

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Learning-based Adaptive Feature Fusion for Sacrificial Anode Detection in Harsh Subsea Conditions

Alexandre Vilhena da Costa



Mestrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Prof. Andry Maykol Pinto

August 1, 2025

Resumo

As operações em turbinas eólicas *offshore* enfrentam desafios críticos na manutenção da sua estrutura subaquática, onde a monitorização de ânodos de sacrifício é essencial para prevenir falhas induzidas pela corrosão. Métodos de inspeção tradicionais que utilizam mergulhadores humanos ou veículos operados remotamente apresentam riscos de segurança significativos e custos operacionais, enquanto as soluções autónomas são afetadas por severa degradação visual subaquática devido à interação da luz com o meio e à presença de matéria orgânica.

Esta investigação aborda estas limitações através de uma arquitetura de fusão adaptativa de características multimodais especificamente desenhada para condições subaquáticas adversas. A metodologia proposta consiste numa *pipeline* de pré-processamento melhorado usando Contrast Limited Adaptive Histogram Equalization (CLAHE) e aumento de dados para contrariar artefactos visuais subaquáticos, uma nova arquitetura de modelo chamada Contextual Anode Locator in Varying Underwater Scenarios (CALVUS), que tira partido da extração de características com consciência de profundidade adaptando características do Transformador Visual do ML-Depth Pro para deteção de objetos subaquáticos, e MF-CALVUS (Multimodal Fusion-CALVUS), que incorpora uma fusão multimodal de características baseadas em profundidade e baseadas em textura usando um mecanismo de atenção multi-cabeça e uma soma ponderada adaptativa de ambas as modalidades.

Validação experimental extensiva nos conjuntos de dados ANODE e NEREON demonstra avanços significativos de desempenho em relação aos métodos de estado da arte. O CALVUS alcança uma melhoria de 12.6 pontos percentuais na Precisão Média (AP) em relação ao YOLO11x e um aumento de 19.9 pontos percentuais no AP@50 (0.979) mostrando os efeitos da utilização de características baseadas em profundidade em oposição às baseadas em textura e uma imagem de maior resolução. Foi realizado um estudo de ablação na arquitetura do bloco de fusão multimodal do MF-CALVUS, com o melhor modelo a apresentar um aumento de 3.3 pontos percentuais no AP sobre o CALVUS (0.782), uma precisão de 99.89% e um recall de 99.48%, enquanto reduziu as deteções de falsos negativos em 58.7%. A validação em ambiente real no *ATLANTIS Test Center* confirmou o desempenho robusto do MF-CALVUS em diversas condições desafiadoras, incluindo iluminação variável, neve marinha e bioincrustação intensa.

Abstract

Offshore wind turbine operations face critical challenges in underwater structural maintenance, where sacrificial anode monitoring is essential for preventing corrosion-induced failures. Traditional inspection methods using human divers or remotely operated vehicles present significant safety risks and operational costs, while autonomous solutions are affected by severe underwater visual degradation due to the light interaction with the medium and the presence of organic matter.

This research addresses these limitations through a multimodal adaptive feature fusion architecture specifically designed to address harsh subsea conditions. The proposed methodology consists on an enhanced pre-processing pipeline using Contrast Limited Adaptive Histogram Equalization (CLAHE) and data augmentation to counteract underwater visual artifacts, a new model architecture named Contextual Anode Locator in Varying Underwater Scenarios (CALVUS), which takes advantage of depth-aware feature extraction by adapting ML-Depth Pro Vision Transformer features for underwater object detection and MF-CALVUS (Multimodal Fusion-CALVUS), incorporating a multimodal fusion of depth-based and texture-based features using a multi-head attention mechanism and an adaptive weighted sum of both modalities.

Extensive experimental validation on the ANODE and NEREON datasets demonstrate significant performance advances over state-of-the-art methods. CALVUS achieves an improvement of 12.6 percentage points in Average Precision over YOLO11x and 19.9 percentage points in AP@50 (0.979) showcasing the effects of using depth-based features as opposed to texture-based and an higher resolution input. An ablation study was done on the multimodal fusion block architecture of MF-CALVUS, with the best model having a 3.3 percentage point increase in AP over CALVUS (0.782), a precision of 99.89% and recall of 99.48%, while reducing false negative detections by 58.7%. Real-world validation at ATLANTIS Test Center confirmed the robust performance of MF-CALVUS across diverse challenging conditions including varying illumination, marine snow, and heavy biofouling.

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework to achieve a better and more sustainable future for all. It includes 17 goals to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice.

The specific Sustainable Development Goals mentioned have the following names:

SDG 7 Ensure access to affordable, reliable, sustainable and modern energy for all

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation

SDG	Target	Contribution	Performance Indicators and Metrics
7	7.2	The development of autonomous monitoring systems for offshore wind turbines improves the reliability and scalability of renewable energy infrastructure, thereby supporting a larger share of renewables in the energy mix.	Increased operational uptime of offshore wind farms; reduction in manual inspection needs; contribution to renewable energy capacity growth.
	7.a	This research advances clean energy technologies by introducing robust computer vision solutions for harsh underwater environments, fostering international cooperation in marine renewable energy innovation.	Number of scientific contributions to offshore renewable tech; adoption or testing of the developed system in real-world projects or testbeds.
9	9.4	By enabling efficient, autonomous maintenance of offshore wind infrastructure, this work supports the upgrade of energy infrastructure using sustainable and resource-efficient technologies.	Reduction in inspection costs; improved maintenance cycles;

Acknowledgements

Este trabalho é o culminar de cinco anos de trabalho árduo e muitas noites sem dormir, mas nada disto teria sido possível sem o apoio e a ajuda incondicional dos meus amigos e da minha família.

Ao meu orientador, Andry Pinto, o meu profundo agradecimento pela oportunidade dada, bem como por todo o conhecimento partilhado ao longo do mestrado. Obrigado pelas suas palavras de confiança nas reuniões, que me levaram a acreditar mais em mim e nas minhas capacidades. Ao meu orientador, Pedro Leite, o meu obrigado pela ajuda inestimável que me deste desde o primeiro dia em que cheguei. O teu apoio constante, as críticas construtivas e piadas, por muito más que algumas fossem, tornaram esta tese possível, mesmo tendo que ter esperado pela última semana para um TDC teu. Tenho um enorme orgulho em dizer que, para além de um orientador, ganhei uma amizade que espero que perdure por muitos anos e muitas cervejas.

À equipa MOTUS, um enorme obrigado pela constante disponibilidade para ajudar. A ética de trabalho e os convívios deste grupo mudaram por completo a minha visão do futuro. Um agradecimento especial ao Celso Pereira pela partilha indispensável de conhecimentos em visão por computador, e à Maria Pereira pela amizade e ajuda contínua, principalmente nos testes em Viana. Aos meus amigos e colegas de tese, Francisco Silva, Tiago Claro e Bruno Costa, o meu obrigado pelas constantes piadas, ajuda, assobios e provocações, mas, principalmente, pela vossa companhia que tornou estes últimos seis meses bem mais fáceis de gerir.

Aos meus amigos do NEEEC, da SEDE e de Erasmus, obrigado por estes cinco anos e por os terem tornado nos melhores da minha vida. Não escolheria ninguém diferente para ter feito esta jornada.

Um obrigado especial aos meus pais, que me incentivaram, apoiaram e ajudaram durante toda a minha vida, sendo sempre os primeiros a celebrar as minhas conquistas e a garantir que nenhum entrave impedisse o meu sucesso. E ao meu irmão, André, que foi o meu *role model* durante toda a minha vida, sempre capaz de lidar com todos os meus defeitos e, mesmo assim, manter um sorriso no rosto.

À pessoa mais importante, ao amor da minha vida, Sofia, foste a minha pedra nestes seis meses. A pessoa a quem recorri sempre que estava mais em baixo ou inseguro de algo e, em todos os momentos acreditaste sempre mais em mim do que eu próprio. O que fizeste por mim não pode ser quantificado e estou eternamente grato por te ter ao meu lado.

Alexandre Costa

Official Acknowledgements

This work is funded by the European Union's Horizon Europe research and innovation programme, under the Grant Agreement No. 101189723 (AEROSUB).

*“Whatever you do in this life,
it’s not legendary unless your friends are there to see it.”*

Barney Stinson

Contents

1	Introduction	1
1.1	Context	1
1.2	Motivation	2
1.3	Objectives	3
1.4	Contributions	3
1.5	Document Structure	4
2	Literature Review	5
2.1	Challenges of Underwater Perception	5
2.2	Underwater Perception Systems	8
2.3	2D Object Detection methodologies	12
2.3.1	CNN-based Architectures	13
2.3.2	Transformer and Attention-Driven Detection Architectures	16
2.3.3	Underwater Object Detection	18
2.4	Critical Review	20
3	Adaptive Feature Fusion for Robust Underwater Anode Detection	22
3.1	Introduction	22
3.2	Experimental Setup and Problem Definition	23
3.2.1	Dataset Creation	23
3.2.2	Problem Definition	23
3.3	CALVUS Architecture Design	25
3.3.1	Pre-processing Pipeline	26
3.3.2	Depth-Aware Feature Extraction Backbone	28
3.3.3	Detection Head Architecture	29
3.4	MF-CALVUS Architecture Design	29
3.5	Training Methodology	32
4	Experimental Results	36
4.1	Performance Metrics for Detection Task	36
4.2	Comparing Uni-modality CALVUS with the State of the art	38
4.3	Multimodal Adaptive Fusion for anode detection	41
4.3.1	Ablation Study of the Multimodal Adaptive Fusion	42
4.3.2	Comparison with CALVUS	43
4.3.3	Real World Validation	46
5	Conclusion	49
5.1	Future Work	50

List of Figures

2.1	Simulation of a light wave traveling through water at different depths.	6
2.2	A visual representation of the two main sub-sea phenomena affecting underwater optical systems.	7
2.3	The MARESyE hybrid imaging sensor.	12
2.4	A typical Convolutional Neural Network structure.	13
2.5	R-CNN structure.	14
2.6	Self-Attention and Multi-Head Attention Mechanisms.	17
3.1	Samples from the ANODE and NEREON datasets.	24
3.2	CALVUS architecture overview.	25
3.3	Comparison of images before and after CLAHE.	26
3.4	Representation of the augmentation techniques.	28
3.5	MF-CALVUS architecture overview.	30
3.6	Fusion block architecture.	32
3.7	CALVUS training loss and validation loss progression.	35
3.8	MF-CALVUS training loss and validation loss progression.	35
4.1	Comparison of CALVUS, YOLO11x and YOLO12x detections in varying scenarios.	40
4.2	Confusion matrices highlighting classification results on the test dataset for YOLO11x, YOLO12x and CALVUS.	41
4.3	MF-CALVUS fusion block architecture with highlighted components evaluated in the ablation study.	42
4.4	Confusion matrices highlighting classification results on the test dataset for CALVUS and MF-CALVUS.	43
4.5	Comparison of CALVUS and MF-CALVUS detections in varying scenarios.	45
4.6	Overview of the real-world validation setup.	47
4.7	Sample images from real-world validation, illustrating varying illumination conditions.	47
4.8	Comparison of CALVUS and MF-CALVUS on the new acquired data.	48

List of Tables

3.1	Training Parameters and Configuration.	34
4.1	Confusion matrix for a binary classification task.	37
4.2	Performance results of the proposed CALVUS model compared to baseline object detection architectures (YOLO and RT-DETR-L) on the real-world dataset. . . .	39
4.3	Performance results extracted from the confusion matrices.	41
4.4	Ablation study results of MF-CALVUS fusion block configurations.	43
4.5	Performance results extracted from the confusion matrices comparing both CALVUS variations.	44

Abbreviations

2D Two Dimensional

3D Three Dimensional

AUV Autonomous Underwater Vehicle

CALVUS Contextual Anode Locator in Varying Underwater Scenarios

CLAHE Contrast Limited Adaptive Histogram Equalization

CNN Convolutional Neural Network

DL Deep Learning

FEUP Faculty of Engineering of the University of Porto

FFN Feed-Forward Network

FPN Feature Pyramid Network

IoU Intersection over Union

GIoU Generalized Intersection over Union

IR Infrared

IoU Intersection over Union

LSR Light Stripe Range

mAP Mean Average Precision

MF-CALVUS Multimodal Fusion - Contextual Anode Locator in Varying Underwater Scenarios

MHA Multi-Head Attention

NN Neural Network

O&M Operations and Maintenance

OWT Offshore Wind Turbine

R-CNN Region-based Convolutional Neural Network

RGB Red Green Blue

RoI Region of Interest

ROV Remotely Operated Vehicle

RPN Region Proposal Network

RT-DETR Real Time DEtection TRansformer

Sonar Sound Navigation and Ranging

SVM Support Vector Machine

ViT Vision Transformer

YOLO You Only Look Once

Chapter 1

Introduction

1.1 Context

The global transition towards sustainable energy has led to the creation and rapid expansion of Offshore Wind Turbines (OWTs), which have become a key factor to reduce the use of fossil fuels and climate change. Offshore wind energy leverages the stronger and more consistent wind patterns over the ocean, making it more efficient and reliable source of renewable energy compared to land-based turbines [1]. Europe has been at the front seat when it comes to OWT expansion¹ with the United States and China following closely as they try to meet the goals set by institutions like the European Union² or the United Nations³.

However, the construction of these turbines presents unique challenges, particularly regarding the durability and maintenance of their underwater foundations. These components are constantly exposed to harsh marine conditions, including strong currents, waves, and varying levels of salinity, which lead to material degradation through processes such as corrosion, biofouling, and mechanical stress. The failure of OWT foundations can result in significant economic losses and operational downtimes, making cost-effective and reliable monitoring systems essential for sustaining offshore wind energy infrastructure [2]. One of the most common methods for mitigating corrosion in the underwater foundations of offshore wind turbines is the use of sacrificial anodes as part of a cathodic protection system. Sacrificial anodes are typically composed of materials such as zinc, aluminum, or magnesium, which have a more negative electrochemical potential compared to the steel foundations they protect [3]. When immersed in seawater, these anodes corrode preferentially, shielding the underwater structures from corrosion. While sacrificial anodes are an effective and widely used solution, their degradation must be monitored to ensure continued

¹Wind Industry in Germany, *Offshore Expansion in Europe at Record Level*, 2023, accessed September 23, 2024, <https://www.windindustry-in-germany.com/announcements/offshore-expansion-in-europe-at-record-level>.

²European Commission, *Renewable Energy Targets*, 2024, accessed September 23, 2024, https://energy.ec.europa.eu/topics/renewable-energy/renewable-energy-directive-targets-and-rules/renewable-energy-targets_en.

³United Nations, *Affordable and Clean Energy - Goal*, 2023, accessed September 24, 2024, <https://www.un.org/sustainabledevelopment/energy/>.

protection [4]. This involves assessing the remaining volume of the anodes and their rate of consumption, which provides indirect insight into the environmental conditions and the overall health of the OWT foundations.

The effective management of offshore wind turbines hinges on robust Operations and Maintenance (O&M) practices, traditional inspection methods, such as deploying human divers or Remotely Operated Vehicles (ROVs), are increasingly inadequate due to their high costs, safety risks, and limitations in accessing deeper and more remote installations [5]. As wind farms grow in scale and move further offshore, these logistical challenges tend to intensify. To address these issues, the adoption of real-time advanced monitoring technologies, such as autonomous systems, is becoming essential [6–9]. However, despite their potential, these technologies also face significant challenges, mainly the degradation on the quality of the images collected on underwater due to the harsh conditions of the environment, the lack of datasets for this applications due to the inherent difficulties of the acquisition and the difficulty of current state of the art to thrive in such conditions [10].

The assessment of sacrificial anodes, which traditionally involves direct observation by human divers or subsequent analysis of images captured by commercial cameras, relies on visual data that is severely affected by the very challenges previously discussed. These factors significantly complicate accurate and timely anode assessment, thereby underscoring the urgent need for more robust and autonomous perception solutions.

1.2 Motivation

One of the primary challenges in developing robust underwater structure detection systems is the critical shortage of high-quality annotated datasets. Existing underwater anode datasets are either non-existent or characterized by poor image quality, which significantly limits the development and validation of advanced computer vision algorithms for marine applications. The scarcity of high-quality annotated underwater data has consequently limited the amount of dedicated research in underwater object detection, particularly for structures such as sacrificial anodes. Existing works in this domain typically employ off-the-shelf models, such as You Only Look Once (YOLO) and Real Time DEtection TRansformer (RT-DETR), without substantial modifications. This often results in suboptimal performance compared to their effectiveness in conventional environments. A key reason for this performance discrepancy lies in the reliance of these algorithms on image texture for feature extraction. In the context of sacrificial anodes, heavy biofouling accumulation obscures the distinct texture of the anodes, making them difficult to differentiate from the surrounding background. Furthermore, the diminished light conditions and the monochromatic nature of acquired underwater images further hinder performance by reducing the inherent contrast between various objects in the medium.

To address this fundamental limitation, this work leverages a comprehensive dataset acquired at the ATLANTIS Test Center [11] and FEUP, using state-of-the-art imaging systems such as

MARESyE [12–15] and the FIFISHv6 ROV⁴. However, even with advanced imaging technology, underwater data acquisition remains inherently challenging due to environmental factors including reduced visibility, water turbidity, and marine snow phenomena, all of which significantly degrade image quality and compromise detection performance. These adverse conditions represent fundamental obstacles that must be overcome to achieve reliable automated inspection systems in real-world marine environments.

Consequently, this dissertation aims to develop and implement a pre-processing pipeline specifically designed to enhance image quality under such challenging conditions, alongside the development of methods to foster the model’s insensitivity to these difficulties. To achieve this, this dissertation proposes an approach that leverages a backbone pre-trained on depth features, diverging from conventional texture-based feature extraction. This work introduces a multimodal framework that integrates both texture-based and depth-based features to enhance detection accuracy. This is achieved through a multi-scale feature fusion strategy between the two distinct backbones, thereby developing a robust method for effectively detecting sacrificial anodes in challenging underwater conditions.

1.3 Objectives

This thesis aims to develop a robust solution for anode detection in harsh underwater environments, thereby enabling the transition from manual to autonomous O&M. The objectives of this work are:

- Explore the advantages of using multimodal features, namely depth and texture, to improve the accuracy and robustness of sacrificial anode detection. Develop and evaluate fusion methodologies, particularly those incorporating attention mechanisms, to effectively integrate these distinct data streams, thereby establishing an approach for reliable sacrificial anode detection in challenging underwater conditions.
- Validate the developed methodology through a dataset acquired at the ATLANTIS Test Center [11] under harsh real-world conditions. This tests will assess the performance of the proposed approach against current state-of-the-art to demonstrate its improved accuracy and real-world applicability.

1.4 Contributions

This section highlights the main contributions of this dissertation for the improvement of the detection of sacrificial anodes in harsh sub-sea conditions. The key contributions consist in:

- A pre-processing pipeline incorporating CLAHE [16] to enhance image quality and mitigate underwater-specific visual artifacts, as well as different augmentation techniques to mimic the challenges of underwater image capture.

⁴<https://store.qysea.com/fifish-v6.html>. Accessed Jun 30, 2025.

- The development of an approach which adapts a depth estimation ViT backbone [17] for underwater object detection through strategic transfer learning and integrates an RT-DETR-based detection head [18] optimized for the high-dimensional feature representations generated by the ViT encoder.
- The development of a multimodal adaptive fusion methodology which leverages both depth-based and texture-based features to improve the detection of sacrificial anodes in harsh underwater conditions.

From this research the following publication was accepted to be presented on the IEEE 30th International Conference on Emerging Technologies and Factory Automation (ETFA)⁵:

- A. V. Costa, P. N. Leite and A. M. Pinto, "Robust Vit-enhanced Detection of Sacrificial Anodes in Harsh Underwater Conditions", IEEE 30th International Conference on Emerging Technologies and Factory Automation (ETFA), 2025.

1.5 Document Structure

The present document is organized into four main chapters, each focusing on different aspects of the research and development of the learning-based adaptive feature fusion approach for sacrificial anode detection. Chapter 2 provides a comprehensive literature review relevant to this thesis. It starts with an overview of the challenges inherent to underwater environments, specifically concerning anode detection. Subsequently, the chapter reviews existing perception systems and current methodologies for 2D object detection, progressively narrowing the focus to prior work in underwater settings. The chapter concludes with a critical review of the current state of the art, thereby identifying the research gaps and challenges addressed by the present work.

Chapter 3 initiates by delineating the dataset characteristics and the precise problem definition addressed by this thesis. The chapter then proceeds to detail the architectural design of both proposed models, CALVUS and MF-CALVUS. Chapter 4 starts by stating the performance metrics employed for evaluating the proposed models. The chapter then presents a comparative analysis of CALVUS's detection results against current state-of-the-art approaches. Furthermore, it analyzes the performance gains achieved by the MF-CALVUS approach relative to CALVUS. Finally, the chapter demonstrates the performance of the proposed models on unseen data acquired during the development of this thesis. Finally, Chapter 5 synthesizes the main achievements and findings of this work. It offers a critical discussion on the advancements made in underwater sacrificial anode detection and their practical relevance for offshore operations, followed by a discussion on potential future research.

⁵<https://etfa2025.ieee-ies.org/>. Accessed Jun 30, 2025.

Chapter 2

Literature Review

This chapter provides a comprehensive review of existing literature relevant to the detection and localization of subsea structures using Two Dimensional (2D) data. It starts by exploring the fundamental challenges of underwater perception on Section 2.1, focusing on the environmental and technological constraints that influence detection accuracy in marine environments. Then, Section 2.2 examines the evolution of underwater perception systems, highlighting key advancements in sensing modalities and computational techniques specifically applied to structural monitoring and inspection tasks. The discussion covers both traditional sensor configurations and emerging multi-modal approaches that have been developed to address the unique requirements of subsea structure assessment.

Section 2.3 focuses on 2D detection methodologies, analyzing computer vision and machine learning approaches for identifying, classifying, and locating subsea structures from visual and sonar data. This section examines both classical image processing techniques and modern deep learning methods, with particular attention to their effectiveness in handling the complexity and environmental variability encountered in subsea structural inspection.

Finally, a critical review in Section 2.4 identifies key findings from the literature, highlighting gaps, limitations, and opportunities for future research in 2D-based subsea structure detection and localization. The chapter establishes a foundation for understanding current capabilities and limitations in this field, underscoring the importance of developing more effective 2D detection approaches for subsea structural applications.

2.1 Challenges of Underwater Perception

Underwater perception presents a range of unique difficulties that extend beyond traditional computer vision challenges, primarily due to the complex physical properties of marine environments. These challenges are particularly pronounced in subsea structure detection and localization tasks, where environmental factors can significantly compromise data quality and system performance.

As discussed and reviewed by Zhou *et al.* (2023) [19], the underwater domain introduces both optical and acoustic limitations that require specialized approaches for effective structural monitoring and inspection.

The most fundamental optical challenges in underwater environments stem from light interaction with water and suspended particles which lead to light absorption and scattering. Light attenuation depends on wavelength, dissolved organic compounds, water salinity, and the concentration of organic matter in the water [20]. Depending on the depth of operation, the different wavelengths are absorbed at different rates, with red light being almost completely absorbed after 5 meters of depth [21], while blue and green wavelengths penetrate deeper, with blue light maintaining visibility up to approximately 60 meters in clear water conditions, which leads to the blue and green hue seen on underwater images [22]. This difference in attenuation can be seen in Figure 2.1, where a graph shows the difference in spectral radiance for the different wavelengths at varying depths. This selective absorption also leads to the poor performance of Infrared (IR)-based technologies as the longer wavelengths of such technologies are attenuated after just 20 cm [23]. The natural illumination in this scenarios is also a concern since after a certain depth it becomes insufficient. This creates a need to use artificial light sources which if the setup is not done correctly can lead to a non-uniform illumination, causing problems such as vignetting, contrast and brightness inconsistencies [24].

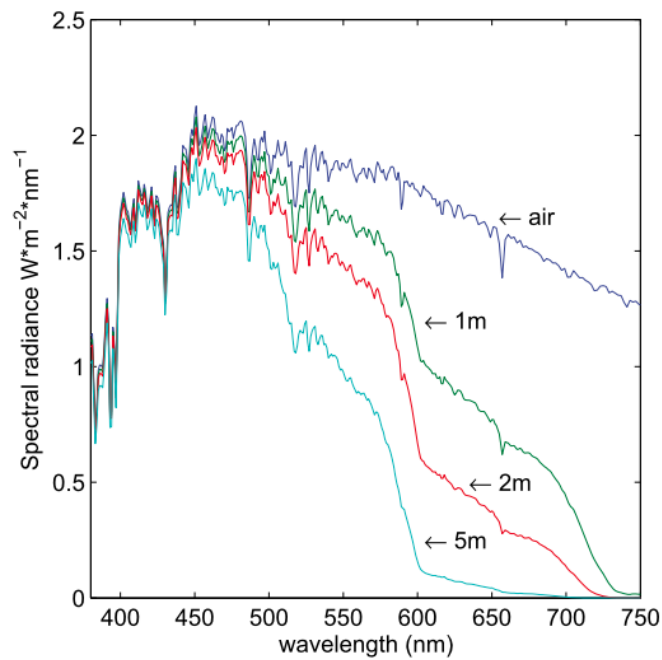


Figure 2.1: Simulation of a light wave traveling through water at different depths. Colors drop one by one, starting with the longest wavelengths. The absorption rate varies with water conditions, accelerating in high-salinity or turbid waters. Both suspended particles and dissolved organic matter further influence the absorption process. Extracted from [21].

Light scattering, caused by suspended particles in the water, known as marine snow, results in reduced contrast and blur on the captured images. The concentration of such particles increases

the water turbidity and leads to light being deviated from the imaging sensor [25]. Scattering can be classified as forward scattering, which happens when the light that reflects on the observed object is scattered which leads to a blurring effect, and backscattering, which is caused by the light reflected on the suspended particles and captured by the sensor which reduces contrast and visibility [26].

In Figure 2.2, it can be seen a visual representation of the aforementioned problems, cementing the problems caused by both absorption and scattering. These challenges are caused by the presence of marine snow which lead to the degradation of the image captured by the optical system.

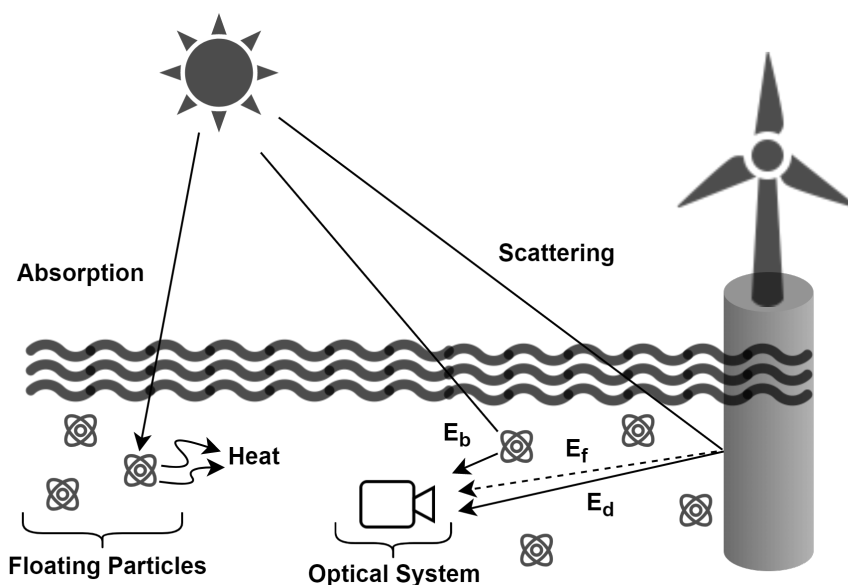


Figure 2.2: A visual representation of the two main sub-sea phenomena affecting underwater optical systems: absorption, where the energy of light is dissipated as heat when interacting with molecules in the medium, and scattering, where light deviates from the target scene and is scattered onto the sensor. Extracted from [27].

Software-based mitigation strategies have emerged to diminish the effect of this sub-sea phenomena. Dehazing and color correction algorithms compensate for selective light absorption by restoring natural color balance and enhancing contrast in underwater images. These techniques adjust blue-green dominated images to recover lost spectral information and improve the effectiveness of detection algorithms originally designed for full-spectrum environments [16, 28, 29]. The presence of biofouling are also a challenge for long-term structural monitoring. Biological growth on subsea structures can obscure original structural features and change the surface texture that complicate the task of detection algorithms. The rate and pattern of biofouling vary significantly with location, depth, season, and structure composition, making it difficult to develop universally applicable detection methodologies [30]. These challenges underscore the complexity of developing reliable underwater perception systems for subsea structure detection and localization. Effective solutions must integrate multiple sensing modalities and robust software approaches to

overcome the inherent limitations of the underwater environment while maintaining the accuracy and reliability required for critical infrastructure monitoring applications.

2.2 Underwater Perception Systems

Underwater perception systems have evolved to address the challenges mentioned in Section 2.1, enabling the acquisition, processing, and interpretation of sensory data for diverse applications, including navigation, object detection, and environmental monitoring [5, 31, 32]. These systems integrate various sensors and algorithms to overcome the limitations of optical and acoustic data in underwater settings. Gathering high-quality, multimodal data is crucial to improve the reliability and precision of detection systems, as it allows for more accurate environmental analysis.

Acoustic Systems

Sound-based sensors are the most common underwater detection modality due to the superior propagation characteristics of sound in water. Acoustic waves can propagate over considerable distances underwater with relatively low energy loss [33].

Sound Navigation and Ranging (Sonar) based systems represent the most established acoustic sensing technology for underwater applications. These active systems emit controlled acoustic pulses and analyze the returned echoes to extract information about structural geometry, material properties, and environmental conditions. The effectiveness of Sonar for structure detection stems from its ability to operate effectively across the full range of underwater conditions encountered in subsea environments. Single-beam Sonar systems provide the simplest acoustic sensing approach, emitting focused acoustic pulses directly beneath the sensor platform. While suitable for basic depth measurements and simple structural profiling, these systems offer limited spatial coverage and are primarily used for preliminary surveys or specific point measurements on structures [34]. Their narrow beam pattern restricts their effectiveness for comprehensive structural assessment. Side-scan Sonar systems address coverage limitations by emitting acoustic energy in a fan-shaped pattern perpendicular to the sensor's direction of travel. This configuration enables wide-area mapping of structural elements and seafloor features, making it particularly effective for detecting large subsea installations, pipeline routes, and overall structural layouts [35]. The acoustic shadows created by structural elements provide valuable information about their three-dimensional characteristics, though resolution limitations restrict detailed surface analysis.

Several recent approaches have demonstrated the effectiveness of side-scan Sonar for structure detection. Hou *et al.* (2022) [35] developed a system capable of identifying underwater structures and detecting certain types of structural defects, while Yu *et al.* (2021) [36] implemented neural network approaches to extract precise localization information from Sonar data. The work by Hou *et al.* shows decent values of apparent damage detection with an average Intersection over Union (IoU) of 0.63 but only focuses on big structures such as bridges where defects have areas between 0.1 and 1 *m*. Yu *et al.* achieves Mean Average Precision (mAP) of 85.6% using a Side-scan Sonar for the detection of shipwrecks but again is only applied to high volume objects. Both

studies achieved decent results for large-scale targets such as bridge substructures and shipwrecks, though their resolution capabilities are insufficient for the object detection required for this work. Multibeam Sonar systems provide the highest resolution acoustic sensing capability by simultaneously emitting multiple precisely controlled acoustic beams across a wide angular range. This approach generates high-resolution three-dimensional bathymetry data that can reveal detailed structural geometry and surface characteristics. The work by Galceran *et al.* (2012) [37] demonstrated multibeam Sonar detection of man-made objects in underwater environments, achieving localization accuracy within several meters. While this precision is adequate for general structural surveys, it remains insufficient for detailed inspection tasks requiring millimetric accuracy for object detection or precise damage assessment.

The Sonar-based systems have very high efficiency and achieve good results for high dimension objects such as bridges and shipwrecks. However these systems lack the necessary resolution for detection tasks for O&M tasks explored by this work where the objects have small dimensions and there is a need for millimetric precision. Most of the current systems based on Sonar can only perform at near metric resolution which renders them useless for underwater inspection of objects such as anodes which only measure a few centimeters.

Imaging Systems

Imaging sensors offer complementary capabilities to acoustic systems, providing high-resolution surface detail and color information essential for detailed structural inspection and precise feature detection [33]. For 2D detection applications, optical systems can achieve spatial resolutions significantly higher than acoustic methods, enabling detection of small-scale structural features, surface defects, and biological fouling patterns. However, their effectiveness is significantly constrained by the optical challenges discussed in Section 2.1, particularly in the turbid conditions commonly encountered around subsea infrastructures.

Optical systems can be classified into three different categories: active, passive, or hybrid, depending on their method of data acquisition [38]. Active imaging systems emit their own signals into the environment and rely on the reflection or return of these signals to gather data about the underwater surroundings. These systems are particularly effective in environments where natural light is insufficient or where high-resolution data is required for tasks such as object detection and underwater mapping. Common examples of active systems include laser-based imaging systems. In contrast, passive imaging systems do not emit any signals, instead, they detect naturally occurring environmental signals. These systems rely on available light, or other ambient data to interpret the underwater environment. Passive systems are ideal for applications where minimal disturbance is necessary. Examples of passive systems include optical imaging that captures natural light or bioluminescence. Hybrid systems combine the benefits of both active and passive imaging techniques, leveraging the strengths of each to improve the overall perception and robustness. By integrating both signal emission and detection of ambient data, hybrid systems can provide more comprehensive and reliable information about underwater environments. These systems are particularly useful for complex underwater tasks, such as navigation in deep or turbid

waters, where a combination of active and passive data is required to create accurate models and perform effective detection and monitoring.

Traditional single-camera optical systems, which are passive systems, capture 2D images of structural surfaces, providing detailed visual information about surface conditions, marine growth patterns, corrosion, and damage. These systems are fundamental for subsea structure inspection as they can reveal surface-level defects and biological fouling that may not be detectable through acoustic methods. For 2D detection applications, underwater cameras must be specifically designed to address the unique challenges of marine environments. Specialized housings need to be designed to protect electronic components from pressure and corrosion, while domes for the camera lenses must be designed to minimize distortion and maintain image quality [13]. Artificial illumination systems are typically required to provide consistent lighting conditions for reliable 2D detection algorithms, as natural light availability decreases rapidly with depth and is often insufficient for detailed structural inspection [24].

However, underwater optical systems face several significant limitations for 2D detection applications, as mentioned in Section 2.1. Stereo vision systems, another type of passive imaging, represent an advanced optical configuration that can enhance underwater structure detection capabilities while maintaining focus on high-quality 2D image acquisition. These systems utilize two cameras positioned at a fixed distance apart, mimicking human binocular vision to capture synchronized image pairs from slightly different perspectives. Each camera independently captures detailed 2D images of structural surfaces, which can be processed using traditional computer vision techniques for surface defect identification, texture analysis, and feature extraction. The primary advantage of stereo systems for 2D detection applications lies in their ability to provide enhanced spatial context and improved detection reliability through dual-image analysis. By comparing corresponding features between left and right camera images, stereo systems can validate detection results and reduce false positives that commonly occur with single-camera systems in challenging underwater conditions. This cross-validation capability is particularly valuable when image quality is compromised by scattering, illumination variations, or biological fouling.

The work by Oleari *et al.* (2015) [39] presents an underwater stereo vision systems with a focus on tackling issues like calibration, lens distortion, and hardware optimization, showcasing their utility in generating accurate Three Dimensional (3D) reconstructions for underwater robotics. This approach offers the advantages of showcasing a stereo setup specifically designed for underwater conditions, along with an evaluation of its power consumption and calibration process. However, the system has notable disadvantages, as it lacks measures to address the challenging visual conditions of the subsea environment. It does not incorporate light sources to mitigate the low-light conditions at depth, nor does it include complementary sensors to augment the information provided by the stereo system. Additionally, all tests are conducted in clear water, which does not accurately represent the challenging conditions encountered in real-world underwater scenarios.

Another system used for imaging is the projection of light patterns into the environment. This active imaging method uses cameras to capture this projections and retrieves 3D information by

segmenting and analyzing the deformation of such projections. These approaches are able to achieve high accuracy information; however, as mentioned in Section 2.1, they are quickly attenuated due to light absorption. An approach by Massot-Campos and Oliver-Codina (2014) [40], uses a two-laser configuration to do underwater inspection tasks. This work resulted in measurements with errors of approximately 0.04m. Another approach by Ou *et al.* (2023) [41], used a rotating laser scanner to develop a system capable of achieving 1.025mm accuracy in clear water, with the error increasing to 3mm in turbid environments. Although these active systems approaches provide millimeter-level precision, this accuracy is only attainable at extremely close range due to the attenuation of the lasers. Furthermore, their operational scope is limited to the small region of the captured image illuminated by the projected lasers.

The limitations inherent to individual sensing modalities have driven the development of hybrid systems that integrate complementary technologies to improve overall underwater perception capabilities for 2D detection applications. These systems leverage the strengths of active and passive techniques while compensating for their individual weaknesses, providing more robust and reliable data for subsea structure detection and localization. By combining these imaging techniques within unified platforms, hybrid systems can maintain consistent detection performance across the challenging range of environmental conditions encountered in subsea operations. The integration of different sensors to achieve these hybrid systems requires data fusion algorithms that can align and correlate information from distinct sources operating at different spatial and temporal scales. Temporal synchronization ensures that data from different sensors corresponds to the same environmental conditions, while spatial registration aligns data from sensors with different viewing geometries and coordinate systems. These fusion processes are essential for creating consistent and reliable 2D detection results from multi-modal sensing platforms.

An advancement in hybrid underwater perception systems is the MARESys, developed by Pinto and Matos (2020) [12] with a later improvement made by Leite *et al.* (2024) [14]. This system represents an integration of multiple sensing modalities specifically designed for underwater structural inspection and monitoring applications. As illustrated in Figure 2.3, the MARESys system combines stereo vision cameras, Light Stripe Range (LSR), and synchronized illumination systems within a unified platform optimized for harsh subsea conditions. A key feature of the MARESys system is its incorporation of range-gated imaging, a technique designed to overcome the challenges posed by the scattering and absorption effects of the underwater environment. Range-gated imaging operates by synchronizing the illumination source with the camera's sensor to capture light returning only from specific distances. By controlling the time window during which the sensor is active, the system selectively captures reflected light from the targeted structures while filtering out backscatter from particles and turbidity in the water. This approach improves the clarity and contrast of the captured images, allowing for improved perception and identification of submerged objects even in conditions where there is a lot of suspended particles. The integration of range-gated imaging ensures that the system maintains reliable performance across a wide range of underwater scenarios, providing a crucial advantage in sub-sea exploration and monitoring applications.

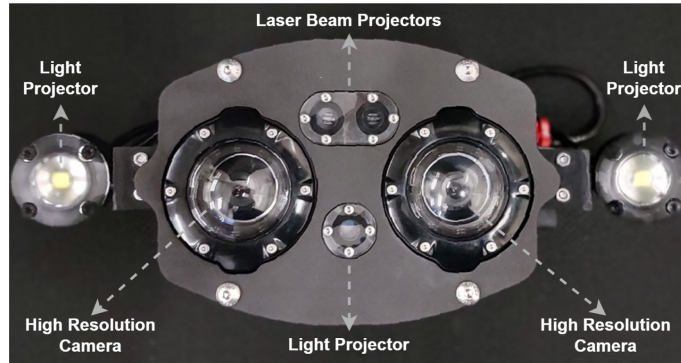


Figure 2.3: The MARESyE hybrid imaging sensor. The system employs a pair of high resolution cameras for stereo vision, two laser beam projectors that allow for active imaging techniques, and multiple light source projectors. A triggering system synchronizes data acquisition via hardware, while an internal processing unit enables edge computing. Extracted from [13].

The subsequent work by Leite and Pinto (2024) [13] improved this approach by implementing data fusion algorithms to merge the precise but sparse depth information from the LSRs and the dense but not as precise information from the stereo system. The depth information captured by the LSR has absolute errors as low as 0.5mm where as the stereo system achieves absolute errors of around 10cm. By fusing both modalities, this work is capable of improving the average absolute error of the dense point cloud captured by the stereo system by up to 52.5%. A different approach by Leite and Pinto (2025) [15], which combined the use of texture-based features with the magnitude of pixel displacement through a pyramidal approach. This approach was able to achieve improvements of 58.7% of metric error against UW_Depth [42] on the FLSea-Stereo dataset [43].

By using the range-gated technique to capture information, this system produces clearer, higher-contrast 2D images that enable more reliable feature extraction and detection algorithms. This enhanced image quality is crucial for computer vision applications that depend on sharp edges and consistent illumination for effective structure detection and classification in challenging underwater environments.

2.3 2D Object Detection methodologies

Underwater object detection plays a crucial role in various marine applications, from environmental monitoring to infrastructure inspection. The rapid advancements in Deep Learning (DL) have profoundly propelled 2D object detection methods forward. Modern approaches take advantage of Neural Network (NN) architectures to automatically learn and identify complex patterns and features within underwater images, leading to significant improvements in detection accuracy and robustness. Due to the inherent challenges in underwater data acquisition, 2D detection in subsea environments is still an area with scarce dedicated research. Therefore, this section will first examine generic architectures and their applications in other contexts, followed by a discussion of existing works within the underwater domain.

2.3.1 CNN-based Architectures

The application of Convolutional Neural Networks (CNNs) to 2D optical images enhanced object detection capabilities by learning to identify patterns and features directly from the data. CNNs represent a major advancement in feature extraction and classification [44]. Unlike conventional image processing approaches, which rely on manually crafted features based on statistical regularities or prior knowledge, CNNs learn hierarchical features through an end-to-end training process.

The structure of a typical CNN architecture can be visualized in Figure 2.4. The process begins with an input image, where a kernel extracts local features through convolution operations. These features are then passed through multiple layers of convolution operations progressively reducing the spatial dimensions while aggregating robust features. This ability to automatically learn rich, discriminative features has made CNNs the backbone of numerous object detection models.

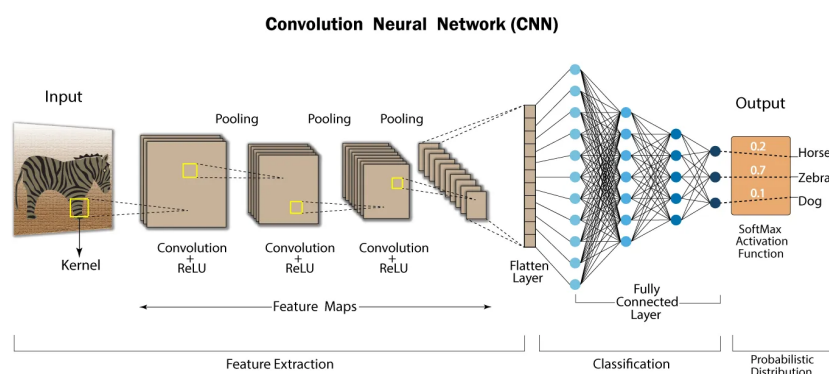


Figure 2.4: A typical CNN structure. The process begins with an input image, where a kernel extracts local features. These features are passed through multiple layers of convolution operations with ReLU activation and pooling, generating feature maps. Extracted from Medium¹.

Following the foundational understanding of CNNs as robust feature extractors, they were integrated into comprehensive frameworks for 2D object detection. These CNN-based architectures represent the backbone of many modern detection systems, leveraging the hierarchical feature learning capabilities of CNNs to localize and classify objects within an image. Broadly, these methods can be categorized into two main paradigms: multi-stage detectors that first propose regions of interest and then classify them, and single-stage detectors that predict object locations and classes simultaneously. This subsection will detail prominent examples from both categories, including the influential Region-based Convolutional Neural Network (R-CNN) family [45–47] and the highly efficient YOLO series [48, 49].

Region-based CNNs

The R-CNN models revolutionized 2D object detection by integrating deep learning with region proposals, offering significant improvements over traditional methods. These models build on the

¹<https://shorturl1.at/YT9h8>. Accessed November 11, 2024

foundation of CNNs but introduce more sophisticated mechanisms to improve detection accuracy and efficiency.

The first model to use this approach was R-CNN which combine dCNNs with selective search techniques for object detection. Introduced by Girshick *et al.* (2014) [45], R-CNN works by first generating a set of potential object proposals from the input image using selective search. These proposals are then passed through a CNN for feature extraction, and the resulting features are classified using a Support Vector Machine (SVM). Finally, a bounding box regression model refines the object locations. Despite its innovative approach, R-CNN suffers from high computational costs, primarily due to the need to run a CNN for each region proposal, which limits its scalability for real-time applications. This structure can be visualized in Figure 2.5.

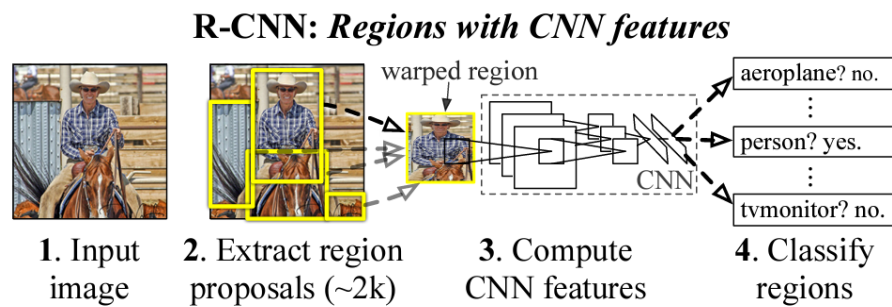


Figure 2.5: Illustration of the R-CNN pipeline. The process begins with the input image (1), from which approximately 2,000 region proposals are extracted using selective search (2). Each proposed region is warped to a fixed size and passed through a convolutional neural network (CNN) to compute features (3). These features are then classified into object categories (such as structures like anodes and mooring lines) or background using a classifier, such as a support vector machine (4). Extracted from [45].

Faster R-CNN, introduced by Ren *et al.* (2015) [47] as an improvement on Fast R-CNN proposed by Girshick (2015) [46], addresses the inefficiencies of R-CNN by replacing the selective search procedure with a Region Proposal Network (RPN). The RPN is a fully convolutional network that generates region proposals directly from the image, enabling the entire pipeline to be trained end-to-end. This eliminates the need for separate region proposal methods and significantly improves both speed and accuracy. The RPN shares convolutional features with the object detection network, which reduces computational redundancy. Faster R-CNN has set new standards in terms of detection speed and accuracy, making it a popular choice for many real-time object detection applications.

Presented as an extension of Faster R-CNN, Mask R-CNN proposed by He *et al.* (2018) [50], further advances the state of the art by introducing the capability for instance segmentation. While Faster R-CNN is designed to detect objects and classify them, Mask R-CNN extends this by adding a branch to the network that generates pixel-wise masks for each detected object. This allows Mask R-CNN to not only classify objects but also delineate their exact boundaries, providing a detailed understanding of the object's shape. Mask R-CNN leverages the same architecture as

Faster R-CNN, but adds a parallel branch for predicting binary masks at each Region of Interest (RoI), improving its versatility for tasks requiring more fine-grained object localization.

All three models, R-CNN, Faster R-CNN, and Mask R-CNN, have contributed to significant advancements in the field of object detection. While R-CNN laid the groundwork by combining CNNs with region proposals, Faster R-CNN introduced a more efficient end-to-end framework with RPNs, and Mask R-CNN expanded the model's capabilities to include segmentation, providing a deeper understanding of the objects in an image. Despite the significant strides made by two-stage detectors such as the R-CNN family, their inherent multi-step process often posed challenges for real-time applications. This led to the development of single-stage detectors, which aim to predict bounding boxes and class probabilities in a single forward pass, drastically increasing inference speed.

Single Stage Detectors

You Only Look Once (YOLO) is a state-of-the-art object detection model introduced by Redmon *et al.* (2016) [48]. Unlike traditional methods like R-CNN, which generate region proposals first and then classify them, YOLO treats object detection as a single regression problem. This means that the entire image is passed through the network at once, and it simultaneously predicts both the class labels and bounding box coordinates for all objects in the image in a single pass. This approach results in much faster detection speeds, making YOLO particularly well-suited for real-time applications. YOLO divides the input image into a grid and assigns each grid cell the task of predicting bounding boxes and class probabilities for objects whose center falls within that grid cell. Each bounding box is described by its center coordinates, width, height, and a confidence score, which indicates the likelihood of the box containing an object. The network uses convolutional layers to extract features and generate these predictions. Unlike methods like R-CNN, which require multiple stages and separate computations, YOLO performs detection in a single, unified framework, significantly reducing computation time and complexity. This unified approach also helps YOLO make fewer mistakes when it comes to background noise compared to R-CNN-based methods. Since YOLO processes the entire image at once, rather than focusing only on RoI, it can leverage the full context of the image. This enables the model to make more accurate judgments about whether a detected object is indeed part of the scene or simply background noise, enhancing its robustness against false positives.

Since its introduction, YOLO has gone through multiple iterations, each improving on the original design to increase both speed and accuracy. The latest version, YOLOv11², introduces several advancements that make it even more powerful and efficient for real-time object detection. Developed by Ultralytics, YOLOv11 builds on the architecture of previous YOLO versions but incorporates state-of-the-art techniques such as enhanced backbone networks and improved training strategies to boost detection performance across a variety of object sizes and types. Additionally,

²<https://docs.ultralytics.com/models/yolov11>. Accessed December 15, 2024.

YOLOv11 provides better adaptability to diverse environments and more accurate handling of complex scenes, solidifying its position as a leading model for high-speed object detection.

2.3.2 Transformer and Attention-Driven Detection Architectures

The transformer architecture has fundamentally revolutionized deep learning, extending its influence from natural language processing to computer vision. This paradigm shift, centered around powerful attention mechanisms, has enabled the development of a new generation of object detection models capable of capturing global dependencies and contextual information more effectively than traditional convolutional approaches. This subsection will first introduce the core principles of Attention Mechanisms and Transformer fundamentals, then explore prominent object detection architectures built upon this framework, including real-time optimized detectors like RT-DETR and those leveraging Vision Transformer (ViT) backbones.

Attention Mechanisms and Transformer Fundamentals

The Transformer architecture, initially introduced by Vaswani *et al.* (2017) [51], fundamentally relies on the concept of attention mechanisms. Unlike CNNs that process local neighborhoods, attention allows a model to weigh the importance of different parts of the input data when processing a particular element, enabling it to focus dynamically on relevant features across arbitrary distances in the input sequence. This innovative paradigm was subsequently adapted for visual data with the introduction of ViTs by Dosovitskiy *et al.* (2020) [52], pioneering the application of these mechanisms to images by treating them as sequences of patches.

The most critical component is self-attention, which allows a model to relate different positions of a single sequence to compute a representation of the same sequence. For visual data, this means each "patch" or pixel can attend to every other patch/pixel in the image, capturing comprehensive global context. This mechanism can be visualized in Figure 2.6. Multi-Head Attention (MHA) further enhances this by performing multiple attention calculations in parallel. Each head learns different aspects of relationships, and their outputs are then concatenated and linearly transformed, allowing the model to jointly attend to information from different representation subspaces at different positions. This parallel processing of attention provides a richer and more robust understanding of the input, as depicted in Figure 2.6. Transformers then combine these attention layers with feed-forward networks, forming powerful blocks that can be stacked to build deep architectures capable of complex visual reasoning.

Real-Time Detection Transformer

Building upon the robust foundation of Transformer architectures and attention mechanisms, the Real Time DEtection TRansformer (RT-DETR) proposed by Zhao *et al.* (2024) [18], represents a recent advancement in 2D object detection, leveraging the power of transformer-based architectures for high-speed and high-accuracy detection. Introduced to address the limitations of traditional CNN-based detectors in complex or cluttered scenes, RT-DETR combines the strengths of

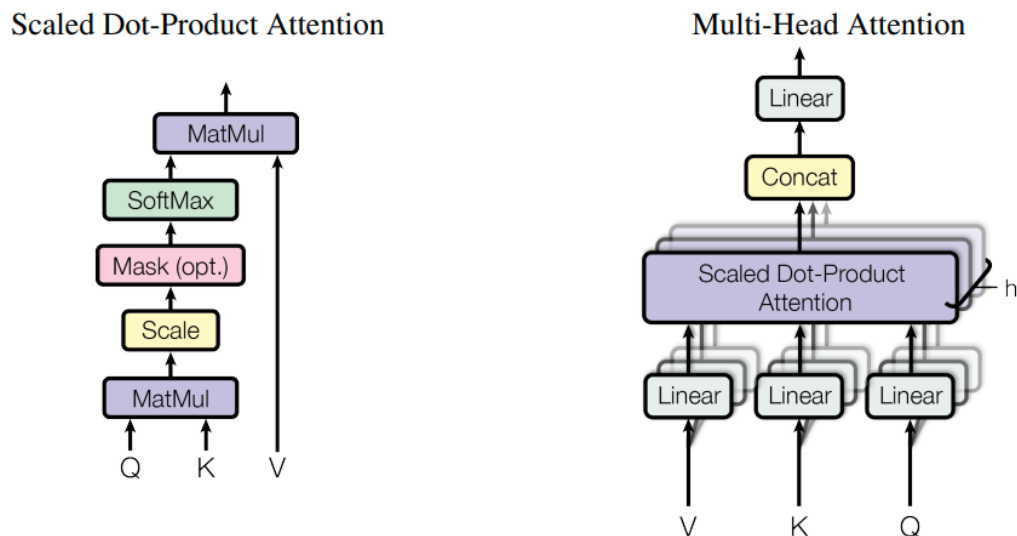


Figure 2.6: Self-Attention and Multi-Head Attention Mechanisms. (left) Self-Attention. (right) Multi-Head Attention consists of several attention layers running in parallel. Extracted from [51].

transformers with an architecture optimized for real-time applications, making it competitive with established methods such as YOLO and Faster R-CNN.

In RT-DETR, object detection is achieved through an end-to-end framework that combines both feature extraction and object proposal generation in a single, integrated process. This is achieved by leveraging the transformer’s ability to attend to key areas of the image directly, without needing a region proposal stage like those found in R-CNN architectures. The attention mechanism allows RT-DETR to dynamically focus on relevant areas of the image, improving detection accuracy while maintaining real-time processing speeds. The model’s architecture is optimized for computational speed, enabling it to achieve real-time performance on standard hardware. This efficiency, combined with its high detection accuracy, makes RT-DETR well-suited for applications that require rapid processing.

This method demonstrates the potential of transformer-based models in object detection, offering an alternative to CNN-based architectures for scenarios where both speed and accuracy are crucial. Its unique approach to leveraging global context through self-attention not only improves object recognition but also addresses some limitations of prior CNN-based methods, particularly in complex scenes with overlapping or occluded objects.

Vision Transformer-Based Detectors

Further leveraging the transformer paradigm, ViTs offer a distinct approach to feature extraction in computer vision by processing images as sequences of patches. ViTs enable models to capture long-range dependencies and potentially learn very fine-grained visual details across the entire image, a capability crucial for tasks requiring comprehensive scene understanding. The ability of ViTs to capture these subtle visual features could be particularly beneficial for identifying early

indicators of anode degradation, for example, where minor textural or structural changes might be critical. However, a known challenge with ViTs is their computational intensity, often demanding substantial processing resources, which could be a limiting factor in real-time underwater applications.

As demonstrated by Li *et al.* (2022) [53], the use of pre-trained ViT models effectively mitigates computational demands and enhances performance. Following this, several approaches have leveraged pre-trained ViT backbones to reduce training time and enhance generalization. An example of this approach is Apple's ML-Depth Pro [17], which, while primarily designed for monocular depth estimation, showcases the power of ViT-based architectures. Its architecture employs pre-trained ViT encoders applied to patches extracted at multiple scales, aiming to learn scale-invariant representations by sharing the weights of the patch encoder across these scales [17]. The success of such transfer learning strategies in achieving complex visual understanding tasks like depth estimation highlights the significant potential of adapting pre-trained ViT features for other challenging domains, including the complexities of underwater vision and object detection.

2.3.3 Underwater Object Detection

The application of 2D object detection techniques to underwater environments represents a specialized and challenging domain that requires adaptation of the architectures described in the previous subsections to address the unique visual characteristics of marine environments. While the fundamental detection frameworks remain applicable, the harsh conditions of underwater imaging require approaches and considerations that differ significantly from terrestrial applications.

The existing literature on underwater object detection reveals a significant focus on marine life identification and monitoring, with limited attention to subsea structural applications. This research emphasis reflects the historical priorities of marine research and ecological monitoring, but highlights a notable gap in addressing the detection requirements for underwater infrastructure and industrial applications [32].

Zhao *et al.* (2022) [54] proposed an enhanced YOLOv4 architecture incorporating an FPN-Attention module specifically designed for underwater object detection. This approach addresses the challenging visual characteristics of underwater vision through enhanced attention mechanisms that can better focus on relevant features despite the typical blue-green color cast and reduced contrast of underwater images. The Feature Pyramid Network (FPN) component enables the model to effectively handle objects at multiple scales, which is particularly important in underwater environments where distance estimation can be challenging due to optical distortions. This work focused on detecting marine life such as fish, crabs and starfish obtaining an overall mAP of 87.88%. Building on similar transformer-based attention concepts discussed in the previous subsection, Liu *et al.* (2023) [55] integrated Transformer self-attention mechanisms into the YOLOv5 encoder architecture to detect holothurian, sea urchins, scallop, and starfish, obtaining and mAP@50 of 83.1% and an mAP averaged across different IoU thresholds ranging from 0.5 to 0.95 at intervals of 0.05 of 45.6%. This work demonstrates the potential for advanced attention mechanisms to improve detection performance in underwater conditions by enabling the model

to focus on relevant image regions while suppressing noise and distortions typical of underwater vision. The integration of transformer attention with CNN-based detection frameworks represents a promising direction for handling the complex visual patterns present in underwater scenes.

The YOLO-Fish approach developed by Muksit *et al.* (2022) [56] specifically focused on improving small object detection capabilities, addressing a critical challenge in underwater environments where objects may appear small due to distance or water clarity limitations. This work also merged two underwater datasets for marine life detection focused on fish detection, DeepFish [57] and OzFish [58]. The emphasis on small object detection is particularly relevant for underwater applications, where limited visibility often requires detection of objects at close range or under challenging optical conditions. This approach achieved 70.34% AP@50 on the OzFish dataset, 96.15% on the DeepFish dataset and 76.56% on the merged dataset which use images from both.

More recent developments have continued to advance marine life detection capabilities. Chen and Er (2024) [59] introduced a dynamic feature fusion framework integrated with YOLO-based architectures, designed to adaptively combine features from different network layers to improve detection accuracy in the variable conditions typical of underwater environments. The dynamic nature of this approach allows the model to adjust its feature utilization based on image quality and conditions, potentially improving robustness across different underwater scenarios. This work was tested on the same dataset as Liu *et al.* (2023) [55] and obtained an mAP@50 of 86.7% and an mAP@50-95 of 68.6% which shows an improvement of 3.6% of mAP@50 and 23.0% of mAP@50-95 compared to the work of Liu *et al.* (2023) [55]. Similarly, Zhou *et al.* (2024) [60] modified the YOLOv8 backbone architecture specifically for enhanced feature extraction in underwater conditions. This work represents the adaptation of the latest YOLO developments to underwater applications, which was trained and tested on the same dataset as [55, 59] obtaining a mAP@50 of 84.0% and an mAP@50-95 of 49.0%, effectively improving on the work by Liu *et al.* (2023) [55], though it does not achieve the performance of Chen and Er (2024) [59].

The application of 2D detection techniques to underwater structural elements represents a significantly underexplored area of research, with very limited work addressing the specific challenges of detecting and localizing subsea infrastructure components. This research gap is particularly notable given the critical importance of structural monitoring for offshore energy installations, underwater pipelines, and other marine infrastructure.

An exception to this trend is the recent work by Fahy *et al.* (2024) [61], which investigated the automated detection of sacrificial anodes on offshore wind turbines using commercially available YOLOv8 models. This study represents one of the few attempts to apply modern 2D detection techniques specifically to subsea structural components, addressing a practical application relevant to offshore wind farm maintenance and inspection operations. The approach demonstrated by Fahy *et al.* provides valuable insights into the potential for automated structural component detection in real-world offshore environments. The study employed YOLOv8, leveraging the single-stage detection capabilities discussed earlier to achieve real-time detection performance suitable for ROV-based inspection operations, and obtained 86% of precision, 88% of recall and 87% of F1-score for the offshore tests and 65% of precision, 21% of recall and 24% of F1-score for the

tank tests. The choice of YOLOv8 reflects the practical requirements of underwater inspection applications, where computational efficiency and real-time performance are critical constraints. However, the work also reveals several limitations that highlight the challenges inherent in underwater structural detection applications. The datasets employed in the study exhibit constraints in both size and diversity, with offshore data collected from only two turbine installations. This limited sampling may restrict the model's generalization capabilities to anodes located in different environmental contexts, potentially characterized by variations in biofouling patterns, illumination conditions, water clarity, or structural configurations that were not represented in the training data.

Furthermore, the validation methodology employed tank-based experiments using 3D-printed anode replicas that exhibit notable visual differences compared to real-world installations. The simplified appearance and surface characteristics of printed replicas may not adequately represent the complex visual features of actual anodes, which typically exhibit corrosion patterns, marine growth, and surface degradation that significantly affect their visual appearance in operational environments. Real-world anodes often display varying degrees of corrosion, irregular surface textures, and biofouling that create significantly different visual signatures compared to clean, geometric replicas used in controlled testing environments.

These limitations in dataset representativeness emphasize the challenges of developing robust and generalizable detection methodologies for underwater structural applications. The transfer from controlled laboratory conditions to operational marine environments presents substantial challenges that have not been adequately addressed in existing literature.

2.4 Critical Review

This literature review reveals significant challenges and research gaps in the application of 2D object detection techniques to subsea structural monitoring. The underwater environment presents unique constraints that fundamentally limit the effectiveness of both sensing systems and detection algorithms developed for terrestrial applications.

The analysis of underwater perception challenges demonstrates that optical systems, while capable of high-resolution surface detail detection, are severely constrained by light attenuation, scattering, and dynamic environmental conditions. Acoustic systems provide reliable long-range detection capabilities but lack the spatial resolution required for detailed structural assessment. These limitations demand hybrid approaches that combine complementary sensing modalities, as demonstrated by systems like MARESyE, which integrate stereo vision, and laser scanning to overcome individual sensor limitations and improve 2D image quality in challenging conditions.

The review of detection architectures reveals that while both CNN-based and Transformer-based approaches have achieved remarkable success in terrestrial applications, their adaptation to underwater environments remains limited. The CNN architectures, particularly the YOLO family, offer computational efficiency suitable for real-time applications, while Transformer-based approaches provide superior global context understanding through attention mechanisms. However, most underwater applications of these architectures have focused exclusively on marine life

detection, with limited exploration of their effectiveness for structural applications. The most significant finding is the substantial research gap in underwater structural detection applications. The few existing structural detection studies reveal critical limitations including restricted dataset diversity, limited environmental representation, and inadequate validation methodologies that fail to capture the complexity of real-world operational conditions.

These findings highlight the need for research that specifically addresses subsea structural detection through the development of robust 2D detection methodologies that can operate effectively in the challenging conditions of marine environments. Furthermore, the advancement of this field crucially depends on the development of specialized datasets that adequately represent the diversity of subsea structural conditions, including varying biofouling patterns, corrosion states, and environmental contexts. The combination of advanced sensing technologies with specialized detection algorithms optimized for structural rather than biological targets, alongside comprehensive datasets, represents a critical research direction for enabling reliable automated inspection of off-shore infrastructure.

Chapter 3

Adaptive Feature Fusion for Robust Underwater Anode Detection

3.1 Introduction

This thesis presents Contextual Anode Locator in Varying Underwater Scenarios (CALVUS), a robust deep learning architecture specifically designed to address the unique challenges of underwater structural detection. The core hypothesis underlying CALVUS is that depth-aware feature representations, originally learned for monocular depth estimation tasks, contain rich spatial and contextual information that can be effectively transferred to enhance object detection performance in challenging underwater environments. By leveraging the pre-trained ViT backbone from ML-Depth Pro [17], CALVUS exploits the inherent capability of depth-aware networks to capture fine-grained spatial relationships and subtle visual distinctions crucial for accurate anode identification in complex subsea conditions. Subsequently, an extension to this architecture is introduced, leveraging multimodal feature fusion between the depth-aware and a texture-based backbone. This yields a more robust method for sacrificial anode detection in challenging underwater environments.

This chapter is structured as follows: Section 3.2 presents the experimental setup and problem definition, including the comprehensive dataset used for model training and evaluation, as well as the mathematical formulation of the underwater anode detection task. Section 3.3 details the CALVUS architecture, encompassing the pre-processing pipeline, depth-aware feature extraction backbone and detection head design. Section 3.4 presents MF-CALVUS, detailing the methods used to fuse the features from the depth-aware backbone and the texture-based backbone. Finally, Section 3.5 details the training methodology used.

3.2 Experimental Setup and Problem Definition

3.2.1 Dataset Creation

The development and evaluation of CALVUS relied on a comprehensive real-world dataset specifically designed for underwater anode detection tasks. This dataset¹ was acquired at the indoor pool of FEUP and at ATLANTIS Test Centre in Viana do Castelo, Portugal, providing diverse perspectives and operational scenarios representative of real-world inspection conditions. The dataset consists of images captured using MARESyE [12], a hybrid underwater imaging platform specifically designed for robust data acquisition in challenging subsea environments. Additionally, images were acquired using the FIFISH V6 underwater vehicle. To enhance dataset diversity and ensure comprehensive coverage of potential deployment scenarios, background images were sourced from the NEREON dataset [10], an underwater imaging dataset captured in similar conditions, that provides additional contextual information and environmental variability characteristic of marine environments.

The final dataset consists of 27,146 high-resolution underwater images, carefully partitioned to ensure robust model training and unbiased evaluation. The dataset composition includes 6,990 background images sourced from the NEREON dataset, which contain no anode annotations and serve to provide environmental diversity and reduce false positive predictions. The remaining 20,156 images contain anode structures, with each image featuring exactly one anode detection to maintain annotation consistency and training stability.

The dataset distribution has the following composition:

- **Training set:** 19,189 images (70% of total dataset)
- **Validation set:** 3,978 images (15% of total dataset)
- **Test set:** 3,979 images (15% of total dataset)

The images include a wide range of underwater conditions representative of real-world offshore wind turbine inspection scenarios, which can be seen in Figure 3.1. These conditions include varying degrees of turbidity, different illumination levels, presence of marine snow and suspended particles, and various stages of biofouling on anode surfaces, providing a comprehensive representation of the structural variations encountered during operational inspections.

3.2.2 Problem Definition

The primary objective of this work is to develop a deep learning model capable of accurately locating and classifying sacrificial anodes within given input images. As detailed in Section 3.2.1, the dataset used in this work comprises only one class, making this a single-class detection problem.

The task is formally defined as taking an input image $I \in \mathbb{R}^{H \times W \times 3}$, where H denotes the height, W the width, and 3 represents the Red Green Blue (RGB) color channels, the model is required

¹https://rdm.inesctec.pt/fa_IR/dataset/nis-2024-005. Accessed June 15, 2025.

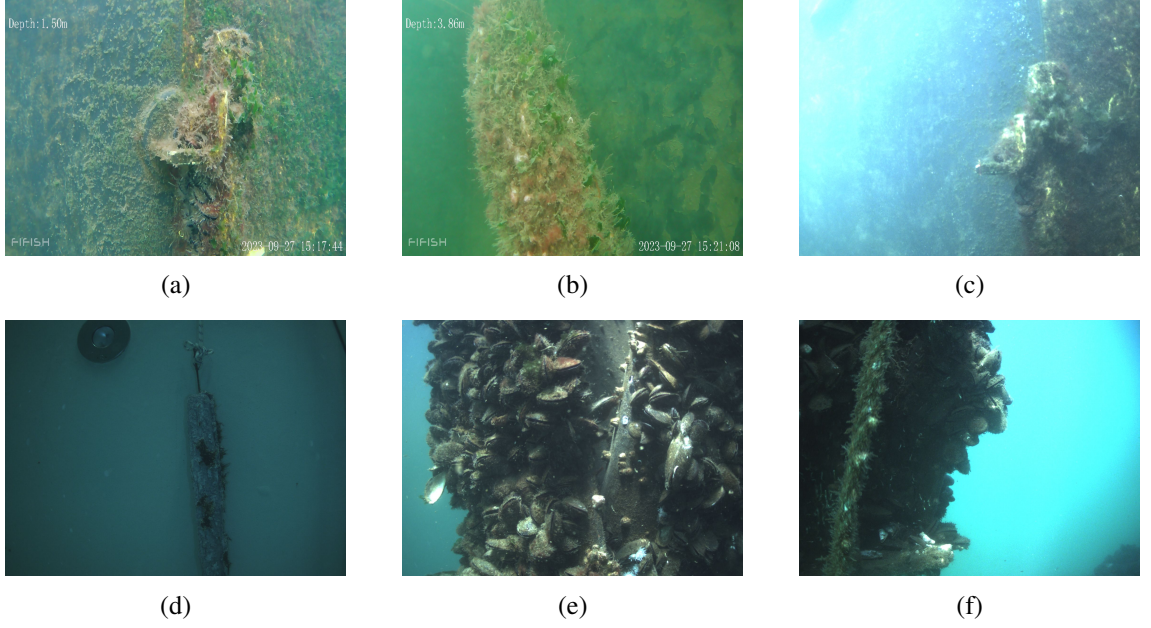


Figure 3.1: Samples from the ANODE and NEREON datasets. Figures (a) and (b) are from the ANODE dataset and were captured in Viana do Castelo with a FifishV6 ROV. Figures (c) and (d) are also from the ANODE dataset and were collected by the MARESyE system in Viana do Castelo and FEUP, respectively. Figures (e) and (f) are from the NEREON dataset and were collected by the MARESyE system.

to predict a set of bounding boxes $B = \{b_1, b_2, \dots, b_n\}$ and corresponding class probabilities $C = \{c_1, c_2, \dots, c_m\}$.

Each predicted bounding box b_i is represented by a array (x_i, y_i, w_i, h_i) , where (x_i, y_i) denote the coordinates of the bounding box center, and w_i and h_i represent its width and height, respectively. The classification output $c_i \in [0, 1]$ for each bounding box b_i denotes the model's confidence that the detected object belongs to the sole target class, which is an anode.

The problem can be formally stated as:

Given $I \in \mathbb{R}^{H \times W \times 3}$, predict (B, C) where:

$$B = \{b_i\}_{i=1}^n, \quad \text{with } b_i = (x_i, y_i, w_i, h_i),$$

$$C = \{c_i\}_{i=1}^m, \quad \text{where } c_i \text{ is the confidence for the classes, anode or background,}$$

$$c_i \in [0, 1], \forall_i \quad \text{and} \quad \sum_i c_i = 1.$$

where n is the number of detected objects, (x_i, y_i) are the center coordinates of the i -th bounding box, (w_i, h_i) are its width and height, and c_i is the confidence score for the i -th detected object belonging to the anode class.

3.3 CALVUS Architecture Design

This section presents the detailed architecture of CALVUS for robust underwater anode detection. CALVUS represents a new approach to underwater object detection to achieve robust anode detection in challenging subsea environments. The system architecture is built upon the hypothesis that features learned for depth estimation tasks contain rich spatial and contextual information that can be effectively transferred to enhance object detection performance in underwater scenarios. The overall CALVUS architecture follows a multi-stage processing pipeline, as illustrated in Figure 3.2. The system receives an input image resizing it to 1536×1536 pixels and processing it through three distinct stages: pre-processing and image enhancement, followed by multi-scale feature extraction using a depth-aware ViT backbone, and finally feature processing and object detection via an RT-DETR-based head.

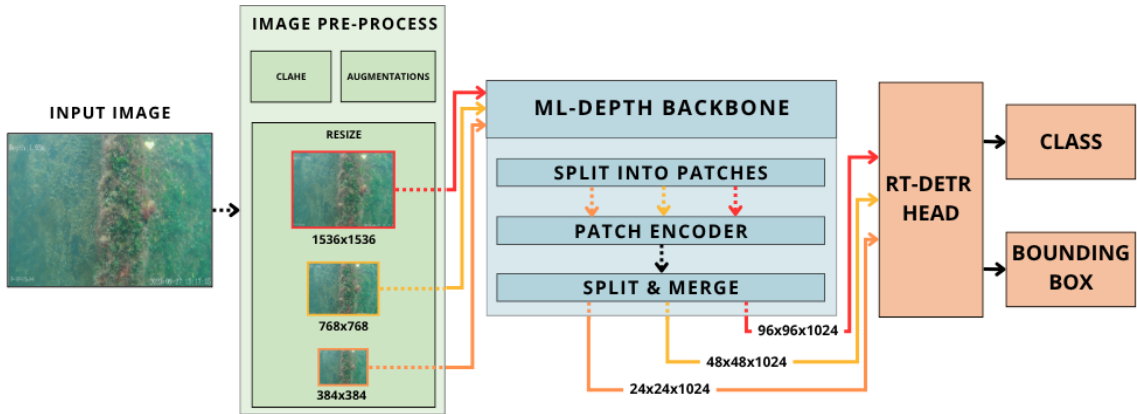


Figure 3.2: CALVUS architecture overview. The diagram illustrates the main components: image pre-processing (green), the ML-Depth-based backbone (blue), and the RT-DETR-based head (orange). Arrows' colors denote information flow from different image resolutions: red for 1536×1536 , yellow for 768×768 , and orange for 384×384 .

The main innovations of CALVUS lies in its pre-processing pipeline and the adaptation of the ML-Depth Pro patch encoder, originally designed for monocular depth estimation, as the primary feature extraction backbone for underwater object detection. This approach exploits the inherent capability of depth-aware networks to capture fine-grained spatial relationships. The pre-trained ViT backbone processes input images through a multi-scale approach, extracting patches at three different resolutions to capture both local details and global context necessary for robust anode detection.

The extracted multi-scale features are subsequently processed by a modified RT-DETR detection head, specifically adapted to handle the high-dimensional feature representations (1024 dimensions) generated by the ViT encoder, in contrast to the standard 256-dimensional features typically used in the RT-DETR architecture. This integration enables the system to leverage the rich contextual information encoded by the depth-aware backbone while maintaining the efficiency and accuracy of transformer-based object detection.

The following subsections provide a comprehensive description of each architectural component, including the detailed analysis of the pre-processing pipeline, the depth-aware feature extraction backbone, the detection head architecture, and the training methodology employed.

3.3.1 Pre-processing Pipeline

The pre-processing pipeline in CALVUS is specifically designed to enhance image quality and mitigate the visual challenges commonly encountered in underwater imaging environments that were explored in Section 2.1. This stage is crucial for improving the subsequent feature extraction capabilities and overall detection performance of the system.

Contrast Limited Adaptive Histogram Equalization (CLAHE)

To address the color distortion and uneven illumination characteristic of subsea environments, CALVUS incorporates a CLAHE algorithm [16] as the primary image enhancement technique. CLAHE operates by dividing the input image into small rectangular tiles and applying histogram equalization to each tile independently. The algorithm then uses bilinear interpolation to combine the results and eliminate artificially induced boundaries between tiles.

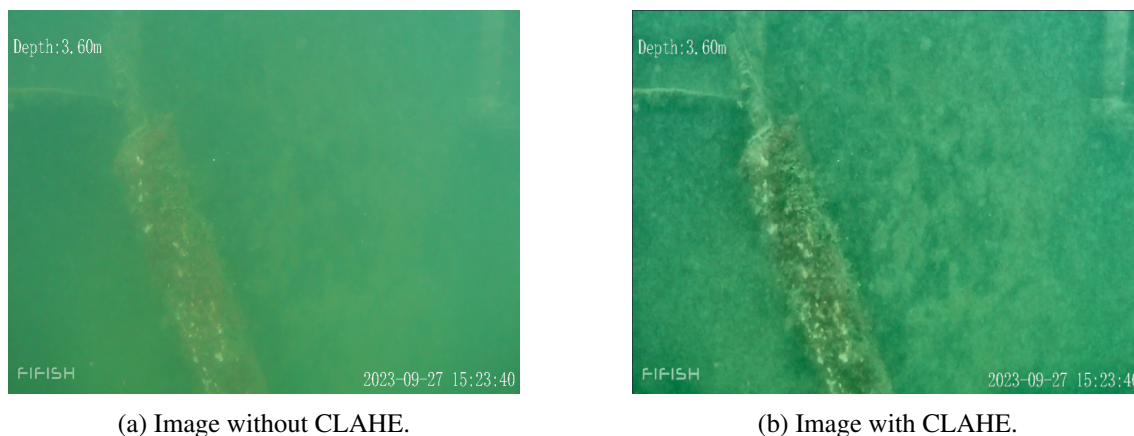


Figure 3.3: Comparison of images before and after CLAHE.

As shown in Figure 3.3, the adaptive nature of CLAHE makes it particularly well-suited for underwater imaging conditions, as it enhances local contrast while preventing over-amplification of noise in uniform regions. This pre-processing step effectively reveals finer details in the underwater images, corrects for uneven illumination patterns, and mitigates color distortion effects, thereby providing the feature extraction network with higher quality input data. The integration of CLAHE results in significantly improved image clarity with reduced noise, leading to enhanced overall system performance.

Data Augmentation Strategies

To enhance the model's robustness against the diverse challenges posed by harsh underwater conditions, CALVUS employs a set of data augmentation techniques during training. These augmentations are specifically designed to simulate the various environmental factors that affect underwater image quality:

- **Motion Blur:** Simulates the image blurring effects resulting from camera movement or vibration during image capture on moving underwater platforms, such as ROVs or Autonomous Underwater Vehicles (AUVs) operating in dynamic marine environments.
- **Salt and Pepper Noise:** Introduces random noise to replicate the visual effects of marine snow and suspended particles commonly present in seawater, which can significantly impact image clarity and visibility.
- **RGB Channel Shift:** Modifies the green and blue color channels to emulate the characteristic color shift and attenuation observed in deeper marine environments, where red wavelengths are absorbed faster by the medium.
- **Illumination Variation:** Applies controlled modifications to image brightness and contrast to simulate the varying illumination conditions encountered at different depths, ranging from well-lit shallow waters to dimly lit deep-sea environments.
- **Geometric Transformations:** Includes horizontal and vertical flips, as well as random rotations, to improve the model's invariance to different viewing angles and orientations commonly encountered during underwater inspections.

Figure 3.4 demonstrates the visual impact of each augmentation technique applied to a representative underwater anode image, clearly showing how each transformation addresses specific underwater imaging challenges. These augmentation techniques collectively ensure that the model develops robustness to the wide range of visual variations and challenges inherent in real-world underwater operational scenarios, ultimately contributing to improved detection reliability and accuracy in deployment conditions.

Multi-Scale Image Resizing

The final step of the pre-processing pipeline involves resizing the enhanced input image to three different scales: 1536×1536 , 768×768 and 384×384 pixels. This multi-scale approach enables the feature extraction backbone to capture both fine-grained local details and broader contextual information at different spatial resolutions. The three scaled images, with CLAHE enhancement and augmentation techniques applied, are then fed into the ML-Depth Pro backbone for feature extraction, where they will be processed through the patch-based encoder to generate multi-scale feature representations.

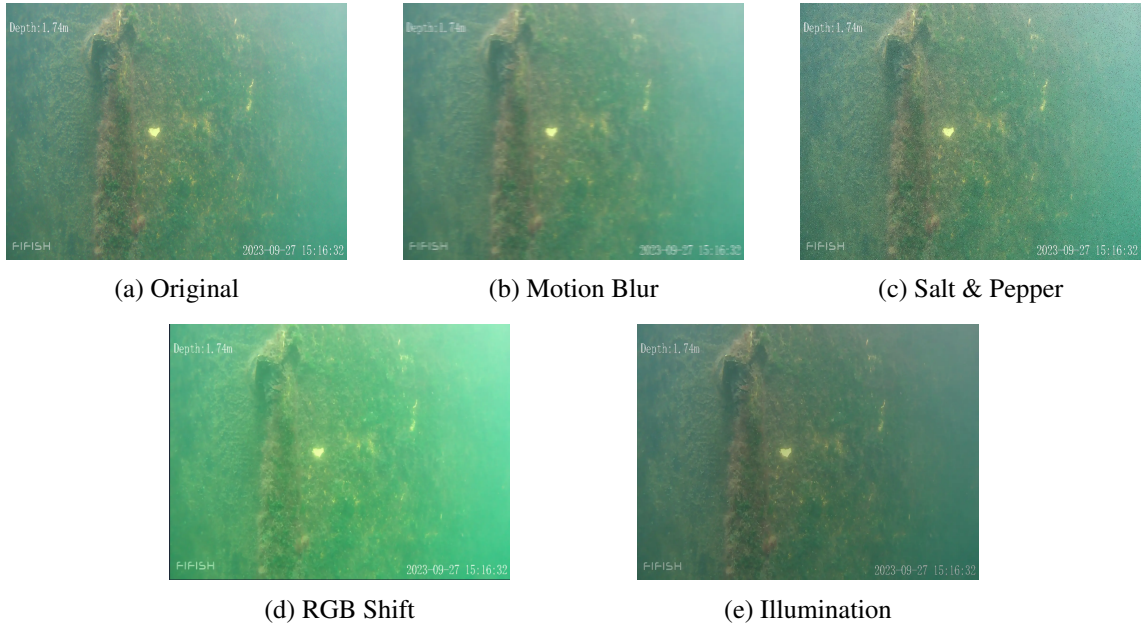


Figure 3.4: Representation of the augmentation techniques. Original image and augmented versions demonstrating the techniques used to simulate the harsh and challenging conditions encountered in sub-sea environments.

3.3.2 Depth-Aware Feature Extraction Backbone

The core innovation of CALVUS lies in its adoption of a depth-aware ViT backbone, specifically leveraging the patch encoder architecture from the ML-Depth Pro network [17]. The ML-Depth Pro patch encoder employs a multi-scale processing approach that operates on image patches extracted at different spatial resolutions. As illustrated in Figure 3.2, the three preprocessed input (1536×1536 , 768×768 and 384×384 pixels) are the inputs of the ML-Depth Backbone. The ViT-based backbone used in the patch encoder architecture operates on 384×384 pixel patches. For the two larger scaled images (1536×1536 and 768×768), the input is divided into overlapping patches of 384×384 pixels. This results in multiple patches from the larger images, while the 384×384 image is processed as a single patch. All extracted patches are then concatenated and fed into the shared patch encoder, which processes each patch through the same ViT architecture. The encoder applies the standard Vision Transformer processing pipeline: patches are flattened into sequences, combined with positional embeddings, and processed through multiple transformer layers with self-attention mechanisms. This shared weight approach across multiple scales enables the network to learn scale-invariant representations while capturing both local details and global context. Following the patch encoder processing, the output features are systematically reorganized back into three distinct feature maps corresponding to the original input scales:

- **High-resolution features:** $96 \times 96 \times 1024$ (derived from 1536×1536 input)
- **Medium-resolution features:** $48 \times 48 \times 1024$ (derived from 768×768 input)
- **Low-resolution features:** $24 \times 24 \times 1024$ (derived from 384×384 input)

This multi-scale feature representation provides the subsequent detection head with rich hierarchical information, enabling the detection of anodes at various scales while maintaining sensitivity to both coarse structural patterns and fine-grained surface details critical for underwater anode identification. The high-dimensional nature of the extracted features (1024 dimensions compared to the typical 256 dimensions in RT-DETR and YOLO) provides richer representational capacity, enabling the capture of subtle visual distinctions that are crucial for robust underwater object detection in challenging environmental conditions.

3.3.3 Detection Head Architecture

The detection head of CALVUS is based on the RT-DETR architecture [18], specifically adapted to process the high-dimensional feature representations generated by the depth-aware ViT backbone, meaning the three layers that originally received the features from the RT-DETR decoder are now going to receive the new feature maps provided by this backbone. In order to do so, those layers were changed to have 1024 input channels opposed to the initial 256 that were used by RT-DETR. Additional changes were made to the RT-DETR network to work for a single-class problem opposed to the standard 80 classes used and trained on the COCO dataset [62].

3.4 MF-CALVUS Architecture Design

Building upon the foundation established by CALVUS, this section introduces Multimodal Fusion - Contextual Anode Locator in Varying Underwater Scenarios (MF-CALVUS), an enhanced architecture that incorporates dual backbone processing and a multimodal feature fusion approach to merge the information from the depth-based features and the texture-based features. While CALVUS demonstrated the effectiveness of depth-aware feature extraction through the ML-Depth Pro backbone, MF-CALVUS extends this approach by integrating both the depth-aware ViT backbone and the texture-based conventional RT-DETR backbone, combining the strengths of both modalities through a multi-head attention-based fusion block.

The overall system architecture of MF-CALVUS, illustrated in Figure 3.5, demonstrates the integration of dual backbone processing with the multimodal fusion block. MF-CALVUS maintains the multi-scale processing approach, handling input images at three different resolutions (1536×1536, 768×768, and 384×384 pixels) and incorporating the same pre-processing pipeline with CLAHE enhancement and comprehensive data augmentation strategies. The RT-DETR detection head architecture remains consistent with the original implementation, but now processes the enhanced features generated by the fusion block rather than features from a single backbone.

MF-CALVUS uses this dual-backbone architecture to process input images through two parallel feature extraction pathways. The first pathway maintains the depth-aware processing capabilities of the original CALVUS system, using the frozen ML-Depth Pro patch encoder to generate rich spatial representations with fine-grained contextual understanding. The second pathway operates on the 768×768 resized image and introduces the RT-DETR backbone, which provides

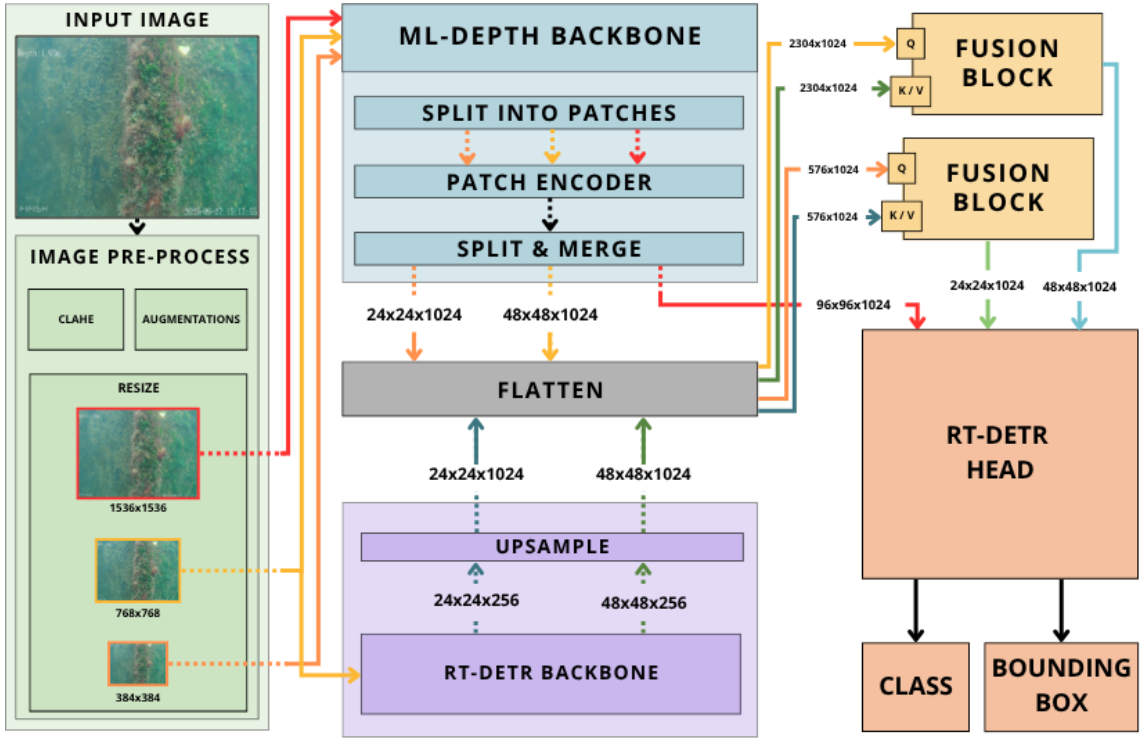


Figure 3.5: MF-CALVUS architecture overview. The diagram illustrates the main components: image pre-processing (green), the ML-Depth-based backbone (blue), RT-DETR backbone (purple), Fusion Block (yellow) and the RT-DETR-based head (orange). Arrows' colors denote information flow from different image resolutions: red for 1536×1536 , yellow, dark blue and dark green for 768×768 , and orange for 384×384 .

conventional object detection features optimized for general detection tasks. This parallel processing approach allows the system to leverage both the depth-aware representations crucial for underwater spatial understanding and the robust general-purpose features that have proven effective in diverse detection scenarios.

The key improvement in MF-CALVUS is the introduction of a multi-head attention-based multimodal feature fusion block that combines the multi-scale representations from both backbones. This multimodal fusion mechanism employs a transformer-style multi-head attention architecture to merge the 1024-dimensional features from the ML-Depth Pro encoder with the 256-dimensional features from the RT-DETR backbone. Specifically, the fusion only occurs between the two smaller feature maps of both backbones, in order to still take advantage of the feature map derived from the bigger size image (1536×1536) of the ML-Depth backbone that is capable of the finer details of the image. This multimodal approach creates an unified feature representation that captures both spatial cues and general object detection patterns through learned attention weights that dynamically emphasize the most relevant features for different detection scenarios.

Feature Fusion Block

The fusion of the depth features from the ML-Depth backbone and the texture features from RT-DETR backbone are made by the fusion block in Figure 3.6. Let $F_{depth} \in \mathbb{R}^{H \times W \times D}$ and $F_{texture} \in \mathbb{R}^{H \times W \times D}$ represent the feature maps from the ML-Depth and RT-DETR backbones, respectively.

This block consists initially of a Multi-Head Attention [51] which processes the features through parallel attention heads, each learning different aspects of the cross-modal relationships between the depth-aware and conventional detection features. Each attention head applies normalization layers to the input feature map of the ML-Depth backbone, corresponding to the query (Q), and to the input feature map of the RT-DETR backbone, corresponding to the key (K) and value (V), in order to stabilize the attention computation:

$$Q = \text{LayerNorm}(F_{depth}); \quad (3.1)$$

$$K = \text{LayerNorm}(F_{texture}); \quad (3.2)$$

$$V = \text{LayerNorm}(F_{texture}). \quad (3.3)$$

The attention mechanism computes scaled dot-product attention between the query and key features, and later between this output and the value features, producing attention weights that determine the relevance of RT-DETR features for each spatial location in the ML-Depth feature map. Following the multi-head attention computation, the outputs from all attention heads are concatenated to form a comprehensive feature representation that captures both modalities. This entire multi-head attention operation is denoted as $\text{MHA}(Q, K, V)$. This concatenated output is then added to the original ML-Depth feature map through a residual connection, preserving the original depth-aware information while incorporating the complementary detection features from the RT-DETR backbone:

$$F_{residual} = F_{depth} + \text{MHA}(Q, K, V). \quad (3.4)$$

The fused features undergo a reshaping operation to ensure proper spatial dimensions for subsequent processing, denoted as $F_{reshaped}$. The final stage of the fusion block employs a weighted sum mechanism that adaptively combines the reshaped multi-head attention output with the interpolated features. This weighted combination is computed through a learnable weighting scheme (feature interpolation) that adaptively balances the contributions from the attention-processed features and the interpolated features based on their relevance for the specific detection task in order to create the best balance between the depth and texture information:

$$F_{output} = \alpha \cdot F_{depth} + (1 - \alpha) \cdot F_{reshaped}. \quad (3.5)$$

where $\alpha \in [0, 1]$ is a learnable parameter that adaptively balances the contributions from the multimodal fused features and the depth features.

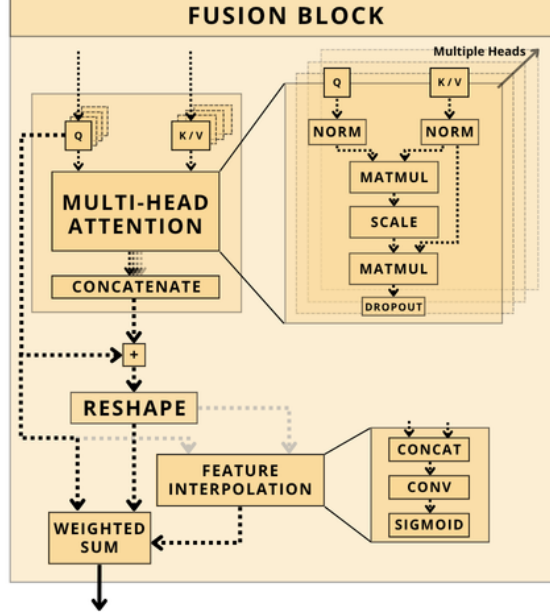


Figure 3.6: Fusion block architecture. The input feature maps from ML-Depth (depth features) and RT-DETR (texture features) undergo a normalization layer to create query (Q), key (K), and value (V) inputs for the Multi-Head Attention mechanism. The MHA processes cross-modal relationships between depth and texture modalities, producing enhanced features that are combined with the original depth features through a residual connection. The resulting fused features are reshaped to maintain proper spatial dimensions and then fed into a Feature Interpolation block alongside the original depth features, where a learnable weighted sum adaptively balances the contributions from both feature representations to produce the final output.

3.5 Training Methodology

The training strategy employed for both CALVUS and MF-CALVUS architectures follows a selective transfer learning approach designed to preserve the valuable depth-aware representations while enabling task-specific adaptation for underwater anode detection. To maximize the utilization of depth-aware features learned during the ML-Depth Pro pre-training phase, a selective freezing strategy was implemented across both architectures. The ML-Depth backbone parameters were frozen throughout the entire training process, ensuring the preservation of the rich spatial understanding capabilities crucial for underwater scene comprehension. The selective training configuration differs between the two proposed architectures based on their respective complexity. For CALVUS, only the RT-DETR detection head parameters were updated during training, focusing the learning process on adapting the pre-trained detection capabilities to the specific characteristics of underwater anode identification. In contrast, MF-CALVUS required training of both the RT-DETR detection head and the added feature fusion block, enabling the system to learn how

to combine depth-aware and conventional detection features while maintaining the frozen state of the feature extraction backbone.

Both architectures employ a initialization strategy that uses the pre-trained weights from their respective source models. The ML-Depth Pro backbone is initialized with weights from the original depth estimation model [17]. Simultaneously, the RT-DETR detection head is initialized with weights from the original RT-DETR model trained on general object detection tasks, retaining valuable detection-specific knowledge that facilitates faster convergence and improved initial performance. This initialization approach provides several advantages such as reducing the training time required for convergence by starting from well-optimized parameter configurations and ensuring that both components begin training with relevant domain knowledge.

The training process uses the official RT-DETR loss function (RTDETRLoss) [18], which incorporates classification loss, bounding box regression with Generalized Intersection over Union (GIoU), and a denoising component specifically designed to enhance robustness to noisy inputs. This loss formulation is particularly suited for underwater applications where environmental artifacts such as marine snow and motion blur frequently contaminate the input images. This loss function can be characterized by the following equations:

$$L_{\text{RT-DETR}} = L_{\text{standard}} + L_{\text{denoising}}; \quad (3.6)$$

$$L_{\text{standard}} = \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{bbox}} L_{\text{bbox}} + \lambda_{\text{giou}} L_{\text{giou}}; \quad (3.7)$$

$$L_{\text{denoising}} = \lambda_{\text{dn_cls}} L_{\text{dn_cls}} + \lambda_{\text{dn_bbox}} L_{\text{dn_bbox}} + \lambda_{\text{dn_giou}} L_{\text{dn_giou}}. \quad (3.8)$$

The standard detection loss L_{standard} consists of three components that collectively optimize object detection performance. The classification loss L_{cls} employs focal loss to address class imbalance between foreground and background regions, which is particularly beneficial in underwater scenarios where anodes often represent a small fraction of the total image area. The bounding box regression loss L_{bbox} uses L1 loss to ensure precise localization of detected objects, specially in underwater conditions, while the GIoU loss L_{giou} provides scale-invariant optimization that handles objects of varying sizes effectively. The denoising component $L_{\text{denoising}}$ represents a key innovation in RT-DETR that significantly accelerates training convergence and improves robustness. During training, this component introduces controlled noise to ground truth bounding boxes and labels, creating additional denoising queries that the model must learn to refine back to their original clean state. This mechanism forces the network to develop robust feature representations that can handle noisy inputs, making it particularly effective for underwater environments where image quality is frequently compromised by turbidity, light variations, and marine snow. The denoising loss mirrors the structure of the standard loss with identical components ($L_{\text{dn_cls}}$, $L_{\text{dn_bbox}}$, $L_{\text{dn_giou}}$) but operates on the noisy query predictions.

$$\lambda_{\text{cls}} = \lambda_{\text{dn_cls}} = 1, \quad \lambda_{\text{bbox}} = \lambda_{\text{dn_bbox}} = 5 \quad \text{and} \quad \lambda_{\text{giou}} = \lambda_{\text{dn_giou}} = 2. \quad (3.9)$$

The loss weighting coefficients are set to emphasize localization accuracy over classification confidence, with bounding box regression receiving the highest weight ($\lambda_{\text{bbox}} = 5$) followed by GIoU loss ($\lambda_{\text{giou}} = 2$) and classification loss ($\lambda_{\text{cls}} = 1$).

The training methodology used the following configuration settings detailed in Table 3.1.

Table 3.1: Training Parameters and Configuration.

Optimizer	AdamW
Learning Rate Scheduler	ReduceLROnPlateau
Batch Size	6
Epochs	50
Hardware	NVIDIA RTX 2080 SUPER 8GB VRAM

The AdamW optimizer was selected due to its established state-of-the-art performance in detection tasks and its common pairing with RT-DETR [63]. The ReduceLROnPlateau scheduler enables adaptive learning rate adjustment, automatically reducing the learning rate when validation loss plateaus, thereby facilitating fine-tuned convergence without manual intervention. The batch size of 6 samples was determined by the available GPU memory constraints, representing the maximum feasible batch size for the RT-DETR model on the RTX 2080 SUPER's 8GB VRAM while maintaining training stability. Although larger batch sizes could potentially improve training dynamics, this configuration ensures efficient memory utilization without compromising model performance. A maximum of 50 epochs was established to provide sufficient training iterations while incorporating early stopping mechanisms based on validation loss monitoring. This approach prevents overfitting and reduces computational overhead by terminating training when no further improvement is observed.

The training and validation loss for the CALVUS model are depicted in Figure 3.7. The observed discrepancy between these loss values stems from the RTDETRLoss function's mechanism. The denoising term, presented in Equation 3.8, is applied exclusively to the training loss. Analysis of both losses reveals that the validation loss begins to plateau around epoch 24, whereas the training loss plateaus closer to epoch 30. However, due to the applied early stopping criterion, training concluded after 10 epochs. This approach prevents the model from overfitting by saving the model from the epoch that yielded the best validation loss.

Figure 3.8 presents the training and validation loss progression for MF-CALVUS. Consistent with the observations for CALVUS, the disparity between the training and validation losses is attributed to the RTDETRLoss function's design. Analysis of these curves indicates that both validation and training loss for MF-CALVUS begin to stabilize around epoch 28. Consequently, training was concluded after 38 epochs due to the implemented early stopping criterion.

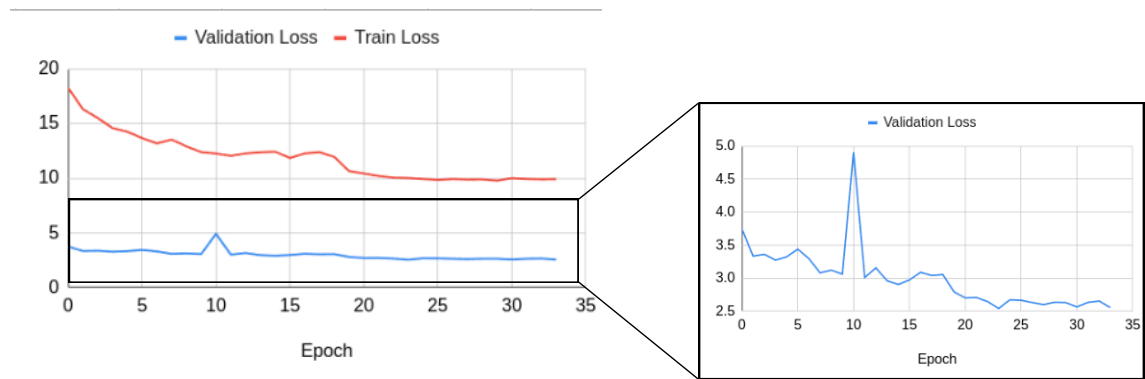


Figure 3.7: CALVUS training loss and validation loss progression.

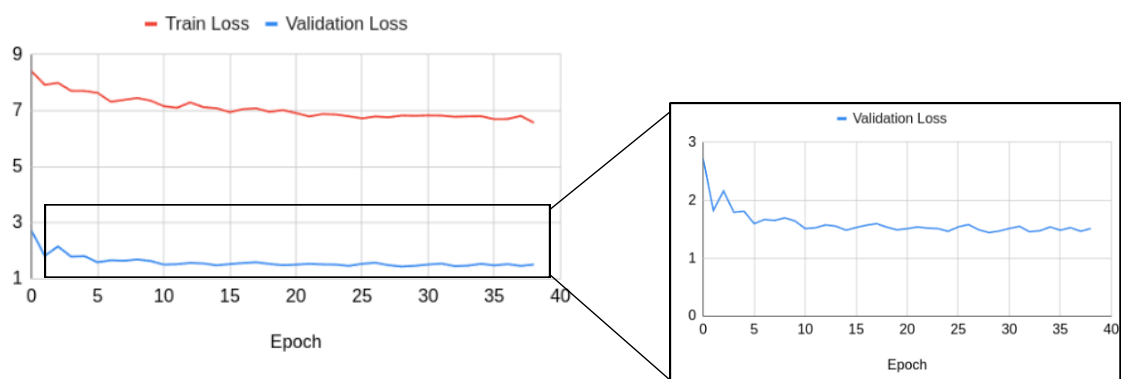


Figure 3.8: MF-CALVUS training loss and validation loss progression.

Chapter 4

Experimental Results

In this chapter, the results obtained by evaluating the proposed architectures will be presented and discussed. Firstly, the performance metrics employed for object detection tasks will be described to establish the evaluation framework used throughout this analysis. Subsequently, a comprehensive comparison between CALVUS and state-of-the-art object detection approaches, specifically RT-DETR and YOLO, will be conducted to assess the relative performance of the proposed method against established baselines. Following this comparative analysis, the performance differences between the standard CALVUS architecture and its enhanced variant, MF-CALVUS, will be examined to demonstrate the improvements of the proposed multimodal fusion approach. Finally, a real world validation is conducted to test both approaches and validate their robustness.

4.1 Performance Metrics for Detection Task

To quantitatively assess the performance of the CALVUS model in detecting of sacrificial anodes, a standard set of object detection evaluation metrics were employed. These metrics provide a comprehensive understanding of the model's accuracy in both localization and classification.

- **IoU@50 (Average Intersection over Union for detections with at least IoU of 0.5):** This metric represents the average Intersection over Union (IoU) calculated only for detected bounding boxes that achieve an IoU of at least 0.5 with their corresponding ground truth boxes. IoU indicates the overlap of the predicted bounding box to the ground truth box. IoU@50 provides a direct measure of the average spatial overlap accuracy for confidently detected objects. A higher IoU@50 indicates better bounding box prediction quality.

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \quad (4.1)$$

- **Average Precision (AP):** This is the primary metric for evaluating object detection models. It represents the average of the Precision-Recall curve calculated across multiple IoU thresholds. In this study, AP will be averaged across IoU thresholds ranging from 0.50 to 0.95, with a step size of 0.05. A higher AP indicates a more accurate and robust model, capable

of both precise localization and correct classification across varying degrees of overlap with ground truth.

$$AP = \int_0^1 P(R) dR, \quad (4.2)$$

where $P(R)$ represents precision as a function of recall.

- **AP@50 (Average Precision at IoU threshold of 0.50):** This metric specifically measures the Average Precision when the IoU between a predicted bounding box and a ground truth bounding box must be at least 0.50 to be considered a true positive. It is a common metric to evaluate overall detection performance, particularly for general object recognition tasks where a reasonable overlap is sufficient.
- **AP@75 (Average Precision at IoU threshold of 0.75):** This metric evaluates the Average Precision with a stricter IoU threshold of 0.75. It is particularly important for applications requiring high localization accuracy, as it penalizes detections that are not tightly aligned with the ground truth. Achieving a high AP@75 signifies the model's ability to precisely delineate the boundaries of the target structures.

In addition to these metrics, confusion matrices were used to provide classification insights into the model's detection capabilities. The confusion matrix is a table that presents the relationship between predicted and actual classifications, serving as the foundation for calculating various performance measures. For the binary detection task of identifying sacrificial anodes, the confusion matrix organizes detection outcomes into four fundamental categories, as illustrated in Table 4.1.

Table 4.1: Confusion matrix for a binary classification task.

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

The confusion matrix elements are defined based on detection outcomes at a specified IoU threshold of 0.5. A True Positive (TP) occurs when the predicted bounding box achieves an IoU ≥ 0.5 with a ground truth bounding box of the same class, indicating successful detection and localization. A False Positive (FP) represents detections that either have insufficient spatial overlap with ground truth objects or correspond to incorrect classifications. False Negatives (FN) occur when ground truth objects are not detected by the model, representing missed detections. True Negatives (TN) correspond to background regions correctly identified as not containing objects of interest, though this metric receives less emphasis in detection evaluation due to the predominance of background regions in typical images.

From these confusion matrix elements, several key performance metrics can be calculated to provide comprehensive evaluation of the detection system. The accuracy of the model, calculated according to Equation 4.3, represents the percentage of correctly classified instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}. \quad (4.3)$$

Precision, defined in Equation 4.4, measures the proportion of detections that were actually correct, indicating the model's ability to avoid false alarms.

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (4.4)$$

Recall, shown in Equation 4.5, quantifies the fraction of actual objects that were successfully detected, representing the model's effectiveness in identifying positive instances.

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (4.5)$$

The F1-score, presented in Equation 4.6, provides a harmonic mean of precision and recall, offering a balanced measure that accounts for both detection accuracy and completeness.

$$\text{F1-Score} = \frac{2 \times TP}{2 \times TP + FP + FN}. \quad (4.6)$$

These confusion matrix-derived metrics complement the Average Precision-based evaluation by providing interpretable measures of the model's detection performance, enabling a comprehensive assessment of the model's effectiveness in detecting sacrificial anodes under varying underwater conditions.

4.2 Comparing Uni-modality CALVUS with the State of the art

A comparison between CALVUS and other state-of-the-art object detection approaches is conducted. Specifically, the proposed architecture is evaluated against RT-DETR and YOLO, more specifically RT-DETR-L, YOLO11x and YOLO12x. To ensure a fair and meaningful comparison, all models were trained under identical conditions using the same transfer learning strategy, where the backbone remains frozen while the head retrain for the task of detecting anodes, CLAHE pre-processing, and evaluation protocols.

The quantitative results presented in Table 4.2 demonstrate the superior performance of the proposed CALVUS model across all evaluated metrics compared to RT-DETR-L and the YOLO architectures tested. When applied to the anode detection task, the RT-DETR-L backbone exhibits limited capability in effectively capturing relevant features, resulting in comparatively poor performance with an AP of only 0.345 and an AP@50 of 0.435. Among the YOLO variants, YOLO11x achieves the best performance with an AP of 0.624 and an AP@50 of 0.780, while YOLO12x shows slightly degraded performance with an AP of 0.573 and an AP@50 of 0.711. Despite the competitive performance of YOLO11x, it still underperforms compared to the proposed

Table 4.2: Performance results of the proposed CALVUS model compared to baseline object detection architectures (YOLO and RT-DETR-L) on the real-world dataset.

Model	Image Size	IoU@50	AP	AP@50	AP@75
YOLO11x	640	0.886	0.624	0.780	0.697
YOLO12x	640	0.891	0.573	0.711	0.640
RT-DETR-L	640	0.887	0.345	0.435	0.398
CALVUS	1536	0.893	0.750	0.979	0.883

model. CALVUS achieves remarkable performance with an AP of 0.750 and an exceptionally high AP@50 of 0.979, representing substantial improvements of 20.2% and 25.5% respectively over the best-performing baseline (YOLO11x). Additionally, CALVUS demonstrates superior precision at higher IoU thresholds, achieving an AP@75 of 0.883 compared to YOLO11x’s 0.697. Looking at the IoU@50, all the models exhibit good results with CALVUS having the highest value at 0.893 beating YOLO11x by 0.7%, YOLO12x by 0.2% and RT-DETR-L by 0.6%. CALVUS operates at an input resolution of 1536×1536, notably higher than the 640×640 used by baseline models. Attempts to train these YOLO and RT-DETR baselines at the elevated 1536×1536 resolution did not result in convergence, suggesting CALVUS’s enhanced detection capabilities are partly due to its effective utilization of higher-resolution inputs. These results strongly validate our hypothesis that CALVUS exhibits superior feature extraction capabilities for anode detection compared to current state-of-the-art object detection architectures.

Figure 4.1 presents a comprehensive comparative analysis of detection outputs from CALVUS, YOLO11x, and YOLO12x across four representative instances, each illustrating distinct challenging conditions encountered in underwater environments. Instance 1 demonstrates the superior robustness of CALVUS and YOLO11x compared to YOLO12x in avoiding false positives. While CALVUS and YOLO11x correctly identify that no anode is present in the underwater structure, YOLO12x incorrectly detects a non-existent anode, resulting in a false positive detection. Instance 2 showcases the models’ comparative performance under marine snow conditions, a prevalent challenge in underwater object detection. CALVUS maintains its detection accuracy with a single correct bounding box, whereas YOLO11x fails to produce any detections. Although YOLO12x correctly identifies an anode in the image, its bounding box precision is significantly inferior to that achieved by CALVUS, demonstrating reduced localization accuracy under challenging visibility conditions.

Instance 3 and 4 provide particularly compelling evidence for the effectiveness of depth-aware feature extraction. In these scenarios, extensive biofouling creates uniform texture patterns across the image, presenting a significant challenge for traditional RGB-based detection approaches. In Instance 3, YOLO12x produces multiple detections due to the visual similarity created by bio-fouling, while YOLO11x fails to generate any detections. Instance 4 exhibits a similar pattern of inconsistent performance, with YOLO11x producing multiple anode detections and YOLO12x failing to detect any objects. Both YOLO variants demonstrate susceptibility to the challenging

visual conditions created by biofouling, resulting in either over-detection or complete detection failure. In contrast, CALVUS leverages its depth-informed feature extraction capabilities to effectively distinguish between background elements and actual anodes in both instances. This depth-aware approach enables CALVUS to maintain consistent detection accuracy despite the uniform texture patterns created by biofouling, resulting in correct single-object detection in both challenging scenarios.

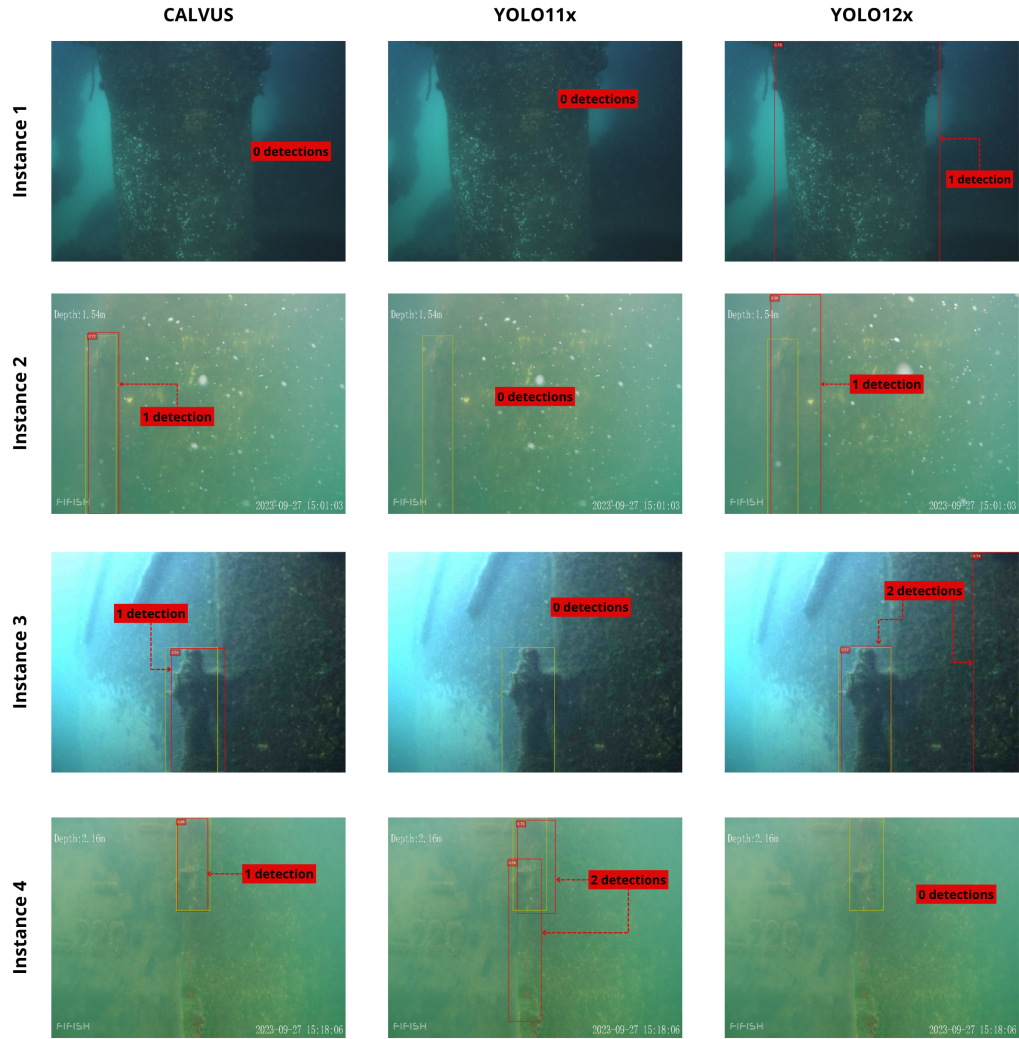


Figure 4.1: Comparison of CALVUS, YOLO11x and YOLO12x detections in varying scenarios.

The confusion matrices presented in Figure 4.2 provide quantitative validation of these qualitative observations, comparing the performance of YOLO11x, YOLO12x, and CALVUS. These matrices offer detailed insights into detection accuracy through the distribution of True Positives, False Negatives, and False Positives. From these confusion matrices, key performance metrics including accuracy, precision, recall, and F1-score were calculated, as presented in Table 4.3.

The quantitative analysis reveals substantial performance differences between the evaluated models. CALVUS demonstrates exceptional performance across all metrics, achieving near-

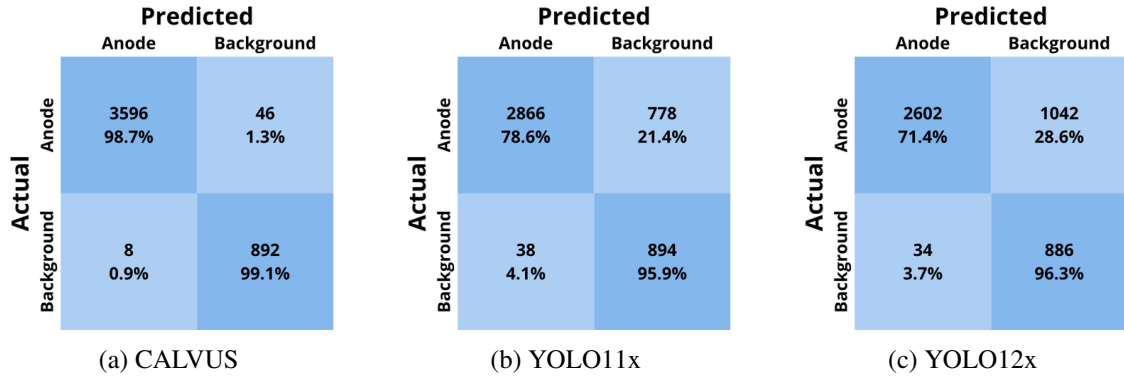


Figure 4.2: Confusion matrices highlighting classification results on the test dataset for YOLO11x, YOLO12x and CALVUS.

perfect scores with an accuracy of 0.988, precision of 0.998, recall of 0.987, and F1-score of 0.992. These results represent significant improvements over the baseline models, with CALVUS outperforming YOLO11x by 16.6 percentage points in accuracy and 21.1 percentage points in recall while maintaining similar precision performance. Notably, all three models exhibit high precision values, indicating their ability to successfully avoid the detection of false positives. However, the models differ significantly in their recall capabilities. YOLO12x demonstrates the poorest overall performance with an accuracy of only 0.764 and recall of 0.714, suggesting that while it correctly identifies the anode class (as seen in its high precision), it misses a notable portion of them, which was in evidence on Instance 2 of Figure 4.1. YOLO11x shows improved performance with an accuracy of 0.822 and recall of 0.786, caused by the reduced number of false negatives compared to YOLO12x, yet still falling considerably short of CALVUS's performance. The superior performance of CALVUS, particularly evident in its F1-score of 0.992, demonstrates its effectiveness in achieving both high recall and precision simultaneously. This combination is crucial for practical underwater inspection applications, where minimizing false alarms while ensuring comprehensive detection coverage is essential for operational efficiency and reliability.

Table 4.3: Performance results extracted from the confusion matrices.

Model	Accuracy	Precision	Recall	F1-Score
YOLO11x	0.822	0.987	0.786	0.875
YOLO12x	0.764	0.987	0.714	0.829
CALVUS	0.988	0.998	0.987	0.992

4.3 Multimodal Adaptive Fusion for anode detection

This section presents a comprehensive evaluation of the MF-CALVUS architecture, beginning with an ablation study to optimize the fusion block design, followed by a detailed performance comparison against the CALVUS model.

4.3.1 Ablation Study of the Multimodal Adaptive Fusion

To determine the most effective fusion block design for the MF-CALVUS architecture, an ablation study was performed comparing four different configurations. The complete fusion block architecture integrates multi-head attention mechanisms with feature processing components, as illustrated in Figure 4.3. The highlighted components show the key elements evaluated in the ablation study: the Feed-Forward Network (FFN) module and the Features Weighted Sum mechanism. These approaches, along with their combination and a baseline configuration, where the fusion is only done by the MHA, were systematically evaluated, with results presented in Table 4.4.

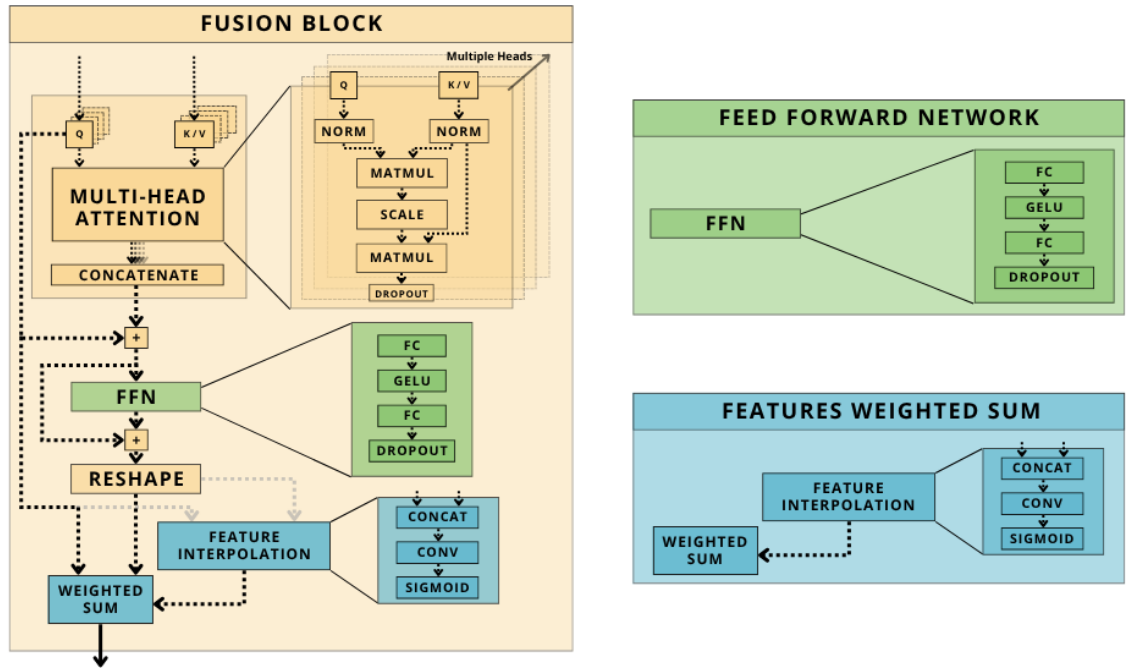


Figure 4.3: MF-CALVUS fusion block architecture with highlighted components evaluated in the ablation study. The Feed-Forward Network (green) and Features Weighted Sum (blue) modules represent the key fusion strategies tested.

The results depicted in Table 4.4 demonstrate significant performance improvements achieved by the MF-CALVUS architecture over the CALVUS model. The baseline MF-CALVUS configuration already shows substantial gains, with AP increasing from 0.750 to 0.7740 (+3.2 percentage points) and AP@75 improving from 0.883 to 0.9106 (+3.1 percentage points), indicating that the feature fusion of the dual backbone architecture yields better results. Among the fusion block variants, the Weighted Sum configuration achieves the best overall performance, delivering the highest scores in three out of four metrics: AP (0.7824), AP@75 (0.9136), and IoU@50 (0.9027). This represents a further improvement of 1.1 percentage points in AP and 0.3 percentage points in AP@75 compared to the baseline configuration. The use of the Weighted Sum approach demonstrates consistent improvements across all evaluation metrics, suggesting robust performance enhancement rather than optimization for specific detection scenarios. The Weighted Sum + FFN combination exhibits interesting trade-off characteristics, achieving the highest AP@50

score (0.9889) but showing decreased performance in AP@75 (0.9004) compared to the baseline configuration. This suggests that while the combined approach excels at detecting objects with lower IoU thresholds, it may struggle with more precise localization requirements. The FFN-only configuration shows minimal improvement over the baseline implementation, with only a marginal 0.5 percentage point gain in AP, while actually decreasing performance in AP@75 and IoU@50 metrics. These findings indicate that the weighted fusion mechanism is the primary driver of performance gains in the MF-CALVUS architecture. The superior performance of the standalone Weighted Sum configuration suggests that feature interpolation and adaptive weighting provide more effective multi-scale feature integration than feed-forward processing alone, making it the optimal choice for the final MF-CALVUS implementation.

Table 4.4: Ablation study results of MF-CALVUS fusion block configurations.

Model	IoU@50	AP	AP@50	AP@75
CALVUS	0.8932	0.7494	0.9786	0.8826
MF-CALVUS (Baseline)	0.9009	0.7740	0.9791	0.9106
MF-CALVUS (FFN)	0.8990	0.7780	0.9791	0.9075
MF-CALVUS (Weighted Sum + FFN)	0.8995	0.7750	0.9889	0.9004
MF-CALVUS (Weighted Sum)	0.9027	0.7824	0.9885	0.9136

4.3.2 Comparison with CALVUS

To quantitatively assess the architectural improvements in MF-CALVUS, confusion matrices are presented in Figure 4.4. These matrices allow for a direct comparison between the CALVUS model and its enhanced, fusion-based version, with corresponding metrics found in Table 4.5.

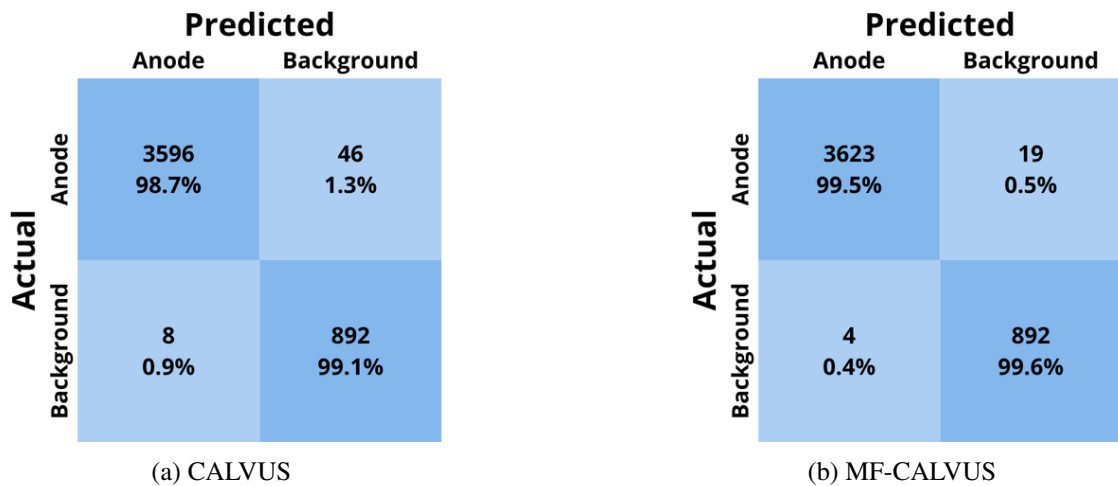


Figure 4.4: Confusion matrices highlighting classification results on the test dataset for CALVUS and MF-CALVUS.

Table 4.5: Performance results extracted from the confusion matrices comparing both CALVUS variations.

Model	Accuracy	Precision	Recall	F1-Score
CALVUS	0.9881	0.9978	0.9874	0.9925
MF-CALVUS	0.9949	0.9989	0.9948	0.9968

The quantitative analysis reveals meaningful performance improvements achieved through the enhanced fusion mechanisms applied. MF-CALVUS demonstrates superior performance across all evaluation metrics, achieving near perfect scores with an accuracy of 0.9949, precision of 0.9989, recall of 0.9948, and F1-score of 0.9968. These results represent consistent improvements over the already strong CALVUS model, with MF-CALVUS outperforming CALVUS by 0.68 percentage points in accuracy and 0.74 percentage points in recall while maintaining enhanced precision performance. The most significant improvement is observed in recall performance, where MF-CALVUS achieves 0.9948 compared to CALVUS’s 0.9874. This enhancement translates to a substantial reduction in false negative detections, with MF-CALVUS missing only 19 actual anodes compared to 46 missed by the baseline model. This improvement demonstrates the effectiveness of the weighted fusion mechanism in capturing subtle features that distinguish actual underwater structures from complex marine environments. Both models exhibit exceptionally high precision values, with MF-CALVUS achieving 0.9989 compared to CALVUS’s 0.9978. This minimal difference reflects both architectures’ robust capability to avoid false positive predictions, with MF-CALVUS generating only 4 false positives compared to 8 from the baseline model. The enhanced precision indicates that the fusion-based feature integration does not compromise the model’s ability to distinguish actual targets from background elements. The superior performance of MF-CALVUS in its F1-score of 0.9968, demonstrates the effectiveness of the weighted fusion approach in achieving optimal balance between precision and recall. This improvement is crucial for practical underwater inspection applications, where both comprehensive detection coverage and minimal false alarms are essential for operational reliability and cost-effectiveness. The enhanced fusion mechanism successfully leverages multi-scale feature information to achieve more robust detection performance in challenging underwater environments.

To further examine the specific instances where MF-CALVUS’s feature fusion approach yielded improvements over CALVUS, Figure 4.5 illustrates a few representative cases. As depicted in Figure 4.5, CALVUS exhibits increased difficulty in detecting anodes within certain scenarios, a challenge not observed with MF-CALVUS. Instance 1 presents a situation where the majority of the anode is occluded, yet its texture remains clearly differentiable from the background. The significant occlusion of the anode hinders the depth backbone’s ability to accurately detect it, leading to a False Negative for CALVUS. In contrast, MF-CALVUS successfully leverages the anode’s textural information to enhance feature extraction and correctly identify the anode. Instance 2 further exemplifies the benefits derived from incorporating an additional texture-based backbone. In this particular instance, the anode is captured at a challenging angle, presenting a complex scenario

for depth-based feature extraction. As the anode is positioned substantially non-perpendicular to the image plane, its projected depth variations become less distinct. This ambiguous depth information leads to a missed detection by CALVUS. However, through the fusion of features from both the texture and depth backbones, MF-CALVUS effectively avoids this issue and correctly identifies the anode. In Instance 3, the considerable distance to the anode combined with the low illumination of the scene results in minimal discernible variations between the anode and its background. The only differentiating information available is the whitish tone of the anode's texture in areas not obscured by biofouling. This subtle textural cue is successfully used and detected exclusively by MF-CALVUS.

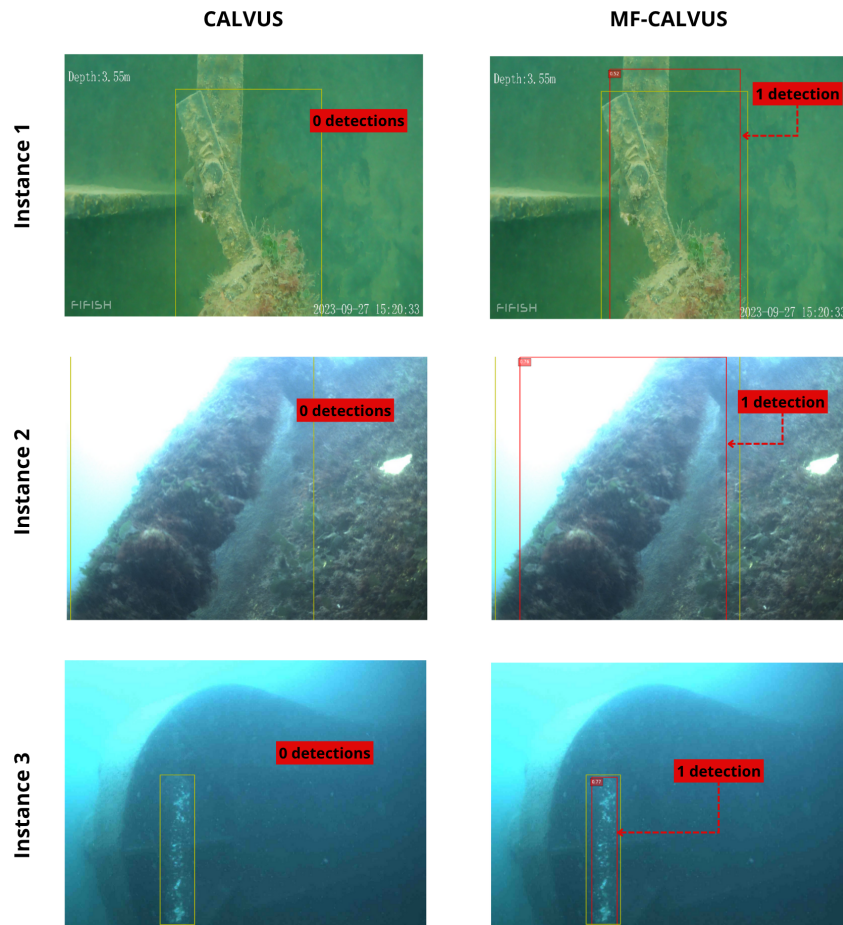


Figure 4.5: Comparison of CALVUS and MF-CALVUS detections in varying scenarios.

The comprehensive evaluation presented in this subsection provides compelling evidence for the effectiveness of the enhanced fusion mechanisms implemented in MF-CALVUS. The quantitative analysis demonstrates consistent and meaningful improvements across all performance metrics, with MF-CALVUS achieving near-perfect detection accuracy (99.49%) while significantly reducing false negative detections by 58.7% compared to the baseline CALVUS model. This substantial reduction in missed detections, from 46 to 19 anodes, shows a major improvement for practical underwater inspection applications. The fusion-based architecture demonstrates superior

robustness to common challenges in underwater environments, including partial occlusion, non-optimal viewing angles, and poor illumination conditions. In each challenging scenario examined, MF-CALVUS successfully leveraged complementary information from both depth and texture modalities to maintain accurate detection capabilities, while the single-backbone CALVUS approach failed to adapt to these adverse conditions.

The multi-modal fusion approach addresses fundamental limitations of single-modality detection systems in underwater environments. By combining depth-aware spatial information with texture-based discriminative features, MF-CALVUS achieves the robustness necessary for reliable deployment in real-world marine inspection scenarios.

4.3.3 Real World Validation

As a final validation step for the MF-CALVUS approach in detecting sacrificial anodes under harsh subsea conditions, real-world testing was conducted. This involved deploying the FIFISHv6 ROV at the ATLANTIS Test Center in Viana do Castelo to acquire new anode images, which was entirely unseen during model training. Additionally, data was collected of some anodes present in the original training dataset, specifically chosen due to their significantly increased biofouling accumulation resulting from an approximate two-year gap between the initial dataset acquisition and this new data collection. The testing setup, with DURIUS, a moored floating structure from which the anode images were acquired, visible in the background, is depicted in Figure 4.6. Representative examples of the images acquired during this operation are presented in Figure 4.7. It is notable that images captured in the morning exhibited significantly lower light levels due to cloudy weather, whereas those taken in the afternoon, under clearer conditions, demonstrated the opposite effect with increased illumination.

To validate the proposed approach, both CALVUS and MF-CALVUS were tested on the newly collected data, with relevant results presented in Figure 4.8. This new set of test images revealed that MF-CALVUS exhibits greater consistency and robustness than CALVUS, particularly in scenarios affected by additional challenges such as marine snow. For example, Instances 1 and 2 in Figure 4.8 were captured during the same acquisition period; however, Instance 1 displays a more pronounced presence of marine snow and a general blurring effect compared to Instance 2. In this specific case, MF-CALVUS successfully detects the anode in both situations, whereas CALVUS struggles with the increased visual clutter in Instance 1, consequently failing to detect the anode. Similarly, in Instance 3, the proximity to the surface introduces image degradation caused by sunlight. This effect blurs the edges of the anode's upper portion, consequently leading to a missed detection by CALVUS, whereas MF-CALVUS successfully detects the anode. Instance 4 serves as a compelling illustration of areas for future model improvement. Given that the training dataset primarily comprised anodes in frontal orientations, both CALVUS variations exhibit difficulty in detecting anodes presented at unconventional angles. Furthermore, the anodes in Instance 4 display an extreme degree of biofouling, rendering them nearly imperceptible even to the human eye. This real-world evaluation confirms MF-CALVUS as a robust approach for practical deployment,

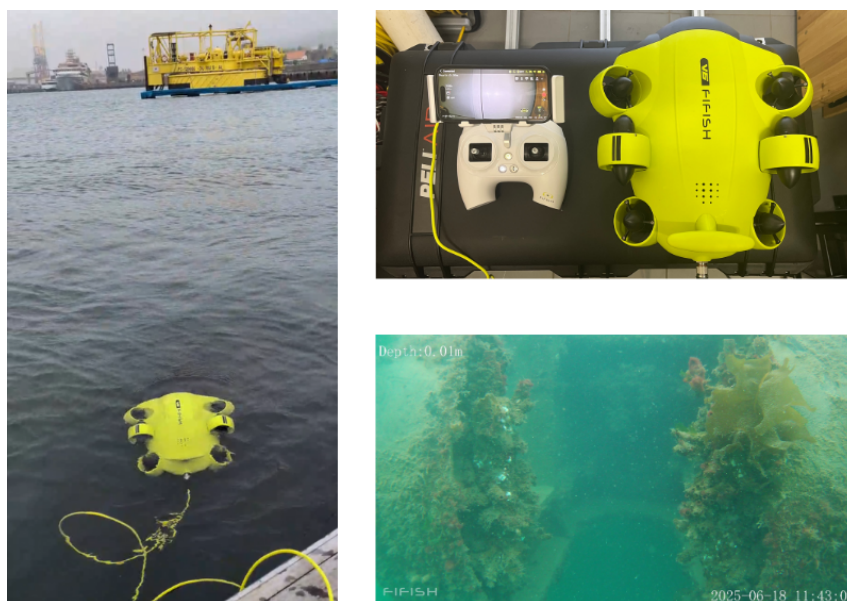


Figure 4.6: Overview of the real-world validation setup. The FIFISHv6 ROV is depicted (left) maneuvering towards the DURIUS floating structure for anode image acquisition. Complementing this, the top-right portion presents the ROV and its remote controller setup, while one of the frames captured is shown in the bottom-right portion.

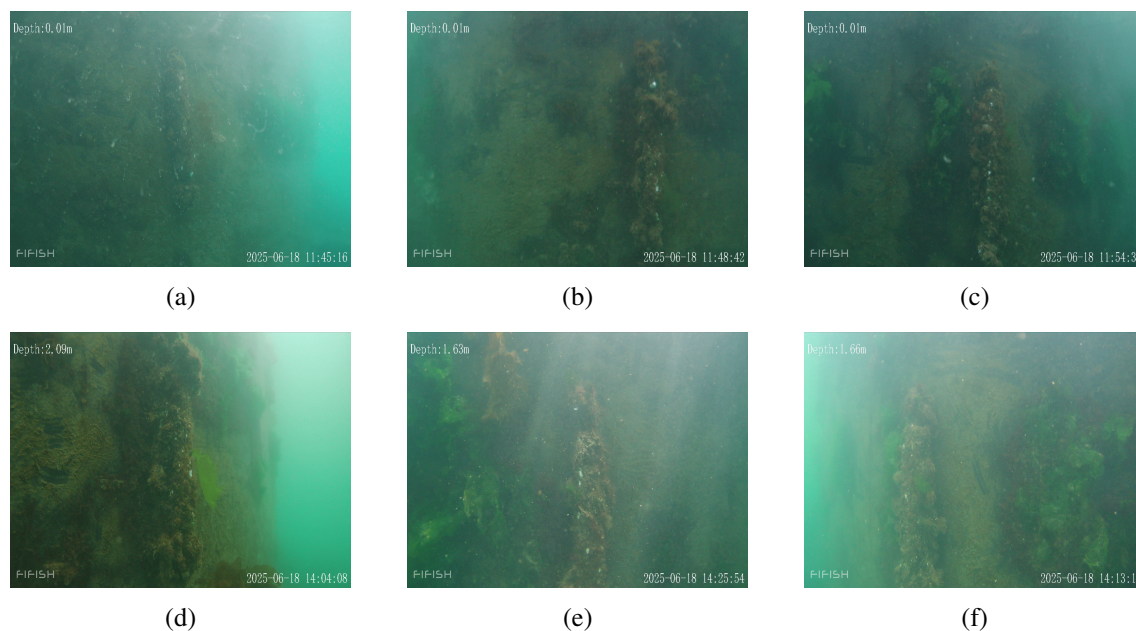


Figure 4.7: Sample images from real-world validation, illustrating varying illumination conditions. Figures (a)-(c) were captured under cloudy weather conditions, resulting in notably darker images, while Figures (d)-(f) were taken under clear conditions, exhibiting some light rays.

showing the advantages of using a multimodal approach opposed to the unimodal CALVUS, while also highlighting specific areas for future performance optimization.

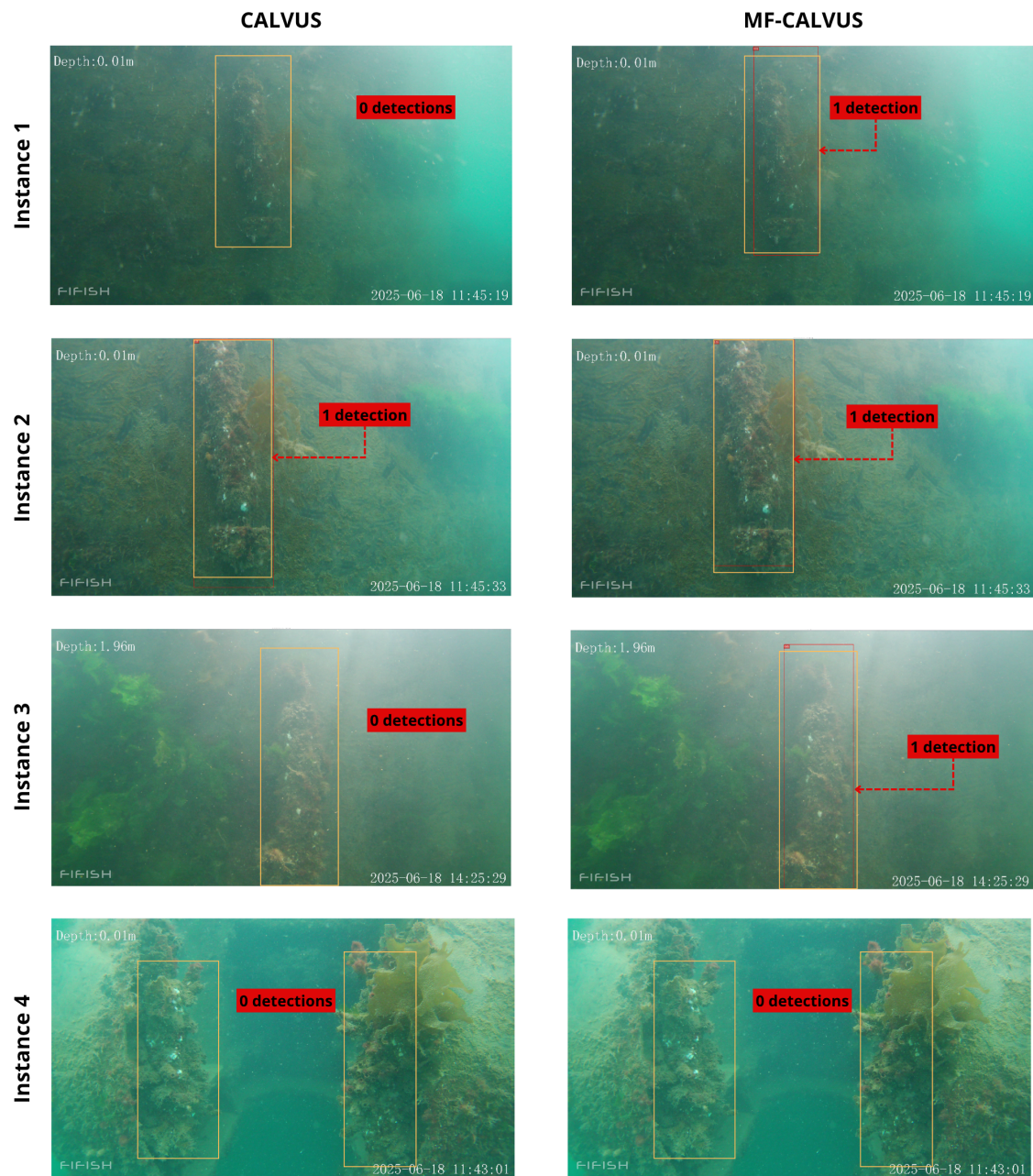


Figure 4.8: Comparison of CALVUS and MF-CALVUS on the new acquired data.

Chapter 5

Conclusion

The offshore wind energy sector has emerged as a critical component of the global transition toward sustainable energy systems. However, the operational success and economic viability of these installations depend on effective operations and maintenance (O&M) practices, particularly for underwater structural components that face constant exposure to harsh marine environments. Current O&M practices rely heavily on manual operations requiring human divers or Remotely Operated Vehicles (ROVs), operations that pose significant safety risks to personnel and incur substantial operational costs, creating an urgent need for autonomous monitoring solutions that can reduce human risk exposure while maintaining inspection reliability. In the context of OWTs, the maintenance of underwater foundations is vital for structural integrity and operational safety. The monitoring of sacrificial anodes, which provide essential corrosion protection for underwater structures, represents a critical aspect of this maintenance regime. However, the development of autonomous solutions for anode detection and assessment is severely hindered by the inherent challenges of underwater environments, where data degradation significantly compromises the effectiveness of conventional computer vision approaches. Additionally, the scarcity of high-quality annotated datasets, resulting from the difficulties associated with underwater data acquisition, further impedes research progress in this domain.

This work addressed this challenges by firstly developing a pre-processing pipeline to mitigate visual degradation in underwater environments. The implementation of Contrast Limited Adaptive Histogram Equalization (CLAHE) effectively addressed color distortion and uneven illumination, while data augmentation strategies simulating motion blur, marine snow, color shifts, and illumination variations enhanced model robustness across diverse underwater conditions encountered in real-world deployments. The development of Contextual Anode Locator in Varying Underwater Scenarios (CALVUS), explored the application of depth-aware feature extraction to underwater structural detection. CALVUS leverages the pre-trained Vision Transformer backbone from ML-Depth Pro, exploiting depth-aware networks' capability to capture fine-grained spatial relationships crucial for underwater anode identification. Through multi-scale processing and integration with a modified RT-DETR detection head, CALVUS feature extraction outperforms state-of-the-art architectures, achieving an Average Precision of 0.750 compared to 0.624

for YOLO11x (12.6% improvement) and an exceptional AP@50 of 0.979 (19.9% improvement over the best baseline).

The primary contribution of this work was Multimodal Fusion - Contextual Anode Locator in Varying Underwater Scenarios (MF-CALVUS), incorporating dual backbone processing with multimodal feature fusion. This architecture combines depth-based spatial information with conventional texture-based features through a multi-head attention-based fusion block, operating through parallel pathways that merge complementary data streams. MF-CALVUS achieved substantial improvements over CALVUS with an AP of 0.782 (3.3% increase), exceptional precision (99.89%) and recall (99.48%), while reducing false negative detections by 58.7% from 46 to 19 missed anodes, a critical improvement for preventing undetected structural degradation that could lead to catastrophic failures. The comprehensive experimental evaluation conducted throughout this research provides compelling evidence for the effectiveness of the proposed approaches. The systematic comparison against established baselines, including YOLO11x, YOLO12x, and RT-DETR-L, demonstrates the superior performance of both CALVUS variants across diverse underwater conditions. The confusion matrix analysis reveals that MF-CALVUS achieves near-perfect detection accuracy (99.49%) while maintaining great precision, minimizing false alarm rates crucial for operational deployment. The real-world validation conducted at the ATLANTIS Test Center using entirely unseen data further confirms the robustness and practical applicability of the proposed methods. The testing revealed that MF-CALVUS exhibits greater consistency and reliability than single-modality approaches, particularly in challenging scenarios involving marine snow, varying illumination conditions, and heavy biofouling.

This dissertation addresses a critical gap in underwater structural monitoring, providing a practical solution that can significantly enhance the safety, efficiency, and economic viability of offshore wind O&M. This work establishes depth-aware feature extraction as a viable and superior approach for underwater object detection, opening new research directions in underwater object detection. The successful use of multimodal features for detection tasks shows their potential in challenging environments where conventional approaches fail.

The transition from manual to autonomous underwater inspection represents a paradigm shift that reduces human exposure to hazardous marine environments while providing more consistent, reliable, and cost-effective monitoring capabilities. This research contributes to that transition by providing validated, deployable solutions that address real-world operational challenges in offshore wind maintenance, ultimately supporting the continued growth and success of renewable energy infrastructure.

5.1 Future Work

While the proposed approaches demonstrate significant performance improvements, several limitations provide opportunities for future research. The current models show reduced performance when detecting anodes at unconventional viewing angles, reflecting the predominance of frontal orientations in the training dataset. Future work should focus on expanding dataset diversity to

include more varied viewing perspectives and structural configurations to improve the model's robustness across different inspection scenarios.

The extension of these approaches to other underwater structural components represents a significant opportunity for broader impact. Adapting the depth-aware feature extraction and multimodal fusion techniques to detect other critical underwater structures such as mooring lines and anchors would substantially expand the practical utility of autonomous inspection systems. This expansion would require the development of additional annotated datasets which combined all this types of structures.

The integration of these detection capabilities with robotic manipulation systems could enable fully autonomous inspection and maintenance operations. Future work could explore the combination of the proposed detection algorithms with underwater robotics platforms equipped with manipulation capabilities, potentially enabling autonomous cleaning of anodes, replacement of degraded components, or real-time structural health assessment. This integration would represent a complete autonomous maintenance solution for offshore wind installations. Furthermore, the development of predictive maintenance capabilities based on the detected structural conditions could provide significant value for offshore operators. By combining the detection results with temporal analysis and degradation analysis, future systems could predict maintenance needs and optimize inspection schedules, leading to more efficient and cost-effective O&M strategies.

References

- [1] Rehana Perveen *et al.* “Off-shore wind farm development: Present status and challenges”. In: *Renewable and Sustainable Energy Reviews* 29 (2014), pp. 780–792. ISSN: 1364-0321. DOI: [10.1016/j.rser.2013.08.108](https://doi.org/10.1016/j.rser.2013.08.108).
- [2] Maria Martinez Luengo and Athanasios Kolios. “Failure Mode Identification and End of Life Scenarios of Offshore Wind Turbines: A Review”. In: *Energies* 8.8 (2015), pp. 8339–8354. ISSN: 1996-1073. DOI: [10.3390/en8088339](https://doi.org/10.3390/en8088339).
- [3] Abdel Salam H. Makhoulf *et al.* “Chapter 6 - Corrosion and protection of the metallic structures in the petroleum industry due to corrosion and the techniques for protection”. In: *Handbook of Materials Failure Analysis*. Ed. by Abdel Salam Hamdy Makhoulf and Mahmood Aliofkhazraei. Butterworth-Heinemann, 2018, pp. 107–122. ISBN: 978-0-08-101928-3. DOI: [10.1016/B978-0-08-101928-3.00006-9](https://doi.org/10.1016/B978-0-08-101928-3.00006-9).
- [4] J. V. Sharp and G. Ersdal. “Damage types in offshore structures”. In: *Underwater Inspection and Repair for Offshore Structures*. John Wiley & Sons Ltd, 2021. Chap. 3, pp. 37–78. DOI: [10.1002/9781119633891.ch3](https://doi.org/10.1002/9781119633891.ch3).
- [5] Renato Silva *et al.* “Multi-criteria Metric to Evaluate Motion Planners for Underwater Intervention”. In: *Autonomous Robots* 46 (2022), pp. 971–983. DOI: [10.1007/s10514-022-10060-x](https://doi.org/10.1007/s10514-022-10060-x).
- [6] Maria Inês Pereira and Andry Maykol Pinto. “Reinforcement learning based robot navigation using illegal actions for autonomous docking of surface vehicles in unknown environments”. In: *Engineering Applications of Artificial Intelligence* 133 (2024), p. 108506. ISSN: 0952-1976. DOI: [10.1016/j.engappai.2024.108506](https://doi.org/10.1016/j.engappai.2024.108506).
- [7] Daniel F. Campos *et al.* “Nautilus: An autonomous surface vehicle with a multilayer software architecture for offshore inspection”. In: *Journal of Field Robotics* 41.4 (2024), pp. 966–990. DOI: [10.1002/rob.22304](https://doi.org/10.1002/rob.22304).
- [8] Rafael Marques Claro *et al.* “A Multimodal Perception System for Precise Landing of UAVs in Offshore Environments”. In: *Journal of Field Robotics* (2025). DOI: [10.1002/rob.22517](https://doi.org/10.1002/rob.22517).
- [9] Rafael Marques Claro *et al.* “ArTuga: A novel multimodal fiducial marker for aerial robotics”. In: *Robotics and Autonomous Systems* 163 (2023), p. 104398. ISSN: 0921-8890. DOI: [10.1016/j.robot.2023.104398](https://doi.org/10.1016/j.robot.2023.104398).
- [10] João M. M. Dionísio *et al.* “NEREON - An Underwater Dataset for Monocular Depth Estimation”. In: *OCEANS 2023 - Limerick*. 2023, pp. 1–7. DOI: [10.1109/OCEANS_Limerick52467.2023.10244675](https://doi.org/10.1109/OCEANS_Limerick52467.2023.10244675).

- [11] Andry Maykol Pinto *et al.* “ATLANTIS - The Atlantic Testing Platform for Maritime Robotics”. English. In: *OCEANS 2021: San Diego - Porto*. Oceans Conference Record (IEEE). United States: IEEE Institute of Electrical and Electronic Engineers, 2022. ISBN: 978-1-6654-2788-3. DOI: [10.23919/OCEANS44145.2021.9706059](https://doi.org/10.23919/OCEANS44145.2021.9706059).
- [12] Andry Maykol Pinto and Anibal C. Matos. “MARESyE: A hybrid imaging system for underwater robotic applications”. In: *Information Fusion* 55 (2020), pp. 16–29. ISSN: 1566-2535. DOI: [10.1016/j.inffus.2019.07.014](https://doi.org/10.1016/j.inffus.2019.07.014).
- [13] Pedro Nuno Leite and Andry Maykol Pinto. “Fusing heterogeneous tri-dimensional information for reconstructing submerged structures in harsh sub-sea environments”. In: *Information Fusion* 103 (2024), p. 102126. ISSN: 1566-2535. DOI: [10.1016/j.inffus.2023.102126](https://doi.org/10.1016/j.inffus.2023.102126).
- [14] Pedro Nuno Leite *et al.* “Hybrid underwater imaging for the tri-dimensional inspection of critical structural elements in offshore platforms”. In: *Ocean Engineering* 314 (2024), p. 119658. ISSN: 0029-8018. DOI: [10.1016/j.oceaneng.2024.119658](https://doi.org/10.1016/j.oceaneng.2024.119658).
- [15] Pedro Nuno Leite and Andry Maykol Pinto. “Multimodal information fusion using pyramidal attention-based convolutions for underwater tri-dimensional scene reconstruction”. In: *Information Fusion* 124 (2025), p. 103339. ISSN: 1566-2535. DOI: [10.1016/j.inffus.2025.103339](https://doi.org/10.1016/j.inffus.2025.103339).
- [16] Stephen M. Pizer *et al.* “Contrast-limited adaptive histogram equalization: speed and effectiveness”. In: *Proceedings of the First Conference on Visualization in Biomedical Computing*. Oct. 1990, pp. 337–345. DOI: [10.1109/VBC.1990.109340](https://doi.org/10.1109/VBC.1990.109340).
- [17] Alexey Bochkovskii *et al.* “Depth Pro: Sharp Monocular Metric Depth in Less Than a Second”. In: *arXiv preprint* (2024). DOI: [10.48550/arXiv.2410.02073](https://doi.org/10.48550/arXiv.2410.02073).
- [18] Yian Zhao *et al.* “DETRs Beat YOLOs on Real-time Object Detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 16965–16974. DOI: [10.48550/arXiv.2304.08069](https://doi.org/10.48550/arXiv.2304.08069).
- [19] J. Zhou *et al.* “Underwater vision enhancement technologies: a comprehensive review, challenges, and recent trends”. In: *Applied Intelligence* 53 (2023), pp. 3594–3621. DOI: [10.1007/s10489-022-03767-y](https://doi.org/10.1007/s10489-022-03767-y).
- [20] Yan-Tsung Peng and Pamela C. Cosman. “Underwater Image Restoration Based on Image Blurriness and Light Absorption”. In: *IEEE Transactions on Image Processing* 26.4 (2017), pp. 1579–1594. DOI: [10.1109/TIP.2017.2663846](https://doi.org/10.1109/TIP.2017.2663846).
- [21] Iuliu Vasilescu *et al.* “Color-accurate underwater imaging using perceptual adaptive illumination”. In: *Autonomous Robots* 31.2 (2011), pp. 285–296. DOI: [10.1007/s10514-011-9245-0](https://doi.org/10.1007/s10514-011-9245-0).
- [22] Kai Hu *et al.* “An Overview of Underwater Vision Enhancement: From Traditional Methods to Recent Deep Learning”. In: *Journal of Marine Science and Engineering* 10.2 (2022). ISSN: 2077-1312. DOI: [10.3390/jmse10020241](https://doi.org/10.3390/jmse10020241).
- [23] Sundara Tejaswi Digumarti *et al.* “Underwater 3D capture using a low-cost commercial depth camera”. In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2016, pp. 1–9. DOI: [10.1109/WACV.2016.7477644](https://doi.org/10.1109/WACV.2016.7477644).
- [24] Yuanjie Zheng *et al.* “Single-Image Vignetting Correction from Gradient Distribution Symmetries”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.6 (2013), pp. 1480–1494. DOI: [10.1109/TPAMI.2012.210](https://doi.org/10.1109/TPAMI.2012.210).

- [25] Risheng Liu *et al.* “Real-World Underwater Enhancement: Challenges, Benchmarks, and Solutions Under Natural Light”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.12 (2020), pp. 4861–4875. DOI: [10.1109/TCSVT.2019.2963772](https://doi.org/10.1109/TCSVT.2019.2963772).
- [26] Raimondo Schettini and Silvia Corchs. “Underwater Image Processing: State of the Art of Restoration and Image Enhancement Methods”. In: *EURASIP Journal on Advances in Signal Processing* 2010 (2010), p. 746052. DOI: [10.1155/2010/746052](https://doi.org/10.1155/2010/746052).
- [27] Pedro Nuno Barbosa Leite. “Underwater Tri-dimensional Scene Understanding for an Eye-in-Hand Perception System”. PhD thesis. Faculty of Engineering of the University of Porto, FEUP, 2025. URL: <https://hdl.handle.net/10216/166694>.
- [28] Chong-Yi Li *et al.* “Underwater Image Enhancement by Dehazing With Minimum Information Loss and Histogram Distribution Prior”. In: *IEEE Transactions on Image Processing* 25.12 (2016), pp. 5664–5677. DOI: [10.1109/TIP.2016.2612882](https://doi.org/10.1109/TIP.2016.2612882).
- [29] Yujie Li *et al.* “Underwater image de-scattering and classification by deep neural network”. In: *Computers & Electrical Engineering* 54 (2016), pp. 68–77. ISSN: 0045-7906. DOI: [10.1016/j.compeleceng.2016.08.008](https://doi.org/10.1016/j.compeleceng.2016.08.008).
- [30] M. Apolinario and R. Coutinho. “6 - Understanding the biofouling of offshore and deep-sea structures”. In: *Advances in Marine Antifouling Coatings and Technologies*. Ed. by Claire Hellio and Diego Yebra. Woodhead Publishing Series in Metals and Surface Engineering. Woodhead Publishing, 2009, pp. 132–147. ISBN: 978-1-84569-386-2. DOI: [10.1533/9781845696313.1.132](https://doi.org/10.1533/9781845696313.1.132).
- [31] Pedro Pereira *et al.* “Raya: A Bio-Inspired AUV for Inspection and Intervention of Underwater Structures”. In: *IEEE Access* 13 (2025), pp. 71253–71274. DOI: [10.1109/ACCESS.2025.3560507](https://doi.org/10.1109/ACCESS.2025.3560507).
- [32] João Carvalho *et al.* “Artificial Intelligence for Automated Marine Growth Segmentation”. In: *Robot 2023: Sixth Iberian Robotics Conference*. Ed. by Lino Marques *et al.* Vol. 978. Lecture Notes in Networks and Systems. Springer, Cham, 2024. DOI: [10.1007/978-3-031-59167-9_13](https://doi.org/10.1007/978-3-031-59167-9_13).
- [33] Wenyu Cai *et al.* “Multi-modality object detection with sonar and underwater camera via object-shadow feature generation and saliency information”. In: *Expert Systems with Applications* 287 (2025), p. 128021. ISSN: 0957-4174. DOI: [10.1016/j.eswa.2025.128021](https://doi.org/10.1016/j.eswa.2025.128021).
- [34] Gary D. Melvin *et al.* “Extraction and Comparison of Acoustic Backscatter from a Calibrated Multi- and Single-beam Sonar”. In: *ICES Journal of Marine Science* 60.3 (2003), pp. 669–677. DOI: [10.1016/S1054-3139\(03\)00055-9](https://doi.org/10.1016/S1054-3139(03)00055-9).
- [35] Shitong Hou *et al.* “Underwater inspection of bridge substructures using sonar and deep convolutional network”. In: *Advanced Engineering Informatics* 52 (2022), p. 101545. ISSN: 1474-0346. DOI: [10.1016/j.aei.2022.101545](https://doi.org/10.1016/j.aei.2022.101545).
- [36] Yongcan Yu *et al.* “Real-Time Underwater Maritime Object Detection in Side-Scan Sonar Images Based on Transformer-YOLOv5”. In: *Remote Sensing* 13.18 (2021). ISSN: 2072-4292. DOI: [10.3390/rs13183555](https://doi.org/10.3390/rs13183555).
- [37] Enric Galceran *et al.* “A Real-time Underwater Object Detection Algorithm for Multi-beam Forward Looking Sonar”. In: *IFAC Proceedings Volumes* 45.5 (2012). 3rd IFAC Workshop on Navigation, Guidance and Control of Underwater Vehicles, pp. 306–311. ISSN: 1474-6670. DOI: [10.3182/20120410-3-PT-4028.00051](https://doi.org/10.3182/20120410-3-PT-4028.00051).

- [38] Gianfranco Bianco *et al.* “A Comparative Analysis between Active and Passive Techniques for Underwater 3D Reconstruction of Close-Range Objects”. In: *Sensors* 13.8 (2013), pp. 11007–11031. ISSN: 1424-8220. DOI: [10.3390/s130811007](https://doi.org/10.3390/s130811007).
- [39] F. Oleari *et al.* “An underwater stereo vision system: From design to deployment and dataset acquisition”. In: *OCEANS 2015 - Genova*. Genova, Italy, 2015, pp. 1–6. DOI: [10.1109/OCEANS-Genova.2015.7271529](https://doi.org/10.1109/OCEANS-Genova.2015.7271529).
- [40] Miquel Massot-Campos and Gabriel Oliver-Codina. “Underwater Laser-based Structured Light System for one-shot 3D reconstruction”. In: *SENSORS, 2014 IEEE*. 2014, pp. 1138–1141. DOI: [10.1109/ICSENS.2014.6985208](https://doi.org/10.1109/ICSENS.2014.6985208).
- [41] Yaming Ou *et al.* “Binocular Structured Light 3D Reconstruction System for Low-Light Underwater Environments: Design, Modeling and Laser-Based Calibration”. In: *IEEE Transactions on Instrumentation and Measurement* PP (Jan. 2023), pp. 1–1. DOI: [10.1109/TIM.2023.3261941](https://doi.org/10.1109/TIM.2023.3261941).
- [42] Luca Ebner *et al.* “Metrically scaled monocular depth estimation through sparse priors for underwater robots”. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 3751–3757. DOI: [10.48550/arXiv.2310.16750](https://doi.org/10.48550/arXiv.2310.16750).
- [43] Yelena Randall. “Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets”. MA thesis. University of Haifa (Israel), 2023. DOI: [10.48550/arXiv.2302.12772](https://doi.org/10.48550/arXiv.2302.12772).
- [44] Wang Zhiqiang and Liu Jun. “A review of object detection based on convolutional neural network”. In: *2017 36th Chinese Control Conference (CCC)*. 2017, pp. 11104–11109. DOI: [10.23919/ChiCC.2017.8029130](https://doi.org/10.23919/ChiCC.2017.8029130).
- [45] Ross Girshick *et al.* “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2014. DOI: [10.48550/arXiv.1311.2524](https://doi.org/10.48550/arXiv.1311.2524).
- [46] Ross B. Girshick. “Fast R-CNN”. In: *CoRR* abs/1504.08083 (2015). arXiv: [1504.08083](https://arxiv.org/abs/1504.08083).
- [47] Shaoqing Ren *et al.* “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *CoRR* abs/1506.01497 (2015). arXiv: [1506.01497](https://arxiv.org/abs/1506.01497).
- [48] Joseph Redmon *et al.* *You Only Look Once: Unified, Real-Time Object Detection*. 2016. arXiv: [1506.02640](https://arxiv.org/abs/1506.02640) [cs.CV].
- [49] Yunjie Tian *et al.* *YOLOv12: Attention-Centric Real-Time Object Detectors*. 2025. arXiv: [2502.12524](https://arxiv.org/abs/2502.12524) [cs.CV].
- [50] Kaiming He *et al.* “Mask R-CNN”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2018. arXiv: [1703.06870](https://arxiv.org/abs/1703.06870).
- [51] Ashish Vaswani *et al.* “Attention is all you need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017. arXiv: [1706.03762](https://arxiv.org/abs/1706.03762).
- [52] Alexey Dosovitskiy *et al.* “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint* (2020). arXiv: [2010.11929](https://arxiv.org/abs/2010.11929).
- [53] Yutong Li *et al.* “Exploring plain vision transformer backbones for object detection”. In: *European Conference on Computer Vision*. Springer, 2022, pp. 280–296. arXiv: [2203.16527](https://arxiv.org/abs/2203.16527).
- [54] Shijia Zhao *et al.* “An Improved YOLO Algorithm for Fast and Accurate Underwater Object Detection”. In: *Symmetry* 14.8 (2022), p. 1669. DOI: [10.3390/sym14081669](https://doi.org/10.3390/sym14081669).
- [55] Kun Liu *et al.* “Underwater Object Detection Using TC-YOLO with Attention Mechanisms”. In: *Sensors* 23.5 (2023), p. 2567. DOI: [10.3390/s23052567](https://doi.org/10.3390/s23052567).

- [56] Aditya Arie Muksit *et al.* “YOLO-Fish: A robust fish detection model to detect fish in realistic underwater environment”. In: *Ecological Informatics* 72 (2022), p. 101847. DOI: [10.1016/j.ecoinf.2022.101847](https://doi.org/10.1016/j.ecoinf.2022.101847).
- [57] Alzayat Saleh *et al.* “A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis”. In: *Scientific reports* 10.1 (2020), p. 14671. DOI: [10.48550/arXiv.2008.12603](https://doi.org/10.48550/arXiv.2008.12603).
- [58] Australian Institute of Marine Science (AIMS) *et al.* *OzFish Dataset - Machine learning dataset for Baited Remote Underwater Video Stations*. Dataset. 2025. DOI: [10.25845/5e28f062c5097](https://doi.org/10.25845/5e28f062c5097).
- [59] Jiawei Chen and Meng Joo Er. “Dynamic YOLO for small underwater object detection”. In: *Artificial Intelligence Review* 57 (2024), p. 165. DOI: [10.1007/s10462-024-10788-1](https://doi.org/10.1007/s10462-024-10788-1).
- [60] Hao Zhou *et al.* “Real-time underwater object detection technology for complex underwater environments based on deep learning”. In: *Ecological Informatics* 82 (2024), p. 102680. DOI: [10.1016/j.ecoinf.2024.102680](https://doi.org/10.1016/j.ecoinf.2024.102680).
- [61] Cian Fahy *et al.* “Automated Detection of Sacrificial Anodes on Offshore Wind Farms for ROV Inspections”. In: *OCEANS 2024 - Halifax*. IEEE. Halifax, NS, Canada, 2024, pp. 1–6. DOI: [10.1109/OCEANS55160.2024.10753752](https://doi.org/10.1109/OCEANS55160.2024.10753752).
- [62] Tsung-Yi Lin *et al.* *Microsoft COCO: Common Objects in Context*. 2015. arXiv: [1405.0312 \[cs.CV\]](https://arxiv.org/abs/1405.0312).
- [63] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. 2019. arXiv: [1711.05101 \[cs.LG\]](https://arxiv.org/abs/1711.05101).