machine learning & knowledge extraction

Systematic Review

State of the Art: Analysis of Deep Learning Techniques in Images Acquired in an Aquatic Environment

Vanesa Lopez-Vazquez [1,2], Geovanny Satama-Bermeo [1], Hasan Issa Raheem [1] and Jose Manuel Lopez-Guede [1,*]

1 Department of Automatic Control and Systems Engineering, University of the Basque Country (UPV/EHU) Nieves Cano, 12, 01006 Vitoria-Gasteiz, Spain; vlopez017@ikasle.ehu.eus (V.L.-V.); geovannydaniel.satamab@ehu.eus (G.S.-B.); halyaqoobi001@ikasle.ehu.eus (H.I.R.)
2 Deusto SEIDOR, 01015 Vitoria-Gasteiz, Spain
* Correspondence: jm.lopez@ehu.es

Abstract

The oceans and other marine ecosystems are indispensable to life, so the understanding and knowledge of their biodiversity is crucial to the use of their resources and exploration. These environments are complex and difficult to access, so different types of remote sensing technologies are used to study them. These intelligent sensors can collect a massive amount of data, which, once reviewed and analyzed, can help to draw conclusions and increase knowledge of these underwater environments. Manually reviewing and organizing through this large amount of information is both time-consuming and costly. Therefore, it is advisable to employ automated techniques from machine learning and deep learning fields. In recent years, these methods have proven to be efficient and have obtained very good results in solving different problems applied to the marine world: image enhancement, image classification, segmentation and object detection. This paper presents a systematic review, conducted in accordance with the PRISMA 2020 guidelines, aimed at summarizing the methods used to address underwater problems and their reported results.

Keywords: deep learning; underwater imagery; PRISMA review; image enhancement; species detection

1. Introduction

Most of our planet is covered by oceans, and many of them are unexplored. In recent decades, advances in sensorial technology have improved the imaging of ocean biodiversity. All of this, as well as the increasing development of robotic vehicles, made progress in the monitoring of the continental waters and deep seas [1].

Comprehensive knowledge of marine ecosystems and their biodiversity is essential for the sustainable utilization of their resources. To effectively explore and preserve the extensive biodiversity present within underwater ecosystems, systematic monitoring and rigorous analysis of collected data are required.

The ocean floor is inaccessible to humans due to its constituent features such as extreme pressure, cold temperatures and darkness. In this way, imaging or video devices prove to be valuable tools for oceanographic surveillance. In addition to underwater cameras or ROVs (Remotely Operated Vehicles), there are wired observatories; platforms connected to the coast that have multiple types of sensors to collect a wide variety of data [2]. These platforms can acquire image and/or video material during consecutive years, from which animals of different species can be identified and counted [3–8].

Academic Editor: Isaac Triguero

Received: 6 March 2026
Revised: 29 April 2026
Accepted: 4 May 2026
Published: 14 May 2026

The acquisition of images of the biodiversity of the oceans has increased dramatically over the last two decades [9], supporting a revolution in the monitoring of marine communities at all depths of the continental margins and on the seabed [10]. However, the large amounts of image and video data generated cannot be processed manually. That can be solved by automatic processes of classification, recognition and labelling. These automatic processes are carried out thanks to artificial intelligence techniques. These techniques are grouped into different branches, such as Computer Vision (CV), Machine Learning (ML) and Deep Learning (DL).

CV techniques have been commonly used for image processing, feature detection and pattern recognition [11]. This branch of artificial intelligence is widely used to process and enhance images, improving their quality and brightness [12,13]. These types of techniques have worked quite well, but sometimes they are not powerful enough to achieve good results. In addition, they are techniques that require multiple manual adjustments.

ML algorithms are quite popular and well-known in the world of image classification and data analysis since these algorithms are able to learn from the data. These methods have been applied to underwater image classification, marine animal identification, and a variety of other underwater tasks [14,15].

Image analysis, pattern recognition, and object detection in DL frequently employ deep neural networks. However, it was not until a few years ago that these techniques were introduced in the marine world, enabling the detection and classification of plants and animals, the improvement and restoration of underwater images and consequently, to gain a greater knowledge about the sea. Applying DL methods to these types of problems presents significant challenges, largely due to the intricate nature of the images, as they are very likely to be noisy or of poor quality. However, if applied correctly, they could help to track and count marine species without the need for constant supervision.

A number of reviews have addressed the topics of fish classification and detection, with some, such as [16], encompassing over 90 cited works. Other papers, despite containing deep state-of-the-art examinations, are not up-to-date because they were published some years ago [17]. The review contained in [18] compares more than 80 articles covering fish classification methods, feature extraction techniques, and classification algorithms. However, despite the inclusion of DL techniques, the article focuses considerably on more traditional methods such as SVM (Support Vector Machine), PCA (Principal Components Analysis), and classification algorithms based on manual features. Other reviews can also be found that focus on aerial images [19], or on more specific species (such as corals [20]) that, despite being very complete reviews, do not provide such a holistic and adaptable view of the field, due to the lack of variety in some aspects. In contrast, the present review extends the scope by providing a detailed and systematic analysis of 132 studies, offering a broader and more comprehensive synthesis of the existing literature.

The structure of this paper is as follows: Section 2 introduces the PRISMA methodology. Section 3 gives a background, setting the stage with fundamental concepts and relevant literature critical to understanding the subsequent sections. Section 4 delineates the various problem types and applications related to the study, explaining how these problems manifest in practical scenarios. Section 5 provides an overview of deep learning approaches applied specifically to underwater and water-related fish images, which is subdivided into detailed discussions, ranging from image processing and enhancement techniques to aquatic species detection, classification, and segmentation techniques, including subsections detailing the applications of different deep learning models to different types of images. In Section 6, we discuss the implications of our findings. Finally, the paper concludes with Section 7, where we draw conclusions from the research conducted.

https://doi.org/10.3390/make8050131

2. Methodology

This systematic review aims to examine the application of DL techniques applied to solve underwater image problems, encompassing convolutional neural networks (CNNs), transformer-based architectures, and hybrid multimodal approaches. The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol (accessed on 11 August 2025: www.prisma-statement.org) [21,22], ensuring a transparent, rigorous, and reproducible process.

The PRISMA methodology offers a rigorous and transparent framework for conducting systematic reviews, ensuring that the search, screening, and selection of studies are replicable and free from unnecessary bias. By requiring explicit documentation of inclusion and exclusion criteria, PRISMA enhances the reproducibility and credibility of the review process. This structure also facilitates the synthesis of large and heterogeneous bodies of literature, enabling researchers to identify trends, gaps, and methodological patterns across studies. For technical domains, such as underwater engineering or deep learning applications, PRISMA provides a clear protocol to handle diverse sources (e.g., conference papers, journal articles, preprints) while maintaining methodological consistency.

Despite its strengths, PRISMA also has limitations when applied to rapidly evolving or highly interdisciplinary technical fields. Its structured nature can reduce flexibility in exploring emerging topics that may not yet have well-established terminology or indexing in academic databases. Additionally, the meticulous documentation and multi-phase screening process can be time-consuming and resource-intensive, particularly when large volumes of literature are retrieved. In technical research, PRISMA's emphasis on predefined criteria may lead to the exclusion of innovative or unconventional studies that could offer valuable insights but fall outside the initial scope. Researchers must, therefore, balance methodological rigour with adaptability to ensure the review remains comprehensive and relevant.

2.1. Introduction

A comprehensive literature search was conducted in IEEE Xplore, Scopus, Web of Science, and arXiv, covering publications from January 2015 to August 2025. This time frame was selected to capture developments from the introduction of AlexNet to the emergence of state-of-the-art architectures such as Vision Transformers (ViT) and Segment Anything Models (SAMs). Search strings combined DL-related terms with underwater detection keywords: ("deep learning" OR "neural network" OR "machine learning") AND ("underwater" OR "fish" OR "coral" OR "plankton" OR "underwater plants") AND ("image classification" OR "object detection" OR "image restoration" OR "image enhancement").

The initial search retrieved 19,275 articles. This search was refined to a final 132 papers. Figure 1 shows an adapted PRISMA flow diagram showing the number of articles in each phase.

This systematic review was guided by the following research questions, designed to structure the search strategy, selection criteria, and synthesis of results:

- **RQ1:** What advanced deep learning architectures have been applied to the analysis of marine species imagery captured both underwater and above water?
- **RQ2:** What datasets and image acquisition methods are most frequently used in these studies?
- **RQ3:** What evaluation metrics are reported, and how do performance levels vary across methods and application contexts?
- **RQ4:** What limitations, challenges, and future research directions are identified in the literature regarding the use of deep learning for marine species image analysis?

These findings will be presented in Section 6 of this paper.

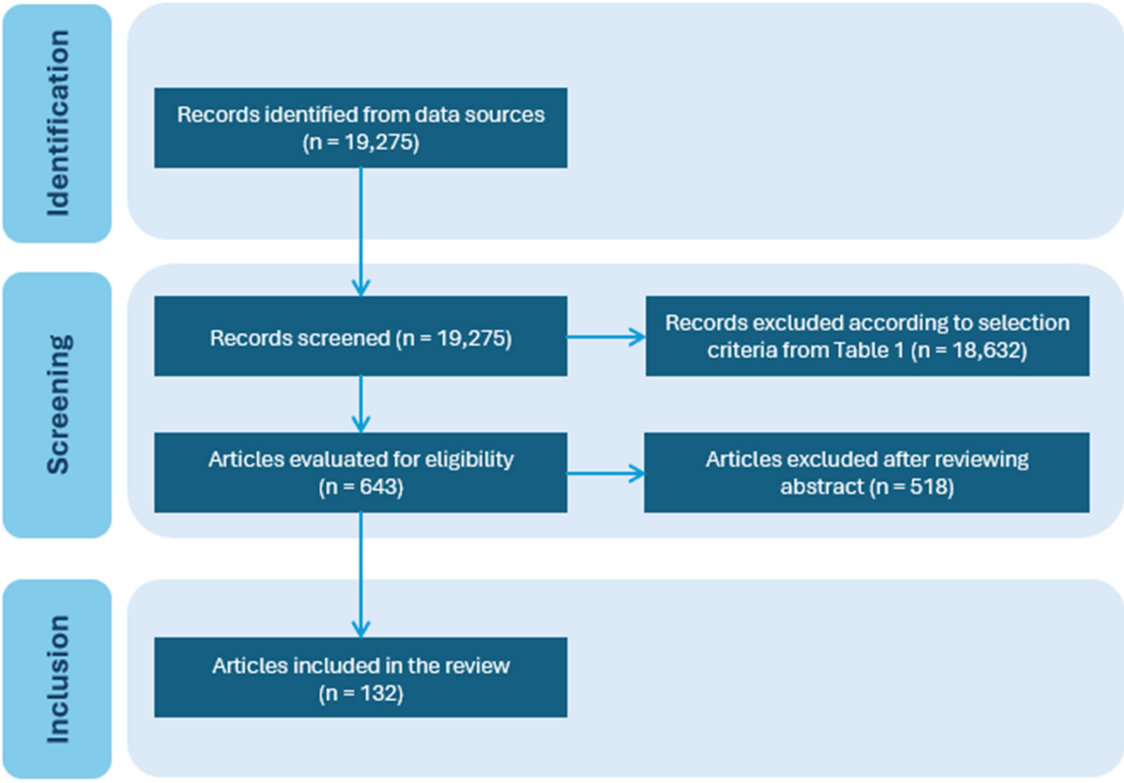

Figure 1. Adapted PRISMA flow diagram.

2.2. Screening

Studies were selected according to predefined inclusion and exclusion criteria designed to ensure relevance, methodological rigour, and reproducibility. The scope of this review was limited to the application of Artificial Intelligence (AI) methods to the analysis of images containing marine species, captured either underwater or above water. Table 1 below presents a summary of the selection criteria.

Table 1. Selection criteria.

Criterion Category	Inclusion Criteria	Exclusion Criteria
Publication type	Peer-reviewed journal articles, conference papers, and peer-reviewed preprints from recognized repositories (e.g., arXiv).	Abstract-only publications, posters, editorials, commentaries, theses, or non-indexed reports.
Methodological scope	Application of AI techniques (deep learning, CNNs, transformers, hybrid approaches) to the analysis of marine species images.	AI methods without application to marine species images.
Domain relevance	Imagery of marine species captured underwater (e.g., coral reefs, open ocean) or above water (e.g., shoreline, aerial surveys).	Studies focusing exclusively on non-marine species or terrestrial ecosystems.
Methodological transparency	Detailed description of model architecture, datasets, and training/testing procedures.	Insufficient methodological detail to enable replication.
Temporal coverage	Published between January 2015 and August 2025.	Publications outside the specified date range.
Language and availability	Written in English (or accessible language for review team) with full-text available.	Languages not accessible to the review team or full text unavailable.

https://doi.org/10.3390/make8050131

Table 1. *Cont.*

Criterion Category	Inclusion Criteria	Exclusion Criteria
Data modality	Image-based studies (still images or video frames) of marine species.	Studies using only non-visual modalities (e.g., acoustic, sonar-only, textual).
Complexity of methods	Use of advanced deep learning methods (e.g., CNNs, Vision Transformers, hybrid architectures).	Use of only traditional machine learning methods (e.g., SVM, random forest, k-NN) without deep learning models.
Access status	Articles available as open access or through freely accessible repositories.	Articles requiring paid subscription or inaccessible through institutional or open channels.

2.3. Inclusion

After applying the inclusion and exclusion criteria, a comprehensive full-text review was conducted on the 132 qualifying studies. This process aimed to capture the breadth of research on advanced deep learning methods applied to marine species imagery, encompassing both underwater and above-water contexts. The review considered variations in neural network architectures, target species, imaging conditions, and dataset composition, as well as reported performance metrics and validation approaches.

Through this systematic assessment, the review seeks to establish a consolidated view of the methodologies currently employed, identify patterns in their application, and highlight the most effective strategies for accurate and efficient species detection. The synthesis of these findings will inform best practices and guide the development of more robust AI-based tools for marine biodiversity research and conservation.

The distribution of articles published over the years is shown in Figure 2. A general upward trend is observed until 2018, followed by fluctuations in subsequent years, mainly attributable to selection criteria and the availability of studies during the analyzed period. The increase recorded in recent years highlights the growing attention to current DL approaches, although it should be noted that the data for the last few years are influenced by the time frame of the literature search.

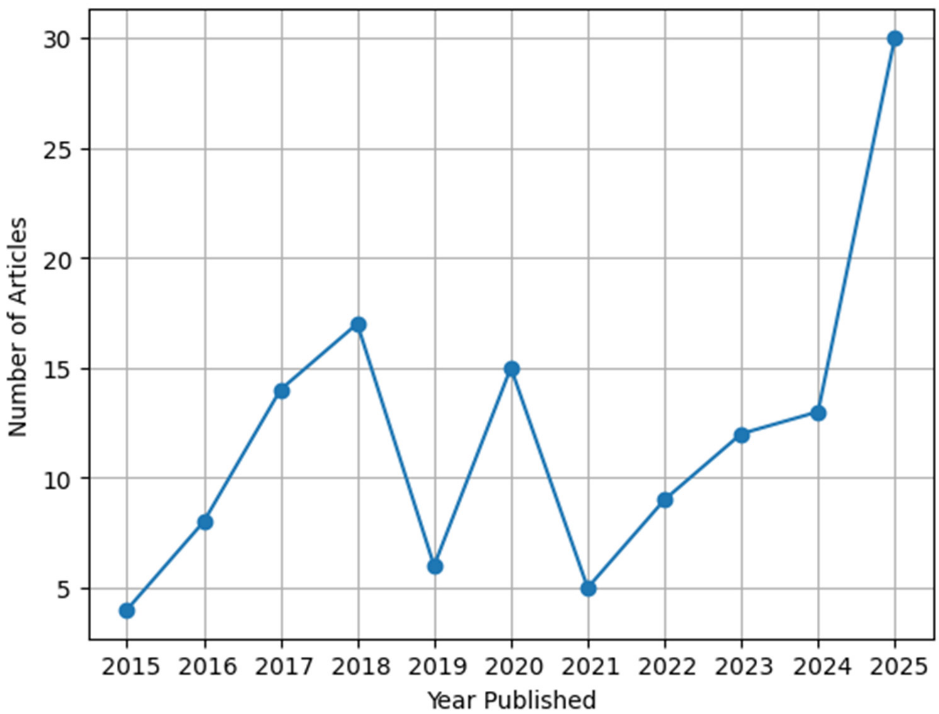

Figure 2. A distribution of the articles published over the years.

2.4. Data Extraction and Synthesis

Each included study was reviewed individually, and its key characteristics were extracted into comparative tables. The extracted variables included the article reference, the name of the dataset used, the type of task or problem it solved, the model architecture, the chosen evaluation metrics, and the values associated with each metric.

The extracted information was organized into tables to facilitate analysis and comparison between studies. These tables summarize the performance of the different models across various datasets and tasks, providing a consistent basis for comparison.

For each study, quantitative performance results were extracted from the models when available. These included accuracy, precision, recall, F1 score, mean average precision (mAP), Intersection Over Union (IoU), and, for image enhancement or restoration tasks, PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index Measure), PCQI (Perception-based Colour Quality Index), UCIQE (Underwater Colour Image Quality Evaluation), UIQM (Underwater Image Quality Measure), UICM (Underwater Image Colourfulness Measure), and other task-specific metrics reported in the original publications, among others, depending on the study.

3. Background

Artificial intelligence (AI) is the branch of computer science that attempts to solve tasks using computers for which human intelligence is required. In turn, this branch is composed of sub-branches that encompass different techniques that can solve problems automatically. Some of the best-known branches are the following: Computer Vision, Machine Learning and Deep Learning.

Computer Vision (CV) is considered a part of the artificial intelligence family, as they share topics such as pattern recognition and learning techniques. Its aim is to imitate human vision perception and reasoning by means of computers. In other words, computer vision is the analysis and processing of videos or images. CV tasks range from methods for acquiring, processing, analyzing, and understanding digital images to extracting high-dimensional data from the real world to produce numerical or symbolic information.

Machine Learning (ML), which is considered as a subset of AI, is the study of algorithms that computer systems use to perform a specific task based on pattern recognition and inference, without the need for explicit instructions. In ML, the programme infers its own rules through a process called training. In the case of supervised learning, this process consists of analyzing a sample of data (known as the training set or training data) in order to predict some properties or labels. Under some specific conditions, the model trained in this way can generalize to previously unseen data and thus produce the predictions for these.

Deep Learning (DL) is a sub-branch of machine learning, where the most basic computational model is an Artificial Neural Network (ANN), which is inspired from the human brain functionality. ANNs were first introduced in 1943 in the paper [23] by W. McCulloch and W. Pitts. They presented a computational model based on how brain neurons work with each other to perform more complex tasks using logic of propositions [24]. Early neural networks played a fundamental role in the development of ML, but they also had significant limitations. The concept of a perceptron was introduced by Frank Rosenblatt in 1958 [25]. A perceptron is a supervised linear classification algorithm designed to identify and select subsets within larger datasets based on learned criteria. However, despite its historical importance, the perceptron could not solve nonlinearly separable problems, such as the XOR problem.

But it was not until the 1980s that the interest in the ANNs was regained due to various advances made in this field, like the contribution of Hopfield in [26], who introduced recurrent neural network models. Hopfield's networks demonstrated that neural systems could reach stable states and function as associative memory. Another fundamental advance came with the development and popularization of the backpropagation algorithm, particularly thanks to the work of Rumelhart, Hinton, and Williams [27]. Backpropagation provided an efficient mechanism for training multilayer perceptron (MLP) networks. These networks consist of an input layer, one or more hidden layers, and an output layer, and are capable of learning complex and nonlinear mappings. By iteratively propagating error gradients from the output layer backwards through the network, backpropagation enables the network parameters to be updated using gradient-based optimization. The breakthrough of backpropagation, together with increased computational power and the availability of data, laid the foundations for modern deep learning. Contemporary deep neural networks extend the MLP paradigm through deeper and more specialized architectures, enabling the learning of more complex feature representations.

These networks with a greater number of layers are called deep neural networks (DNN). These networks are considered hard to train, besides being a costly process, but they can also have better results and reduce error [28]. These models are capable of learning hierarchical feature representations directly from raw data, leading to significant performance improvements over classical machine learning approaches in many tasks. In CV, this evolution has resulted in a diversity of DL architectures, such as Convolutional Neural Networks (CNNs), autoencoders, and, more recently, Transformer-based models; each addressing different aspects of visual perception and representation.

4. Types of Problems in Images and DL Applications

During the last few years, artificial intelligence has been used to solve different problems related to images in multiple fields, like agriculture [29–32], medicine [33–36] or meteorology [37–39], to name a few. However, in a general way, we can differentiate the following applications that are generally used in all areas: image processing, image classification, object detection and multimodal understanding and reporting. We can see the breakdown of these applications in Figure 3.

Digital Image Processing (DIP) is the technical analysis of a digital image by applying computer vision or machine/deep learning techniques. DIP involves low-level operations such as noise reduction, contrast enhancement and sharpening or blurring. These kinds of operations are made in order to improve further classification results, object detection results or simply achieve increased image enhancement [40].

Image classification is the process of taking an input (in this case an image) and outputting a class or a probability that the input belongs to a class or category. In other words, image classification refers to the labelling of images into one of the already predefined categories.

In order to carry out this labelling, it is necessary to extract information or characteristics from the images. This is performed by pre-processing the images, usually using computer vision techniques.

The next steps would be training and classification, both of which can be performed using machine learning or deep learning techniques [41].

Object detection is an area of computer vision and is related to image processing that deals with detecting instances of semantic objects in digital images and videos. Object Detection is modelled as a classification problem where windows of fixed sizes slide through an input image to find the possible locations of the objects and then feed these patches to an image classifier. The first object detection framework was proposed by Paul Viola and Michael Jones in 2001 [42]. It was motivated by the problem of face detection, even though it can be trained to detect a variety of elements. The framework was composed of a Haar-like feature Selector and an Adaboost algorithm for classification.

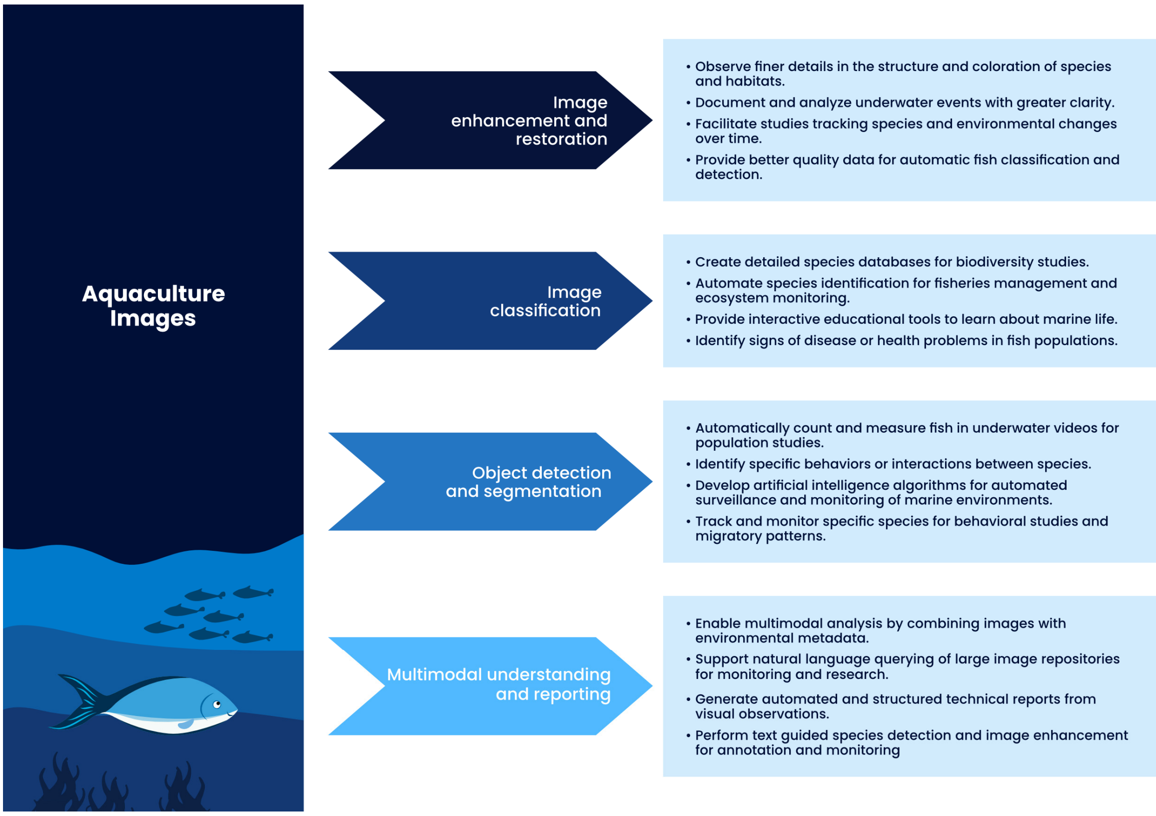

Figure 3. Different tasks that can be performed using AI techniques on images related to aquaculture.

Even though object detection can be carried out using HOG Features, methods for object detection can generally be categorized into two main groups, based on deep learning techniques: Region-based Convolutional Neural Networks (R-CNNs) and regression-based object detectors. The first group contains detectors like R-CNN [43]; the other group contains detectors that model the object detection problem as a regression problem, like YOLO (You Only Look Once) [44] or SSD (Single Shot Multibox Detector) [45].

Recent advances in Vision-Language Models (VLMs) and Large Language Models (LLMs) have made it possible to understand and generate multimodal reports as a higher-level task in the analysis of aquatic imagery. This paradigm allows images to be modelled in conjunction with natural language, sensor metadata or other relevant information. In underwater and aquatic scenarios, approaches based on these multimodal language models enable applications such as scene-level understanding and description, automated report generation and semantic summarisation of detected visual observations.

5. Results

CV techniques and ML algorithms fit well when performing image processing, item classification or object detection tasks, and therefore they have been used together to solve different problems in different areas [46,47], among them, the analysis of aquatic or underwater species field [48–53]. DL techniques are more advanced than CV and ML, and although computationally they tend to be more expensive, the results are often better. Recently, these techniques have been used in the underwater world to classify or detect species and even monitor them and better understand their behaviour [54,55].

5.1. Image Processing, Enhancement and Restoration

The processing of an underwater image is quite challenging, due to the underwater conditions, which include lack of light, water currents and other phenomena that cause distortion and noise. Even though computer vision techniques are more widespread (probably because they are older), DL has proven to be a highly effective approach for the enhancement and processing of underwater images, with neural networks playing a central role in this field. A wide range of neural network architectures have been designed and implemented to tackle the specific challenges associated with the underwater environment. Similar image quality degradation issues have also been examined in other application domains, such as medical imaging, further reinforcing the general importance of enhancement and denoising as preprocessing steps under challenging acquisition conditions [56].

Some reference datasets exist in the literature for underwater image enhancement and restoration research. The dataset called Underwater Image Enhancement Benchmark (UIEB) [57] provides a comprehensive set of degraded underwater images paired with reference-enhanced counterparts, enabling the evaluation of colour correction and dehazing methods. The Underwater Video Enhancement Benchmark (UVEB) [58] extends this concept to video sequences, supporting the assessment of temporal consistency in enhancement algorithms. The U45 dataset [59] offers 45 particularly challenging underwater scenes for testing robustness under extreme visibility, turbidity, and lighting conditions. Finally, the Enhancement of Underwater Visual Perception (EUVP) dataset [60] contains paired and unpaired underwater images collected across varying depths and environments, facilitating both supervised and unsupervised training. Together, these datasets form a comprehensive foundation for developing and benchmarking underwater image enhancement techniques.

Although CNNs are primarily used for image classification and detection, they can also be used to improve and restore image quality, like in [61], where they train a CNN to dehaze and perform image enhancement. Compared to other state-of-the-art methods, their approach produces results that closely resemble the ground truth images. They carried out two experiments in which they picked up the error obtained when classifying the treated images, in turn comparing them with Automatic Colour Enhancement (ACE) [62] and histogram equalization techniques. The error obtained with their technique in the validation set was 14.1%, while with ACE it was 15.7%, and with the histogram 20.5%.

While considered distinct networks, there are other types of networks that contain a CNN-type architecture, like **Generative Adversarial Networks (GANs)**. Generative modelling uses unsupervised learning to find patterns in data, allowing the model to generate new, unseen examples [63]. In [64], they perform underwater image enhancement by a Conditional Generative Adversarial Network (cGAN), achieving a clear image by a multi-scale generator. Although they achieved the results of the current state of the art, there still exist several limitations, as they achieved an SSIM (Structural Similarity Index Measure) value of 0.6552.

Another experiment based on GANs is performed in [65]. They built two networks called UGAN (Underwater Generative Adversarial Network) and UGAN-P (with Gradient Difference Loss), and they compared their results to CycleGAN (Cycle Generative Adversarial Network) [66]. Both UGAN and UGAN-P achieved better results. In addition, they performed the detection of swimmers both in original images and in images enhanced by the Mixed-Domain Periodic Motion (MDPM) tracker [67]. Out of 500 frames of each dataset, in the original images, there were only 42 correct detections, while in the generated images, there were 147 detections.

In Li et al. [68], the Water Generative Adversarial Network (WaterGAN) is used to produce realistic underwater images and the Underwater Image Restoration Network for correcting the colour. WaterGAN learns a realistic representation from real unlabelled underwater images; it leverages terrestrial images along with associated depth information to generate artificial underwater images. They compared their colour restoration method to other kinds of methods like histogram equalization or the Shin et al. approach [69], among others. Although the proposed method obtained very good results in terms of colour accuracy and colour consistency, the histogram equalizer also obtained high and sometimes better results. They tested the presented model on multiple datasets, including B3DO (The Berkeley 3D Object Dataset), UW RGB-D Object (Underwater RGB-Depth Object dataset) and NYU Depth (New York University Depth dataset), among others. An **autoencoder** is a type of neural network that learns without supervision and uses backpropagation for training. The aim of an autoencoder is that the output data is the same as the input data [70]. For that purpose, they learn the input data features so they can represent it later in the output layer. Autoencoders comprise three principal components: the encoder, the latent-space representation (also referred to as the code), and the decoder. In [71], authors developed an Underwater Denoising Autoencoder (UDAE) model using a denoising autoencoder with a U-Net [33] architecture as a CNN architecture for underwater image restoration. They compared their results with the UGAN [65], and concluded that their method performed better than the network based on a GAN, obtaining a lower value of MSE (Mean Squared Error). They also obtained an SSIM value of 0.9653, while the UGAN achieved a value of 0.9186.

Li et al. [72] introduced a method that eliminates scatter while maintaining colour in underwater images. They found that their method and colour correction of images influence the classification results. In addition, they propose a more comprehensive image quality assessment [73] called index Qu as a rule to compare results of the algorithm, which achieved a value of 0.7929.

Variational autoencoders are also a resource used for image quality enhancement, such as in [74], where a novel probabilistic network named PUIE-Net (Probabilistic Network for Underwater Image Enhancement) is introduced to model the enhancement distribution of degraded underwater images. Built upon the U-Net framework, the PUIE-Net architecture improves underwater images by modifying aspects like colour and contrast, while keeping the original content intact. It uses a technique called probabilistic adaptive instance normalization (PAdaIN), which combines a conditional variational autoencoder (CVAE) [75] with adaptive instance normalization (AdaIN) [76], enabling effective extraction of image features for enhancement. The network has two branches: the upper estimates the prior for a raw underwater image, and the lower constructs posterior distributions using the original image and a reference. They tested the proposed method on two datasets: images from UIEB (Underwater Image Enhancement Benchmark) and RUIE (Real-world Underwater Image Enhancement dataset) [77], which only contain raw underwater images. The network was trained with the modified set from the UIEB dataset. Evaluating on the RUIE dataset, they reached a value of 4.512 and 3.7 in Mean Opinion Score (MOS) and Natural Image Quality Evaluator (NIQE) [78], respectively. MOS was used to quantify the subjective evaluation.

Although the use of reinforcement learning is more common in solving other types of problems, it has also been used for image enhancement. In [79], they made use of a Markov decision process (MDP), where states are represented by feature maps, actions are represented by enhancement methods (basic adjustment, colour tuning, correction, and deblurring), and rewards are represented as improvements in image quality (determined by a pixelwise loss and a perceptual loss). A **deep Q network** is responsible for selecting an image enhancement action, which causes a state to change from one to another at each step of the MDP. To train and test the network, data obtained from the Underwater Image Enhancement Benchmark (UIEB) dataset was used, a dataset containing 950 authentic underwater images. They evaluated the images qualitatively (visual inspection) and quantitatively. For the quantitative evaluation, they chose 3 metrics: NIQE, Underwater Colour Image Quality Evaluation (UCIQE) [80], and Underwater Image Quality Measure (UIQM) [81], with which they achieved the values 36.5574, 0.5962, and 4.34436, respectively, on the challenging image set they preselected.

The study in [82] proposes a Multiscale Dense Generative Adversarial Network (MDGAN) to enhance underwater images. The researchers worked with a dataset that included both synthetic and real underwater images. Their proposed method achieved better performance on both metrics than other methods such as FE (Fusion Enhance), RB (retinex-based), UDCP (Underwater Dark Channel Prior), CycleGAN, WSCT (Weakly Supervised Colour Transfer), and UGAN. In qualitative tests, the method showed significant improvements in colour correction and detail recovery in underwater images compared to other existing methods.

The work in [83] proposes DNnet (Dynamic range and Normalization framework), a compact neural network for enhancing underwater images, especially in 4K. It introduces novel modules like FAN (Fast Average Normalization) and CDR (Channel Dynamic Range) to efficiently restore colour and detail. DNnet achieves real-time performance with high accuracy, outperforming larger models on datasets like UIEB (Large Scale Underwater Image Dataset) and LSUI (Large Scale Underwater Image Dataset).

Another underwater image enhancement method called UIEVUS (Underwater Image Enhancement method designed for Various Underwater Scenes) is presented in [84]. This method addresses colour distortion and uneven illumination. It decomposes images into illumination and reflection maps via Retinex theory, enhances them using GAN-based modules, and fuses results for high-quality outputs. The proposed method demonstrates state-of-the-art results on metrics like PSNR (Peak Signal-to-Noise Ratio) (23.55 dB) and UIQM (3.26), improving visual quality and downstream tasks (e.g., object detection).

This paper [85] proposes CCL-Net, a two-stage network based on cascaded contrastive learning (CCL) for underwater image enhancement. Stage 1 (CC-Net) corrects colour distortion in Lab space, while Stage 2 (HR-Ne, a Haze Removal Network) removes haze using multi-scale features. Cascaded contrastive learning progressively refines results by using raw images (Stage 1) and intermediate outputs (Stage 2) as negative samples. The CCL-Net achieved the best value of UIQM (3.021) and the second-best value of UCIQE (0.464) on the UIEB-T90 dataset.

This work [86] proposes a lightweight U-Net variant for underwater image enhancement, using simplified channel attention and SK fusion to reduce model complexity. Using the EUVP paired dataset for training, it delivers state-of-the-art performance with 9.27 million parameters and 2.74 billion FLOPS (Floating-point Operations)—outperforming nine baselines in PSNR (21.565 dB), SSIM (0.879), and efficiency. Ablation studies validate design choices like LayerNorm and GELU (Gaussian Error Linear Unit) activation. The method excels in restoring colour/contrast in challenging underwater scenes (tested on UIEB), demonstrating practical deployment potential.

UDNet (Uncertainty Distribution Network) proposes an unsupervised framework for underwater image enhancement, eliminating the need for paired training data [87]. It integrates uncertainty modelling via SGMCSS (Statistically Guided Multicolour Space Stretch module for reference generation), CVAE (feature extraction), and PAdaIN (adaptive normalization). Evaluated on 8 datasets, UDNet outperformed 10 methods in quantitative metrics (PSNR, SSIM, UIQM) and qualitative results. It excels in generalizing to unseen/unpaired data, addressing challenges like colour distortion and low contrast.

The **Transformer** model, proposed by Google in 2017 [88] and has achieved successful results in Natural Language Processing (NLP) tasks, due to its multi-attention mechanism. Later, the **Vision Transformer (ViT)** [89] was introduced, which is capable of extracting image features, achieving outstanding results in various computer vision tasks. ViTs divide images into fixed-size patches, use them as tokens, and apply self-attention to model global context efficiently. The Swin Transformer, introduced in 2021 [90], was designed as an enhanced version of ViT, offering better performance and greater computational efficiency. In [91] they propose a hybrid UNet-MaxViT (Unet with Multi-axis Vision Transformer) architecture for underwater image enhancement, integrating UNet's local feature extraction with MaxViT's global multi-axis attention to address colour distortion, low contrast, and detail loss. Evaluated on UIEB, EUVP, and UFO-120 (Dataset for Simultaneous Enhancement and Super-Resolution (SESR) of underwater imagery) datasets, it achieves state-of-the-art results: PSNR 22.91 (UIEB) and 26.12 (EUVP), outperforming 12+ methods including PhISH-Net (Physics-Inspired Network) and URSCT-SESR (U-Net-based reinforced Swin-Convs Transformer for simultaneous enhancement and super-resolution). The model's efficiency (10.82 GFLOPs (Giga FLOPS), 0.015 s/image) enables real-time deployment. Visual results demonstrate enhanced colour fidelity and structural preservation in complex underwater conditions.

Multi-scale transformers have proven to be an effective approach for image enhancement and restoration tasks by effectively capturing information at various spatial resolutions. By processing images through hierarchical representations, these models combine fine-grained local details with broader global context, which is crucial for addressing complex degradations such as noise, blur, and colour distortion. This multi-scale architecture allows for more accurate and coherent enhancement compared to single-scale methods, improving the fidelity and visual quality of restored images. This paper introduces UWFormer, a new multi-scale Transformer network for enhancing underwater images [92]. UWFormer uniquely combines a Nonlinear Frequency-aware Attention (NFA) module and a Multi-Scale Fusion Feed-forward Network (MSFN) within a semi-supervised learning framework. A key innovation is the Subaqueous Perceptual Loss (SPL), used to generate reliable pseudo-labels from unlabelled data. UWFormer surpasses leading methods in several benchmarks (such as EUVP, UIEB, U45 and RUIE (Retrieval-based Unified Information Extraction), restoring colour balance, contrast, and detail without typical artefacts such as colour casts or over-enhancement.

The **Segment Anything Model (SAM)** is a prompt-driven framework designed for versatile, large-scale image segmentation. SAM is able to precisely separate different objects and areas without needing to retrain for specific tasks. Its ability to separate foreground and background with high precision has made it a powerful tool across diverse fields, from medical imaging to remote sensing. In underwater vision, SAM offers a unique advantage by isolating scene elements that are often degraded differently due to scattering, turbidity, and colour distortion. The following work proposes a discriminative underwater image enhancement method using SAM to segment foreground/background regions [93]. It applies adaptive colour compensation and pyramid-based fusion separately to each region, eliminating foreground-background crosstalk. High-frequency edge fusion restores blurred details.

Table 2 summarizes the most relevant deep learning-based approaches for underwater image enhancement, including the corresponding references discussed above.

Table 2. Deep Learning techniques for underwater image enhancement and processing.

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[61]	Images obtained by an autonomous underwater vehicle	Underwater image enhancement	CNN	Visual inspection and error rate in posterior classification	14.1%
[64]	UIEB dataset and synthetic underwater image dataset	Underwater image enhancement	Conditional generative adversarial network (cGAN)	PSNR/SSIM (Structural Similarity Index)	17.715 dB/0.6552
[65]	Subsets from Imagenet, YouTube videos and images from Flicr™	Underwater image enhancement	GAN (U-Net) called UGAN and UGAN-P (with Gradient Difference Loss)	Gradient difference loss metrics in red, blue, green and orange patch/distances in image space	9.39, 5.50, 3.25, 5.79/94.91 (mean)
[68]	B3DO [94], UW RGB-D Object [95], NYU Depth [96] and Microsoft 7-scenes [97] datasets and 3 underwater datasets collected in a laboratory, Jamaica and Australia	Underwater image generation and restoration	WaterGAN	Euclidean distance/variance of intensity-normalized colour in RGB-space (red, green, blue)	(0.0484, 0.2132, 0.1431)/ (0.0005, 0.0007, 0.0006)
[71]	Images from the Internet (aquariums, underwater images with artificial light and processed images)	Underwater image enhancement and restoration	Underwater Denoising Autoencoder (UDAE)	MSE/SSIM	0.0028/0.9653
[72]	JAMSTEC database	Underwater image enhancement	Contrast enhancement method	Index Qu/SSIM	0.7929/0.6027
[74]	Modifed UIEB dataset and RUIE dataset	Underwater image enhancement	U-Net, CVAE, AdaIN	PSNR/SSIM/DeltaE/ NIQE/MOS	UIEB dataset: 21.86 dB/0.870/9.556/ 3.626/4.2
[79]	UIEB dataset	Underwater image enhancement	MDP and deep Q network	NIQE/UCIQE/UIQM	41.334, 0.6412, 4.5935
[82]	Images from ImageNet and SUN (Scene Understanding) database [98]	Underwater image enhancement	Multiscale dense generative adversarial network (GAN)	UCIQE/UIQM	0.6028 ± 0.0282/ 5.0973 ± 0.4163
[83]	UIEB dataset, LSUI [99] and UVEB	Underwater image enhancement	DNnet	PSNR/SSIM/ MSE/UIQM	UIEB dataset 26.335 dB/0.910/ 0.368×10^3/2.994
[84]	UIEB and LSUI datasets for training, UIEB190, LSUI850, OceanEx (full-reference); C60, RUIE dataset Color90, UPoor200, U45 (non-reference).	Underwater image enhancement	UIEVUS Framework: Integrates Retinex decomposition with GAN-based enhancement	Full-reference: PSNR/SSIM. Non-reference: UIQM	Full-reference: UIEB190 = 23.55 dB/0.90 Non-reference: U45 = 3.26
[85]	UIEB-T90 (90 images from UIEB), UIEB-C60 (60 challenging images from UIEB), EUVP-T515 (515 images from EUVP), SQUID-T16 (16 images from SQUID), UIE-T78 (78 images from RUIE dataset).	Underwater image enhancement	CCL-Net	PSNR, SSIM, UIQM, UCIQE	RUIE-T78: UIQM = 3.168, UCIQE = 0.447 UIEB-T90: PSNR = 20.181 dB, SSIM = 0.866, UIQM = 3.021, UCIQE = 0.464
[86]	EUVP and UIEB datasets	Underwater image enhancement	Improved U-Net	NIQE, UCIQE, PSNR, SSIM	4.393/0.430/ 21.565 dB/0.879

Table 2. *Cont.*

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[87]	UIEBD (Unsupervised, no ground truth). For testing: EUVP, UFO, UIEBD (paired); DeepFish [100], FISHTRAC, FishID, RUIE dataset, SUIM (Segmentation of Underwater Imagery dataset) (unpaired).	Underwater image enhancement	UDNet framework integrates SGMCSS Module, CVAE Module and PAdaIN Block.	Full-Reference: PSNR, SSIM, MAD (Most Apparent Distortion), GMSD (Gradient Magnitude Similarity Deviation). No-Reference: UIQM, MUSIQ (Multi-scale Image Quality Transformer), NIQE, UCIQE. Comparison: PSNR, SSIM, UIQM, UCIQE.	Full-Reference: EUVP: PSNR = 22.96 dB, SSIM = 0.771, UIQM = 3.265, UCIQE = 0.749 No-Reference: UCCS: UIQM = 3.974, UCIQE = 0.713
[91]	UIEB, EUVP and UFO-120	Underwater image enhancement and restoration	Hybrid UNet-MaxViT	PSNR, SSIM, PCQI (Perception-based Colour Quality Index), UCIQE, UIQM, UICM, UIConM (Underwater Image Contrast Measure), CCF (Colourfulness Contrast Fog density index).	UIEB: 22.91 dB/23.5286/ 0.9341/0.6460/1.602/ 9.6405/1.1865/37.9325.
[92]	UIEB, EUVP, EUVPUN, RUIE dataset, U45	Underwater Image Enhancement	UWFormer	PSNR, SSIM, LPIPS, UIQM, UCIQE	EUVP: PSNR = 24.40, SSIM = 0645, LPIPS = 0.129, UCIQE = 0.431 U45: UIQM = 3.227/ UCIQE = 0.440
[93]	U45, SIUM (Segmentation of Underwate Imagery) [101], UIEB, SQUID [102]	Underwater image enhancement	SAM	UIQM, CCF, AG, BC(e) (Blind Contrast Restoration Assessment in edge visibility), BC(r) (Blind Contrast Restoration Assessment in edge pixel gradient values), URanker (Ranking-based underwater image quality assessment)	UIEB: 3.544/48.508/ 27.165/1.168/3.889/1.964

5.2. Underwater Species Detection and Classification

The classification of fish is the most recurrent problem, and it has also been solved by deep learning techniques. The images used, again, can vary greatly, as the background can be static, where the photos have been obtained in an aquarium or out of water (in a controlled environment), or dynamic background in an uncontrolled environment, where the images have been obtained on the sea or river bottom and are much more complex. As shown in Figure 4, classification and detection are the final stages in a typical underwater image processing workflow, which begins with image acquisition and involves several intermediate steps such as enhancement and feature extraction.

Deep learning techniques usually obtain better results than machine learning, both in controlled and uncontrolled environments. Again, items using datasets obtained in controlled environments such as tanks or aquariums are many [103–108], as are works using images of fish out of water [109–112], in markets or on fishing boats, for example. However, datasets generated in uncontrolled environments, such as the open sea, are more popular in deep learning articles [113–120]. In such environments, factors like water currents, particle clouds, and low light make accurate detection and classification challenging.

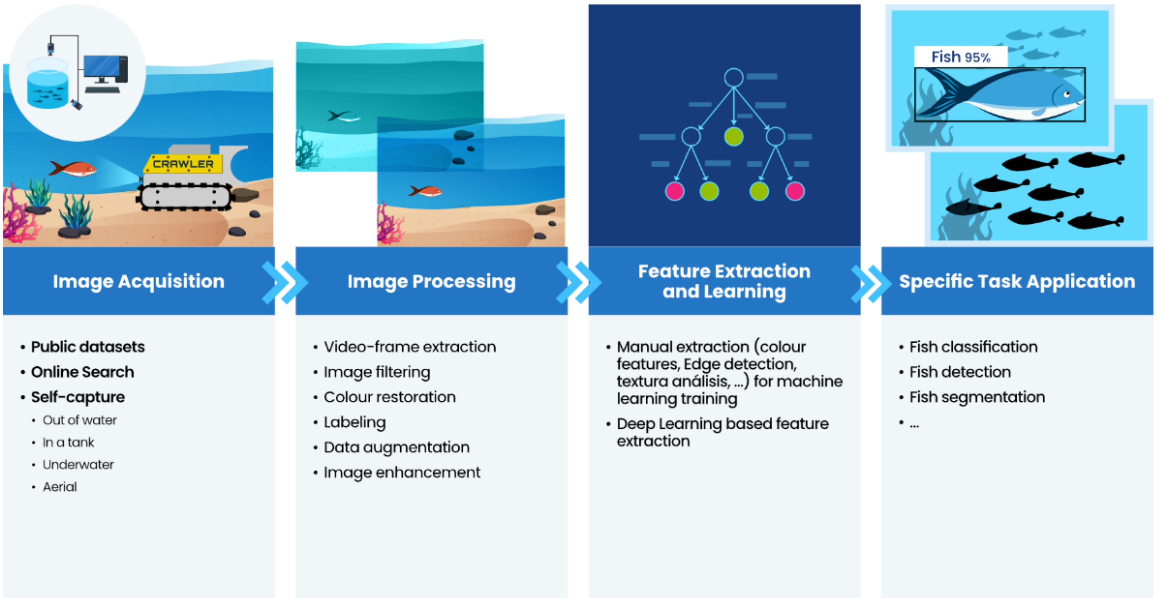

Figure 4. Tasks contained in a typical process involving images of aquatic environments and AI techniques.

Many papers use datasets provided by fixed long-term underwater observatories (FUOs), which use multiple sensors to collect data (videos, images, etc.) from the marine world around them. In these datasets, the problem of counting and tracking species is usually addressed in order to analyze the marine environment to which the images belong.

The LoVe (Lofoten–Vesterålen) Observatory is a FUO situated in Norway that provides a large dataset of images and multiple data from different sensors (like temperature or turbidity). This observatory's data has been used in some papers [121,122].

At present, there are few public datasets with underwater fish images, and most are relatively simple in terms of complexity. SeaCLEF (Sea Content-based Conference and Labs of the Evaluation Forum, https://www.imageclef.org/lifeclef/2017/sea, 5 March 2026) is a benchmark dataset developed as part of the CLEF (Conference and Labs of the Evaluation Forum) initiative, focusing on marine biodiversity monitoring. It typically includes images and videos of underwater species, habitats, and environmental conditions, aiming to support research in species recognition, habitat classification, and environmental assessment in marine contexts. Fish4Knowledge (https://homepages.inf.ed.ac.uk/rbf/fish4knowledge/, 3 May 2026) is a large-scale dataset containing millions of underwater video frames collected from coral reefs in Taiwan. It provides annotations for fish detection, tracking, and species classification, enabling research in underwater computer vision, ecological monitoring, and automated marine life surveys. The NOAA Fisheries datasets (https://www.st.nmfs.noaa.gov/aiasi/DataSets.html, 3 May 2026) contain images and videos collected by the U.S. National Oceanic and Atmospheric Administration for fishery monitoring and management. It includes various fish species in different environments, both in situ and on deck, supporting research in species identification, size estimation, and population assessment. LifeCLEF (Life-based Conference and Labs of the Evaluation Forum) is an annual challenge and dataset series dedicated to biodiversity identification, covering multiple domains such as plants, birds, mammals, and marine species. The marine branch includes underwater imagery for species recognition and habitat mapping, serving as a benchmark for AI-driven biodiversity assessment.

5.2.1. Aerial, Plankton, Aquatic Plants and Coral Detection and Classification

Although the use of **aerial image** datasets is not so widespread, some researchers have used them, obtaining high detection and classification results. In [123], two different Faster R-CNN models, one with a ZF (Zeiler and Fergus model) and the other VGG (Visual Geometry Group model)-based, are used for stingray detection on aerial images. The experiments showed that the VGG model achieved higher average precision, reaching the 99.7% in one of the three test datasets used.

The study in [124] presents results on two different datasets, the first containing underwater images and the second composed of aerial images. They chose a pretrained RetinaNet as the object detector. For tracking, they applied the Simple Online Realtime Tracker (SORT) algorithm over the detections performed by the detection method. For the aerial dataset, the average precision for each of the three classes was 0.4 for Ray, 0.25 for Diver and 0.75 for Shark. The lower results were due to varying annotation counts for each class. In the initial dataset, there were more false positives because certain fish were detected but not annotated.

In [125], the authors developed a solution for whale detection and counting in aerial images. The system was composed of two steps: first, GoogleNet-Inception v3 CNN architecture was applied for whale detection; second, a Faster R-CNN model built on the Inception-Resnet v2 CNN architecture performed the counting. They obtained high results in both detection and counting.

Although they used a low-complexity CNN, in [126] they used aerial dataset to perform species detection and classification over aquatic areas. For detection, they initially halved the size of each frame to reduce the number of pixels to be processed. Subsequently, they identified a homogeneous background by calculating a fine-grained saliency map, where low-value areas represent the background and high-value areas represent possible detections. Finally, the coordinates of the local maxima in high-saliency areas were identified, extracting a window around each coordinate to classify it and determine whether it contained an animal or was part of the background. The performance of the CNN differed depending on the species and factors like morphology, spacing, behaviour, and habitat uniformity. The overall precision of the model for detecting seal and sea turtles was below 0.27 for both classes, while for detecting gannets it was 0.74.

In [127], the authors proposed a human-in-the-loop method for cetacean detection in images. First, generic deep learning models produce a binary land cover map to filter irrelevant images and identify key samples for annotation. Next, an active learning strategy refines a segmentation model with selected images undergoing manual and AI-assisted annotation, iteratively improving performance. Finally, a human reviewer verifies all model detections, correcting them, if necessary, to improve the quality of the final analysis. The selected model was a U-Net architecture with EfficientNet-b3 as the encoder. The trained model managed to detect 84% of the detections on the whole dataset observed by one of the experts who analyzed the dataset.

Most studies using aerial datasets over the sea aim to detect whales, probably because they are large mammals that can be seen from afar. The dataset used in [128] was collected during an aerial survey over Cumberland Sound Bay, Nunavut, Canada, in August 2014. A pre-trained and fine-tuned Faster-RCNN model was used for all experiments. To evaluate the effect of the sliding window approach on object detection, patches of four different sizes (256×256, 512×512, 768×768, and 1024×1024) were cropped with a 20% overlap of the full images. Experimental results indicate that increasing the input image size improves the model performance.

We can find a study that presents a solution for detection in aerial images based on YOLOv8 in [129]. This paper proposes PGG-YOLO (P2P5hGhostConv-G_Ghostbottleneck-

YOLO), a lightweight fish detection model for identifying turned white belly fish in pond environments using UAV (Unmanned Aerial Vehicle) imagery. It enhances YOLOv8 with GhostConv, G-Ghost bottleneck, and a small object detection layer. The model achieves a high mean average precision (mAP50 = 99.4%) and runs efficiently (10.3 ms/frame) with a compact model size (8.62 MB). A custom dataset was built from 3000 annotated UAV images. The method shows excellent detection and counting accuracy, aiding practical aquaculture management and insurance evaluation.

The methods and studies cited in this section are organized and detailed in Table 3.

Table 3. Deep Learning techniques for underwater object detection in aerial images.

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[123]	Aerial images	Stingray detection and classification	Two different Faster R-CNN models (one with a ZF and other VGG based)	Average precision	99.7%
[124]	Underwater images and Aerial images	Marine species classification	RetinaNet as the object detector/Simple Online Realtime Tracker (SORT) for tracking	Average precision	Aerial: 0.4 for Ray, 0.25 for Diver and 0.75 for Shark
[125]	Three datasets combining images from Google Earth38, free Arkive83, NOAA Photo Library84, and NWPU-RESISC45 dataset	Whale detection and counting	GoogleNet Inception v3 CNN architecture for detection and Faster R-CNN based on Inception-Resnet v2 CNN architecture for counting	F1-measure	Detection: 81%, Counting: 94%
[126]	Aerial images	Gannets, seals sea turtles	CNN	Average Recall/average precision	0.826/0.403
[127]	Aerial images	Cetacean detection and classification	U-Net with EfficientNet-b3	Accuracy/f1-score/recall/precision	91.37%/95.49%/98.96%/92.26%
[128]	Aerial images	Beluga whale detection	Faster-RCNN	Intersection over Union	0.79
[129]	Custom dataset built from UAV imagery	Detection and counting of turned white belly fish	PGG-YOLO based on YOLOv8	Precision/Recall/F1-score/mAP50	99.52%/97.66%/98.58%/99.4%

As **plankton** is fundamental in the marine ecosystem, it is very important to carry out monitoring tasks to understand its behaviour and to know its habits, and several papers focus on studying them.

In 2016, Lee et al. [130] used the WHOI-Plankton database [131] containing about 3 million images of approximately 100 plankton species. However, 90% of the images were for only 5 types of plankton. To solve this, they proposed a method using CNN and transfer learning (training with CIFAR10, the Canadian Institute For Advanced Research dataset). For all classes, they obtained an accuracy value of 92.80%, while for the 5 main classes it was 94.77%.

In [132], they used another databased called Planktonset [133]. They developed different versions of a deep convolutional neural network, with different input image sizes. Even though the loss value was over 0.6, they showed effectiveness in plankton classification.

In the paper [134], they proposed ZooplanktonNet, which is a deep convolutional network for Zooplankton classification. For training the net, they used a zooplankton dataset with 9460 images that involves 13 classes. The final accuracy value obtained by this network was 93.7%.

The use of transfer learning is a currently recurrent and successful technique. In [135] it is also used to improve the classification of plankton. They carried out experiments in three datasets (WHOI, ZooScan and one by Kaggle), comparing different structures of the state of the art, such as AlexNet [136], VGG16 or ResNet50 (Residual Network with 50 layers) among others, with their ensemble method. They obtained a maximum F-measure value of 0.953, 0.897 and 0.926 on the WHOI, ZooScan and Kaggle datasets, respectively.

The authors in [137] propose a novel detection strategy by combining a cycle adversarial network and a densely connected YOLOV3 model (YOLOV3-dense), which improves rare taxa detection by increasing the data volume and reducing feature loss. Starting from the WHOI-Plankton dataset, they generated two augmented datasets using a CycleGAN to produce fake data from the original unpaired data to increase the data volume of rare taxa. The results show that this strategy achieves a mean precision (mAP) of 97.21% and 97.14% on the two datasets, outperforming models such as YOLOV3-tiny, YOLOV3, and Faster-RCNN, especially in rare taxa detection, where the precision improves by 4.02% on average. Moreover, the detection time is competitive, being significantly lower than that of the Faster-RCNN model, indicating that the proposed model is efficient and effective for real-time monitoring of the plankton ecosystem.

Ref. [138] analyses the performance of Transformer architectures on a plankton dataset, comparing their classification accuracy and computational speed with various CNN and Transformer neural networks. They propose a novel inter-class similarity distillation algorithm based on feature prototypes, allowing a smaller network to improve on plankton recognition guided by a larger network. The feature extraction capability of Swin-B was chosen after comparing a selection of Transformer neural networks and CNNs, which was afterwards transferred to five lighter networks and evaluated on a taxonomic dataset of in situ plankton images. Finally, the four models that performed better and the Swin-B were tested on images of the coastal waters of Guangdong, China.

In [139], five experiments were carried out to assess the impact of data augmentation, preprocessing, the use of different models, the combination of models, and class clustering. Several deep learning algorithms were compared, and different methods of data augmentation, fine-tuning, and training of all layers were evaluated. Finally, the results of the networks were combined using ensemble learning, showing significant improvements in average classification rates. Models were trained using two Kaggle plankton datasets from Oregon State University's Hatfield Marine Science Centre. DenseNet achieved success rates of 78% for the larger ensemble (118 classes) and 92% for the smaller one (38 classes).

To facilitate comparison, the main features of the previously mentioned studies are listed in Table 4.

Table 4. Deep Learning techniques for plankton classification.

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[130]	WHOI-Plankton database	Plankton classification	CNN and transfer learning (training with CIFAR10)	Accuracy	92.80%
[132]	Planktonset	Plankton classification	Deep convolutional neural network	Loss	60%
[134]	Zooplankton dataset	Zooplankton classification	ZooplanktoNet	Accuracy	93.7%
[135]	WHOI, ZooScan and one by Kaggle	Plankton classification	Their ensemble method compared to AlexNet, VGG16 or ResNet50 and more	F-measure	0.953, 0.897 and 0.926

Table 4. *Cont.*

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[137]	Datasets generated from the WHOI-Plankton database	Plankton classification	YOLOV3-dense	mAP	97.21%
[138]	Dataset collected by PlanktonScope in the coastal area of Guangdong	Plankton classification	Transformers (Swin-T, ViT-B, and Swin-B) and CNNs (ResNet50, ResNet101, ResNet152, MobileNet V2, ShuffleNet)	Precision/recall	92.38%, 91.73%
[139]	Dataset captured by Oregon State University's Hatfield Marine Science Centre	Plankton classification	Inceptionv3, InceptionResNetv2, DenseNet, ResNet, and VGG-16	Precision/recall/f1-score/accuracy/loss	92%, 92%, 92%, 0.93, 0.31

Other papers, like [140], focused on detecting and identifying **underwater plants** situated on the ocean bottom, like Posidonia. They compare different ML and DL techniques' performance. The selected methods include, on the one hand, combinations of techniques like SVM algorithm and artificial neural network (ANN), with Gabor filters, texture descriptors and co-occurrence matrix, and on the other hand, a CNN. Even though the training time was higher with the CNN, the detection hit ratio was better than with the rest of the methods.

The paper [141] goes further, as it aims to identify and segment Posidonia Oceanica semantically. To do this, they use a network architecture that can be divided into an encoder and decoder, called VGG16-FCN8 (for which good results have been seen above [142]). They used six datasets to train and test the network and, in turn, different case studies. All of them obtained an AUC (Area Under the Curve) value of over 95%.

The study in [143] presents a deep learning-based segmentation system that analyses surveillance camera images to detect Posidonia, as well as providing information on coastal features such as shoreline changes, erosion and sediment accumulation. The system was tested on images of Torre Canne beach, in Puglia, Italy, provided by the "former Puglia Interregional Basin Authority". The convolutional neural networks used in the analysis include Detectron2, Mask R-CNN and a proprietary CNN model, which obtained the highest results.

This paper [144] proposes an unsupervised learning framework to classify underwater lake vegetation without labelled data. It applies ConvNeXt for features, reduces dimensions with UMAP (Uniform Manifold Approximation and Projection), and clusters using multi-algorithm voting. The approach enables high-precision classification (up to 97.32% accuracy) and supports the construction of unbiased, large-scale ecological datasets. It was validated on public and private datasets from different lake environments. The method reduces labeling efforts, offering a cost-effective solution for lake ecosystem monitoring.

Within the study of aquatic vegetation, algae represent a key component in both freshwater and marine ecosystems. Their detection and monitoring have gained special relevance in the context of water quality control. In this regard, the authors of [145] propose AlgaeNet, a scalable deep learning model designed to detect floating algae Ulva prolifera in MODIS (Moderate Resolution Spectroradiometer) and SAR (Synthetic Aperture Radar) images. AlgaeNet builds upon the established U-Net architecture with two specific modifications: multi-channel physical information input and a new loss function designed to address the problem of imbalanced samples. The model achieved better results for the SAR image dataset, and as for the MODIS dataset, it barely achieved an Intersection over Union (IoU) value of 48.57%. Adding high-resolution SAR images increases the algae detection by 63.66% compared to MODIS images.

The study in [146] introduces SFD-YOLO (Seafloor-Debris-YOLO), an optimized deep learning framework combining super-resolution reconstruction (SRR) and object detection to automate seafloor debris monitoring in turbid waters. Using a custom dataset from Thailand's Koh Tao Island, the RDN SRR (Residual Dense Network with Super-Resolution Reconstruction) model achieved top image enhancement (PSNR: 41.02 dB, SSIM: 95.08%). SFD-YOLO, enhanced with attention mechanisms and dynamic convolution, detected debris at 91.2% mAP on high-resolution images and 89.7% on RDN-reconstructed images. The approach significantly outperformed traditional methods, offering a cost-effective solution for marine conservation.

As part of this research [147], deep learning is used to automate eelgrass monitoring by classifying its presence or absence in underwater videos. A custom dataset (8324 images) from Danish transects was annotated via the SeagrassFinder platform. **Vision Transformers (ViT)** achieved top performance (96% Receiver Operating Characteristic Area Under the Curve, commonly called AUROC) with UW-enhanced images. A novel temporal coverage estimation method provided ecological insights beyond pixel-based approaches. The pipeline reduces manual annotation costs and supports scalable marine conservation efforts, demonstrating ViT's efficacy in challenging underwater conditions.

Table 5 presents a comparative summary of the studies discussed above.

Table 5. Deep Learning techniques for aquatic plants identification.

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[140]	Dataset obtained in the coastal areas of Mallorca	Detecting and identifying underwater plants (Posidonia)	SVM and ANN, with Gabor filters, texture descriptors and co-occurrence matrix, and on the other hand, a CNN	Mean of the detection hit ratio	Dataset1: 96.37%, Dataset2: 93.35%
[141]	Self-created with an AUV in Palma Bay, Cala Blava and Valdemossa port	Identify and segment Posidonia Oceanica semantically	Encoder and decoder, called VGG16-FCN8	AUC	95%
[143]	Images of Torre Canne beach in Puglia	Posidonia detection	Mask R-CNN, Detectron2, custom CNN	Accuracy/loss/IoU	90.85%/0.18/0.68
[144]	DeepSeagrass datasets and two private datasets: Erhai Lake and Wuhan East Lake datasets (China)	Underwater vegetation classification	ConvNeXt. UMAP and clustering (K-Means, Agglomerative Clustering and Birch)	Accuracy/precision	DeepSeagrass: 97.32 ± 0.36%/90%; Erhai Lake 92.43 ± 0.97%/93%; Wuhan East Lake: 96.15 ± 0.60%/95%
[145]	SAR and MODIS images	Ulva prolifera detection	AlgaeNet (U-Net based model)	Accuracy/precision/recall/f1-score/IoU	99.83%/95.46%/92.32%/93.86%/88.43%
[146]	Custom seafloor debris dataset from Koh Tao (Thailand), COCO and TrashCan dataset	Automated seafloor debris detection and classification	SFD-YOLO (enhanced YOLOv8)	mAP@0.5	91.2% (TrashCan pretraining)
[147]	Custom dataset from Lynetteholm project (Denmark)	Presence/absence classification of eelgrass and temporal coverage estimation in underwater videos	Different versions of ResNet, InceptionV3, DenseNet and ViT	Accuracy, AUROC, Calibration Error (CE)	ViT: 0.902/0.959/0.087

Since **corals** are vitally important to marine life, several jobs involve the task of monitoring and detecting corals.

Mahmood et al. [148] proposed a method that combined hand-crafted features and CNN features to perform coral classification. The MLC dataset was used to train and validate the method, which contains 2055 images collected over three years and annotated with nine different labels, five of the labels being corals. This automatic method is based on a pre-trained VGGnet for feature extraction and a two-layer Multilayer Perceptron (MLP) for classification. Three experiments were performed with different combinations of data and different feature representations. The highest achieved results were an accuracy value of 84.5% and an Average Class Precision (ACP) of 0.69.

The same year, they used the system proposed in their previous work [148] to detect coral reefs in images [106] with a CNN based on VGGnet. A subset of the Australian benthic Benthoz15 dataset [149] was used to fine-tune the weights of the network. They performed three different experiments considering different compositions of data from the years 2011, 2012 and 2013. For all experiments, the achieved accuracy was higher than 92%. They also analyzed unlabelled coral data from three sites of the coral reef of the Abrolhos Islands from two years, 2010 and 2013. The results concluded that the area covered by coral decreased during this period.

The presented workflow on [150] makes possible the analysis of the activity of cold-water coral polyps over a certain period of time. The CNN they use achieved an accuracy value of 0.96 on the test and validation set. The network was applied to one of the datasets obtained by the LoVe Observatory.

In [151], they demonstrated the effectiveness of using drone imagery combined with deep neural networks to monitor coral reefs, providing accurate, high-resolution information in a cost-effective manner. The study used RGB drone images captured over the North Bay coral reef on Lord Howe Island, Australia. Five sets of orthomosaic images were collected over one year using a DJI Phantom 4 Pro drone. The images were segmented into 4096×4096 pixel blocks for model training. A deep neural network architecture, mRES-uNet (Multi Resolution U-net architecture network), was used, which is an adaptation of the classic U-Net model for image segmentation. This improved version incorporates "MultiRes blocks" to learn features at different scales and "Res paths" to improve the connection between the encoder and decoder layers. For unbleached corals they achieved a precision of 0.96, recall of 0.92, and a Jaccard index of 0.89, whereas for the bleached corals they gained a precision of 0.28, recall of 0.58, and a Jaccard index of 0.23.

This study introduces CoralClassify, a deep learning framework using modified ResNet50 to automate coral health monitoring [152]. A dataset of 1700 images (balanced healthy/bleached classes) was compiled from Flickr, StructureRSMAS (Structure Rosenstiel School of Marine and Atmospheric Science dataset), and ReefBase. Custom augmentation resolved class imbalance, while hyperparameter optimization ensured efficiency. The model achieved 87.6% accuracy in just 15 epochs, outperforming existing methods in speed and performance. This approach enables scalable reef monitoring for conservation, with potential for mobile deployment in future work.

The HKCoral benchmark introduces a densely annotated dataset for coral growth form segmentation in challenging underwater environments [153]. It benchmarks 17 semantic segmentation algorithms, revealing significant performance gaps (best baseline: SegFormer (Segmentation framework which unifies Transformers with lightweight MLP decoders), 72.88 mean Intersection over Union; mIoU). The proposed complementary architecture, fusing original and enhanced images, achieved a mIoU value of 73.54. This approach enables accurate coral cover estimation and 3D distribution mapping, reducing sampling bias in ecological surveys. The dataset and methods advance automated coral reef monitoring for conservation.

For clarity, the works cited above are categorized and detailed in Table 6.

Table 6. Deep Learning techniques for coral analysis.

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[106]	A subset of the Australian benthic Benthoz15 dataset	Detect coral reefs in images	A CNN based on VGGnet	Accuracy	92%
[148]	MLC (Moorea Labelled Coral) dataset	Coral classification	For feature extraction: Pre-trained VGGnet; for classification: a two-layer Multilayer Perceptron (MLP)	Accuracy	84.5%
[150]	LoVe Observatory dataset	Analysis of the activity of cold-water coral polyps in a certain period of time	CNN	Accuracy	96%
[151]	RGB drone images captured over the North Bay coral reef on Lord Howe Island, Australia	Coral classification	mRES-uNet	Overall accuracy/average Jaccard index	85.74%/0.56
[152]	Dataset created with images taken from Flickr, StructureRSMAS and ReefBase	Coral health classification	CoralClassify framework (Modified ResNet50)	Accuracy/Precision/Recall/F1-Score	87.6%/87.99%/87.06%/87.37%
[153]	HKCoral Dataset (collected mostly in Hong Kong)	Coral growth form segmentation	Complementary Architecture (based on a CNN): Fuses original and enhanced underwater images to improve segmentation.	mIoU	73.54

5.2.2. Classification of Underwater Animal Species

Deep learning methods have demonstrated strong performance in classifying underwater images. Most of these approaches rely on CNN-based neural networks, as convolutional neural networks remain the leading model for image recognition and classification tasks. These networks contain many layers that transform their input with convolution filters of a small extent [154]. In 1998, LeCun et al. [155] presented the first CNNs called LeNet-5, even though they were already working with them in 1989 [156]. However, as with papers using classical machine learning algorithms, many other papers choose not to focus on a single technique and to compare various deep learning techniques in terms of ranking and performance results.

CNNs are extensively utilized in deep learning for the analysis of visual data, especially in tasks such as image classification and object detection. Their capacity to autonomously learn hierarchical features has made CNNs the predominant choice for classifying various species of marine animals.

Rimavicius et al. [120] performed a comparison between a DNN, DBN (Deep Belief Network) and CNN classifying Norwegian seabed species. The study utilized data gathered in the Norwegian Sea by a remotely operated vehicle (ROV). Each of the images was segmented and labelled for further classification. They generate four datasets. The first one (DS1) was composed of the patches extracted for every region of each of the images and was augmented from 4589 to 18356 samples. As the first dataset was composed of five imbalanced classes, the second dataset (DS2) was composed of the dataset DS1 and some additional samples of the less numerous classes. The dataset DS3 contained information on 111 features of each of the images. The fourth dataset (DS4) was created by reshaping image patches into data arrays. They use the CNN on the datasets DS1 and DS2, while the DNN and DBN were used on the datasets DS3 and DS4. The highest overall accuracy value they obtained was 88.74% by region and 92.78% by pixel, with the CNN on the dataset DS1.

In the paper [105], the DeCAF (Deep Convolutional Activation Feature) is applied on two datasets: Taiwan sea fish and the Monterey Bay Aquarium Research Institute (MBARI) benthic animal. Hand-designed features and CNN features classification result's errors are compared. They concluded that even though CNN features provided satisfying results, hand-designed features worked better with low-quality or noisy images.

Liang et al. [104] present a classification system for ornamental fish, which also helps the user by controlling the temperature of the aquarium depending on the species, as well as calculating if the fish is moving correctly, in order to know if it is sick. They chose a CNN for the classification, which obtained an accuracy value of 98.5%.

In [111] they created a dataset by taking pictures of fishes they took out of the water. They trained a CNN network to be able to differentiate four classes of carp, obtaining a 100% accuracy value.

Qiu et al. [157] explore other transfer learning strategies, which consisted of pre-training the network, first on the ImageNet dataset and then pre-training it again on the Fish4-Knowledge (F4K) dataset. They implement three types of Bilinear CNNs [158] and evaluate them with different combinations of techniques for enhanced data augmentation. The implementation of B-CNN plus refined squeeze-and-excitation blocks obtained the highest results on both datasets, Croatian fish dataset [159] and QUT (Queensland University of Technology) fish dataset [160], with values of 83.92% and 71.80% of accuracy, respectively.

Another approach that employs CNN for fish recognition can be seen in [161]. They designed three different structures of CNNs and evaluated their accuracy results over several iterations. They achieved similar values of accuracy (over 96%) with two of the models, but the other model had a problem of overfitting.

Many papers like [162] examine how deep learning approaches can be an efficient solution for underwater imagery analysis. In the fish recognition task, the researchers used a CNN optimized with SGD (Stochastic Gradient Descent) on the Fish4Knowledge dataset and achieved a test accuracy of 98.57%.

Qin et al. [163] created a deep learning approach to identify live fish within the Fish4Knowledge dataset, primarily utilizing a ConvNet and then classifying with a linear SVM. The foreground of the images was extracted for input into the network using a technique grounded in sparse and low-rank matrix decomposition [164]. They attained state-of-the-art performance, achieving an accuracy of 98.57% on the test dataset.

Rathi et al. [165] employed the same dataset and developed a method integrating CNNs with advanced preprocessing methods, such as Gaussian blurring, morphological operations, and Otsu's thresholding. They evaluate their method on the Fish4Knowledge dataset, using Adam optimizer and three different activation function: ReLU (Rectified Linear Unit), Softmax and tanh. They achieved an accuracy value of 96.29% using the ReLu activation function, while with the tanh and Softmax did not reach the 75% of accuracy.

Understanding the behaviour of animal species can help discover if there are problems or abnormalities in the environment that lead to changes in their routines. Related to this, in article [166] they prepared a water tank to monitor a small school of fish: to watch their movement, their speed when eating or resting, etc. With the help of a CNN, they performed the classification of six types of fish behaviour. They achieved a test accuracy value of 0.825.

Although not directly related to diseases, paper [167] does seek to create a solution to improve the quality of life and health of certain fish. In fishing industry one of the main challenges is production loss. This problem is usually caused by bad handling of the fish; thus, it can be very stressful, it may end in fish death. In order to detect that kind of stressful situations, Jovanović and his fellows presented a novel algorithm based on using of CNNs for splash.

A paper dealing with this topic is [168], in which they present an improved version of convolutional network based tracker (CNT), called *Fast-CNT2*. One of the improvements of this network is that can perform multi-target tracking. For background modelling they use the improved GMM as the processing technique that does not require training.

Although it is not a very common task, probably due to the complexity it can entail, some articles investigate the genus classification of specimens using DL techniques. In [169], they introduce a convolutional neural network architecture that is tailored for classifying the gender of Chinese mitten crabs using images of their shell and abdomen. Their network obtained higher accuracy values than other models such as BPNN (Back Propagation Neural Network) and CNN-SVM.

The studies cited and analyzed in the previous paragraphs and whose methods are based on CNNs are organized and detailed in Table 7.

Table 7. Application of CNNs in underwater images.

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[104]	Images from the internet	Ornamental fish classification	CNN model	Accuracy	98.5%
[105]	Taiwan sea fish and Monterey Bay Aquarium Research Institute (MBARI) benthic animal	Underwater species classification	The DeCAF (Deep Convolutional Activation Feature) framework (CNN)	Overall error	2.09 ± 0.51%/ 0.83 ± 1.78%
[111]	Self-constructed	Carp species classification	CNN-based method	Accuracy	100%
[120]	Data collected in the Norwegian sea by a remotely operated vehicle (ROV)	Classify Norwegian seabed species	DNN, DBN and CNN	Accuracy	88.74% by region and 92.78% by pixel
[157]	Train on ImageNet and F4K, Test on the Croatian fish dataset and QUT fish	Fish classification	Three types of Bilinear CNNs	Accuracy	83.92% and 71.80%
[161]	SeaCLEF	Fish species classification	CNN	Accuracy	96%
[162]	Fish4Knowledge	Fish recognition	CNN with SGD as the optimizer	accuracy	98.57%
[163]	Fish4Knowledge	Live fish recognition	Deep architecture for live fish recognition composed principally of a ConvNet and a linear SVM	Accuracy	98.57%
[165]	Fish4Knowledge	Fish classification	CNNs and several preprocessing techniques like Gaussian blurring, morphological operations and Otsu's thresholding	Accuracy	96.29%
[166]	Own dataset (water tank)	Fish behaviour classification	CNN	Accuracy	82.5%

Table 7. *Cont.*

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[167]	Videos collected in a water tank	Anomaly detection in carp and koi ponds by analyzing behaviours	Method based on CNN	Accuracy	99.9%
[168]	Videos obtained at the culture pond of the Department of Aquaculture of National Taiwan Ocean University	Detect stressful situations when fishing fish in industry	An improved version of convolutional network-based tracker CNT (Fast-CNT2)	-	-
[169]	Self-created dataset	Chinese mitten crab's gender classification	Custom CNN	Accuracy	98.90%

VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman [170] widely used in image classification tasks. In [171], authors use a CNN based on a VGG-16 for the classification of freshness in dead fish, which can be an interesting issue in the fishing industry. They achieved high accuracy value: 98.21%. This method could be applied to other fish diseases that can be seen with the naked eye. Figure 5 shows the architecture of a VGG-16.

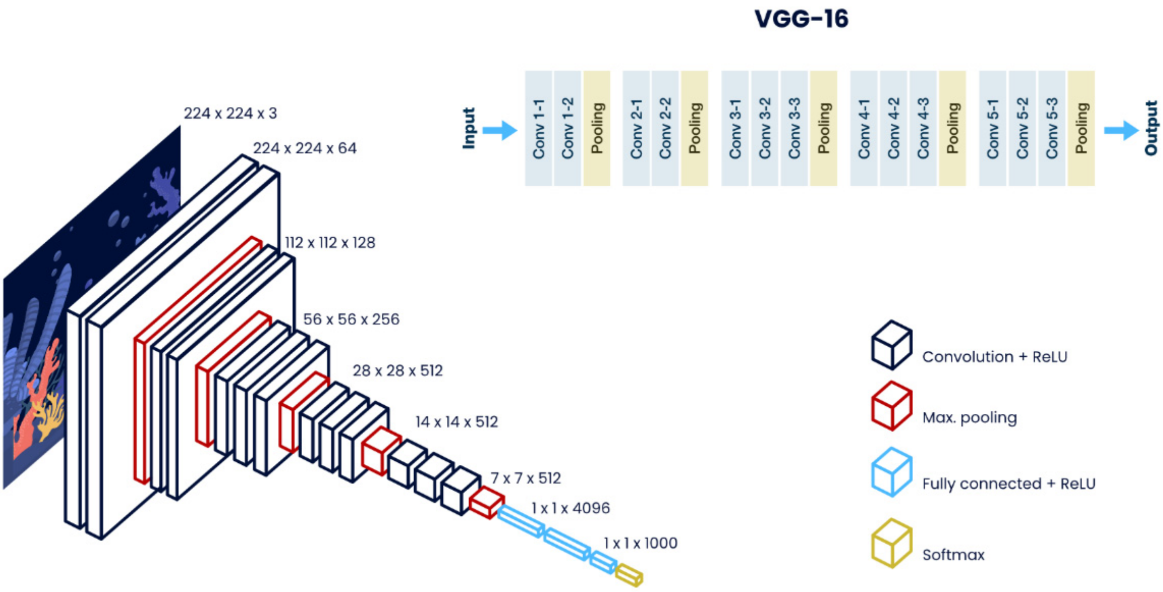

Figure 5. VGG-16 network structure.

Becken et al. performed a research with two objectives related to the aesthetic value of the Great Barrier Reef (GBR) [172]. The first objective has the aim of determine the aesthetic value using eye tracking over photos from the GBR, while the second objective consist in the recognition of objects in underwater images of the GBR and automatically assessing their aesthetic value. The images they used were taken from different sources, and many of them were photoshopped to add species or remove them. For the first objective, 21 images were used, and 705 usable surveys were performed, but for the second one they created a dataset composed of 4909 images with 50 different species to detect and classify. Three different CNNs were compared in the experiments carried out to detect and classify species: ZF, CNN-M and VGG-16. The VGG-16 achieved the highest value of mAP (0.824) in the whole sample.

The work in [103] is part of the development of the FishCam monitoring system [173] that is designed for semi-automatic monitoring of fish migration, and it describes a second classification step of the study that classifies the different fish species. As training data, they used the images extracted from the videos their collected in a controlled environment created in a river for three years, as well as the measures of the length of the fish and the date of the migration. A pretrained VGG-16 in the ImageNet dataset [174,175] was used for classification. The highest accuracy was obtained by the network configuration where the classification is based on the images, in addition to the length and the date; they achieved an accuracy of 89.4%.

In [176], the authors developed a system for fish detection and classification, composed of two branches, one for image-level classification and the other for instance-level classification. The image-level classification estimates a probability value for each of the classes the image may belong to, considering all the elements found in the image. The instance-level classification performs fish detection, object pose estimation and horizontal alignment before class prediction is made. The final prediction is the average value of the classification made in the two branches. Object detection was performed by SSD and YOLOv2. They evaluated three different CNNs in the instance classification problem: ResNet50, VGG-16 and InceptionV3. In the image-level classification, just a fully convolutional network (FCN), which was a modified VGG-16, is used. The adaptive prediction average achieved the lowest loss value, which was 0.604.

In [177], authors use a dataset provided by Kaggle, which contains 3777 images extracted from video footage of a fishing boat. They generated other datasets, one with annotated images and the other two applying transformations to the original and annotated images, creating two datasets with noisy images, containing 12,275 images each one. For classification, two VGG-16 networks were used, one of them with transfer learning from a pre-trained network on the ImageNet dataset. Better results were obtained by the simple VGG-16 model for the dataset composed of noisy original images, achieving 99.38% of accuracy, and values of precision, recall and F-score above 0.91. One reason why transfer learning did not work so well could be because of the difference between the images contained in ImageNet and those used for testing.

In [178] they propose a combination of VGG-16 with Darknet for fish classification. The training session lasted 72 min for 500 epochs covering the 951 photographs. The annotated dataset was downloaded from roboflow.ai. These annotations included bounding boxes to locate fish within the images and labels to classify them into one of three classes: "Adipose," "Non-Adipose," or "Unknown" (representing fish that were out of view). The precision they achieved was 0.7.

The study in [179] developed and compared two CNN architectures for fish species classification: VGG-16 and a reduced version of VGG-16, VGG-8. The latter is a simplified version of VGG-16, specifically designed to reduce the time and memory resources required for training, without significantly compromising accuracy. VGG-8 has a total of 8 layers, with 6 convolutional layers and 2 fully connected layers, making it more efficient in terms of training time and memory usage. The VGG-16 architecture achieved high classification accuracy with micro-average, recall, and F1-score accuracy. However, it required more training time and memory resources due to its higher depth and number of parameters. The scaled-down version, VGG-8, also showed competitive performance with micro-average, recall, and F1-score accuracy. Although the accuracy was slightly lower than that of VGG-16, VGG-8 was significantly more efficient in terms of time and resource usage.

The researchers in [180] developed a modified version of the VGGNet model, named MLR-VGGNet. It incorporates multi-level residual (MLR) blocks to improve feature extraction capabilities. Several models were compared, including VGG16, VGG19, ResNet50, Inception V3, Xception, and the newly proposed MLR-VGG16 and MLR-VGG19 (Multi-Level Residual VGG) models. Both models achieved high accuracy values on both datasets, outperforming existing models such as ResNet50 and Inception V3. The proposed MLR-VGG16 model achieved the best performance with a test accuracy of 98.46% on the Fish-gres dataset [181], outperforming other state-of-the-art models.

In [182] we can find another of the few studies related to fish disease detection. The authors compared the performance of multiple algorithms, both ML and DL. Neural networks obtained better precision, recall, and f-score values for each class, in addition to achieving higher accuracy values. The ResNet-50 model obtained an accuracy of 99.28%, making it one of the best pre-trained models for this task, although the VGG16+VGG19 combination achieved the highest precision.

Table 8 summarizes the most relevant VGG-based approaches for underwater image classification, including the corresponding references discussed above.

Table 8. Application of VGGs in underwater images.

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[103]	Self-constructed dataset from Austrian river	River fish classification	Pretrained VGG-16	Accuracy	89.4%
[171]	Dataset obtained from a fish farm (Khorramabad, Lorestan province, Iran)	Freshness diagnosis of common carp	Method based on CNN (VGG-16)	Accuracy	98.21%
[172]	Images from Flickr	Aesthetic value of underwater images and species recognition	ZF, CNN-M and VGG-16 Conditional generative adversarial network (cGAN)	mAP	82.4%
[176]	Kaggle fishing boat dataset	Fish detection, object pose estimation and horizontal alignment before class prediction	SSD and YOLOv2 as detectors, VGG-16 as pose estimation, VGG-16 and Inception V3 as classification	Loss	60.4%
[177]	Kaggle fishing boat dataset	Fish classification	Two VGG-16 networks (one with transfer learning)	Accuracy	99.38%
[178]	Fish Image dataset from roboflow.ai	Fish classification	VGG-16 and Darknet	Precision/recall	70%/80%
[179]	Fish4Knowledge	Fish classification	VGG-8 and VGG-16	Micro average: Precision/recall/ f1-score	99%/99%/99%
[180]	Fish4Knowledge and Fish-gres dataset (out of water fish images)	Fish classification	MLR-VGG16 and MLR-VGG19	Accuracy	Fish-gres: 98.46%/ Fish4Knowledge: 97.09%
[182]	Dataset collected from various public sources	Fish disease classification	ML algorithms, VGG16, VGG19, ResNet-50, VGG16+VGG19 and VGG16+Inception V3	Accuracy	99.64%

AlexNet, GoogLeNet (Inception), and ResNet represent landmark convolutional neural network architectures that have significantly advanced image analysis tasks.

Sungbin Choi describes the work done by his team in the LifeCLEF fish task 2015 [183]. They performed foreground detection and fish species classification with the **GoogleNet** (pretrained on ImageNet). As the images used are temporarily connected (since they have been obtained from videos), the classification result has been refined by comparing the contiguous frames of each k-frame. The best precision result obtained was 0.81.

Jäger et al. described in their paper the results obtained in the recognition task in SeaCLEF 2016 [184]. For the generation of bounding boxes, they use adaptive background mixture models, followed by erosion and a blob detection method to filter the smallest blobs. The generated proposals are utilized to extract CNN features, specifically from AlexNet

(pretrained on Large Scale Visual Recognition Challenge 2012 dataset, better known as ILSVRC 2012). Finally, they used an SVM for discriminating between fish and background. They obtained a precision of 0.66, a lower value than in 2015, which was of 0.81. In 2017, they introduced a multi-object tracking approach that built upon the improved method developed by Mothes and Denzler [185], as well as their own earlier research. The method combines CNN activations with a two-stage graph-based tracker, which deals with the problem of object's occlusion [186]. To compare their method results, they used a subset of the SeaCLEF 2016 dataset. Even their approach achieved a Multi-Object Tracking Accuracy (MOTA) value of 87.6, they did not exceed the result obtained by the method of Mothes and Denzler.

The filtering deep convolutional network (FDCNet) is a proposed model for classifying underwater species [119]. They also build a dataset called Kyutech10K, which contains seven classes, 10,728 images, and 1489 videos derived from the database of the Japan Agency for Marine–Earth Science and Technology (JAMSTEC). The proposed structure overcomes the underwater degraded images descattering problem due to the use of the Underwater Dark Channel Prior (UDCP) estimator and deep convolutional neural fields (DCNF), which calculate the disparity and remove the noise from the images. Then, a modified GoogleNet performs classification. With that combination of techniques, they achieved an accuracy rate exceeding 92%.

Even though aerial drones are useful for recording images of the sea from the top, the development of underwater drones enables recording images of organisms under the sea, directly from their environment, which could be more representative. The authors in [107] developed an underwater drone that has a 360° panoramic camera embedded. However, they performed classification on a dataset collected from a Google search, composed of 100 images of each of the four kinds of fish, which they augmented up to 86.400 images. The recognition accuracy rate they achieved was 87%, 85% and 67% for the three networks they used, **AlexNet**, **GoogleNet** and **LeNet**, respectively. Although data augmentation is a common strategy for addressing the problem of limited data availability, extreme data augmentation can have adverse consequences. Operations such as rotation and blurring can improve training stability, but they do not generate entirely new samples, increasing the risk of overfitting. Furthermore, when the original images are obtained from controlled sources (for example, in aquariums or in properly lit areas) and high accuracy scores are achieved, this does not necessarily mean that the model will generalise correctly in real-world aquatic conditions.

Many papers have shown that transfer learning can obtain better results when classifying, like [187], where they compare two networks, **AlexNet** and **GoogleNet**, following this strategy. The dataset they used was provided by The Nature conservancy through the Kaggle competition "The Nature Conservancy Fisheries Monitoring" *(https://www.kaggle.com/c/the-nature-conservancy-fisheries-monitoring, 3 May 2026.)*. Without transfer learning, they barely reached a success rate higher than 80%, whereas with it, they achieved an accuracy of at least 96% with both networks.

This paper [112] introduces SuperFish, a mobile app for recognizing 38 Mauritian fish species using dual approaches: traditional computer vision (kNN, 96% accuracy) and deep learning (Inception-v3, 98% accuracy). The custom dataset (1520 images) was processed via contour extraction and geometric/colour features for traditional methods, while deep learning used transfer learning.

The most recent studies on underwater image classification and detection tend to employ more advanced deep learning architectures than those used in earlier work. Whilst more classical networks such as AlexNet provided a fundamental starting point for the adoption of convolutional networks in this domain, subsequent research has exploited deeper and more specialised architectures, incorporating attention mechanisms, transfer learning and optimised design for complex underwater scenes. Consequently, the performance differences observed between older and more recent studies should be interpreted as the result of the natural evolution of architectures and available resources, rather than as a direct comparison between methods evaluated under equivalent conditions.

Among these more recent studies is the one presenting DAMNet, wich introduces a dual-attention mechanism, CBAM (Convolutional Block Attention Module) with multi-stage stacking to classify underwater biological images [188]. It achieves 96.93% accuracy on a 7-class dataset, outperforming benchmarks like ResNet50 (91.11%) and GoogLeNet (95.24%). The Gravity Optimizer reduces loss (0.1860) and accelerates convergence. Though effective overall, fish classification lags (91.84%) due to intra-class diversity. The model addresses challenges like turbidity and low contrast in marine imagery.

The authors of [189] address the task of distinguishing raw underwater images from enhanced ones, which is relevant for underwater image processing pipelines. The authors propose a modified ResNet-18 that augments a pre-trained backbone with additional fully connected layers to better model underwater degradations. Transfer learning is applied by freezing early layers, while newly added layers focus on noise, colour distortion, and illumination variability. Class imbalance is addressed through data augmentation and downsampling. Experiments on SAUD (Subjectively Annotated UIE benchmark Dataset) and MSRB (Marine Snow Removal Benchmarking) datasets show that the approach outperforms standard CNN and Transformer-based baselines while remaining computationally lightweight.

Table 9 presents a comparative summary of the studies that apply complex and deep neural networks (such as ResNet or Inception) analysed previously.

Table 9. Application of deeper CNNs in underwater images for classification and detection.

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[107]	Images from Google Search (86.400 with data augmentation)	Marine species classification	AlexNet, GoogleNet and LeNet	Accuracy	87%
[112]	Dataset collected mostly from open fish markets	Identification of fish species which are found in Mauritian waters	InceptionV3 model	Accuracy	98%
[119]	Kyutech10K dataset (10,728 images and 1489 videos)	Underwater species classification	A modified GoogleNet	Accuracy	92%
[183]	Pretrained on ImageNet, SeaCLEF	Fish species classification	GoogleNet	Precision	81%
[184]	SeaCLEF 2015	Fish species classification	AlexNet, SVM	Precision	66%
[186]	SeaCLEF 2016	Fish detection and classification	A method based on the refined method of Mothes and Denzler [185]	MOTA value	87.6%
[187]	Kaggle fishing boat dataset—The Nature conservancy	Fish classification	AlexNet and GoogleNet	Accuracy	+96% (with transfer learning)
[188]	Dataset created from multiple sources: OceanDark, RUIE, UIEB, UFO-120, EUVP	Multiclass classification: Fish, Turtles, Sea Urchins, Sea Cucumbers, Corals, Humans, Wreckage	DAMNet	Overall Accuracy/ Precision/Recall/ F1-Score/Loss	96.93%/96.70%/ 96.78%/96.74%/ 0.1860
[189]	SAUD [190], PHISMID [191], MSRB [192]	Binary underwater image classification (raw vs. enhanced)	Modified ResNet-18	Accuracy, Precision, F1-score, AUC-ROC	SAUD: 96%/99%/ 95%/96%

5.2.3. Object Detection and Segmentation Applied on Aquatic Images

Object detection presents greater complexity and challenge compared to classification, as it requires not only identifying the presence of an item within an image but also accurately classifying it. In marine environments it is even more complicated because it is a hostile environment, and the images used usually contain a lot of noise and can be blurry, and they are often not well lit.

Although object detection can be accomplished with HOG features, contemporary methods are typically divided into two main categories, both leveraging deep learning approaches: region-based convolutional neural networks and regression-based object detectors. Region-based methods include models such as R-CNN [43] and its faster versions, Fast R-CNN [193] and Faster R-CNN [194]. Spatial Pyramid Pooling (SPP-net) and Mask R-CNN [195], among other object detectors, can also be included in this family.

In [196] they propose PVANet, a lightweight deep neural network designed specifically for detecting fish in underwater images. It integrates C.ReLU (to reduce early-stage computation), Inception modules (for multi-scale features), and HyperNet (to fuse hierarchical features). Evaluated on the ImageCLEF dataset (24 k images, 12 species), it achieves 89.95% mAP, surpassing Faster R-CNN by 7.25% with faster inference (0.089 s/image). The design addresses missed small-fish detection in prior work through optimized multi-scale feature extraction.

Region-Based Convolutional Neural Networks (R-CNN) generate a collection of bounding boxes from an input image, with each bounding box enclosing an object and indicating its classification. This family of nets is slower compared to others, but obtains high classification and detection results, so it has been widely used in the underwater field. There are other faster variants, such as Fast R-CNN, where the convolution operation is performed only once per image and a feature map is generated. Faster R-CNN is considerably faster than its predecessors, since instead of using a selective search algorithm to identify region proposals as R-CNN and Fast R-CNN do, a separate network called RPN (Region Proposal Network) is used.

Controlling the population of many species of fish in some areas is a necessity, because many of them have adapted to other habitats in which they do not have predators, thus their population would increase heavily. In [114], they designed a real-time detection system using an underwater robot (called OpenROV) and implementing a red lionfish detection system, for dealing with the problem of this specie's invasion. The detection is made by a Matlab implementation of a R-CNN, which was trained with 1000 frames. They obtained a true positive detection rate of 93%.

Li et al. [197] employed the Fast R-CNN framework for fish detection and identification on a subset of the SeaCLEF 2014 dataset. The model architecture was based on AlexNet, which had been pretrained on the ImageNet **(ILSVRC2012)** dataset. Performance analyses were conducted by comparing the Deformable Part Model (DPM), Region-based CNN (R-CNN), and two implementations of Fast R-CNN; one incorporating truncated Singular Value Decomposition (SVD) to compress fully connected layers. Fast R-CNN yielded the highest mean Average Precision (mAP) at 81.4%, followed by R-CNN at 81.2%, and Fast R-CNN with SVD at 78.9%. Although the Fast R-CNN with SVD had slightly lower precision than R-CNN, it demonstrated substantially greater training and testing efficiency.

The study in [114,198] presents an enhanced Faster R-CNN model for identifying marine organisms, including sea cucumbers, sea urchins, scallops, starfish, and algae. They improve the model replacing the VGG16 network by Res2Net101 in the feature extraction module; the Online Hard Example Mining (OHEM) algorithm is used to address positive/negative bounding box imbalance; and the use of IoU was replaced by Generalized IOU (GIOU) and implementing Soft-NMS (non-maximum suppression) over standard

NMS, which is simpler and easily adapted to different object detection algorithms. Initially, 2372 image samples of underwater environments (such as sea cucumbers, sea urchins, scallops, starfish and algae) were collected. Data augmentation was performed using the Mosaic technique, which consists of randomly combining four images into one by random arrangement, zooming and cropping. Through ablation experiments (removal tests), the impact of each improvement was evaluated independently. The results showed that the improved model achieves a mAP@0.5 of 71.7%, 3.3% higher than the original Faster RCNN model.

The research in [199] proposed two multi-stream fusion approaches based on Faster R-CNN: Shared RPN Fusion, which uses a single region proposal network (RPN) shared between two CNNs to improve the detection of moving objects; and Shared Classifier Fusion which uses two RPNs, one for appearance and one for motion, but shares a single classifier. Using the LifeClef 2015 Fish (LCF-15) dataset, the Shared RPN Fusion architecture achieved an F-score of 83.16%, surpassing other leading fusion methods for underwater fish detection.

In [115] they present the Video and Image Analytics for Marine Environments (VIAME) open-source computer vision library. This platform contains several modules to detect and classify fishes or other organisms; while some of them are specific to individual applications, others are more general. Faster R-CNN was also added to the system as an object detector due to its generality. They performed the network training on data collected at the Monterey Bay Aquarium Research Institute. In [116] they also used a Faster R-CNN for object detection and fish abundance estimation. They used three classification models from [170,200,201] (which are CNNs) in combination with the RPN (Region Proposal Network). The second experiment performed, in which only species with acceptable training samples were considered, obtained better results. The model in [170], a modified VGG-16, achieved a mAP value of 0.824.

The aim of the study in [202] is to detect and analyse the behavioral trajectory of fish in a simulated aquatic environment with different levels of ammonium chloride to observe how these conditions affect fish behavior and vitality. To do so, they compare the performance of Faster R-CNN and YOLOv3. The experimental results indicate that fish movement decreases significantly in the presence of ammonia, with fish remaining motionless for longer as the ammonia concentration increases. Faster R-CNN achieved a higher average recognition success rate, but YOLOv3 showed superior efficiency in detection speed.

Detecting small animals, such as shrimp, is more complex, as they can sometimes blend into the background if they do not stand out too much and models are not able to generalize well. The dataset used in [203] includes 20,000 images of shrimp captured in various aquatic conditions. The images were obtained in RGB and grayscale, with different formats and lighting conditions, both during the day and at night, to increase the robustness of the model. They built a CNN using a transfer learning technique based on the Faster R-CNN and InceptionV2 architecture. The model showed good real-time performance, although with a slight delay in frames per second (FPS), resulting in a slow-motion video effect.

This study [204] uses Faster R-CNN to automatically detect, identify, and count 11 deep-water snapper species in BRUVS footage from New Caledonia. Using 12,100 annotated fish images, the model achieved strong performance for well-represented species (F-measure up to 0.87). A semi-automatic protocol (expert verification of AI detections) significantly improved accuracy (F-measure up to 1.0). Automatic abundance estimates correlated highly with manual counts (*r* = 0.85), rising to 0.96 in semi-automatic mode, demonstrating feasibility for fisheries monitoring. The approach addresses data scarcity in deep-sea fisheries management while reducing manual processing time.

Residual Neural Network (ResNet) was presented in [205] with an innovative architecture designed to address the issue of vanishing or exploding gradients. This network uses the technique called skip connections, which skips training from a few layers. Several papers have used this network to solve problems in underwater images. Some papers have focused on the detection and counting of jellyfish, such as in [118]. This paper introduces a system called Jellytoring, which uses a deep object detection neural network to automatically identify and measure various jellyfish species. The researchers chose to implement the Faster R-CNN model with Inception ResNet v2. They reached an F1 score of 93.8% in the jellyfish detection task. In 2017, Labao et al. [117] proposed a 152-layer Fully Convolutional Residual Network (ResNet-FCN) [206] for fish segmentation and identification. Six videos were chosen from various locations within the Verde Island Passage (Philippines). Each video presents unique challenges, including abrupt changes in illumination, reduced visibility, or the presence of marine snow; a phenomenon resulting from suspended particles in underwater environments. Due to the quantity of frames, they used weakly labeled ground truth derived from the adaptative Gaussian mixture-based background subtraction [207]. The average precision obtained was 65.91%, since although the accuracy obtained for most of the sites was greater than 70%, for some of the sites it was quite low. However, the achieved average recall was higher, about 84%.

The articles in which different types of R-CNNs are applied in underwater images are listed in Table 10.

Table 10. Application of region-based networks in underwater images.

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[114]	OpenROV images	Marine species detection	R-CNN	True positive detection rate	93%
[115]	Data from Monterey Bay Aquarium Research Institute	Detect and classify fishes or other organisms	Video and Image Analytics for Marine Environments (VIAME) open-source computer vision library	ROC (AUC)	70%
[116]	Videos obtained in marine waters of beaches and estuaries across southeast Queensland	Object detection and fish abundance estimation	Faster R-CNN	mAP	82.4%
[117]	Six videos recorded at various locations within the Verde Island Passage, Philippines	Fish segmentation and identification	Fully Convolutional Residual Network (ResNet-FCN)	Precision/ accuracy/recall	65.91%/70%/84%
[118]	Self-created with a GoPro at the seafloor	Jellyfish detection	Faster R-CNN-based implementation of the Inception ResNet v2	F1-score	93.8%
[196]	ImageCLEF	Fish detection and classification	PVANet to DPM, R-CNN, Fast R-CNN and Faster R-CNN	mAP	90%
[197]	SeaCLEF 2014	Fish detection and classification	Fast R-CNN based on an AlexNet	mAP	81.4%
[198]	Self-created underwater dataset	Underwater marine species detection and classificarion	Faster RCNN with Res2Net101	Average precision/ mAP/f1-score	43%/71.7%/55.3%
[199]	LifeClef 2015	Fish detection and segmentation	Two multi-stream fusion approaches based on Faster R-CNN	F1-score/mAP	83.16%/73.69%

Table 10. *Cont.*

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[202]	Self-created in a tank at laboratory	Fish detection and trajectory analysis	Faster R-CNN and YOLO-V3	Accuracy/proportion of lost points	98.13%/1.87%
[203]	Images downloaded randomly from various sources	Shrimp detection	Faster R-CNN InceptionV2	Precision/recall/ f1-score/accuracy	97%/97%/ 96%/96%
[204]	Underwater baited video footage from New Caledonia (South Pacific)	Automated detection, species identification, and counting of deep-water snapper species	Faster R-CNN with Inception-ResNet V2 backbone	Recall/Precision/ F-measure	Best specie (*Etelis coruscans*): 0.91/0.84/0.87

The second category of detectors approaches object detection as a direct regression problem, simultaneously predicting object classes and bounding box coordinates. Representative models in this group such as **YOLO (You Only Look Once)** [44] (as well as the improved versions, like YOLO9000 [208], YOLOv4 [209] and the latest YOLOv9 [210]) and **SSD (Single Shot Multibox Detector)** [45] detect objects by processing information through the network just once. These detectors are faster than the R-CNN family, but they are not as accurate detecting small objects.

SSDs perform object detection in a single forward pass through a neural network by simultaneously predicting bounding boxes and class probabilities at multiple scales. Unlike two-stage detectors (like R-CNN family), SSDs eliminate the need for a separate region proposal step, resulting in faster inference while maintaining competitive accuracy, especially for real-time applications. Figure 6 shows the operation of the YOLO network detection operation.

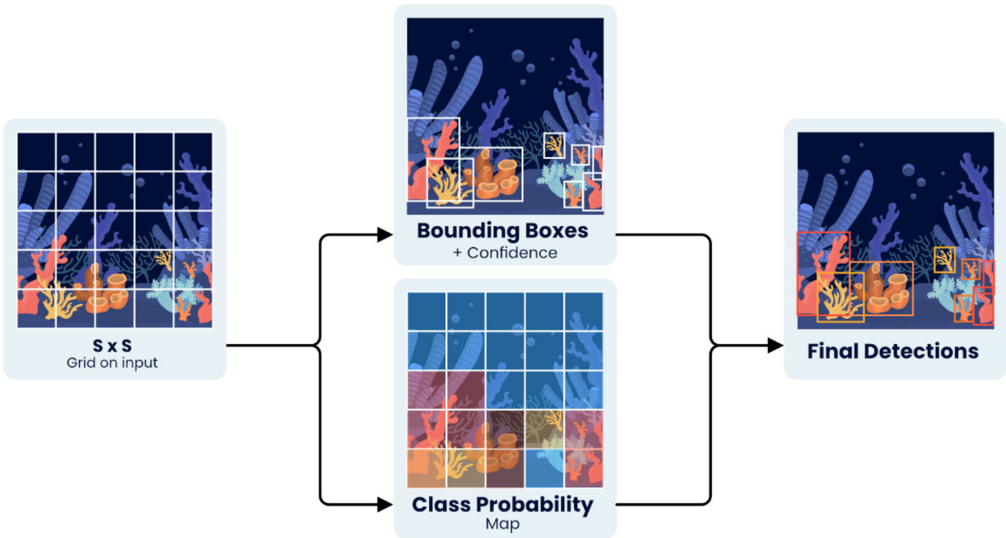

Figure 6. Diagram of the detection operation of the YOLO network. It divides the image into an S $\times$ S grid and for each cell in the grid it predicts bounding boxes, the confidence for those boxes, and the class probabilities.

The Kyutech10K dataset was partially used in [211], for training a system for underwater real-time recognition and tracking of four different organisms using their previous underwater image enhancement method [72] and YOLO. They compared the performance of YOLO with other two methods: Tracking Learning Detection (TLD) [212] and MedianFlow [213]. YOLO and Medianflow obtained good results, but YOLO showed many advantages, such as the adaptative bounding boxes in different frames for the same object.

While YOLO achieved relatively high precision and recall scores for each species, it failed to detect some objects and could not continue with tracking in those cases. There are other jobs that use tracking to monitor fish health, like work done in [214]. This paper introduces a framework for detecting anomalies in carp and koi pond health monitoring by analyzing the behaviours of these species, quantifying changes in their movement patterns, and identifying unusual behaviours through tracking. The framework is based on YOLOv3.

The authors in [176] developed a fish detection and classification system that relies primarily on object detection techniques, specifically SSD and YOLOv2. The system features two branches: the image-level branch provides overall class probabilities based on all elements in the image, while the instance-level branch focuses on detecting individual fish, estimating their pose, and aligning them horizontally before classification. Three CNN architectures were tested for instance-level classification, whereas the image-level classification used a fully convolutional network based on a modified VGG-16. By combining predictions from both branches through adaptive averaging, the system achieved a minimum loss of 0.604, demonstrating the effectiveness of SSD and YOLOv2 for accurate fish localization and subsequent classification.

In [215] they present two improved models (YOLO-Fish-1 and YOLO-Fish-2) that are based on YOLOv3, which is much lighter than its next version, YOLOv4. YOLO-Fish1 and YOLO-Fish2 achieved an average accuracy of 76.56% and 75.70%, respectively, for fish detection in real marine environments, which is significantly better than YOLOv3. However, they did not manage to surpass the results of YOLOv4, since it reached 81.02% average precision.

In [216] they adapted YOLOv3 by training it with both real and synthetic images, using lobster parts placed on varied backgrounds from the Benthoz15 underwater set. The synthetic data approach allowed to address the variability and complexity of the underwater environment and improve the model's capacity to detect lobsters even when they are partially hidden in images.

In [217], a streamlined model was introduced using MobileNetv2 in combination with YOLOv4. They tested their network on three datasets: PASCAL VOC (Pattern Analysis, Statistical Modelling, and Computational Learning Visual Object Classes dataset), Brackish [218] and Underwater Robot Picking Contest **(URPC 2020)** datasets. They reached a mAP of 81.67% and 92.65% on the PASCAL VOC dataset and brackish dataset, respectively. On the URPC dataset the mAP achieved was lower: 79.54%. In [219] they have chosen the latest version of YOLO optimized with Res2Net to perform fish detection on a self-built fish dataset, reaching 95.4% mAP.

The authors of [220] investigate the detection and classification of seven species of jellyfish and fish: *Cyanea purpurea*, *Rhizostoma pulmo*, *Phacellophora camtschatica*, *Agalma okeni*, *Aurelia aurita*, *Phyllorhiza punctata*, *Rhopilema esculentum*, and fish. The dataset used in this study contains 11,926 images that were obtained through tracking technology and laboratory captures, and were divided into training, verification, and test sets. The developed model is an improved version of the YOLOv4-tiny algorithm, which is adapted for real-time jellyfish detection. Improvements include the addition of a CBAM [221] attention mechanism module to improve the feature extraction capability, especially for small and hidden targets. The improved YOLOv4-tiny algorithm achieved a detection accuracy of 95.01% and a detection speed of 223 FPS (frames per second), outperforming other compared algorithms such as YOLOv4, YOLOv5, YOLOv6, YOLOv7 and YOLOv8.

Other research on a version of a YOLO algorithm applied to jellyfish detection and classification can be found in [222]. The model developed in this study is JF-YOLO, based on an improved version of the YOLOv4 model for jellyfish detection. The dataset used presents challenges such as turbid water, high jellyfish density, pixel blur, transparency, scale variations, and illumination. Improvements to the developed model include, among others, residual dilated convolution (RDC) and super feature creation for feature fusion, and modification of PANet by removing SPP (Spatial Pyramid Pooling) to better handle scale variation and high target density. The JF-YOLO model achieved a recall of 85.74%, outperforming various versions of YOLOv4 and YOLOv3. In addition, JF-YOLO showed real-time performance with a speed of 41 FPS.

The study in [223] used a dataset composed of 1462 images of green turtles (Chelonia mydas) taken on Little Liuqiu Island, Taiwan. The images were collected using drones (UAVs) and underwater cameras between 2016 and 2021, along with additional images obtained from a local Facebook group. The study evaluated and compared three object detection models based on deep learning algorithms: YOLOv3, YOLOv5s, and YOLOv5l. The study concluded that, despite its more complex architecture and higher number of layers, YOLOv5l underperformed compared to YOLOv5s and YOLOv3, which stood out for the high accuracy achieved. YOLOv3 performed best on the smaller dataset of 781 images, reaching 98.25% precision, 97.68% recall, and a 97.96% F1-score.

The authors of [224] seek to determine how many shrimps are present in culture tanks by analyzing images, and they assess the precision and efficiency of various deep learning methods and models. The main dataset includes 1379 images of shrimp in RAS culture tanks, captured with an iPhone 11 mini. Two variants of the Faster R-CNN model and two variants of YOLOv5m6 were used, with the latter achieving the best mean absolute percentage error (MAPE) value.

In [225] they developed a model called NAM-YOLOv7, which is an improvement of the YOLOv7 algorithm that incorporates a normalization-based attention module (NAM) [226] that improves detection accuracy by focusing on relevant features of fish images. The aim of the study is the rapid detection of fish with Spring Viremia of Carp (SVC) symptoms in aquaculture. To do so, they created a dataset of 1814 images obtained from videos of fish, specifically zebrafish, infected with the SVC virus in a controlled environment.

The following study also focused on the detection of different types of turtles, although this time the images were obtained using a binocular camera in a controlled environment with a blue PVC background [227]. The dataset utilized in this research comprises more than 11,000 images of Chinese soft-shelled turtles. An enhanced object detection model, referred to as YOLOv7-SS, was developed based on the YOLOv7 architecture. This model integrates both the SE (Squeeze-and-Excitation) and SimAM (Simultaneous Attention Mechanism) attention mechanisms. These attention mechanisms improve the model's ability to focus on relevant features in turtle images, thus improving the accuracy of plastron and shell detection.

Diffusion models are a class of generative models that create data by progressively transforming noise into structured outputs through a learned denoising process. Inspired by thermodynamic diffusion, these models iteratively reverse a gradual noising procedure applied to training data, enabling high-quality image, audio, and video synthesis. Diffusion models have recently gained popularity due to their ability to generate diverse and realistic samples, often surpassing traditional GANs in stability and image fidelity. Their flexible framework supports conditional generation, inpainting, and super-resolution, making them powerful tools in modern generative modelling. This research introduces AIT-YOLOv7 (Advanced imaging technique-YOLOv7), a novel framework combining diffusion models, CBAM, and MSTB (Modified Swin Transformer Block) with YOLOv7 to enhance underwater object detection [228]. It addresses challenges like turbidity and low visibility using the TrashCan dataset (22 classes, 7212 images). The system achieves 81.4% mAP@0.5, outperforming other methods by up to 24.21%. This advancement supports marine conservation by enabling the precise identification of marine debris and biological organisms. The integration of denoising, resolution enhancement, and real-time tracking significantly improves reliability in underwater environmental monitoring.

A fairly updated version of YOLO is used in [229]. The model developed in the study is YOLOv8m, a medium-sized variant of the YOLOv8 model, which is based on the Darknet architecture. In order to improve underwater object detection, image enhancement techniques such as contrast stretching and gamma correction were applied to the images. The dataset employed in this study comprises 1434 images representing diverse fish species, including manta rays, jellyfish, and penguins, among others, organized into seven distinct classes for detection purposes. The YOLOv8m model reached an F1 score of 64.31%, outperforming Faster-RCNN (52.8%) and SSD (54.5%) in precision and recall balance.

The paper [230] introduces YOLOv8-TF (Transformer-enhanced YOLOv8), an advanced object detection model designed for identifying fish species in underwater environments. By incorporating transformer blocks into YOLOv8, the model effectively captures global context, and its class-aware loss mechanism addresses significant class imbalance found in the SEAMAPD21 dataset **(Southeast Area Monitoring and Assessment Programme Dataset 2021)**. Achieving a state-of-the-art performance of 87.9% mAP@0.5 on SEAMAPD21, YOLOv8-TF surpasses previous YOLO variants. While computationally heavier (30.56 M parameters), it maintains real-time speed (116 FPS). Validated on Pascal VOC and MS COCO, it demonstrates broad applicability for underwater ecological monitoring.

Accurate detection of abnormal fish behaviour in aquaculture is essential. The investigation in [231] proposes an improved detection algorithm based on YOLOv9, named DDEYOLOv9 (an enhanced high-precision detection algorithm based on YOLOv9), to detect abnormal fish behaviours early in industrial aquaculture settings, allowing farmers to take swift action to safeguard fish health and prevent economic losses. They also created the Takifugu rubripes Abnormal Behaviour Dataset, which includes five categories of fish behaviours. Experimental results achieved precision, recall and mAP values of 91.7%, 90.4%, and 94.1%, respectively.

Due to the increasing interest in the fishery industry and activities, many approaches focus on monitoring fishing vessels and analyzing images/videos obtained during those activities. In [232], fish tracking and segmentation are performed on videos recorded in a vessel. They introduced a tracking approach that integrates a deep convolutional neural network with an object detector and applies a Kalman filter in 3D. The object segmentation is updated and refined with the result of the RGB-depth image derived from stereoscopic images. The RGB-D images derived from the stereo videos are useful for obtaining the length of the detected objects. The proposed rescoring and segmentation method, in combination with the SSD model, obtained a Multiple Object Tracking Accuracy (MOTA) of 96.3%, which was quite better than the result obtained by the combination of FCNT (Faster Convolutional Network based Tracker) +SSD and the CFNet (Correlation Filter Network) +SSD, which did not reach the 90%. In [113], the authors present a real-time and robust underwater live crab detector called **Faster MSSDLite** (An SSD with MobileNetV2 as its backbone), which performed better than traditional SSD in the test results. As the backbone of this single-shot multi-box detector, they selected MobilNetV2. They also applied a denoising and enhancement method for underwater images effectively. An average precision of 99.01% was achieved by the network.

The paper [233] presents Foc_YOLOXn_ASFF, an enhanced version of YOLOX-nano with Adaptively Spatial Feature Fusion for detecting and classifying underwater fish. The system incorporates focal loss to address class imbalance and utilizes Adaptively Spatial Feature Fusion (ASFF) to enhance multi-scale feature learning. The model was trained on a custom dataset featuring four fish species from marine farms. It achieved an AP50 score of 98.9%.

In the line of research on fish diseases, we can find the work in [234]. This investigation introduces DCW-YOLO, an enhanced YOLOv10-based model for detecting skin diseases in Asian croaker fish. It replaces CIoU loss with NWD (Normalized Gaussian Wasserstein Distance) loss and adds C2f-D-LKA (Deformable Large Kernel Attention integrated into a C2f module (Cross-Stage Partial with full concatenation)) and DySample modules to improve accuracy and efficiency. A custom underwater dataset was created and augmented to train the model. DCW-YOLO achieves 96.87% mAP50 and runs at 90.91 FPS, outperforming other models. It enables fast, accurate, and lightweight fish health monitoring in aquaculture environments.

RetinaNet is a one-stage object detector that uses focal loss to address class imbalance. It combines high accuracy with efficient inference, making it suitable for detecting small and densely packed objects in images. Levy et al. [124] performed detection and tracking with RetinaNet [235] on two different datasets, the first containing underwater images [236] and the second is composed of aerial images. They concluded that the use of transfer learning could improve classification results. The average precision (AP) of fish detection was 0.74 for the underwater dataset.

An overview of research that applied different versions of YOLO and other types of single-shot networks is presented in Table 11.

Table 11. Application of single-shot detectors in underwater images.

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[113]	Self-created by the cruise muddy underwater video monitoring system at Changzhou	Live crabs' detector	Faster MSSDLite	Average precision/ f-score	99.01%/98.94%
[211]	Kyutech10K	Real-time recognition and tracking of four different organisms	YOLO, Tracking Learning Detection (TLD) and MedianFlow	AUC	53.1%
[124]	Underwater images and Aerial images	Marine species classification	RetinaNet as the object detector/Simple Online Realtime Tracker (SORT) for tracking	Average precision	Aerial: 40% for Ray, 25% for Diver and 75% for Shark
[176]	Kaggle fishing boat dataset	Fish detection, object pose estimation and horizontal alignment before class prediction	SSD and YOLOv2 as detectors, VGG-16 as pose estimation, VGG-16 and Inception V3 as classification	Loss	60.4%
[214]	Self-created dataset in a water tank	Analysis of motion patterns for koi fish health monitoring	YOLOv3	Heatmap visualization of fish locations and different plots comparison	-
[215]	DeepFish and OzFish [237] datasets	Fish detection	YOLO-Fish-1 and YOLO-Fish-2 (based on YOLOv3)	Precision/recall/ f1-score/AP	95%/96%/ 94%/96.15%
[216]	Images of lobsters captured in Western Australia using AUV and synthetic images	Western rock lobster detection	YOLOv3	mAP	46.9%
[217]	PASCAL VOC, Brackish and URPC Dataset	Underwater animal detection	A model based on MobileNet v2 and YOLOv4	mAP	92.65%

Table 11. *Cont.*

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[219]	Self-created underwater dataset in a lake, in a laboratory tank and gathered from the internet	Fish detection	A model based on Res2Net and YOLOv5	Precision/recall/mAP	95.7%/88%/95.4%
[220]	Self-created with crawler technology and on laboratory	Jellyfish species and fish detection	An improved version of the YOLOv4-tiny	Precision/recall/mAP/f1-score	92.62%/89.69%/95.01%/0.91
[222]	Images from videos about jellyfish through the web crawler	Jellyfish detection	JF-YOLO (based on YOLOv4)	mAP/recall	92.67%/85.74%
[223]	Self-created dataset collected by UAVs and images gathered from Facebook	Turtle detection	YOLOv3, YOLOv5s, and YOLOv5l	Precision/recall/f1-score	97.21%/97.82%/97.51%
[224]	Images of shrimp in Recirculating Aquaculture System (RAS) culture tanks	Shrimp detection	Faster RCNN and YOLOv5m6	MAPE	5.48
[225]	Self-created dataset in a tank	Healthy and sick fish with SVC detection	NAM-YOLOv7	Precision/recall	97.3%/93.8%
[227]	Self-created; images obtained in a controlled environment (out of water)	Chinese soft-shelled turtle detection	YOLOv7-SS	Precision/recall/mAP	95.38%/94.68%/89.82%
[228]	TrashCan dataset	Underwater object detection and classification (marine debris, biological organisms and submerged artefacts)	AIT-YOLOv7	mAP@0.5	81.40%
[229]	Not specified	Fish (sharks, jellyfish and other species) detection	YOLOv8m	Precision/recall/mAP/f1-score	68.6%/61.24%/66.7%/64.31%
[230]	Pascal VOC, SEAMAPD21 and MS COCO datasets	Identification of underwater fish Species	YOLOv8-TF	mAP50/mAP50:95	Pascal VOC: 94.60/SEAMAPD21: 61.2%
[231]	Self-created underwater dataset in a tank in a laboratory	Abnormal fish behaviour detection and classification	A model based on YOLOv9	Precision/recall/mAP	91.7%/90.4%/94.1%
[232]	Self-created dataset on a fishing vessel	Fish tracking and segmentation	A deep convolutional neural network and an object detector with a Kalman filter in 3D + SSD/YOLOv2	Multiple Object Tracking Accuracy (MOTA)	96.3%
[233]	Custom underwater fish dataset	Fish detection and classification	Foc_YOLOXn_ASFF (base model YOLOX-nano)	AP50/AP75/AP50-95/AR	98.2%/93.3%/84.6%/87.0%
[234]	Self-created dataset: images captured in a semi-submerged underwater cage	Fish disease detection	DCW-YOLO (base model YOLOv10)	Precision/recall/mAP50/mAP50:95	95.46%/90.12%/96.87%/75.04%

Transformers, originally developed for sequence modelling, have been adapted for computer vision tasks through architectures like **Vision Transformers (ViT)**. Their ability to capture long-range dependencies makes them a powerful alternative to traditional CNNs.

A paper featuring a model based on ViTs is [238]. This paper proposes the ADANSE ViT, featuring Amended Locale Self Attention (ALSA) and External Attention (EA) mechanisms, for classifying underwater fish and shrimp species, particularly on small datasets. It achieved high accuracy on a small proprietary dataset (701 images, 4 species) and the large WildFish benchmark. The ADANSE ViT significantly outperforms existing ViT models, achieving accuracies up to 92.3% (proprietary) and 93.9% (WildFish) across various image resolutions. Key innovations include learnable temperature parameters in ALSA, dual memory units in EA, and Layer Normalization for stability on small data. This demonstrates the potential for efficient and accurate underwater species classification even with limited data resources.

Some of the same authors of the previous work introduced ADANSE-TL [239], an integrated model combining a novel ADANSE ViT (for semantic features) with DenseNet-169 Transfer Learning (for local features) to classify marine species in challenging underwater images. On a self-collected dataset (701 images, 4 species) and benchmark sets (Croatian, Blue Bot), it delivered state-of-the-art accuracy: 96.21% on self-collected data and 95.09% on Blue Bot, outperforming other CNN variants and classifiers. Key innovations include a hybrid image enhancement pre-processing step and a feature fusion mechanism unifying local and semantic representations. The model demonstrates robustness across diverse underwater conditions and datasets.

In [240], IMViT, an improved Vision Transformer, is introduced for detecting and classifying fish species in challenging underwater images. To overcome ViT's limitations on smaller datasets, IMViT integrates convolutional layers and residual units from ResNet *before* the transformer, enhancing its ability to extract low-level features and inductive biases. Evaluated on the "Fish-Dataset" (~9k images, 9 species), IMViT achieved state-of-the-art results: 95.73% accuracy, 95.31% precision, 95.14% recall, and 94.92% F1-score, significantly outperforming CNNs (ResNet, VGG) and the standard ViT. The hybrid architecture demonstrates strong potential for fish monitoring tasks in complex underwater environments.

Table 12 lists previous articles that use transformers for object detection.

Table 12. Application of Transformers for image classification.

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[238]	Proprietary Dataset (4 categories) and WildFish Dataset	Underwater Species Classification (specifically fish and shrimp species)	ADANSE ViT	Accuracy	Propietary: 92.3%/ WildFish: 93.9%
[239]	Self-collected dataset, Croatian Fish dataset and Blue Bot dataset	Self-collected dataset	ADANSE-TL (ADANSE ViT and DenseNet-169)	Accuracy/Loss/ Precision/ Recall/ F1-Score	96.21%/ 0.174/96.32%/ 96.17%/96.20%
[240]	A Large-Scale Dataset for Segmentation and Classification [241]	Fish detection and fish species classification	Improved Vision Transformer (IMViT)	Accuracy/ Precision/ Recall/ F1-Score	95.73%/ 95.31%/95.14%/ 94.92%

Autoencoders constitute a category of unsupervised neural networks that are intended to extract concise representations of input data by means of an encoding–decoding process. Beyond their traditional applications in image reconstruction and denoising, autoencoders have also proven effective in supporting object detection and segmentation tasks. By learning salient feature representations in an unsupervised manner, they can enhance downstream models or serve as building blocks within more complex architectures, especially when labelled data is scarce.

In the paper [242], the authors propose a novel **classification convolution autoencoder (CCAE)** to improve accuracy classification. They trained the net on ImageNet and Fish4Knowledge and obtained quite high results: 0.7375 of accuracy on ImageNet, 0.9928 on Fish4Knowledge.

O'Byrne et al. [243] trained a deep encoder–decoder network called **SegNet** (which they combined with an SVM for better results), using 2500 images. Even though the images were rendered in a virtual underwater environment, they obtained good results that could be extrapolated to a real environment, as they achieved an MIoU of 87% in a real environment.

Another approach based on the SegNet encoder–decoder can be seen in [244]. They performed fish segmentation on images of fish out of water. They achieved high values of IoU with the Nile tilapia fish, but when they applied the same method to other species, the mean IoU went down to 31.8%.

Object segmentation is a further step than object detection, since, in addition to detecting the location of the object and classifying it, the exact shape of the object must be segmented. Although this type of task can be performed with VC and ML techniques, DL approaches usually have better and more accurate results. In [245], **DeepLabv3+** was adapted for underwater image segmentation by adding an unsupervised colour correction method (UCM) unit [246] to the encoder, enhancing both image quality and segmentation performance. The Mean Intersection over Union (MIoU) score they reached was 64.65%.

Mask R-CNN extends the Faster RCNN network by adding a branch that includes a binary mask that aims to predict whether the image pixel contributes to the given part of the object or not. This framework is made up of essential elements such as a backbone network, a feature extraction network called FPN (Feature Pyramid Network), a region proposal network called RPN and another RoI (Region of Interest) proposal network. Figure 7 shows the different layers that make up the architecture of the Mask R-CNN.

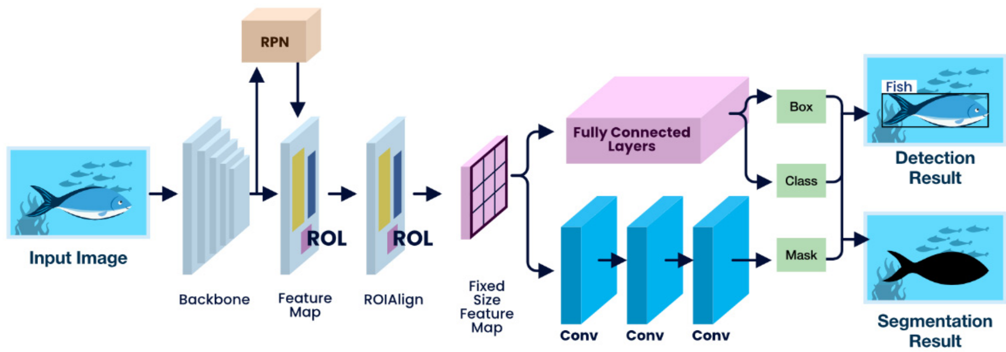

Figure 7. Architecture of Mask R-CNN for image detection and segmentation.

In [247] they replace the baseline ResNet50/101 network in Mask R-CNN with multiple residual blocks, constructing a 32-layer feature extraction network to decrease the network's training parameters without compromising detection performance. Using transfer learning, they pre-trained the network with the MS-COCO (Microsoft Common Objects in Context) dataset. To test the network, they used a dataset acquired using a Tritech Gemini 720i sonar inside a water tank.

U-Net is a convolutional neural network from the University of Freiburg, designed to work with limited training data and deliver precise image segmentation by extending fully convolutional architectures [33]. Although initially developed for biomedical image segmentation, it has been used in multiple fields, as well as in the classification of aquatic species. The research in [248] employs an enhanced deep neural network (DNN) design for fish segmentation in underwater videos, utilizing distributed learning and edge computing to boost both energy efficiency and real-time processing capabilities. Their primary model is a customized U-Net, which integrates Pix2Pix blocks—commonly found in generative adversarial networks—to handle upsampling. To decrease the number of trainable parameters and accelerate training, they apply transfer learning by incorporating pre-trained weights from MobileNetV2, originally trained on the ImageNet dataset. The model was trained using a Distributed Computing System (DCS) on AWS (Amazon Web Services) SageMaker, leveraging 20 identical computers to distribute the workload. In distributed training with 20 nodes, the Sparse Categorical Cross Entropy (SCCE) loss was 0.053, improving by 18% compared to training on a single computer (loss of 0.065).

The classification model in [249] is based on the U-Net architecture. For training and testing the network, they created two datasets from internet videos. The first dataset featured clownfish and the second featured cavefish. They achieved an mIoU value of 88.19%.

The behaviour of the fish is not the only thing to be monitored. Möller et al. [250] tracked the size and behaviour of a sponge. They used images from a fixed underwater observatory (FUO) called Lofoten-Vesterålen (LoVe). They performed the sponge segmentation by U-Net successfully.

Some studies only focus on studying a single species, as can be seen in [251]. This study uses a dataset of images featuring roman seabream (*Chrysoblephus laticeps*), a fish native to southern Africa. The developed model is an implementation of Mask R-CNN designed for localization, classification, counting and tracking of Roman bream in uncontrolled underwater environments. The Mask R-CNN model achieved an mAP70 of 80.28% on the previously unseen test dataset.

The dataset used in the experiment in [252] is composed of 1824 images obtained from different underwater environments. The images come from several existing datasets, such as Semantic Segmentation of Underwater Imagery (SUIM) [101], DeepFish, and RockFish, as well as videos captured by underwater drones in Peru. The study compared two deep neural network architectures for semantic segmentation: DeepLabV3+ and U-Net. Optimal performance was achieved with a U-Net-scSE (U-net with Spatial Channel Squeeze-and-Excitation module) model that incorporated a ResNeSt-269e encoder. The training process used a combined loss function featuring Chan-Vese-based learning active contour loss, cross-entropy loss, and Dice loss.

This paper [253] introduces a neural network for underwater fish segmentation, addressing challenges like colour distortion and blur using ResNet50 with an enhanced feature pyramid (PAFE) and attention mechanisms. The model attains 95.1% MIoU and 0.901 F1-Score, surpassing U-Net and PSPNet (Pyramid Scene Parsing Network). Data augmentation (Mixup, CutMix) and a new deep-sea dataset validate robustness in real-world conditions. The PAFE module improves multi-scale feature extraction, balancing accuracy and computational efficiency.

AASNet (Agricultural Aqua Segmentation Network) is an advanced deep learning framework developed to provide precise and efficient underwater fish instance segmentation for use in smart fisheries [254]. It tackles key challenges like lighting/colour variations and extreme class imbalance through two core innovations: the Linear Correlation Attention (LCA) module for robust feature correlation and the Dynamic Adaptive Focal Loss (DAFL) for handling data imbalance. Evaluated on UIIS (Underwater Image Instance Segmentation dataset) [255] and USIS10K (Underwater Salient Instance Segmentation 10K dataset) [256] datasets, AASNet achieves state-of-the-art performance (31.7 mAP on UIIS, 47.4 mAP on USIS10K) while maintaining high inference speed (28.9 ms/image). Its accuracy and efficiency make it ideal for real-time agricultural uses, such as monitoring fish farms and assessing health.

This research [257] introduces a self-supervised approach to fish segmentation, eliminating the requirement for manual annotation by employing feature alignment and spatiotemporal consistency. It employs a CoaT Transformer (Co-Scale Conv-Attentional Image Transformer) to capture multi-scale features—combining local details and global context—for improved robustness in low-visibility underwater scenes. The method achieves 50.0 $\mathcal{J}\&\mathcal{F}_{mean}$ (*Jaccard Index & Dice Coefficient mean*) on the Seagrass dataset [258] and 63.3 on YouTube-VOS [259], outperforming previous self-supervised approaches. The approach is computationally efficient thanks to parallel processing and anchor sampling, although it still faces challenges in scenarios with heavy occlusion and high background similarity.

Swin Transformers are hierarchical vision transformers designed to efficiently model images at multiple scales using shifted windows for self-attention. This architecture enables better capture of both local and global context, making it well-suited for dense prediction tasks like segmentation. Swin Transformers have demonstrated state-of-the-art performance in semantic and instance segmentation benchmarks, often outperforming traditional CNN-based methods. Thanks to their flexibility and strong feature representation, Swin Transformers have become a popular backbone choice in modern segmentation models.

Following this line, we can find works like [260], where they present SwinConvMixerUNet, a novel hybrid model combining Swin Transformer's attention mechanisms and ConvMixer's efficient feature mixing within a U-Net structure for underwater image semantic segmentation. It addresses challenges like low visibility, light scattering, and particulate matter in underwater environments, crucial for tasks like fish habitat monitoring and marine ecosystem preservation. On the SUIM dataset, the model achieves an mIoU of 84.83%, surpassing DeepLabV3, PSPNet, UNet, and other models. The architecture effectively handles complex underwater conditions through patch partitioning, hierarchical Swin Transformer blocks, and ConvMixer refinement. This advancement enables more accurate environmental monitoring, resource management, and pollution tracking in marine applications.

In [261] they propose **UISFormer**, a ViT-based model for segmenting aquatic organisms (fish, coral) in complex underwater aquaculture scenes. It addresses challenges like limited data via CutStitch augmentation and enhances feature fusion using RFEM (Residual Feature Enhancement) and MFRF (Multiscale Feature Residual Fusion) modules. Evaluated on three datasets (FishData, Underwater Images Segmentation dataset better known as UISD, large-scale fish data), UISFormer outperforms CNNs and hybrid models (e.g., 92.20% MIoU on UISD). They performed practical tests in Shandong marine ranches to confirm its robustness for real-world aquaculture monitoring.

As seen before, **SAMs** are large-scale, prompt-driven segmentation models used to perform general-purpose image segmentation without task-specific training. They are trained on billions of masks across diverse images, enabling strong zero-shot performance on unseen objects and domains. Its architecture combines a powerful image encoder (ViT) with a lightweight prompt encoder and a mask decoder. In [262] they introduce UIIS10K, a large underwater instance segmentation dataset to date, containing 10,048 images across 10 categories. They also propose UWSAM (Segment Anything Model Guided Underwater Instance Segmentation), a method that leverages MG-UKD (Mask GAT-based Underwater Knowledge Distillation) to transfer knowledge from SAM into a lightweight ViT-Small encoder, using GAT-based (Graph Attention Network- based) feature reconstruction. They also design EUPG (End-to-End Underwater Prompt Generator), an automatic prompt generation mechanism that enables end-to-end segmentation without relying on external object detectors. The approach achieves a mAP value of 44.6 on UIIS10K for UWSAM-Teacher and 38.7 mAP for UWSAM-Student. Additionally, the UWSAM-Student model demonstrates efficiency (47M parameters) and robustness against common underwater challenges such as turbidity and occlusion.

A more detailed account of the methods cited can be found in Table 13.

Table 13. Application of segmentation networks in underwater images.

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[242]	ImageNet and Fish4Knowledge	Fish classification	Classification convolution autoencoder (CCAE)	Accuracy	73.75% and 99.28%
[243]	Images containing virtual underwater scenery	Underwater object semantic segmentation	Deep encoder–decoder network called SegNet	MIoU/mean accuracy	87%/94%
[244]	Self-created	Semantic segmentation of tilapia fish body parts	Encoder–decoder based on the SegNet approach	MIoU	60-85%
[245]	Self-made underwater image dataset	Underwater species Semantic segmentation	DeepLabv3 +	MIoU	64.65%
[247]	MS-COCO dataset for pre-training and images acquired by sonar (self-created)	Underwater object detection and semantic segmentation	Mask RCNN	For detection: Precision/recall/mAP For segmentation: AP/AP_{50}	For detection: 95.73%/97.21%/96.97% For segmentation: 53.23%/95.15%
[248]	DeepFish dataset	Underwater fish segmentation	A modified version of the U-Net	SCCE loss/SCF loss/Sparse Categorical Crossentropy Accuracy (SCCA)/IoU	0.053/3.48/98.81%/87.6%
[249]	"Coralfish" and "cavefish" dataset obtained from YouTube videos	Fish detection and classification	U-Net	F1-score/mIoU	95.04%/88.19%
[250]	LoVe Observatory dataset	Sponge size and behaviour tracking	U-Net	Pearson's r	0.98
[251]	Dataset collected using submerged action cameras mounted on a baited underwater remote video (BRUV) rig.	Roman seabream detection and tracking	Mask R-CNN	mAP_{50}/mAP_{75}	81.45%/80.28%
[252]	Images collected from multiple existing datasets	Bottom water, seafloor/obstacles, and fish segmentation	DeepLabV3+ and U-Net-based model variations	IoU/Hausdorff distance (HD)	78.76%/19.17
[253]	Fish4Knowledge and Real deep-sea fish images	Underwater fish segmentation	ResNet50 backbone with an enhanced Feature Pyramid Convolutional Architecture (PAFE)	F1-score/MioU/Pixel Accuracy (PA)	90.1%/95.1%/92.1%
[254]	UIIS and USIS10K	Underwater Fish Instance Segmentation for smart fisheries	AASNet framework (GELAN Backbone from YOLOv9)	mAP/AP/AP_{75}	31.7%/49.5%/35.1%
[257]	DeepFish (training), Seagrass and YouTube-VOS dataset	Underwater fish segmentation	CoaT Transformer backbone: Combines Conv-Attentional Image Transformer (CAIT) and Co-Scale Feature Attention Network (CFAN)	$\mathcal{J}\&\mathcal{F}_{mean}$/ $\mathcal{J}_{mean}$/$\mathcal{J}_{recall}$/ $\mathcal{F}_{mean}$/$\mathcal{F}_{recall}$/	YouTube-VOS: 63.3%/63.9%/74.0%/62.7%/69.6%

Table 13. *Cont.*

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[260]	SUIM dataset	Underwater Image Semantic Segmentation	SwinConvMixerUNet	mIoU	84.83%
[261]	FishData, UISD, Large-scale fish dataset	Underwater image segmentation	UISFormer	mIoU/Accuracy/ CPA (Category Pixel Accuracy)/ F1-score	Large-scale Fish Data: 96.7%/99.34%/98.97%/ 98.32%
[262]	UIIS10K (proposed), UIIS, USIS10K	Underwater instance segmentation	UWSAM Framework	Bounding Box metrics (mAP^b/ AP^b_{50}/AP^b_{75}) Mask metrics (mAP^s/AP^s_{50}/ AP^s_{75})	UWSAM-Teacher on USIS10K: mAP^b = 45.8%/ AP^b_{50} = 64.1%/ AP^b_{75} = 55.1%) Mask metrics (mAP^s = 46.0%/ AP^s_{50} = 61.7%/ AP^s_{75} = 51.7%)

5.2.4. Multimodal Understanding and Reporting

Recent advances in multimodal models have enabled underwater species detection and classification through vision-language integration. Unlike traditional CNN or Transformer-based detectors, these approaches leverage natural-language prompts, zero-shot reasoning, and cross-modal alignment, enabling robust performance even under scarce annotations or unseen species. Vision-Language Models (VLMs) and Large Language Models (LLMs) have thus emerged as a new family of techniques applicable to underwater environments, warranting a dedicated subsection.

The UVLM (Underwater Video Language Model) [263] targets the lack of video–language benchmarks tailored to underwater environments, where visual degradation, complex temporal dynamics, and domain-specific biological knowledge significantly challenge existing Video-Language Models (VidLMs). The authors introduce the first underwater video-language benchmark built through a human–AI collaborative annotation pipeline, combining expert frame-level annotations with GPT-4o–assisted text generation and strict manual verification. The dataset includes 2109 videos covering diverse scenes and 419 marine organism classes and defines 20 subtasks related to biological and environmental understanding. Extensive evaluations show that current VidLMs suffer significant performance drops underwater compared to terrestrial settings. Fine-tuning on UVLM notably improves results, enabling smaller models to approach large proprietary ones, although fine-grained taxonomic reasoning remains challenging. The results also highlight a persistent performance gap on fine-grained taxonomic classification, indicating intrinsic limitations of smaller models in knowledge-intensive underwater tasks.

State-of-the-art multimodal large language models (MLLMs) show strong general vision–language capabilities, yet their effectiveness in fine-grained marine species recognition remains unclear. The work in [264] presents a large-scale diagnostic study revealing that even the best open-source MLLMs perform poorly on fish species identification, achieving below 10% accuracy in zero-shot settings. To investigate the causes of this failure, the authors introduce FishNet++, a comprehensive multimodal benchmark extending FishNet with rich textual, spatial, and morphological annotations. The analysis disentangles errors arising from missing taxonomic knowledge, weak visual domain understanding, and limited fine-grained perception. Experiments show that MLLMs can localize fish reasonably well but struggle with subtle morphological cues critical for species-level recognition. Fine-tuning on FishNet++ substantially improves performance, and explainable training further enhances interpretability, highlighting the need for domain-specific multimodal benchmarks in aquatic science.

This work [265] proposes a cloud–edge framework for real-time semantic understanding in AUVs, combining lightweight underwater image enhancement and frame sampling at the edge with cloud-based inference using the Qwen2.5-VL model. The edge pipeline performs colour correction, dehazing, and white-balance normalization, followed by low-rate frame sampling to meet bandwidth constraints. A latency model incorporating edge, communication, and cloud processing enables end-to-end performance analysis. Simulations across multiple underwater scenarios report stable latencies around 1.0–1.2 s per frame and reliable cloud inference times. The VLM produces natural language scene descriptions capturing objects, spatial relations, and environmental cues, with object detection recall up to 0.94 and hallucination rates as low as 4% depending on the scene. Results indicate that enhancement substantially reduces hallucination and improves descriptive fidelity under turbidity and occlusion.

The work of [266] introduces AquaVLM, the first underwater communication system leveraging mobile Vision-Language Models (VLMs) to generate context-rich, semantically meaningful messages for scuba divers. It combines multimodal perception (images and sensor data) with hierarchical, intent-guided message generation. To address the lack of underwater knowledge, a diving conversation dataset was created, and the model was trained for message generation, response generation, and message retrieval. It includes error-resilient fine-tuning to handle corrupted acoustic transmissions and was validated with a VR simulator and an iOS prototype. Lake experiments show high semantic similarity (>90%) and low error rates (<3%), outperforming traditional OFDM-based (Orthogonal Frequency Division Multiplexing) methods.

Ref. [267] introduces AquaOV255, a large and detailed underwater segmentation dataset, with over 20,700 images across 255 categories, addressing the limited diversity of previous benchmarks. The authors combine five more datasets to create UOVSBench, the first unified benchmark for open-vocabulary underwater segmentation. To handle domain shifts from terrestrial VLMs, they propose Earth2Ocean, a training-free adaptation framework. The method incorporates two key modules: a Geometric guided Visual Mask Generator (GMG) leveraging DINO (Distillation with No Labels) based self-similarity priors to correct CLIP (Contrastive Language-Image Pre-Training) visual features, and a Category visual Semantic Alignment (CSA) module that enhances textual embeddings using multimodal LLM reasoning and underwater specific templates. Without underwater fine-tuning, it improves average mIoU by 6–8 points while remaining efficient. Ablations show both geometric and semantic modules are crucial for recognizing fine-grained and rare underwater categories, enabling practical marine monitoring applications.

FishDetectLLM [268] is a fish detection framework using a lightweight multimodal LLM, extending TinyLLaVA (a smaller version of the Large Language and Vision Assistant model) to perform both classification and bounding box prediction. It treats fish detection as a visual question answering task, combining visual features with textual reasoning, unlike traditional CNN/Transformer detectors. A multi-turn instruction dataset is created from FishNet (94,532 images, 17,357 species) linking taxonomy descriptions with bounding boxes. The model achieves high classification accuracies (up to 99.09% at the class level) and 84.2 mAP50, outperforming YOLO, DINO, Faster/Mask RCNN, CLIP, and LLaVA, while generalizing to unseen species. Limitations include inference speed, with future work aimed at model compression and broader ecological integration.

This study [269] addresses the limitations of current Vision-Language Models (VLMs), such as CLIP, in fine-grained coral species recognition, where domain-specific morphology and hierarchical taxonomy play crucial roles. The authors introduce HSCR16K, a large-scale hierarchical stony coral dataset containing 16,659 images from 16 species, annotated from kingdom to species and enriched with expert-crafted morphological descriptions. Building on this dataset, they propose CORAL Adapter, a lightweight plug-and-play module designed to augment CLIP with coral-specific knowledge via two complementary components: a morphological adapter that captures texture, structure, and colour cues, and a taxonomic adapter that encodes biological hierarchy relations. The model is optimized via a combination of cross entropy, similarity alignment, and a reversed Jensen–Shannon divergence to balance prior CLIP knowledge with newly learned coral-specific representations. The method demonstrates strong robustness in recognizing unseen bleaching-prone species and transferring across ocean regions. Overall, the work showcases how structured biological knowledge can significantly enhance VLMs for ecological monitoring and coral health assessment.

Table 14 provides a more detailed overview of the methods referenced.

Table 14. Applications of VLMs and LLMs in underwater images.

Ref	Dataset	Problem Type	Algorithms	Metric	Value (Highest)
[263]	UVLM (proposed), WebUOT (re-annotated)	Underwater video-language understanding	General VidLMs with supervised fine-tuning (VideoLLaMA3, Qwen2.5VL)	Overall accuracy	Qwen2.5VL-72B: 75.49% VideoLLaMA3-7B fine-tuned: 73.04% (+10.34)
[264]	FishNet++ (proposed), FishNet	Fine-grained fish species recognition (open-vocabulary classification), detection, keypoint localization	CLIP, BioCLIP, SigLIP, Qwen2.5-VL, Gemma-3, GPT-4o, YOLO-based baselines	Accuracy/IoU50/ IoU90	GPT-4o: 17.9%/ YOLO-12: 95.2%; Qwen2.5-VL: 91.5%/ YOLO-12: 35.2%; Qwen2.5-VL: 26.7% (Frequent Species)
[265]	Simulated underwater videos	Real-time VLM semantic scene analysis for AUVs	Qwen2.5-VL (72B)	Object Detection Recall (ODR)/ Spatial Relationship Accuracy (SRA)/ Hallucination Rate /Output Accuracy	0.94/0.91/0.04/0.88
[266]	Five public scuba diving videos captured in different locations and simulated sensor data	Context-aware message generation & recovery for diver communication (mobile VLM)	MobileVLM-3B	Purpose-Align Rate /Semantic Similarity (received vs. original message)	80%/90%
[267]	AquaOV255, UOVSBench (AquaOV255 + USIS16K, SUIM, MAS3K, USIS10K, DUT-USEG)	Training-free open-vocabulary segmentation in underwater scenes	Earth2Ocean (GMG + CSA; transfers terrestrial VLMs to underwater)	mIoU	55.24
[268]	FishNet + LLaVA-1.5 datasets; extra 1100 DeepFish images	LLM-based detection and classification	FishDetectLLM (TinyLLaVA + SigLIP + StableLM-2; instruction conversations)	Accuracy	FishNet: 99.09%
[269]	HSCR16K	Fine-grained coral recognition (species/ genera); zero-/few-shot VLM adaptation	CORAL-Adapter (morphological + taxonomic adapters on CLIP)	Accuracy	77.27%

6. Discussion

The application of deep learning (DL) to the analysis of marine species imagery has grown rapidly over the past decade, driven by advances in computer vision and the increasing availability of underwater and above-water image data. These methods hold significant potential for biodiversity monitoring, fisheries management, and ecological research, yet the field faces unique challenges due to the variability of aquatic environments and the limited availability of labelled datasets. The following research questions guide this study's review of existing literature and methods:

RQ1: Advanced deep learning architectures

DL applications in marine imagery have leveraged a variety of architectures. Regarding image enhancement and restoration, the most widely used architectures are GANs and improved U-Net versions. CNN-based models such as ResNet, DenseNet, and VGGNet are widely used for classification, while region-based methods (e.g., Faster R-CNN) and real-time detectors (YOLO, SSD) address detection tasks. Segmentation is addressed using specific architectures such as Mask R-CNN. Vision Transformer (ViT) models are emerging for complex, cluttered environments and video analysis, benefiting from their ability to model long-range dependencies. Large Vision-Language Models and multimodal LLMs bring zero-/few-shot flexibility to underwater species analysis, but their operational costs and reliability constraints are significant. Model scale has become a fundamental aspect of the architecture of multimodal language and vision models used to study underwater environments. The approaches reviewed range from relatively compact models such as MobileVLM (≈3 billion parameters) to mid-scale architectures such as VideoLLaMA3-7B, and even very large models such as Qwen2.5-VL-72B. Beyond the contrast with traditional CNN-based architectures, such as VGG16 (with approximately 138 million parameters), the gap within the multimodal family itself is particularly pronounced, raising significant questions regarding efficiency, implementation feasibility and fair comparison across studies. Although larger models generally offer more robust reasoning without the need for fine-tuning, their high computational consumption and inference costs represent substantial limitations for real-time underwater applications or those with limited resources.

RQ2: Datasets and acquisition methods

Research draws on both public datasets, such as Fish4Knowledge or LifeCLEF, and custom collections targeting specific taxa or habitats. Image acquisition methods include underwater cameras mounted on ROVs and AUVs, diver-operated photography, surface-level imaging for mammals, and, in some cases, artificially created environments in tanks. Video data is often decomposed into frames for training. It is worth noting the repeated use of certain datasets, such as Fish4Knowledge. Several studies published prior to 2018 report very high and closely clustered accuracy values (above 95%), suggesting that this dataset may not offer sufficient discriminatory power to distinguish meaningfully between methods. This highlights a significant limitation of the extensive use of a specific dataset over an extended period and underscores the need for new datasets that better capture environmental variability and pose a greater challenge to modern DL models.

RQ3: Evaluation metrics and performance trends

Performance is assessed using metrics such as accuracy, precision, recall, F1-score, mAP, and IoU. CNN classifiers often exceed 85% accuracy on clear imagery, while detection in turbid or cluttered conditions yields more variable results (mAP 50–80%). Transformer-based models show superior performance in complex scenes, especially with large datasets, due to their enhanced contextual reasoning. It should be noted that, as discussed in the dataset, a baseline performance close to saturation can also limit the interpretability of evaluation metrics.

RQ4: Limitations, challenges, and future directions

Key challenges include limited labelled datasets, inconsistent image quality, class imbalance, and poor generalization across domains. One challenge associated with image quality assessment is the lack of real reference images in underwater image enhancement. In real-world aquatic environments, it is generally not easy to obtain an undistorted version of an image, which limits the applicability of full-reference metrics such as PSNR or SSIM. Consequently, many studies resort to reference-free metrics, such as UIQM and UCIQE, to quantify the performance of the enhancement. However, these metrics do not always correlate with perceived visual realism or performance in downstream tasks, such as object detection or species classification. For this reason, several studies supplement objective scores with a subjective visual assessment. These differences complicate comparisons between studies and highlight the need for more robust evaluation methodologies. Future research priorities include creating large, standardized datasets, integrating multimodal detection (e.g., optical imaging and sonar), advancing self-supervised methods, and developing lightweight, green AI models for real-time implementation. Explainable AI is also identified as essential for decision-making and user trust. Despite its promising capabilities across multiple domains, current approaches based on foundational models, VLMs, and LLMs remain limited due to their high computational demands, latency, and difficulty of real-time implementation. Their sensitivity to prompt design and underwater visual artifacts can also introduce instability in practical environments. Regarding future work, it would be interesting to investigate the development of lighter multimodal architectures, strategies for the use of more robust language models, and improved adaptation mechanisms tailored to murky, low-light, or species-rich environments.

The results of this study demonstrate the significant potential of artificial intelligence techniques in addressing the unique challenges associated with underwater imaging and marine species analysis. The application of deep learning-based enhancement algorithms markedly improved image clarity and colour fidelity, mitigating the effects of light absorption, scattering, and turbidity commonly observed in subaquatic environments. These enhancements facilitated more accurate downstream tasks, including species classification, object detection, and segmentation, highlighting the interdependence between image preprocessing and model performance.

Overall, the findings of this study highlight the transformative potential of AI-driven methodologies for marine ecology and conservation, offering scalable solutions for monitoring biodiversity and supporting informed environmental management decisions. The integration of image enhancement and automated species analysis represents a critical step toward real-time, high-fidelity assessment of underwater ecosystems.

7. Conclusions

The aim of this paper is to review the state-of-the-art applied to underwater imaging, examining computer vision techniques, which are usually applied to image preprocessing and to extract image features, and researching machine learning and deep learning approaches, which are used for classification and object detection.

The core of the paper is dedicated to reviewing DL approaches specifically tailored to underwater fish and water-related imagery. It is divided into two main areas: underwater image processing and enhancement, and aquatic species detection, classification, and segmentation. The first part discusses techniques to improve image quality in challenging underwater conditions, while the second part examines how various DL models are used to identify and analyze aquatic life, with detailed examples of model applications.

Compared to other reviews on the topic, this article is distinguished by its comprehensiveness and depth in its coverage of the existing literature. While previous reviews have focused on a limited set of studies, this review includes a broader analysis covering 132 recent and relevant articles, providing a more complete overview of the current state of knowledge. Furthermore, a particular effort has been made to include a greater diversity of sources, considering research from different methodological approaches and geographical contexts. This breadth and diversity not only enrich the discussion but also offer a more robust and balanced perspective, making this article an essential reference for any researcher interested in the topic.

Author Contributions: Conceptualization, V.L.-V. and J.M.L.-G.; methodology, V.L.-V., G.S.-B. and H.I.R.; formal analysis, V.L.-V.; investigation, V.L.-V., G.S.-B. and H.I.R.; resources, G.S.-B. and H.I.R.; writing—original draft preparation, V.L.-V., J.M.L.-G.; writing—review and editing, V.L.-V., G.S.-B. and H.I.R.; visualization, V.L.-V.; supervision, J.M.L.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Centro para el Desarrollo Tecnológico Industrial (CDTI) (Grant No. EXP 00108707/SERA-20181020).

Acknowledgments: This work was developed at Deusto Seidor S.A. (01015, Vitoria-Gasteiz, Spain) within the framework of the Tecnoterra (ICM-CSIC/UPC) and the following project activities: ARIM (Autonomous Robotic sea-floor Infrastructure for benthopelagic Monitoring); MarTERA ERA-Net Cofund; Centro para el Desarrollo Tecnológico Industrial, CDTI; and RESBIO (TEC2017-87861-R; Ministerio de Ciencia, Innovación y Universidades).

Conflicts of Interest: Author Vanesa Lopez-Vazquez was employed by Deusto SEIDOR. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AASNet	Agricultural Aqua Segmentation Network
ACE	Automatic Colour Enhancement
ACP	Average Class Precision
AdaIN	Adaptive Instance Normalization
ADANSE ViT	Amended Dual Attention oN Self-locale and External Visual Transformer
AG	Average Gradient
AI	Artificial Intelligence
AIT-YOLO	Advanced imaging technique YOLO
ALSA	Amended Locale Self Attention
ANN	Artificial Neural Network
ASFF	Adaptively Spatial Feature Fusion
AUC	Area Under the Curve
AUROC	Receiver Operating Characteristic Area Under the Curve
AWS	Amazon Web Services
B3DO	The Berkeley 3D Object Dataset
BC	Blind Contrast Restoration Assessment (BCRA)
BC(e)	BCRA to assess the increase in edge visibility
BC(r)	BCRA to assess the increase in edge pixel gradient values
BPNN	Back Propagation Neural Network
C2f-D-LKA	Deformable Large Kernel Attention integrated into a C2f module (Cross-Stag Partial with full concatenation)
CAIT	Conv-Attentional Image Transformer
CBAM	Convolutional Block Attention Module
CCAE	Classification Convolution Autoencoder
CCF	Colourfulness Contrast Fog density index
CCL-Net	A tailored UIE approach based on cascaded contrastive learning (CCL)
CDR	Channel Dynamic Range
CE	Calibration Error
CFAN	Co-scale Feature Attention Network
CFNet	Correlation Filter Network

cGAN	Conditional Generative Adversarial Network
CIFAR10	Canadian Institute For Advanced Research dataset
CLIP	Contrastive Language–Image Pretraining
CNN	Convolutional Neural Network
CNT	Convolutional Network based Tracker
CoaT	Co-Scale Conv-Attentional Image Transformer
CSA	Category-visual Semantic Alignment
CV	Computer Vision
CVAE	Conditional Variational Autoencoder
CycleGAN	Cycle Generative Adversarial Network
DAFL	Dynamic Adaptative Focal Loss
DAMNet	Dual Attention Mechanism based network
DBN	Deep Belief Network
DCS	Distributed Computing System
DDEYOLOv9	An enhanced high-precision detection algorithm based on YOLOv9
DeCAF	Deep Convolutional Activation Feature
DeltaE	Colour difference metric
DINO	Distillation with No Labels
DIP	Digital Image Processing
DL	Deep Learning
DNN	Deep Neural Network
DNnet	Dinamic range and Normalization framework
DPM	Deformable Part Model
EUPG	End-to-End Underwater Prompt Generator
EUVP	Enhancement of Underwater Visual Perception
FAN	Fast Average Normalization
FCN	Fully Convolutional Network
FCNT	Faster Convolutional Network based Tracker
FDCNet	Filtering Deep Convolutional Network
FE	Fusion Enhance
FLOPS	Floating-point Operations
FUO	Fixed Underwater Observatory
GAN	Generative Adversarial Networks
GAT	Graph Attention Network
GBR	Great Barrier Reef
GELU	Gaussian Error Linear Unit
GFLOPS	Giga FLOPS
GMG	Geometric guided Visual Mask Generator
GMSD	Gradient Magnitude Similarity Deviation
HR-Net	Haze Removal Network
ILSVRC	Large Scale Visual Recognition Challenge dataset
IMViT	Improved Vision Transformer
IoU	Intersection over Union
J&Fmean	Jaccard Index & Dice Coefficient mean
JAMSTEC	Japan Agency for Marine-Earth Science and Technology
LCA	Linear Correlation Attention
LifeCLEF	Life-based Conference and Labs of the Evaluation Forum
LLaVA	Large Language and Vision Assistant
LLM	Large Language Model
LoVe	Lofoten – Vesterålen
LPIPS	Learned Perceptual Image Patch Similarity
LSUI	Large Scale Underwater Image dataset
MAD	Most Apparent Distortion
mAP	Mean Average Precision

MaxViT	Multi-axis Vision Transformer
MBARI	Monterey Bay Aquarium Research Institute
MDGAN	Multiscale Dense Generative Adversarial Network
MDP	Markov Decision Process
MDPM	Mixed-Domain Periodic Motion
MFRF	Multiscale feature Residual Fusion module
MG-UKD	Mask GAT-based Underwater Knowledge Distillation
mIoU	Mean Intersection over Union
ML	Machine Learning
MLC	Moorea Labelled Coral dataset
MLLM	Multimodal Large Language Model
MLP	Multilayer Perceptron
MLR-VGG16	Multi-Level Residual VGG16
MODIS	Moderate Resolution Spectroradiometer
MOS	Mean Opinion Score
MOTA	Multi Object Tracking Accuracy
mRES-uNet	Multi Resolution U-net architecture network
MS-COCO	Microsoft Common Objects in Context dataset
MSE	Mean Squared Error
MSFN	Multi-Scale Fusion Feed-forward Network
MSRB	Marine Snow Removal Benchmarking Dataset
MSSDLite	An SSD with MobileNetV2 as its backbone
MSTB	Modified Swin Transformer Block
MUSIQ	Multi-scale Image Quality Transformer
NFA	Nonlinear Frequency-aware Attention
NIQE	Natural Image Quality Evaluator
NLP	Natural Language Processing
NMS	Non-maximum suppression
NOAA	U.S. National Oceanic and Atmospheric Administration
NWD	Normalized Gaussian Wasserstein Distance
NYU Depth	New York University Depth dataset
OFDM	Orthogonal Frequency Division Multiplexing
OHEM	Online Hard Example Mining
PAdaIN	Probabilistic Adaptive Instance Normalization
PAFE	Feature Pyramid Convolutional Architecture
PASCAL VOC	Pattern Analysis, Statistical Modelling, and Computational Learning Visual Object Classes dataset
PCA	Principal Component Analysis
PCQI	Perception-based Colour Quality Index
PGG-YOLO	P2P5hGhostConv-G_Ghostbottleneck-YOLO
PhISH-Net	Physics Inspired Network
PHISMID	Physics-Inspired Synthesized Marine Snow Image Dataset
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSNR	Peak Signal-to-Noise Ratio
PSPNet	Pyramid Scene Parsing Network
PUIE-Net	Probabilistic Network for Underwater Image Enhancement
PVANet	Lightweight Deep Neural Networks for Real-time Object Detection
QUT	Queensland University of Technology dataset
RB	Retinex-based (approach)
R-CNN	Region-based Convolutional Network
RDN	Residual Dense Network
ReLU	Rectified Linear Unit
ResNet	Residual Network
RFEM	Residual Feature Enhancement Module

RGB	Red Green Blue
RoI	Region of Interest
ROV	Remotely operated Vehicle
RPN	Region Proposal Network
RUIE	Real-world Underwater Image Enhancement dataset
SAM	Segment Anything Models
SAR	Synthetic Aperture Radar
SAUD	Subjectively Annotated UIE benchmark dataset
SCCE	Sparse Categorical Cross Entropy
SeaCLEF	Sea Content-based Conference and Labs of the Evaluation Forum
SEAMAPD21	Southeast Area Monitoring and Assessment Program Dataset 2021
SegFormer	Segmentation framework which unifies Transformers with lightweight MLP decoders
SFD-YOLO	Seafloor-Debris-YOLO
SGD	Stochastic Gradient Descent
SGMCSS	Statistically Guided Multicolour Space Stretch module
SIUM	Segmentation of Underwater Imagery
SPL	Subaqueous Perceptual Loss
SRR	Super-Resolution Reconstruction
SSD	Single Shot Detector
SSIM	Structural Similarity Index Measure
StructureRSMAS	Structure Rosenstiel School of Marine and Atmospheric Science dataset
SUIM	Segmentation of Underwater Imagery dataset
SUN	Scene Understanding database
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TLD	Tracking Learning Detection
U45	Public underwater test dataset with 45 images
UAV	Unmanned Aerial Vehicle
UCIQE	Underwater Colour Image Quality Evaluation
UCM	Unsupervised Colour Correction
UDAE	Underwater Denoising Autoencoder
UDCP	Underwater Dark Channel Prior
UDnet	Uncertainty Distribution Network
UFO-120	Dataset for Simultaneous Enhancement and Super-Resolution (SESR) of underwater imagery
UGAN	Underwater Generative Adversarial Network
UGAN-P	UGAN with Gradient Difference Loss
UICM	Underwater Image Colourfulness Measure
UIConM	Underwater Image Contrast Measure
UIEB	Underwater Image Enhancement Benchmark
UIEVUS	Underwater Image Enhancement method designed for Various Underwater Scenes
UIIS	Underwater Image Instance Segmentation dataset
UIIS10K	Underwater Image Instance Segmentation 10K dataset
UIQM	Underwater Image Quality Measure
UISD	Underwater Images Segmentation dataset
UISFormer	Underwater Image Segmentation Transformer model
UMAP	Uniform Manifold Approximation and Projection
U-Net-scSE	U-net with Spatial Channel Squeeze-and-Excitation module
UOVSBench	Underwater Open-Vocabulary Segmentation Benchmark
URanker	Ranking-based underwater image quality assessment
URSCT-SESR	U-Net-based reinforced Swin-Convs Transformer for simultaneous enhancement and superresolution

USIS10K	Underwater Salient Instance Segmentation 10K dataset
UVEB	Underwater Video Enhancement Benchmark
UVLM	Underwater Video Language Model
UW RGB-D Object	Underwater RGB-Depth Object dataset
UWFormer	Multi-scale Transformer-based Network
UWSAM	Segment Anything Model Guided Underwater Instance Segmentation
VGG	Visual Geometry Group model
VIAME	Video and Image Analysis for Marine Environments
VidLM	Video-Language Model
ViT	Vision Transformers
VLM	Vision-Language Model
WaterGan	Water Generative Adversarial Network
WSCT	Weakly Supervised Colour Transfer
YOLO	You Only Look Once
YOLOv8-TF	Transformer-enhanced YOLOv8
ZF	Zeiler and Fergus model

References

1. Aguzzi, J.; Chatzievangelou, D.; Marini, S.; Fanelli, E.; Danovaro, R.; Flögel, S.; Lebris, N.; Juanes, F.; De Leo, F.C.; Del Rio, J.; et al. New High-Tech Flexible Networks for the Monitoring of Deep-Sea Ecosystems. *Environ. Sci. Technol.* **2019**, *53*, 6616–6631. [CrossRef]
2. Favali, P.; Beranzoli, L.; De Santis, A. *SEAFLOOR OBSERVATORIES: A New Vision of the Earth from the Abyss*; Springer Science & Business Media: New York, NY, USA, 2015.
3. Schoening, T.; Bergmann, M.; Ontrup, J.; Taylor, J.; Dannheim, J.; Gutt, J.; Purser, A.; Nattkemper, T.W. Semi-Automated Image Analysis for the Assessment of Megafaunal Densities at the Arctic Deep-Sea Observatory HAUSGARTEN. *PLoS ONE* **2012**, *7*, e38179. [CrossRef] [PubMed]
4. Aguzzi, J.; Doya, C.; Tecchio, S.; De Leo, F.; Azzurro, E.; Costa, C.; Sbragaglia, V.; Del Río, J.; Navarro, J.; Ruhl, H.; et al. Coastal Observatories for Monitoring of Fish Behaviour and Their Responses to Environmental Changes. *Rev. Fish Biol. Fish.* **2015**, *25*, 463–483. [CrossRef]
5. Widder, E.; Robison, B.H.; Reisenbichler, K.; Haddock, S. Using Red Light for In Situ Observations of Deep-Sea Fishes. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **2005**, *52*, 2077–2085. [CrossRef]
6. Chauvet, P.; Metaxas, A.; Hay, A.E.; Matabos, M. Annual and Seasonal Dynamics of Deep-Sea Megafaunal Epibenthic Communities in Barkley Canyon (British Columbia, Canada): A Response to Climatology, Surface Productivity and Benthic Boundary Layer Variation. In *Proceedings of the Progress in Oceanography*; Elsevier: Amsterdam, The Netherlands, 2018; Volume 169, pp. 89–105.
7. Leo, F.D.; Ogata, B.; Sastri, A.R.; Heesemann, M.; Mihály, S.; Galbraith, M.; Morley, M. High-Frequency Observations from a Deep-Sea Cabled Observatory Reveal Seasonal Overwintering of *Neocalanus* Spp. in Barkley Canyon, NE Pacific: Insights into Particulate Organic Carbon Flux. *Prog. Oceanogr.* **2018**, *169*, 120–137. [CrossRef]
8. Aguzzi, J.; Costa, C.; Matabos, M.; Azzurro, E.; Lázaro, A.; Menesatti, P.; Sarda, F.; Canals, M.; Delory, E.; Cline, D.; et al. Challenges to the Assessment of Benthic Populations and Biodiversity as a Result of Rhythmic Behaviour: Video Solutions from Cabled Observatories. *Oceanogr. Mar. Biol.* **2012**, *50*, 235–286.
9. Bicknell, A.W.; Godley, B.J.; Sheehan, E.V.; Votier, S.C.; Witt, M.J. Camera Technology for Monitoring Marine Biodiversity and Human Impact. *Front. Ecol. Environ.* **2016**, *14*, 424–432. [CrossRef]
10. Danovaro, R.; Aguzzi, J.; Fanelli, E.; Billett, D.; Gjerde, K.; Jamieson, A.; Ramirez-Llodra, E.; Smith, C.; Snelgrove, P.; Thomsen, L.; et al. An Ecosystem-Based Deep-Ocean Strategy. *Science* **2017**, *355*, 452–454. [CrossRef] [PubMed]
11. Szeliski, R. *Computer Vision: Algorithms and Applications*; Springer Science & Business Media: New York, NY, USA, 2010.
12. Garcia, R.; Nicosevici, T.; Cufí, X. On the Way to Solve Lighting Problems in Underwater Imaging. In *Proceedings of the OCEANS'02 MTS/IEEE*; IEEE: New York, NY, USA, 2002; Volume 2, pp. 1018–1024.
13. Prabhakar, C.; Kumar, P. An Image Based Technique for Enhancement of Underwater Images. *arXiv* **2012**, arXiv:1212.0291. [CrossRef]
14. Raj, M.V.; Murugan, S.S. Underwater Image Classification Using Machine Learning Technique. In *Proceedings of the 2019 International Symposium on Ocean Technology (SYMPOL)*; IEEE: New York, NY, USA, 2019; pp. 166–173.
15. Lippmann, R.P. Pattern Classification Using Neural Networks. *IEEE Commun. Mag.* **1989**, *27*, 47–50. [CrossRef]

16. Liu, H.; Ma, X.; Yu, Y.; Wang, L.; Hao, L. Application of Deep Learning-Based Object Detection Techniques in Fish Aquaculture: A Review. *J. Mar. Sci. Eng.* **2023**, *11*, 867. [CrossRef]
17. Zhao, S.; Zhang, S.; Liu, J.; Wang, H.; Zhu, J.; Li, D.; Zhao, R. Application of Machine Learning in Intelligent Fish Aquaculture: A Review. *Aquaculture* **2021**, *540*, 736724. [CrossRef]
18. Alsmadi, M.K.; Almarashdeh, I. A Survey on Fish Classification Techniques. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 1625–1638. [CrossRef]
19. Xu, Z.; Wang, T.; Skidmore, A.K.; Lamprey, R. A Review of Deep Learning Techniques for Detecting Animals in Aerial and Satellite Images. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *128*, 103732. [CrossRef]
20. Arsad, T.; Awalludin, E.; Bachok, Z.; Yussof, W.; Hitam, M. A Review of Coral Reef Classification Study Using Deep Learning Approach. In *Proceedings of the AIP Conference Proceedings*; AIP Publishing: New York, NY, USA, 2023; Volume 2484.
21. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, 71. [CrossRef]
22. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [CrossRef]
23. McCulloch, W.S.; Pitts, W. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133. [CrossRef]
24. Yegnanarayana, B. *Artificial Neural Networks*; PHI Learning Pvt. Ltd.: Delhi, India, 2009.
25. Rosenblatt, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychol. Rev.* **1958**, *65*, 386. [CrossRef]
26. Hopfield, J.J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. Natl. Acad. Sci. USA* **1982**, *79*, 2554–2558. [CrossRef]
27. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
28. Ciregan, D.; Meier, U.; Schmidhuber, J. Multi-Column Deep Neural Networks for Image Classification. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2012; pp. 3642–3649.
29. Chung, C.-L.; Huang, K.-J.; Chen, S.-Y.; Lai, M.-H.; Chen, Y.-C.; Kuo, Y.-F. Detecting Bakanae Disease in Rice Seedlings by Machine Vision. *Comput. Electron. Agric.* **2016**, *121*, 404–411. [CrossRef]
30. Nguyen, T.T.; Hoang, T.D.; Pham, M.T.; Vu, T.T.; Nguyen, T.H.; Huynh, Q.-T.; Jo, J. Monitoring Agriculture Areas with Satellite Images and Deep Learning. *Appl. Soft Comput.* **2020**, *95*, 106565. [CrossRef]
31. Yamamoto, K.; Guo, W.; Yoshioka, Y.; Ninomiya, S. On Plant Detection of Intact Tomato Fruits Using Image Analysis and Machine Learning Methods. *Sensors* **2014**, *14*, 12191–12206. [CrossRef] [PubMed]
32. Haug, S.; Ostermann, J. A Crop/Weed Field Image Dataset for the Evaluation of Computer Vision Based Precision Agriculture Tasks. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 105–116.
33. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
34. Shi, F.; Wang, J.; Shi, J.; Wu, Z.; Wang, Q.; Tang, Z.; He, K.; Shi, Y.; Shen, D. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19. *IEEE Rev. Biomed. Eng.* **2020**, *14*, 4–15. [CrossRef]
35. Liu, S.; Liu, S.; Cai, W.; Pujol, S.; Kikinis, R.; Feng, D. Early Diagnosis of Alzheimer's Disease with Deep Learning. In *Proceedings of the 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*; IEEE: New York, NY, USA, 2014; pp. 1015–1018.
36. Criminisi, A. Machine Learning for Medical Images Analysis. *Med. Image Anal.* **2016**, *33*, 91–93. [CrossRef]
37. Jang, H.S.; Bae, K.Y.; Park, H.-S.; Sung, D.K. Solar Power Prediction Based on Satellite Images and Support Vector Machine. *IEEE Trans. Sustain. Energy* **2016**, *7*, 1255–1263. [CrossRef]
38. Taravat, A.; Del Frate, F.; Cornaro, C.; Vergari, S. Neural Networks and Support Vector Machine Algorithms for Automatic Cloud Classification of Whole-Sky Ground-Based Images. *IEEE Geosci. Remote Sens. Lett.* **2014**, *12*, 666–670. [CrossRef]
39. Wardah, T.; Bakar, S.A.; Bardossy, A.; Maznorizan, M. Use of Geostationary Meteorological Satellite Images in Convective Rain Estimation for Flash-Flood Forecasting. *J. Hydrol.* **2008**, *356*, 283–298. [CrossRef]
40. Schalkoff, R.J. *Digital Image Processing and Computer Vision*; Wiley: New York, NY, USA, 1989; Volume 286.
41. Lu, D.; Weng, Q. A Survey of Image Classification Methods and Techniques for Improving Classification Performance. *Int. J. Remote Sens.* **2007**, *28*, 823–870. [CrossRef]
42. Viola, P.; Jones, M. Rapid Object Detection Using a Boosted Cascade of Simple Features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; CVPR 2001*; IEEE: New York, NY, USA, 2001; Volume 1, p. I–I.

43. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2014; pp. 580–587.
44. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 779–788.
45. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single Shot Multibox Detector. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
46. Gordan, M.; Dancea, O.; Stoian, I.; Georgakis, A.; Tsatos, O. A New SVM-Based Architecture for Object Recognition in Color Underwater Images with Classification Refinement by Shape Descriptors. In *Proceedings of the 2006 IEEE International Conference on Automation, Quality and Testing, Robotics*; IEEE: New York, NY, USA, 2006; Volume 2, pp. 327–332.
47. Bishop, C.M. *Pattern Recognition and Machine Learning*; Information Science and Statistics; Springer: Secaucus, NJ, USA, 2006.
48. Zion, B. The Use of Computer Vision Technologies in Aquaculture—A Review. *Comput. Electron. Agric.* **2012**, *88*, 125–132. [CrossRef]
49. Osterloff, J.; Nilssen, I.; Nattkemper, T.W. Computational Coral Feature Monitoring for the Fixed Underwater Observatory LoVe. In *Proceedings of the OCEANS 2016 MTS/IEEE Monterey*; IEEE: New York, NY, USA, 2016; pp. 1–5.
50. Villon, S.; Chaumont, M.; Subsol, G.; Villéger, S.; Claverie, T.; Mouillot, D. Coral Reef Fish Detection and Recognition in Underwater Videos by Supervised Machine Learning: Comparison between Deep Learning and HOG+ SVM Methods. In *Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 160–171.
51. Kitasato, A.; Miyazaki, T.; Sugaya, Y.; Omachi, S. Automatic Discrimination between Scomber Japonicus and Scomber Australasicus by Geometric and Texture Features. *Fishes* **2018**, *3*, 26. [CrossRef]
52. Saberioon, M.; Císař, P.; Labbé, L.; Souček, P.; Pelissier, P.; Kerneis, T. Comparative Performance Analysis of Support Vector Machine, Random Forest, Logistic Regression and k-Nearest Neighbours in Rainbow Trout (Oncorhynchus Mykiss) Classification Using Image-Based Features. *Sensors* **2018**, *18*, 1027. [CrossRef]
53. Freitas, U.; Gonçalves, W.N.; Matsubara, E.T.; Sabino, J.; Borth, M.R.; Pistori, H. Using Color for Fish Species Classification. In Proceedings of the Workshop of Industry Applications (WIA), SIBGRAPI, São José dos Campos, Brazil, 4–7 October 2016.
54. Moniruzzaman, M.; Islam, S.M.S.; Bennamoun, M.; Lavery, P. Deep Learning on Underwater Marine Object Detection: A Survey. In *Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 150–160.
55. Sun, M.; Yang, X.; Xie, Y. Deep Learning in Aquaculture: A Review. *J. Comput.* **2020**, *31*, 294–319.
56. Dong, G.; Ma, Y.; Basu, A. Feature-Guided CNN for Denoising Images from Portable Ultrasound Devices. *IEEE Access* **2021**, *9*, 28272–28281. [CrossRef]
57. Li, C.; Guo, C.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; Tao, D. An Underwater Image Enhancement Benchmark Dataset and Beyond. *IEEE Trans. Image Process.* **2019**, *29*, 4376–4389. [CrossRef]
58. Xie, Y.; Kong, L.; Chen, K.; Zheng, Z.; Yu, X.; Yu, Z.; Zheng, B. Uveb: A Large-Scale Benchmark and Baseline towards Real-World Underwater Video Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 22358–22367.
59. Li, H.; Li, J.; Wang, W. A Fusion Adversarial Underwater Image Enhancement Network with a Public Test Dataset. *arXiv* **2019**, arXiv:1906.06819. [CrossRef]
60. Islam, M.J.; Xia, Y.; Sattar, J. Fast Underwater Image Enhancement for Improved Visual Perception. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3227–3234. [CrossRef]
61. Perez, J.; Attanasio, A.C.; Nechyporenko, N.; Sanz, P.J. A Deep Learning Approach for Underwater Image Enhancement. In *Proceedings of the International Work-Conference on the Interplay Between Natural and Artificial Computation*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 183–192.
62. Getreuer, P. Automatic Color Enhancement (ACE) and Its Fast Implementation. *Image Process. Line* **2012**, *2*, 266–277. [CrossRef]
63. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 8–12 December 2014; pp. 2672–2680.
64. Yang, M.; Hu, K.; Du, Y.; Wei, Z.; Sheng, Z.; Hu, J. Underwater Image Enhancement Based on Conditional Generative Adversarial Network. *Signal Process. Image Commun.* **2020**, *81*, 115723. [CrossRef]
65. Fabbri, C.; Islam, M.J.; Sattar, J. Enhancing Underwater Imagery Using Generative Adversarial Networks. In *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2018; pp. 7159–7165.
66. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2017; pp. 2223–2232.
67. Islam, M.J.; Sattar, J. Mixed-Domain Biological Motion Tracking for Underwater Human-Robot Interaction. In *Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2017; pp. 4457–4464.

68. Li, J.; Skinner, K.A.; Eustice, R.M.; Johnson-Roberson, M. WaterGAN: Unsupervised Generative Network to Enable Real-Time Color Correction of Monocular Underwater Images. *IEEE Robot. Autom. Lett.* **2017**, *3*, 387–394. [CrossRef]
69. Shin, Y.-S.; Cho, Y.; Pandey, G.; Kim, A. Estimation of Ambient Light and Transmission Map with Common Convolutional Architecture. In *Proceedings of the OCEANS 2016 MTS/IEEE Monterey*; IEEE: New York, NY, USA, 2016; pp. 1–7.
70. Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
71. Hashisho, Y.; Albadawi, M.; Krause, T.; von Lukas, U.F. Underwater Color Restoration Using U-Net Denoising Autoencoder. In *Proceedings of the 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*; IEEE: New York, NY, USA, 2019; pp. 117–122.
72. Li, Y.; Lu, H.; Li, J.; Li, X.; Li, Y.; Serikawa, S. Underwater Image De-Scattering and Classification by Deep Neural Network. *Comput. Electr. Eng.* **2016**, *54*, 68–77. [CrossRef]
73. Sheikh, H.R.; Bovik, A.C. 8.4—Information Theoretic Approaches to Image Quality Assessment. In *Handbook of Image and Video Processing*, 2nd ed.; BOVIK, A., Ed.; Communications, Networking and Multimedia; Academic Press: Burlington, MA, USA, 2005; pp. 975–989, ISBN 978-0-12-119792-6.
74. Fu, Z.; Wang, W.; Huang, Y.; Ding, X.; Ma, K.-K. Uncertainty Inspired Underwater Image Enhancement. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 465–482.
75. Rezende, D.J.; Mohamed, S.; Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1278–1286.
76. Huang, X.; Belongie, S. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2017; pp. 1501–1510.
77. Liu, R.; Fan, X.; Zhu, M.; Hou, M.; Luo, Z. Real-World Underwater Enhancement: Challenges, Benchmarks, and Solutions under Natural Light. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4861–4875. [CrossRef]
78. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Process. Lett.* **2012**, *20*, 209–212. [CrossRef]
79. Sun, S.; Wang, H.; Zhang, H.; Li, M.; Xiang, M.; Luo, C.; Ren, P. Underwater Image Enhancement with Reinforcement Learning. *IEEE J. Ocean. Eng.* **2022**, *49*, 249–261. [CrossRef]
80. Yang, M.; Sowmya, A. An Underwater Color Image Quality Evaluation Metric. *IEEE Trans. Image Process.* **2015**, *24*, 6062–6071. [CrossRef]
81. Panetta, K.; Gao, C.; Agaian, S. Human-Visual-System-Inspired Underwater Image Quality Measures. *IEEE J. Ocean. Eng.* **2016**, *41*, 541–551. [CrossRef]
82. Guo, Y.; Li, H.; Zhuang, P. Underwater Image Enhancement Using a Multiscale Dense Generative Adversarial Network. *IEEE J. Ocean. Eng.* **2020**, *45*, 862–870. [CrossRef]
83. Cao, T.; Yu, Z.; Zheng, B. DNnet: A Lightweight Network for Real-Time 4K Underwater Image Enhancement Using Dynamic Range and Average Normalization. *Expert Syst. Appl.* **2025**, *270*, 126561. [CrossRef]
84. Ren, S.; Bao, X.; Wang, T.; Xu, X.; Ma, T.; Yu, K. UIEVUS: An Underwater Image Enhancement Method for Various Underwater Scenes. *Signal Process. Image Commun.* **2025**, *270*, 117264. [CrossRef]
85. Liu, Y.; Jiang, Q.; Wang, X.; Luo, T.; Zhou, J. Underwater Image Enhancement with Cascaded Contrastive Learning. *IEEE Trans. Multimed.* **2025**, *27*, 1512–1525. [CrossRef]
86. Zhu, S.; Geng, Z.; Xie, Y.; Zhang, Z.; Yan, H.; Zhou, X.; Jin, H.; Fan, X. New Underwater Image Enhancement Algorithm Based on Improved U-Net. *Water* **2025**, *17*, 808. [CrossRef]
87. Saleh, A.; Sheaves, M.; Jerry, D.; Azghadi, M.R. Adaptive Deep Learning Framework for Robust Unsupervised Underwater Image Enhancement. *Expert Syst. Appl.* **2025**, *268*, 126314. [CrossRef]
88. Vaswani, A. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
89. Dosovitskiy, A. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
90. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2021; pp. 10012–10022.
91. Ji, J.; Man, J. UNet–Transformer Hybrid Architecture for Enhanced Underwater Image Processing and Restoration. *Mathematics* **2025**, *13*, 2535. [CrossRef]
92. Chen, W.; Lei, Y.; Luo, S.; Zhou, Z.; Li, M.; Pun, C.-M. Uwformer: Underwater Image Enhancement via a Semi-Supervised Multi-Scale Transformer. In *Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN)*; IEEE: New York, NY, USA, 2024; pp. 1–8.
93. Wang, H.; Köser, K.; Ren, P. Large Foundation Model Empowered Discriminative Underwater Image Enhancement. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5609317. [CrossRef]

94. Janoch, A.; Karayev, S.; Jia, Y.; Barron, J.T.; Fritz, M.; Saenko, K.; Darrell, T. A Category-Level 3d Object Dataset: Putting the Kinect to Work. In *Consumer Depth Cameras for Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 141–165.
95. Lai, K.; Bo, L.; Fox, D. Unsupervised Feature Learning for 3d Scene Labeling. In *Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2014; pp. 3050–3057.
96. Silberman, N.; Fergus, R. Indoor Scene Segmentation Using a Structured Light Sensor. In *Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*; IEEE: New York, NY, USA, 2011; pp. 601–608.
97. Shotton, J.; Glocker, B.; Zach, C.; Izadi, S.; Criminisi, A.; Fitzgibbon, A. Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2013; pp. 2930–2937.
98. Xiao, J.; Hays, J.; Ehinger, K.A.; Oliva, A.; Torralba, A. Sun Database: Large-Scale Scene Recognition from Abbey to Zoo. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2010; pp. 3485–3492.
99. Peng, L.; Zhu, C.; Bian, L. U-Shape Transformer for Underwater Image Enhancement. *IEEE Trans. Image Process.* **2023**, *32*, 3066–3079. [CrossRef]
100. Saleh, A.; Laradji, I.H.; Konovalov, D.A.; Bradley, M.; Vazquez, D.; Sheaves, M. A Realistic Fish-Habitat Dataset to Evaluate Algorithms for Underwater Visual Analysis. *Sci. Rep.* **2020**, *10*, 14671. [CrossRef]
101. Islam, M.J.; Edge, C.; Xiao, Y.; Luo, P.; Mehtaz, M.; Morse, C.; Enan, S.S.; Sattar, J. Semantic Segmentation of Underwater Imagery: Dataset and Benchmark. In *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2020; pp. 1769–1776.
102. Berman, D.; Levy, D.; Avidan, S.; Treibitz, T. Underwater Single Image Color Restoration Using Haze-Lines and a New Quantitative Dataset. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 2822–2837. [CrossRef] [PubMed]
103. Kratzert, F.; Mader, H. Fish Species Classification in Underwater Video Monitoring Using Convolutional Neural Networks. *Earth ArXiv* **2018**. [CrossRef]
104. Liang, J.; Fu, Z.; Lei, X.; Dai, X.; Lv, B. Recognition and Classification of Ornamental Fish Image Based on Machine Vision. In *Proceedings of the 2020 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS)*; IEEE Computer Society: Washington, DC, USA, 2020; pp. 910–913.
105. Cao, Z.; Principe, J.C.; Ouyang, B.; Dalgleish, F.; Vuorenkoski, A. Marine Animal Classification Using Combined CNN and Hand-Designed Image Features. In *Proceedings of the OCEANS'15 MTS/IEEE Washington*; IEEE: New York, NY, USA, 2015; pp. 1–6.
106. Mahmood, A.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F.; Hovey, R.; Kendrick, G.; Fisher, R.B. Automatic Annotation of Coral Reefs Using Deep Learning. In *Proceedings of the Oceans 2016 MTS/IEEE Monterey*; IEEE: New York, NY, USA, 2016; pp. 1–5.
107. Meng, L.; Hirayama, T.; Oyanagi, S. Underwater-Drone with Panoramic Camera for Automatic Fish Recognition Based on Deep Learning. *IEEE Access* **2018**, *6*, 17880–17886. [CrossRef]
108. Rahmat, B.; Waluyo, M.; Rachmanto, T.A.; Afandi, M.I.; Widyantara, H.; Harianto. Video-Based Tancho Koi Fish Tracking System Using CSK, DFT, and LOT. *J. Phys. Conf. Ser.* **2020**, *1569*, 022036. [CrossRef]
109. Rossi, F.; Benso, A.; Carlo, S.D.; Politano, G.; Savino, A.; Acutis, P.L. FishAPP: A Mobile App to Detect Fish Falsification through Image Processing and Machine Learning Techniques. In Proceedings of the 20th IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR 2016), Cluj-Napoca, Romania, 19–21 May 2016; pp. 1–6.
110. Kutlu, Y.; Altan, G.; İşçimen, B.; Doğdu, S.A.; Turan, C. Recognition of Species of Triglidae Family Using Deep Learning. *J. Black Sea/Mediterr. Environ.* **2017**, *23*, 56–65.
111. Banan, A.; Nasiri, A.; Taheri-Garavand, A. Deep Learning-Based Appearance Features Extraction for Automated Carp Species Identification. *Aquac. Eng.* **2020**, *89*, 102053. [CrossRef]
112. Pudaruth, S.; Nazurally, N.; Appadoo, C.; Kishnah, S.; Vinayaganidhi, M.; Mohammoodally, I.; Ally, Y.A.; Chady, F. SuperFish: A Mobile Application for Fish Species Recognition Using Image Processing Techniques and Deep Learning. *Int. J. Comput. Digit. Syst.* **2020**, *10*, 1157–1165. [CrossRef]
113. Cao, S.; Zhao, D.; Liu, X.; Sun, Y. Real-Time Robust Detector for Underwater Live Crabs Based on Deep Learning. *Comput. Electron. Agric.* **2020**, *172*, 105339. [CrossRef]
114. Naddaf-Sh, M.; Myler, H.; Zargarzadeh, H. Design and Implementation of an Assistive Real-Time Red Lionfish Detection System for AUV/ROVs. *Complexity* **2018**, *2018*, 5298294. [CrossRef]
115. Dawkins, M.; Sherrill, L.; Fieldhouse, K.; Hoogs, A.; Richards, B.; Zhang, D.; Prasad, L.; Williams, K.; Lauffenburger, N.; Wang, G. An Open-Source Platform for Underwater Image and Video Analytics. In *Proceedings of the Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2017; pp. 898–906.
116. Mandal, R.; Connolly, R.M.; Schlacherz, T.A.; Stantic, B. Assessing Fish Abundance from Underwater Video Using Deep Neural Networks. *arXiv* **2018**, arXiv:1807.05838. [CrossRef]

117. Labao, A.B.; Naval, P.C. Weakly-Labelled Semantic Segmentation of Fish Objects in Underwater Videos Using a Deep Residual Network. In *Proceedings of the Asian Conference on Intelligent Information and Database Systems*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 255–265.
118. Martin-Abadal, M.; Ruiz-Frau, A.; Hinz, H.; Gonzalez-Cid, Y. Jellytoring: Real-Time Jellyfish Monitoring Based on Deep Learning Object Detection. *Sensors* **2020**, *20*, 1708. [CrossRef] [PubMed]
119. Lu, H.; Li, Y.; Uemura, T.; Ge, Z.; Xu, X.; He, L.; Serikawa, S.; Kim, H. FDCNet: Filtering Deep Convolutional Network for Marine Organism Classification. *Multimed. Tools Appl.* **2017**, *77*, 21847–21860. [CrossRef]
120. Rimavicius, T.; Gelzinis, A. A Comparison of the Deep Learning Methods for Solving Seafloor Image Classification Task. In *Proceedings of the International Conference on Information and Software Technologies*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 442–453.
121. Osterloff, J.; Nilssen, I.; Nattkemper, T.W. A Computer Vision Approach for Monitoring the Spatial and Temporal Shrimp Distribution at the LoVe Observatory. *Methods Oceanogr.* **2016**, *15*, 114–128. [CrossRef]
122. Møller, T.; Nillsen, I.; Nattkemper, T.W. Active Learning for the Classification of Species in Underwater Images from a Fixed Observatory. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2017.
123. Chen, C.-H.; Liu, K.-H. Stingray Detection of Aerial Images with Region-Based Convolution Neural Network. In *Proceedings of the Consumer Electronics-Taiwan (ICCE-TW), 2017 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*; IEEE: New York, NY, USA, 2017; pp. 175–176.
124. Levy, D.; Belfer, Y.; Osherov, E.; Bigal, E.; Scheinin, A.P.; Nativ, H.; Tchernov, D.; Treibitz, T. Automated Analysis of Marine Video with Limited Data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*; IEEE: New York, NY, USA, 2018; pp. 1385–1393.
125. Guirado, E.; Tabik, S.; Rivas, M.L.; Alcaraz-Segura, D.; Herrera, F. Whale Counting in Satellite and Aerial Images with Deep Learning. *Sci. Rep.* **2019**, *9*, 14259. [CrossRef]
126. Dujon, A.M.; Ierodiaconou, D.; Geeson, J.J.; Arnould, J.P.; Allan, B.M.; Katselidis, K.A.; Schofield, G. Machine Learning to Detect Marine Animals in UAV Imagery: Effect of Morphology, Spacing, Behaviour and Habitat. *Remote Sens. Ecol. Conserv.* **2021**, *7*, 341–354. [CrossRef]
127. Boulent, J.; Charry, B.; Kennedy, M.M.; Tissier, E.; Fan, R.; Marcoux, M.; Watt, C.A.; Gagné-Turcotte, A. Scaling Whale Monitoring Using Deep Learning: A Human-in-the-Loop Solution for Analyzing Aerial Datasets. *Front. Mar. Sci.* **2023**, *10*, 1099479. [CrossRef]
128. Patel, M.; Chen, X.; Xu, L.; Cantu, F.J.P.; Turnes, J.N.; Brubacher, N.C.; Clausi, D.A.; Scott, K.A. The Influence of Input Image Scale on Deep Learning-Based Beluga Whale Detection from Aerial Remote Sensing Imagery. In *Proceedings of the IGARSS 2023—2023 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: New York, NY, USA, 2023; pp. 5732–5734.
129. Kong, M.; Liu, Y.; Li, B.; Duan, Q. A Lightweight Method for Detecting Turned White Belly Fish in Ponds Using Unmanned Aerial Vehicle Imagery. *Eng. Appl. Artif. Intell.* **2025**, *144*, 110111. [CrossRef]
130. Lee, H.; Park, M.; Kim, J. Plankton Classification on Imbalanced Large Scale Database via Convolutional Neural Networks with Transfer Learning. In *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*; IEEE: New York, NY, USA, 2016; pp. 3713–3717.
131. Orenstein, E.C.; Beijbom, O.; Peacock, E.E.; Sosik, H.M. Whoi-Plankton-a Large Scale Fine Grained Visual Recognition Benchmark Dataset for Plankton Classification. *arXiv* **2015**, arXiv:1510.00745.
132. Py, O.; Hong, H.; Zhongzhi, S. Plankton Classification with Deep Convolutional Neural Networks. In *Proceedings of the 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference*; IEEE: New York, NY, USA, 2016; pp. 132–136.
133. Cowen, R.K.; Sponaugle, S.; Robinson, K.; Luo, J. Planktonset 1.0: Plankton Imagery Data Collected from Fg Walton Smith in Straits of Florida from 2014–06-03 to 2014–06-06 and Used in the 2015 National Data Science Bowl (Ncei Accession 0127422) 2015. Available online: https://github.com/Planktos/PlanktonSet-1.0 (accessed on 5 March 2026).
134. Dai, J.; Wang, R.; Zheng, H.; Ji, G.; Qiao, X. Zooplanktonet: Deep Convolutional Network for Zooplankton Classification. In *Proceedings of the OCEANS 2016-Shanghai*; IEEE: New York, NY, USA, 2016; pp. 1–6.
135. Lumini, A.; Nanni, L. Deep Learning and Transfer Learning Features for Plankton Classification. *Ecol. Inform.* **2019**, *51*, 33–43. [CrossRef]
136. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 25 (NeurIPS 2012), Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
137. Li, Y.; Guo, J.; Guo, X.; Hu, Z.; Tian, Y. Plankton Detection with Adversarial Learning and a Densely Connected Deep Learning Model for Class Imbalanced Distribution. *J. Mar. Sci. Eng.* **2021**, *9*, 636. [CrossRef]
138. Yue, J.; Chen, Z.; Long, Y.; Cheng, K.; Bi, H.; Cheng, X. Toward Efficient Deep Learning System for In-Situ Plankton Image Recognition. *Front. Mar. Sci.* **2023**, *10*, 1186343. [CrossRef]
139. Sömek, B.; Yuksel, S.E. Plankton Classification with Deep Learning. In *Proceedings of the 2023 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*; IEEE: New York, NY, USA, 2023; pp. 118–123.

140. Gonzalez-Cid, Y.; Burguera, A.; Bonin-Font, F.; Matamoros, A. Machine Learning and Deep Learning Strategies to Identify Posidonia Meadows in Underwater Images. In *Proceedings of the OCEANS 2017-Aberdeen*; IEEE: New York, NY, USA, 2017; pp. 1–5.
141. Martin-Abadal, M.; Guerrero-Font, E.; Bonin-Font, F.; Gonzalez-Cid, Y. Deep Semantic Segmentation in an AUV for Online Posidonia Oceanica Meadows Identification. *IEEE Access* **2018**, *6*, 60956–60967. [CrossRef]
142. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2015; pp. 3431–3440.
143. Sabato, G.; Scardino, G.; Kushabaha, A.; Chirivì, M.; Luparelli, A.; Scicchitano, G. Deep Learning-Based Segmentation Techniques for Coastal Monitoring and Seagrass Banquette Detection. In *Proceedings of the 2023 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea)*; IEEE: New York, NY, USA, 2023; pp. 524–527.
144. Liu, L.; Bao, Z.; Liang, Y.; Deng, H.; Zhang, X.; Cao, T.; Zhou, C.; Zhang, Z. Unsupervised Learning for Lake Underwater Vegetation Classification: Constructing High-Precision, Large-Scale Aquatic Ecological Datasets. *Sci. Total Environ.* **2025**, *958*, 177895. [CrossRef]
145. Gao, L.; Li, X.; Kong, F.; Yu, R.; Guo, Y.; Ren, Y. AlgaeNet: A Deep-Learning Framework to Detect Floating Green Algae from Optical and SAR Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 2782–2796. [CrossRef]
146. Elsäßer, J.; Weihl, L.; Cheplygina, V.; Nielsen, L.T. SeagrassFinder: Deep Learning for Eelgrass Detection and Coverage Estimation in the Wild. *Ecol. Inform.* **2025**, *90*, 103200. [CrossRef]
147. Zhao, F.; Huang, B.; Wang, J.; Shao, X.; Wu, Q.; Xi, D.; Liu, Y.; Chen, Y.; Zhang, G.; Ren, Z.; et al. Seafloor Debris Detection Using Underwater Images and Deep Learning-Driven Image Restoration: A Case Study from Koh Tao, Thailand. *Mar. Pollut. Bull.* **2025**, *214*, 117710. [CrossRef]
148. Mahmood, A.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F.; Hovey, R.; Kendrick, G.; Fisher, R.B. Coral Classification with Hybrid Feature Representations. In *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*; IEEE: New York, NY, USA, 2016; pp. 519–523.
149. Bewley, M.; Friedman, A.; Ferrari, R.; Hill, N.; Hovey, R.; Barrett, N.; Marzinelli, E.M.; Pizarro, O.; Figueira, W.; Meyer, L.; et al. Australian Sea-Floor Survey Data, with Images and Expert Annotations. *Sci. Data* **2015**, *2*, 150057. [CrossRef]
150. Osterloff, J.; Nilssen, I.; Järnegren, J.; Buhl-Mortensen, P.; Nattkemper, T.W. Polyp Activity Estimation and Monitoring for Cold Water Corals with a Deep Learning Approach. In *Proceedings of the Computer Vision for Analysis of Underwater Imagery (CVAUI), 2016 ICPR 2nd Workshop on Computer Vision for Analysis of Underwater Imagery (CVAUI)*; IEEE: New York, NY, USA, 2016; pp. 1–6.
151. Giles, A.B.; Ren, K.; Davies, J.E.; Abrego, D.; Kelaher, B. Combining Drones and Deep Learning to Automate Coral Reef Assessment with RGB Imagery. *Remote Sens.* **2023**, *15*, 2238. [CrossRef]
152. Anwarul, S.; Tanwar, R. CoralClassify: Advancing Coral Health Monitoring Using Deep Learning. *Procedia Comput. Sci.* **2025**, *259*, 88–97. [CrossRef]
153. Zheng, Z.; Liang, H.; Wut, F.H.; Wong, Y.H.; Chui, A.P.-Y.; Yeung, S.-K. Hkcoral: Benchmark for Dense Coral Growth Form Segmentation in the Wild. *IEEE J. Ocean. Eng.* **2025**, *50*, 697–713. [CrossRef]
154. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
155. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
156. LeCun, Y.; Jackel, L.D.; Boser, B.; Denker, J.S.; Graf, H.P.; Guyon, I.; Henderson, D.; Howard, R.E.; Hubbard, W. Handwritten Digit Recognition: Applications of Neural Network Chips and Automatic Learning. *IEEE Commun. Mag.* **1989**, *27*, 41–46. [CrossRef]
157. Qiu, C.; Zhang, S.; Wang, C.; Yu, Z.; Zheng, H.; Zheng, B. Improving Transfer Learning and Squeeze-and-Excitation Networks for Small-Scale Fine-Grained Fish Image Classification. *IEEE Access* **2018**, *6*, 78503–78512. [CrossRef]
158. Lin, T.-Y.; RoyChowdhury, A.; Maji, S. Bilinear Cnn Models for Fine-Grained Visual Recognition. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2015; pp. 1449–1457.
159. Jäger, J.; Simon, M.; Denzler, J.; Wolff, V.; Fricke-Neuderth, K.; Kruschel, C. Croatian Fish Dataset: Fine-Grained Classification of Fish Species in Their Natural Habitat. *Swans. Bmvc* **2015**, *2*, 6.1–6.7.
160. Anantharajah, K.; Ge, Z.; McCool, C.; Denman, S.; Fookes, C.; Corke, P.; Tjondronegoro, D.; Sridharan, S. Local Inter-Session Variability Modelling for Object Classification. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*; IEEE: New York, NY, USA, 2014; pp. 309–316.
161. Ding, G.; Song, Y.; Guo, J.; Feng, C.; Li, G.; He, B.; Yan, T. Fish Recognition Using Convolutional Neural Network. In *Proceedings of the OCEANS–Anchorage*; IEEE: New York, NY, USA, 2017; pp. 1–4.
162. Qin, H.; Li, X.; Yang, Z.; Shang, M. When Underwater Imagery Analysis Meets Deep Learning: A Solution at the Age of Big Visual Data. In *Proceedings of the OCEANS'15 MTS/IEEE Washington*; IEEE: New York, NY, USA, 2015; pp. 1–5.
163. Qin, H.; Li, X.; Liang, J.; Peng, Y.; Zhang, C. DeepFish: Accurate Underwater Live Fish Recognition with a Deep Architecture. *Neurocomputing* **2016**, *187*, 49–58. [CrossRef]

164. Qin, H.; Peng, Y.; Li, X. Foreground Extraction of Underwater Videos via Sparse and Low-Rank Matrix Decomposition. In *Proceedings of the 2014 ICPR Workshop on Computer Vision for Analysis of Underwater Imagery*; IEEE: New York, NY, USA, 2014; pp. 65–72.
165. Rathi, D.; Jain, S.; Indu, D.S. Underwater Fish Species Classification Using Convolutional Neural Network and Deep Learning. *arXiv* **2018**, arXiv:1805.10106. [CrossRef]
166. Han, F.; Zhu, J.; Liu, B.; Zhang, B.; Xie, F. Fish Shoals Behavior Detection Based on Convolutional Neural Network and Spatiotemporal Information. *IEEE Access* **2020**, *8*, 126907–126926. [CrossRef]
167. Jovanović, V.; Svendsen, E.; Risojević, V.; Babić, Z. Splash Detection in Fish Plants Surveillance Videos Using Deep Learning. In *Proceedings of the 2018 14th Symposium on Neural Networks and Applications (NEUREL)*; IEEE: New York, NY, USA, 2018; pp. 1–5.
168. Huang, R.-J.; Lai, Y.-C.; Tsao, C.-Y.; Kuo, Y.-P.; Wang, J.-H.; Chang, C.-C. Applying Convolutional Networks to Underwater Tracking without Training. In *Proceedings of the 2018 IEEE International Conference on Applied System Invention (ICASI)*; IEEE: New York, NY, USA, 2018; pp. 342–345.
169. Cui, Y.; Pan, T.; Chen, S.; Zou, X. A Gender Classification Method for Chinese Mitten Crab Using Deep Convolutional Neural Network. *Multimed. Tools Appl.* **2020**, *79*, 7669–7684. [CrossRef]
170. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
171. Taheri-Garavand, A.; Nasiri, A.; Banan, A.; Zhang, Y.-D. Smart Deep Learning-Based Approach for Non-Destructive Freshness Diagnosis of Common Carp Fish. *J. Food Eng.* **2020**, *278*, 109930. [CrossRef]
172. Becken, S.; Connolly, R.; Stantic, B.; Scott, N.; Mandal, R.; Le, D. *Monitoring Aesthetic Value of the Great Barrier Reef by Using Innovative Technologies and Artificial Intelligence*; Griffith Institute for Tourism Research Report; Griffith Institute for Tourism, Griffith University: Brisbane, Australia, 2018.
173. Mader, H.; Kratzert, F. The Fishcam Migration Monitoring System for Fish Passes. In *Proceedings of the 11th International Symposium on Ecohydraulics*; Webb, J., Costelloe, J., Casas-Mulet, R., Lyon, J., Stewardson, M., Eds.; University of Melbourne: Melbourne, Australia, 2016.
174. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A Large-Scale Hierarchical Image Database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2009; pp. 248–255.
175. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
176. Chen, G.; Sun, P.; Shang, Y. Automatic Fish Classification System Using Deep Learning. In *Proceedings of the Tools with Artificial Intelligence (ICTAI), 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*; IEEE: New York, NY, USA, 2017; pp. 24–29.
177. Ali-Gombe, A.; Elyan, E.; Jayne, C. Fish Classification in Context of Noisy Images. In *Proceedings of the International Conference on Engineering Applications of Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 216–226.
178. Sirigineedi, M.; Jagan Mohan, R.; Sahu, B. Improving Fish Image Detection Speed with Hybrid VGG16 and Darknet. *Multimed. Tools Appl.* **2024**, *84*, 10551–10566. [CrossRef]
179. Thorat, P.; Tongaonkar, R.; Jagtap, V. Towards Designing the Best Model for Classification of Fish Species Using Deep Neural Networks. In *Proceedings of the International Conference on Computational Science and Applications: ICCSA 2019*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 343–351.
180. Prasetyo, E.; Suciati, N.; Fatichah, C. Multi-Level Residual Network VGGNet for Fish Species Classification. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 5286–5295. [CrossRef]
181. Prasetyo, E.; Suciati, N.; Fatichah, C. Fish-Gres Dataset for Fish Species Classification. *Mendeley Data* **2020**, *10*, 12. [CrossRef]
182. Mamun, M.R.I.; Rahman, U.S.; Akter, T.; Azim, M.A. Fish Disease Detection Using Deep Learning and Machine Learning. *Int. J. Comput. Appl.* **2023**, *975*, 8887. [CrossRef]
183. Choi, S. Fish Identification in Underwater Video with Deep Convolutional Neural Network: SNUMedinfo at LifeCLEF Fish Task 2015. Available online: https://ceur-ws.org/Vol-1391/110-CR.pdf (accessed on 5 March 2026).
184. Jäger, J.; Rodner, E.; Denzler, J.; Wolff, V.; Fricke-Neuderth, K. SeaCLEF 2016: Object Proposal Classification for Fish Detection in Underwater Videos. In *Proceedings of the CLEF 2016 Working Notes, Évora, Portugal, 5–8 September 2016*; CEUR-WS.org: Aachen, Germany, 2016; pp. 481–489.
185. Mothes, O.; Denzler, J. Anatomical Landmark Tracking by One-Shot Learned Priors for Augmented Active Appearance Models. In Proceedings of the VISIGRAPP (6: VISAPP), Porto, Portugal, 27 February–1 March 2017; pp. 246–254.
186. Jäger, J.; Wolff, V.; Fricke-Neuderth, K.; Mothes, O.; Denzler, J. Visual Fish Tracking: Combining a Two-Stage Graph Approach with CNN-Features. In *Proceedings of the OCEANS 2017-Aberdeen*; IEEE: New York, NY, USA, 2017; pp. 1–6.
187. Pelletier, S.; Montacir, A.; Zakari, H.; Akhloufi, M. Deep Learning for Marine Resources Classification in Non-Structured Scenarios: Training vs. Transfer Learning. In *Proceedings of the 2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)*; IEEE: New York, NY, USA, 2018; pp. 1–4.

188. Qu, P.; Li, T.; Zhou, L.; Jin, S.; Liang, Z.; Zhao, W.; Zhang, W. DAMNet: Dual Attention Mechanism Deep Neural Network for Underwater Biological Image Classification. *IEEE Access* **2022**, *11*, 6000–6009. [CrossRef]
189. Mehrunnisa; Leszczuk, M.; Juszka, D.; Zhang, Y. Improved Binary Classification of Underwater Images Using a Modified ResNet-18 Model. *Electronics* **2025**, *14*, 2954. [CrossRef]
190. Jiang, Q.; Gu, Y.; Li, C.; Cong, R.; Shao, F. Underwater Image Enhancement Quality Evaluation: Benchmark Dataset and Objective Metric. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5959–5974. [CrossRef]
191. Kaneko, R.; Ueda, T.; Higashi, H.; Tanaka, Y. PHISWID: Physics-Inspired Underwater Image Dataset Synthesized from RGB-D Images. *APSIPA Trans. Signal Inf. Process.* **2025**, *15*, 1–25. [CrossRef]
192. Kaneko, R.; Sato, Y.; Ueda, T.; Higashi, H.; Tanaka, Y. Marine Snow Removal Benchmarking Dataset. In *Proceedings of the 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*; IEEE: New York, NY, USA, 2023; pp. 771–778.
193. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
194. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28 (NeurIPS 2015)*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; pp. 91–99.
195. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
196. Li, X.; Tang, Y.; Gao, T. Deep but Lightweight Neural Networks for Fish Detection. In *Proceedings of the OCEANS 2017-Aberdeen*; IEEE: New York, NY, USA, 2017; pp. 1–5.
197. Li, X.; Shang, M.; Qin, H.; Chen, L. Fast Accurate Fish Detection and Recognition of Underwater Images with Fast R-CNN. In *Proceedings of the OCEANS'15 MTS/IEEE Washington*; IEEE: New York, NY, USA, 2015; pp. 1–5.
198. Wang, H.; Xiao, N. Underwater Object Detection Method Based on Improved Faster RCNN. *Appl. Sci.* **2023**, *13*, 2746. [CrossRef]
199. Ben Tamou, A.; Benzinou, A.; Nasreddine, K. Multi-Stream Fish Detection in Unconstrained Underwater Videos by the Fusion of Two Convolutional Neural Network Detectors. *Appl. Intell.* **2021**, *51*, 5809–5821. [CrossRef]
200. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; pp. 818–833.
201. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the Devil in the Details: Delving Deep into Convolutional Nets. *arXiv* **2014**, arXiv:1405.3531. [CrossRef]
202. Xu, W.; Zhu, Z.; Ge, F.; Han, Z.; Li, J. Analysis of Behavior Trajectory Based on Deep Learning in Ammonia Environment for Fish. *Sensors* **2020**, *20*, 4425. [CrossRef]
203. Isa, I.S.; Norzrin, N.N.; Sulaiman, S.N.; Hamzaid, N.A.; Maruzuki, M.I.F. CNN Transfer Learning of Shrimp Detection for Underwater Vision System. In *Proceedings of the 2020 1st International Conference on Information Technology, Advanced Mechanical and Electrical Engineering (ICITAMEE)*; IEEE: New York, NY, USA, 2020; pp. 226–231.
204. Baletaud, F.; Villon, S.; Gilbert, A.; Côme, J.-M.; Fiat, S.; Iovan, C.; Vigliola, L. Automatic Detection, Identification and Counting of Deep-Water Snappers on Underwater Baited Video Using Deep Learning. *Front. Mar. Sci.* **2025**, *12*, 1476616. [CrossRef]
205. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
206. Wu, Z.; Shen, C.; van den Hengel, A. High-Performance Semantic Segmentation Using Very Deep Fully Convolutional Networks. *arXiv* **2016**, arXiv:1604.04339. [CrossRef]
207. Zivkovic, Z. Improved Adaptive Gaussian Mixture Model for Background Subtraction. In Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), Cambridge, UK, 23–26 August 2004; pp. 28–31.
208. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 7263–7271.
209. Bochkovskiy, A. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
210. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. Yolov9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616. [CrossRef]
211. Lu, H.; Uemura, T.; Wang, D.; Zhu, J.; Huang, Z.; Kim, H. Deep-Sea Organisms Tracking Using Dehazing and Deep Learning. *Mob. Netw. Appl.* **2018**, *25*, 1008–1015. [CrossRef]
212. Kalal, Z.; Mikolajczyk, K.; Matas, J. Tracking-Learning-Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1409–1422. [CrossRef]
213. Kalal, Z.; Mikolajczyk, K.; Matas, J. Forward-Backward Error: Automatic Detection of Tracking Failures. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*; IEEE: New York, NY, USA, 2010; pp. 2756–2759.
214. Hofmann, A.C.; Bergmann, S.M.; Reulke, R. Analysis of Motion Patterns in Video Streams for Automatic Health Monitoring in Koi Ponds. In Proceedings of the Image and Video Technology: 9th Pacific-Rim Symposium, PSIVT 2019, Sydney, Australia, 18–22 November 2019; p. 27.

215. Al Muksit, A.; Hasan, F.; Emon, M.F.H.B.; Haque, M.R.; Anwary, A.R.; Shatabda, S. YOLO-Fish: A Robust Fish Detection Model to Detect Fish in Realistic Underwater Environment. *Ecol. Inform.* **2022**, *72*, 101847. [CrossRef]
216. Mahmood, A.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F.; Hovey, R.; Kendrick, G. Automatic Detection of Western Rock Lobster Using Synthetic Data. *ICES J. Mar. Sci.* **2019**, *77*, 1308–1317. [CrossRef]
217. Zhang, M.; Xu, S.; Song, W.; He, Q.; Wei, Q. Lightweight Underwater Object Detection Based on YOLO v4 and Multi-Scale Attentional Feature Fusion. *Remote Sens.* **2021**, *13*, 4706. [CrossRef]
218. Pedersen, M.; Bruslund Haurum, J.; Gade, R.; Moeslund, T.B. Detection of Marine Animals in a New Underwater Dataset with Varying Visibility. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*; IEEE: New York, NY, USA, 2019; pp. 18–26.
219. Li, L.; Shi, G.; Jiang, T. Fish Detection Method Based on Improved YOLOv5. *Aquac. Int.* **2023**, *31*, 2513–2530. [CrossRef]
220. Gao, M.; Li, S.; Wang, K.; Bai, Y.; Ding, Y.; Zhang, B.; Guan, N.; Wang, P. Real-Time Jellyfish Classification and Detection Algorithm Based on Improved YOLOv4-Tiny and Improved Underwater Image Enhancement Algorithm. *Sci. Rep.* **2023**, *13*, 12989. [CrossRef] [PubMed]
221. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
222. Zhang, W.; Rui, F.; Xiao, C.; Li, H.; Li, Y. JF-YOLO: The Jellyfish Bloom Detector Based on Deep Learning. *Multimed. Tools Appl.* **2024**, *83*, 7097–7117. [CrossRef]
223. Chen, V.Y.; Wu, Y.-W.; Hu, C.-W.; Han, Y.-S. Enhancing Green Sea Turtle (Chelonia Mydas) Conservation for Tourists at Little Liuqiu Island, Taiwan: Application of Deep Learning Algorithms. *Ocean Coast. Manag.* **2024**, *252*, 107111. [CrossRef]
224. Bukas, C.; Albrecht, F.; Ur-Rehman, M.S.; Popek, D.; Patalan, M.; Pawłowski, J.; Wecker, B.; Landsch, K.; Golan, T.; Kowalczyk, T.; et al. Robust Deep Learning Based Shrimp Counting in an Industrial Farm Setting. *J. Clean. Prod.* **2024**, *468*, 143024. [CrossRef]
225. Cai, Y.; Yao, Z.; Jiang, H.; Qin, W.; Xiao, J.; Huang, X.; Pan, J.; Feng, H. Rapid Detection of Fish with SVC Symptoms Based on Machine Vision Combined with a NAM-YOLO v7 Hybrid Model. *Aquaculture* **2024**, *582*, 740558. [CrossRef]
226. Liu, Y.; Shao, Z.; Teng, Y.; Hoffmann, N. NAM: Normalization-Based Attention Module. *arXiv* **2021**, arXiv:2111.12419. [CrossRef]
227. Jin, Y.; Xiao, X.; Pan, Y.; Zhou, X.; Hu, K.; Wang, H.; Zou, X. A Novel Method for the Object Detection and Weight Prediction of Chinese Softshell Turtles Based on Computer Vision and Deep Learning. *Animals* **2024**, *14*, 1368. [CrossRef]
228. Pachaiyappan, P.; Chidambaram, G.; Jahid, A.; Alsharif, M.H. Enhancing Underwater Object Detection and Classification Using Advanced Imaging Techniques: A Novel Approach with Diffusion Models. *Sustainability* **2024**, *16*, 7488. [CrossRef]
229. Bajpai, A.; Tiwari, N.; Yadav, A.; Chaurasia, D.; Kumar, M. Enhancing Underwater Object Detection: Leveraging YOLOv8m for Improved Subaquatic Monitoring. *SN Comput. Sci.* **2024**, *5*, 793. [CrossRef]
230. Shah, C.; Nabi, M.; Alaba, S.Y.; Ebu, I.A.; Prior, J.; Campbell, M.D.; Caillouet, R.; Grossi, M.D.; Rowell, T.; Wallace, F.; et al. Yolov8-Tf: Transformer-Enhanced Yolov8 for Underwater Fish Species Recognition with Class Imbalance Handling. *Sensors* **2025**, *25*, 1846. [CrossRef]
231. Li, Y.; Hu, Z.; Zhang, Y.; Liu, J.; Tu, W.; Yu, H. DDEYOLOv9: Network for Detecting and Counting Abnormal Fish Behaviors in Complex Water Environments. *Fishes* **2024**, *9*, 242. [CrossRef]
232. Huang, T.-W.; Hwang, J.-N.; Romain, S.; Wallace, F. Fish Tracking and Segmentation from Stereo Videos on the Wild Sea Surface for Electronic Monitoring of Rail Fishing. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 3146–3158. [CrossRef]
233. Li, M.; Li, X.; Chen, S.; Huang, H. A High-Precision and Lightweight Underwater Fish Detection and Recognition Approach Based on the Improved YOLOX-Nano Algorithm. *Aquac. Eng.* **2025**, *110*, 102533. [CrossRef]
234. Wang, D.; Wu, M.; Zhu, X.; Qin, Q.; Wang, S.; Ye, H.; Guo, K.; Wu, C.; Shi, Y. Real-Time Detection and Identification of Fish Skin Health in the Underwater Environment Based on Improved YOLOv10 Model. *Aquac. Rep.* **2025**, *42*, 102723. [CrossRef]
235. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2017; pp. 2980–2988.
236. Cutter, G.; Stierhoff, K.; Zeng, J. Automated Detection of Rockfish in Unconstrained Underwater Videos Using Haar Cascades and a New Image Dataset: Labeled Fishes in the Wild. In *Proceedings of the Applications and Computer Vision Workshops (WACVW), 2015 IEEE Winter*; IEEE: New York, NY, USA, 2015; pp. 57–62.
237. OzFish Dataset—Machine Learning Dataset for Baited Remote Underwater Video Stations. Available online: https://doi.org/10.25845/5e28f062c5097 (accessed on 5 March 2026).
238. Manikandan, D.L.; Santhanam, S.M. Underwater Species Classification Using Deep Learning Technique. *Rev. Română Informatică Autom.* **2024**, *34*, 7–20. [CrossRef]
239. Manikandan, D.L.; Santhanam, S.M. Parallel Desires: Unifying Local and Semantic Feature Representations in Marine Species Images for Classification. *Mar. Geophys. Res.* **2024**, *45*, 16. [CrossRef]
240. Ji, D.; Hussain, A.F.; Hussain, S.; Ogbonnaya, S.G.; Zhu, S.; Wang, X. Fish Detection and Classification Based on Improved ViT. In *Proceedings of the 2023 2nd International Conference on Automation, Robotics and Computer Engineering (ICARCE)*; IEEE: New York, NY, USA, 2023; pp. 1–5.

241. Ulucan, O.; Karakaya, D.; Turkan, M. A Large-Scale Dataset for Fish Segmentation and Classification. In *Proceedings of the 2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*; IEEE: New York, NY, USA, 2020; pp. 1–5.
242. Irfan, M.; Zheng, J.; Iqbal, M.; Arif, M.H. A Novel Feature Extraction Model to Enhance Underwater Image Classification. In *Proceedings of the International Symposium on Intelligent Computing Systems*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 78–91.
243. O'Byrne, M.; Pakrashi, V.; Schoefs, F.; Ghosh, B. Semantic Segmentation of Underwater Imagery Using Deep Networks Trained on Synthetic Imagery. *J. Mar. Sci. Eng.* **2018**, *6*, 93. [CrossRef]
244. Fernandes, A.F.A.; Turra, E.M.; de Alvarenga, É.R.; Passafaro, T.L.; Lopes, F.B.; Alves, G.F.O.; Singh, V.; Rosa, G.J.M. Deep Learning Image Segmentation for Extraction of Fish Body Measurements and Prediction of Body Weight and Carcass Traits in Nile Tilapia. *Comput. Electron. Agric.* **2020**, *170*, 105274. [CrossRef]
245. Liu, F.; Fang, M. Semantic Segmentation of Underwater Images Based on Improved Deeplab. *J. Mar. Sci. Eng.* **2020**, *8*, 188. [CrossRef]
246. Kareem, H.H.; Daway, H.G.; Daway, E.G. Underwater Image Enhancement Using Colour Restoration Based on YCbCr Colour Model. In *Proceedings of the IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2019; Volume 571, p. 012125.
247. Fan, Z.; Xia, W.; Liu, X.; Li, H. Detection and Segmentation of Underwater Objects from Forward-Looking Sonar Based on a Modified Mask RCNN. *Signal Image Video Process.* **2021**, *15*, 1135–1143. [CrossRef]
248. Jahanbakht, M.; Xiang, W.; Waltham, N.J.; Azghadi, M.R. Distributed Deep Learning and Energy-Efficient Real-Time Image Processing at the Edge for Fish Segmentation in Underwater Videos. *IEEE Access* **2022**, *10*, 117796–117807. [CrossRef]
249. Lin, H.-Y.; Tseng, S.-L.; Li, J.-Y. SUR-Net: A Deep Network for Fish Detection and Segmentation with Limited Training Data. *IEEE Sens. J.* **2022**, *22*, 18035–18044. [CrossRef]
250. Møller, T.; Nilssen, I.; Nattkemper, T.W. Tracking Sponge Size and Behaviour with Fixed Underwater Observatories. In *Proceedings of the International Conference on Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 45–54.
251. Conrady, C.R.; Er, Ş.; Attwood, C.G.; Roberson, L.A.; de Vos, L. Automated Detection and Classification of Southern African Roman Seabream Using Mask R-CNN. *Ecol. Inform.* **2022**, *69*, 101593. [CrossRef]
252. Chicchon, M.; Bedon, H.; Del-Blanco, C.R.; Sipiran, I. Semantic Segmentation of Fish and Underwater Environments Using Deep Convolutional Neural Networks and Learned Active Contours. *IEEE Access* **2023**, *11*, 33652–33665. [CrossRef]
253. Yang, G.; Yang, J.; Fan, W.; Yang, D. Neural Network for Underwater Fish Image Segmentation Using an Enhanced Feature Pyramid Convolutional Architecture. *J. Mar. Sci. Eng.* **2025**, *13*, 238. [CrossRef]
254. Kong, J.; Tang, S.; Feng, J.; Mo, L.; Jin, X. AASNet: A Novel Image Instance Segmentation Framework for Fine-Grained Fish Recognition via Linear Correlation Attention and Dynamic Adaptive Focal Loss. *Appl. Sci.* **2025**, *15*, 3986. [CrossRef]
255. Lian, S.; Li, H.; Cong, R.; Li, S.; Zhang, W.; Kwong, S. Watermask: Instance Segmentation for Underwater Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2023; pp. 1305–1315.
256. Lian, S.; Zhang, Z.; Li, H.; Li, W.; Yang, L.T.; Kwong, S.; Cong, R. Diving into Underwater: Segment Anything Model Guided Underwater Salient Instance Segmentation and a Large-Scale Dataset. *arXiv* **2024**, arXiv:2406.06039. [CrossRef]
257. Saleh, A.; Sheaves, M.; Jerry, D.; Azghadi, M.R. Overcoming Annotation Bottlenecks in Underwater Fish Segmentation: A Robust Self-Supervised Learning Approach. *Signal Image Video Process.* **2025**, *19*, 270. [CrossRef]
258. Ditria, E.M.; Connolly, R.M.; Jinks, E.L.; Lopez-Marcano, S. Annotated Video Footage for Automated Identification and Counting of Fish in Unconstrained Seagrass Habitats. *Front. Mar. Sci.* **2021**, *8*, 629485. [CrossRef]
259. Xu, N.; Yang, L.; Fan, Y.; Yue, D.; Liang, Y.; Yang, J.; Huang, T. Youtube-Vos: A Large-Scale Video Object Segmentation Benchmark. *arXiv* **2018**, arXiv:1809.03327.
260. Pavithra, S.; Cicil Melbin Denny, J. An Efficient Approach to Detect and Segment Underwater Images Using Swin Transformer. *Results Eng.* **2024**, *23*, 102460. [CrossRef]
261. Li, D.; Zhao, S.; Hu, J.; Yang, Y.; Ding, J. An Underwater Image Segmentation Model for Complex Scenes in Aquaculture Using Vision Transformer. *Comput. Electron. Agric.* **2025**, *238*, 110764. [CrossRef]
262. Li, H.; Lian, S.; Li, Z.; Cong, R.; Li, C. Taming SAM for Underwater Instance Segmentation and Beyond. *arXiv* **2025**. [CrossRef]
263. Xue, X.; Zhou, Y.; Yan, D.; Tao, L.; Li, J.; Li, Y.; Zhang, H.; Xiao, R. UVLM: Benchmarking Video Language Model for Underwater World Understanding. *arXiv* **2025**, arXiv:2507.02373. [CrossRef]
264. Khan, F.F.; Radwan, Y.; Abdelrahman, E.; Felemban, A.; Mir, A.; Michiels, N.K.; Temple, A.J.; Berumen, M.L.; Elhoseiny, M. FishNet++: Analyzing the Capabilities of Multimodal Large Language Models in Marine Biology. *arXiv* **2025**, arXiv:2509.25564.
265. Li, W.; Zhang, F. Real-Time Vision–Language Analysis for Autonomous Underwater Drones: A Cloud–Edge Framework Using Qwen2. 5-VL. *Drones* **2025**, *9*, 605. [CrossRef]
266. Tian, B.; Zhao, L.; Chen, B.; Zheng, H.; Yang, J.; Wu, M.; Vasisht, D.; Nahrstedt, K. AquaVLM: Improving Underwater Situation Awareness with Mobile Vision Language Models. *arXiv* **2025**, arXiv:2510.21722. [CrossRef]
267. Li, B.; Huo, T.; Zhang, D.; Zhao, Z.; Gao, J.; Li, X. Exploring the Underwater World Segmentation without Extra Training. *arXiv* **2025**, arXiv:2511.07923.

268. Zhu, J.; Yin, S.; Liu, X.; Wang, X.; Yang, Y.-H. FishDetectLLM: Multimodal Instruction Tuning with Large Language Models for Fish Detection. *Knowl.-Based Syst.* **2025**, *318*, 113418. [CrossRef]
269. Han, H.; Wang, W.; Zhang, G.; Li, M.; Wang, Y. Enhancing Vision-Language Models with Morphological and Taxonomic Knowledge: Towards Coral Recognition for Ocean Health. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 28052–28060. [CrossRef]