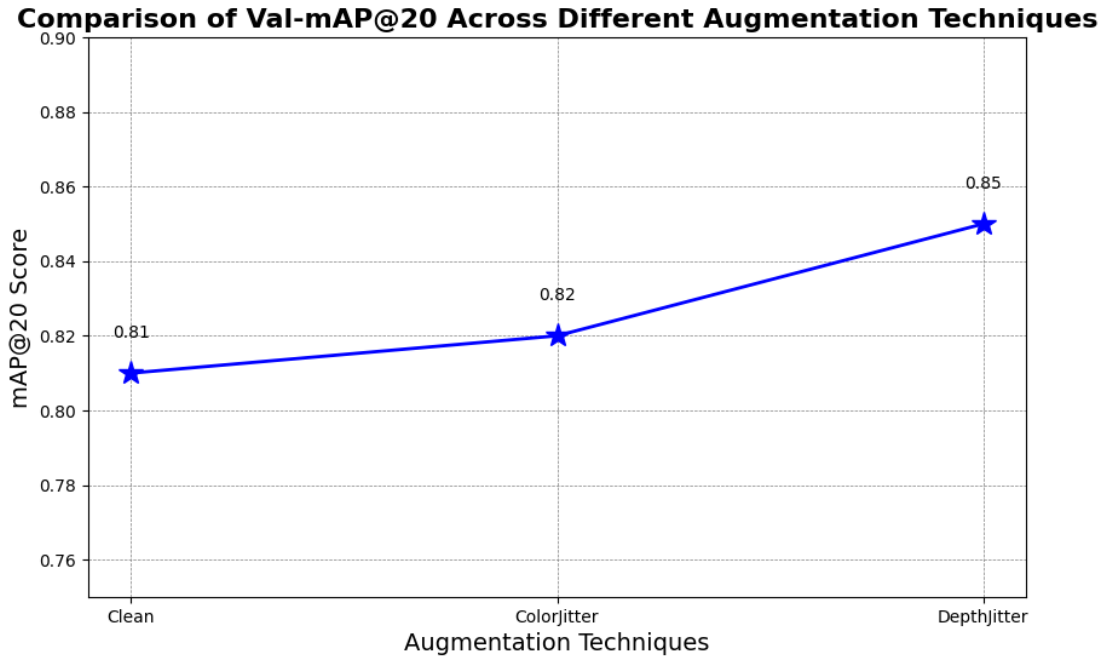


Figure 5.1: Comparison of Val-mAP@20 Scores Across Different Augmentation Techniques: This graph illustrates the validation mAP@20 scores for three augmentation techniques: Clean, ColorJitter, and DepthJitter. The mAP@20 score, which measures the model's average precision at an intersection over union 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 train Query2Label image classification model for category classification beside object detection models. Query2Label is currently one of the best performing models for image classification. It performs better than the object detection models we tried. As our one of the objectives is to identify categories in images we use Query2Label classification model which seems more reliable than object detection models. We use this model in different settings with different backbones, and augmentation techniques. The training image size is 384×384 pixels and the test image size is 640×640 pixels. The reason for different size in images during training and testing is because this gives better result and this is one of the criteria

of the backbones like ResNest and TResnet we used from timm(huggingface). Each configuration is tested five(5) times with different seeds and then averaged to make experiment more 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 Clean augmentation 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 used a learning rate of 0.0001 and a weight decay of 0.05. These hyperparameters were chosen to ensure a balanced optimization process, where the learning rate dictates the step size during gradient descent, and the weight decay helps in preventing overfitting by adding a regularization term to the loss function. However, we observed that in some instances, the training and validation loss could explode, indicating instability in the training process. To mitigate this, it is advisable to reduce the learning rate further. Lowering the learning rate allows the model to make smaller, more controlled updates to the weights, which can help stabilize training.

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	0.089	0.762	0.827
Depth Jitter(Ours)	ResNest101e	384X384	640X640	ASL	0.165	0.223	**0.803**	**0.855**

Table 5.2: Performance of Different Multi-label Classification Models on the FathomNet Dataset

Alternatively, gradient clipping can be employed to address the issue of exploding gradients. Gradient clipping involves setting a threshold value for the gradients; if the gradients exceed this value, they are scaled down to a manageable level. This technique ensures that the model's parameters are updated in a stable and consistent manner, preventing drastic changes that could destabilize training.

We also utilized a one-cycle learning rate scheduler with a cosine anneal strategy. The one-cycle learning rate policy initially increases the learning rate from a minimum to a maximum value and then decreases it back to the minimum. This approach helps the model to escape local minima early in training and fine-tune the parameters more effectively towards the end. The cosine anneal strategy ensures a smooth transition in the learning rate changes, which aids in maintaining training stability.

For optimization, we employed the AdamW optimizer, which combines the benefits of Adam's adaptive learning rate with weight decay regularization. We set the momentum parameter to 0.9, which helps in accelerating gradient vectors in the right directions, thus leading to faster converging.

Furthermore, the superior classification results achieved with our proposed augmentation technique also positively impact the out-of-distribution (OOD) score. The OOD score measures the model's ability to

identify data that significantly deviates from the training distribution, which is crucial for robust performance in real-world applications. We will delve deeper into the implications of these results on the OOD score later in this section.

In summary, our proposed augmentation technique significantly enhances the performance of the Query2Label model on the FathomNet dataset, leading to better generalization and more reliable classification results. The combination of appropriate learning rate adjustments, gradient clipping, and advanced optimization strategies like the one-cycle learning rate scheduler and AdamW optimizer contributes to the robustness and effectiveness of the training process. These improvements not only boost the validation mAP@20 but also strengthen the model's ability to handle out-of-distribution data effectively.

5.1.7 OOD Score Performance

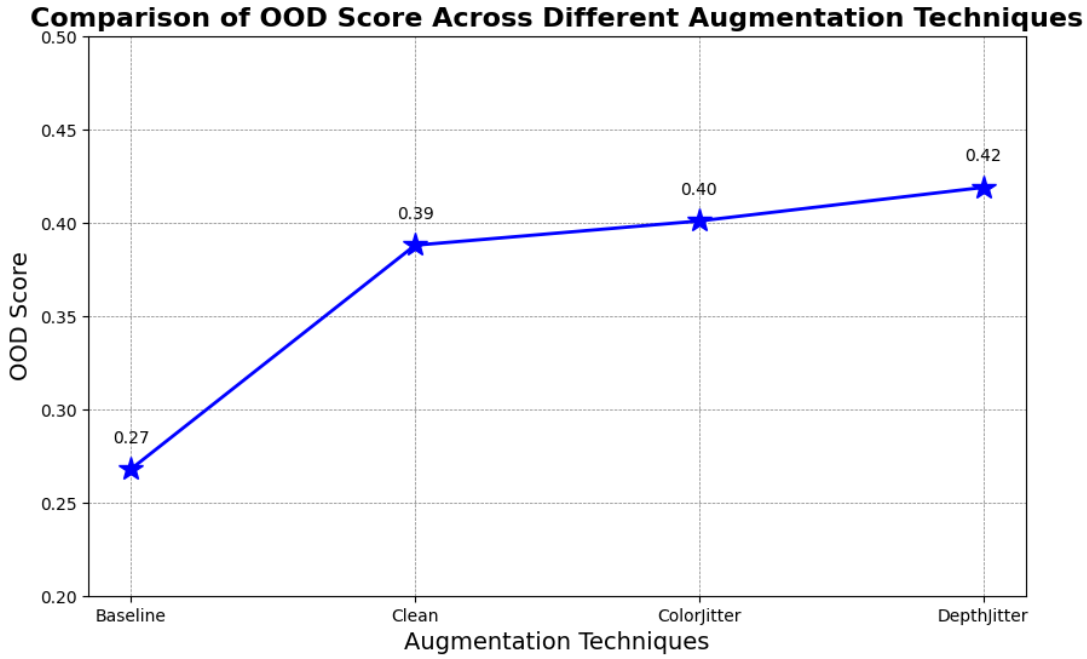


Figure 5.2: Comparison of Out-of-Distribution (OOD) Scores Across Different Augmentation Techniques: This graph depicts the OOD scores for four augmentation techniques: Baseline, Clean, ColorJitter, and DepthJitter. The OOD score, representing the model's ability to handle out-of-distribution data, is 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. The upward trend indicates 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 [43] used YOLOv8 for classification and the average confidence score technique for calculating ood score. The other techniques in the above figure use the Query2label [15] for classification and maximum softmax probability technique for ood score calculation. Both of these ood score calculation techniques have been discussec 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

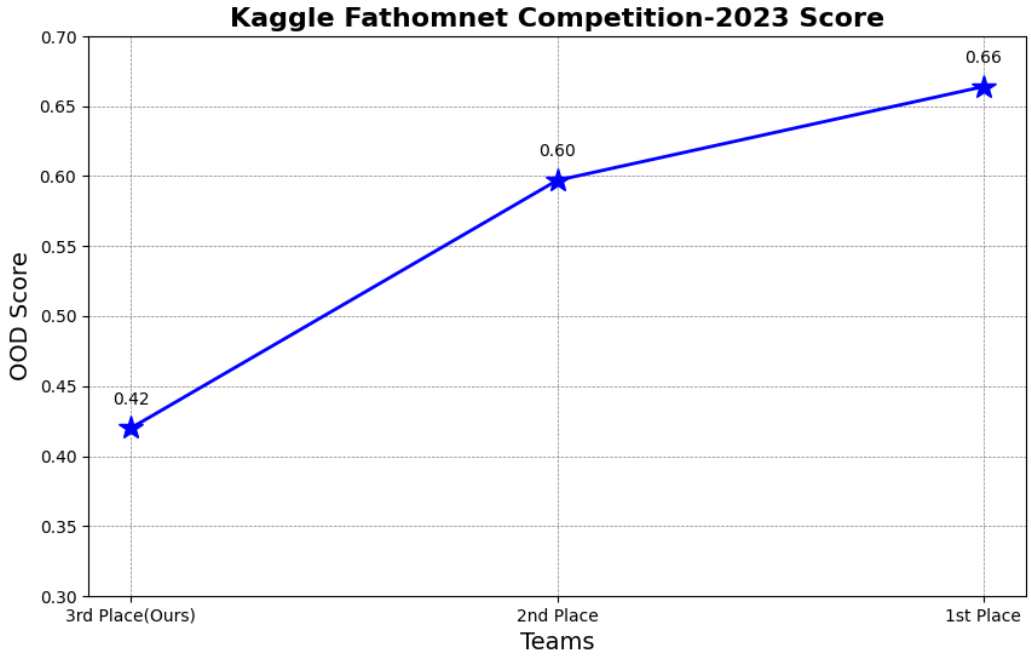


Figure 5.3: Comparison of OOD Scores for Top Teams in Kaggle Fathomnet Competition-2023: This graph presents the out-of-distribution (OOD) scores for the top three teams in the Kaggle Fathomnet Competition-2023. The OOD score, displayed on the y-axis, is a measure of the model's ability to identify out-of-distribution 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 to make their datasets more balanced, resulting in higher scores. In contrast, we worked exclusively with the provided dataset without incorporating external data, and our score is quite respectable considering this constraint.

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)

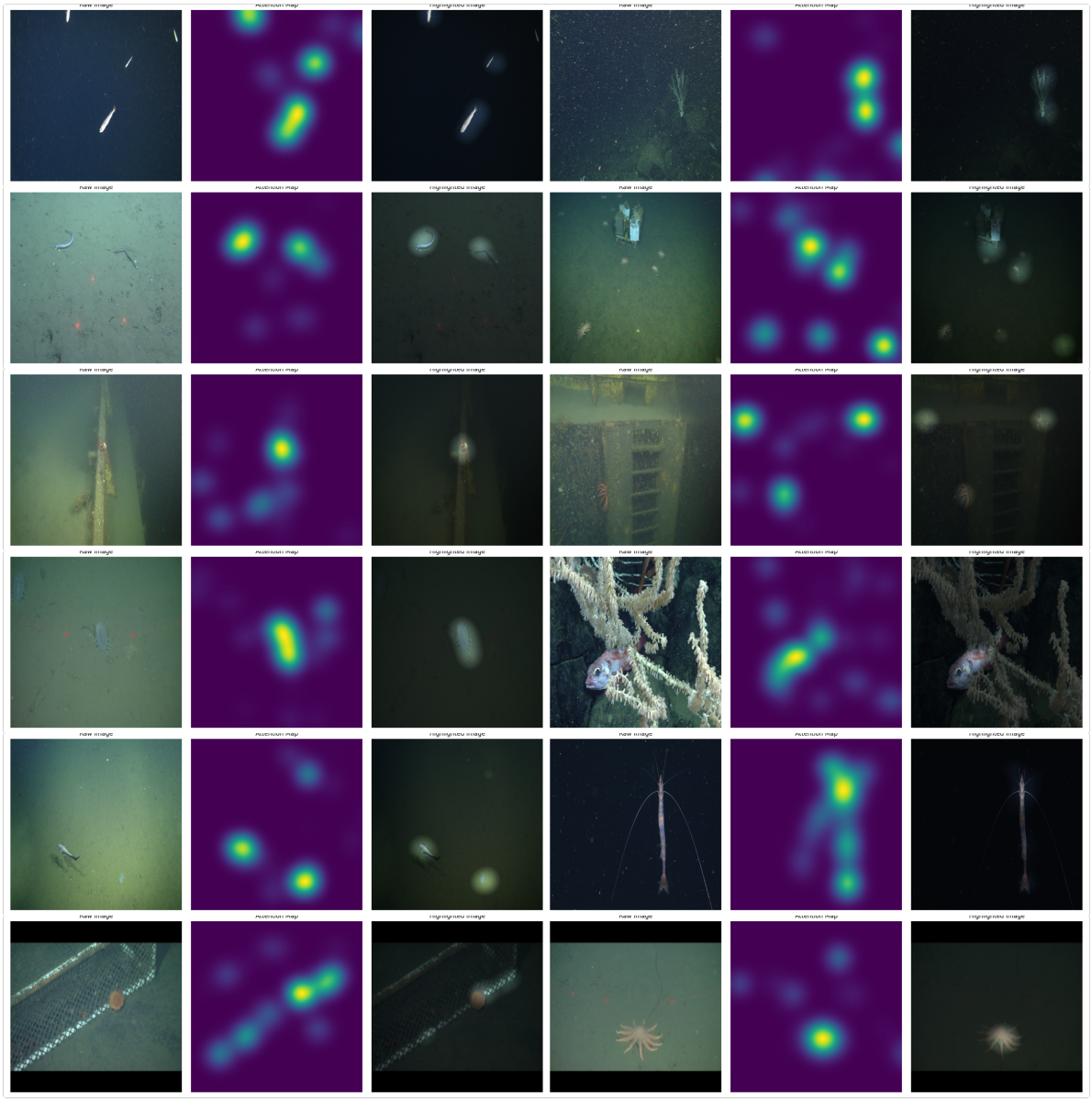


Figure 5.5: Attention maps generated by Query2label [15].

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, specially during the training phases. The deep learning models such as Query2Label and YOLOv9 requires high performance hardwares which may not be always available in all deployment environments. Additionally, the training process is time consuming and the finetuning process may take substantial computational resources.

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 stuck in the lower and upper bound. Because of that, if we add depth offset there is no change in the color of the image. 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 Works

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

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

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

[16] 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

[17] 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

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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/

[25] 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

[26] 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

[27] 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/

[28] 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/

[29] 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

[30] 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

[31] 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

[32] 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

[33] 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

[34] 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

[35] 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.

[36] 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

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

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

[39] 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

[40] 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

[41] 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

[42] 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.

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

[44] 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

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

[46] 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.

[47] ——, "A revised underwater image formation model - akkaynak treibitz cvpr 2018," 06 2018.

[48] 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.

[49] 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/

[50] 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

[51] 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/

[52] 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.

[53] 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

Appendix

More visualization of the Depth Jitter

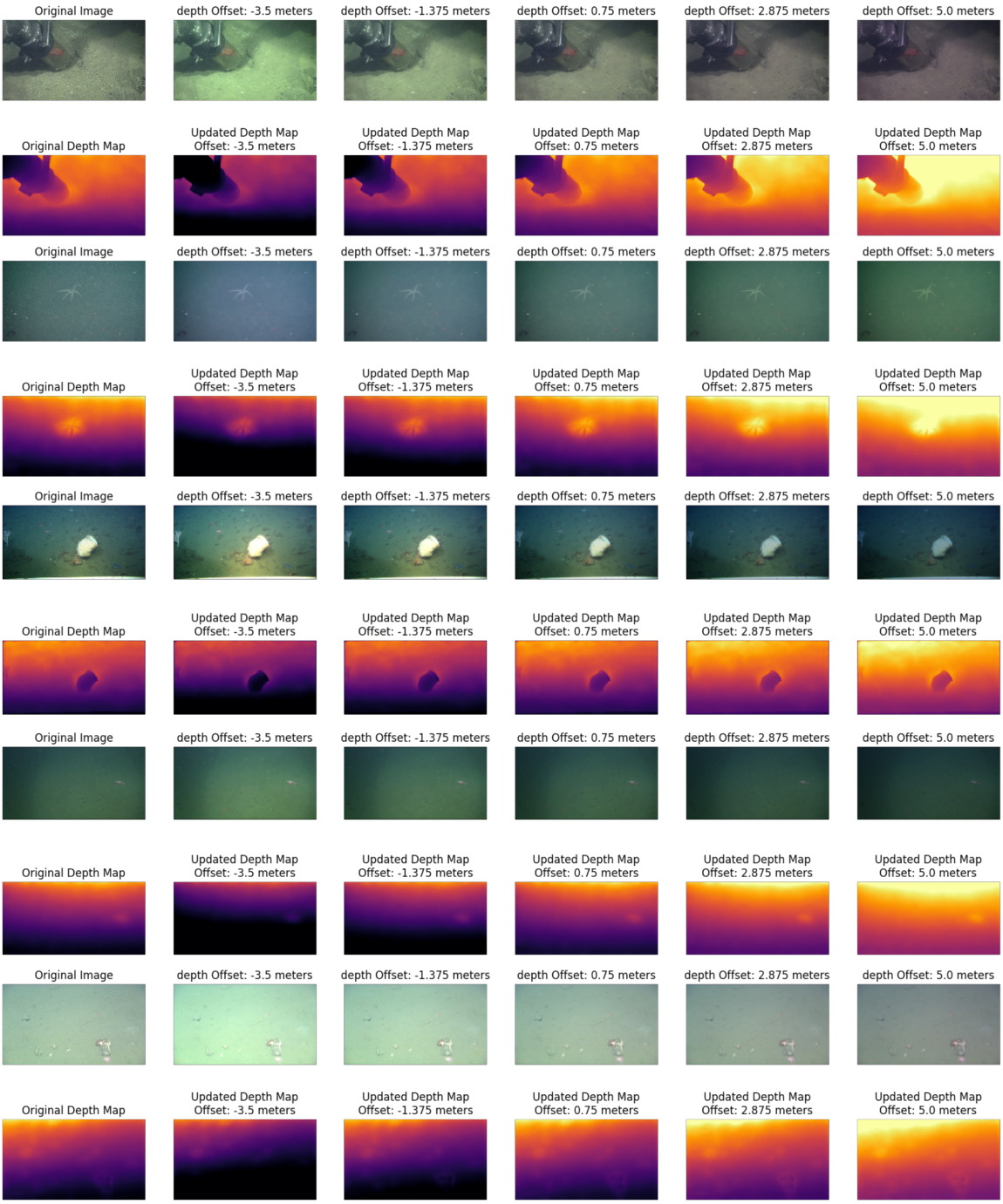


Figure A.1: Visualization of the Depth Jitter Augmentation Technique.