sensors

Review

Research Challenges, Recent Advances, and Popular Datasets in Deep Learning-Based Underwater Marine Object Detection: A Review

Meng Joo Er *,†, **Jie Chen** †, **Yani Zhang and Wenxiao Gao**

Institute of Artificial Intelligence and Marine Robotics, College of Marine Electrical Engineering, Dalian Maritime University, Dalian 116026, China

* Correspondence: mjer@dlmu.edu.cn

† These authors contributed equally to this work.

Abstract: Underwater marine object detection, as one of the most fundamental techniques in the community of marine science and engineering, has been shown to exhibit tremendous potential for exploring the oceans in recent years. It has been widely applied in practical applications, such as monitoring of underwater ecosystems, exploration of natural resources, management of commercial fisheries, etc. However, due to complexity of the underwater environment, characteristics of marine objects, and limitations imposed by exploration equipment, detection performance in terms of speed, accuracy, and robustness can be dramatically degraded when conventional approaches are used. Deep learning has been found to have significant impact on a variety of applications, including marine engineering. In this context, we offer a review of deep learning-based underwater marine object detection techniques. Underwater object detection can be performed by different sensors, such as acoustic sonar or optical cameras. In this paper, we focus on vision-based object detection due to several significant advantages. To facilitate a thorough understanding of this subject, we organize research challenges of vision-based underwater object detection into four categories: image quality degradation, small object detection, poor generalization, and real-time detection. We review recent advances in underwater marine object detection and highlight advantages and disadvantages of existing solutions for each challenge. In addition, we provide a detailed critical examination of the most extensively used datasets. In addition, we present comparative studies with previous reviews, notably those approaches that leverage artificial intelligence, as well as future trends related to this hot topic.

Keywords: underwater marine object detection; vision; image quality degradation; small object detection; poor generalization; popular datasets

check for updates

Citation: Er, M.J.; Chen, J.; Zhang, Y.; Gao, W. Research Challenges, Recent Advances, and Popular Datasets in Deep Learning-Based Underwater Marine Object Detection: A Review. *Sensors* **2023**, *23*, 1990. https://doi.org/10.3390/s23041990

Academic Editor: Petros Daras

Received: 24 November 2022
Revised: 18 January 2023
Accepted: 20 January 2023
Published: 10 February 2023

1. Introduction

With constant exploration and usage, natural resources on land have been gradually depleted, driving the hunt for new alternatives. The oceans, which cover about 70% of the planet, have become the next destinations for exploration, as they are magnificent treasure troves of precious resources, providing humans with food, medicine, minerals, and other necessities [1–3].

In recent years, the development of marine robots has opened up new opportunities for ocean exploration. When combined with advanced machine vision techniques [4], marine robots have been demonstrated to have significant potential for exploring the underwater environment. In ocean exploration, object detection plays an important role; it is capable of detecting instances of visual objects in digital images, which provide essential information for many downstream tasks. The fundamental question that object detection resolves is: *What objects are where?* [5]. The goal of underwater object detection, as shown in Figure 1, is to predict a bounding box (position) and category label for each object of interest.

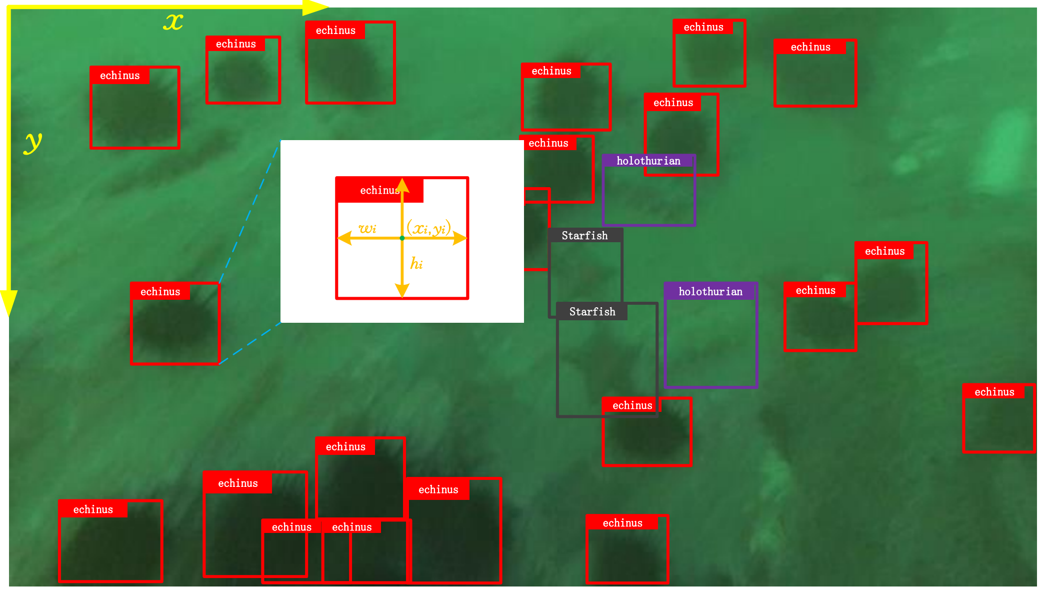


Figure 1. Illustration of vision-based underwater object detection. The detection result is presented as a bounding box with a label, where (x_i, y_i) denotes the ith object's center coordinates and (w_i, h_i) denotes the box's width and height. Here, (x, y) are the axis frames, with the origin in the image's upper left corner (image from DUO dataset [6]).

Equipped with underwater object detection, marine robots have been widely applied in many real-world applications. For example, in monitoring of marine ecosystems, information about species, size, population density, health state, and other characteristics of marine organisms can be gathered by appropriate underwater object detection techniques, which is significant for decision-making [7]. In management of commercial fisheries, it can be applied to extract important information for cultivation, status surveillance, and early warning of diseases [8]. Underwater object detection is an essential technique for robot grabbing tasks, e.g., picking holothurians, echinus, scallops, and other marine products [9]. Furthermore, underwater object detection plays an important role in the operation of self-driving marine robots, supporting the activities of path planning, collision avoidance and control, etc. These pieces of evidence clearly show that underwater object detection plays a critical role in exploration of the ocean.

In recent years, deep learning techniques, which are capable of learning feature representation from data directly, have attracted great attention in many fields, including underwater object detection. However, due to interference from the underwater environment, such as the complex background structures and characteristics of marine objects and the limitations imposed by exploration equipment, underwater object detection is an extremely challenging task. For instance, the attenuation and scattering effects that occur when light passes through the water can result in substantial degradation in the raw pictures collected by marine robots. On the other hand, most underwater objects, especially marine organisms, are usually very small and tend to congregate in dense distributions, making underwater object detection more difficult. Low-resolution underwater images due to limitations imposed by exploration equipment can result in a loss of information about underwater marine objects. These issues impose great challenges when working with conventional underwater object detection techniques.

In this context, we offer a review of deep learning-based underwater marine object detection techniques. To facilitate a thorough understanding of this subject, we first conduct a rigorous analysis of the research challenges involved in detecting marine objects in the underwater environment. Next, we survey detection techniques that are based on deep learning and present a comprehensive taxonomy within the framework of identified research challenges. In addition, we present a comprehensive discussion of popular datasets and provide a prediction of futuristic trends in underwater marine object detection. Through this paper, readers will be able to arrive at a good grasp of the state of the art in underwater object detection and to appreciate the pros and cons of existing techniques when conducting their own research and development work.

The remainder of this paper is organized as follows. Section 2 discusses related preliminary issues and research challenges. In Section 3, typical deep learning-based object detection techniques in the underwater environment are systematically reviewed within the framework of existing research challenges. Popular datasets and a critical analysis thereof are presented in Section 4. A comparison with previous reviews is provided in Section 5, along with a discussion of future directions in underwater marine object detection. Finally, our conclusions are drawn in Section 6.

2. Preliminaries and Research Challenges

2.1. Preliminaries

Two kinds of sensors, namely, sonar and cameras, are commonly used in marine exploration. As is well known, it is often more convenient to perform object detection by relying on sound reflections rather than optical information in underwater scenarios. Both sonars and cameras have distinct advantages and disadvantages. Sonar is an acoustic-based exploration device with a range of hundreds of meters [10]. However, only a little information can be carried by images captured by sonar. As illustrated in Figure 2a, it is only possible to identify a vague sketch of the shipwreck shown in the image. On the other hand, optical images captured by a camera contain a great deal of semantic information, which is beneficial for recognition of target objects. The echinus and starfish in the image can be easily recognized from the texture information, as shown in Figure 2b. However, the exploration range of cameras is severely limited by interference in the underwater environment. Another major benefit of optical cameras is their low cost, which has led to their dominance over other sensors.

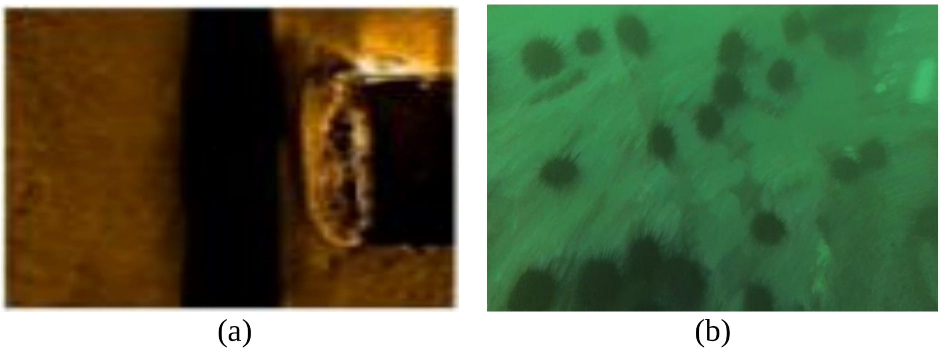


Figure 2. Illustration of (**a**) sonar image and (**b**) camera image (images from [11] and DUO dataset [6]), respectively.

Therefore, in this review we focus on vision-based underwater marine object detection. In fact, in deep learning, object detection based on sonar images and optical images share the same technology stack except for certain specialized preprocessing methods. We do not dwell on the distinction between sonar images and optical images in this review. On the other hand, marine organisms, shipwrecks, other man-made things, and so forth are among the most interesting objects in the underwater object detection task. In this work, we primarily concentrate on the detection of marine organisms because of their enormous economic worth. Next, we formally define underwater marine object detection and the accompanying evaluation matrices.

2.1.1. Definition of Underwater Marine Object Detection

In underwater object detection, it is necessary to recognize all the objects of interest as well as to locate their positions in the image. As illustrated in Figure 1, position information is generally represented by a rectangular bounding box defined by (x_i, y_i, w_i, h_i), where (x_i, y_i) denotes the center coordinates of the *i*th object, which measures the top left corner of image as 0-*indexed*, and (w_i, h_i) denotes the width and height of the box. Formally, the underwater object detection problem can be formulated as follows:

$$X \xrightarrow{f(\theta)} \{(p_i, c_i, x_i, y_i, w_i, h_i) \mid i \in (1, ..., N)\} \tag{1}$$

where $f(\theta)$ indicates an object detector that is based on some neural networks parameterized by θ. It takes as its input an image X, and finally outputs N predictions for objects in that image. Each prediction contains a confidence indicator p_i, the category label c_i that the object belongs to, and the position information encoded in the bounding box, i.e., (x_i, y_i, w_i, h_i). Underwater object detection can provide valuable information for semantic understanding of the underwater environment, and as indicated, it is fundamentally important in the community of marine science and engineering.

2.1.2. Evaluation Metrics

Most of the underwater object detection algorithms are inherited from the community of generic object detection, as are the evaluation metrics. Precision and recall rate are the most widely used evaluation metrics, and are computed based on a confusion matrix, as follows:

$$\begin{aligned} Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \end{aligned} \tag{2}$$

where TP, FP, and FN denote 'True Positive', 'False Positive', and 'False Negative', respectively, and are defined by the Intersection over Union (IoU) between the prediction bounding box and the ground truth. Different thresholds of IoU result in different precision and recall rates.

In the MS COCO dataset [12], three kinds of thresholds are employed. Precision as averaged across all 10 IoU thresholds ($IoU = 0.50{:}0.05{:}0.95$) and all categories of objects is called AP or mAP. Accordingly, we have AP_{50} and AP_{75} when the thresholds are set to $IoU = 0.50$ and $IoU = 0.75$. General speaking, prediction can be evaluated across different scales of objects, such as AP_S $(area < 32^2)$, AP_M $(32^2 < area < 96^2)$, and AP_L $(96^2 < area)$. Recall rates are similarly defined.

2.2. Research Challenges

Conventional object detection methods often fail to provide accurate results due to a number of research challenges faced in the underwater environment [13]. To facilitate a comprehensive understanding of this issue, we organize the research challenges into four main categories.

2.2.1. Image Quality Degradation

Raw underwater images captured in the underwater environment usually suffer from the effect of quality degradation, which is mainly caused by selective absorption and scattering of light in a water body [14].

In the transmission of light in water, red light has a longer wavelength than blue and green light, and as such has faster attenuation. This is a phenomenon known as selective absorption, making most underwater images have a bluish or aqua tone, as illustrated in Figure 3. Because the color is distorted, this phenomenon is referred the color distortion problem, and is well known in underwater imagery analysis.

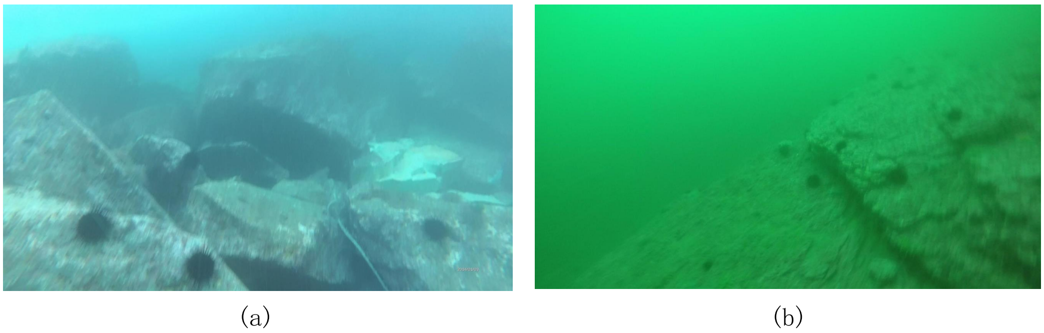


Figure 3. Most underwater images have (**a**) a bluish or (**b**) an aqua tone, which is due to the selective absorption of light in open water (images from DUO dataset [6]).

On the other hand, sediments and particles in water can cause a strong scattering phenomenon, which results in blurry and low-contrast images, as illustrated in Figure 4. To highlight this phenomenon, the image can be converted to grayscale to eliminate other sources of interference. However, even in the absence of color distortion, underwater images suffer from strong haziness. Color distortion and blurring lead to serious quality degradation in underwater images.

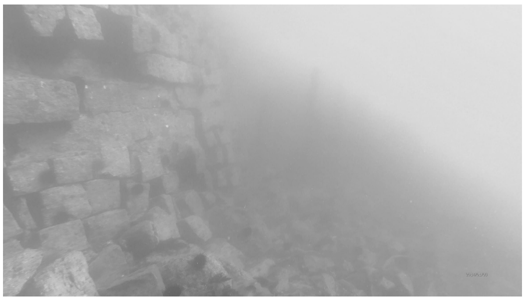


Figure 4. Illustration of blurry image caused by scattering (image from DUO dataset [6]).

2.2.2. Small Object Detection

In underwater marine object detection, most of the objects of interest, such as fish schools [15] and benthic organisms [16], are usually very small, and tend to congregate in dense distributions due to their natural habits, as illustrated in Figure 5. This fact leads to the canonical "small object detection trap", whereby objects only occupy a tiny fraction of an image. In the DUO dataset [6], the vast majority of objects occupy only 0.3% to 1.5% of the image area. Most underwater object detection datasets contain a massive number of small instances.

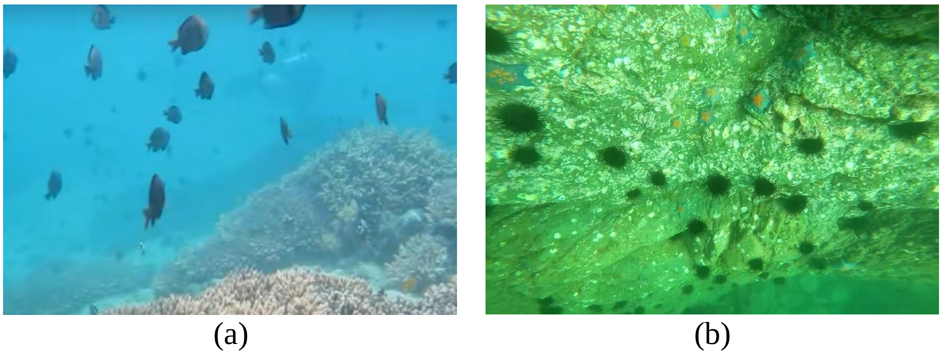


Figure 5. (**a**) Fish school and (**b**) benthic organisms (images from Fish4knowledge dataset [17] and DUO dataset [6]), respectively.

Small object detection has been a long-standing and challenging problem [18,19]. The low resolution of small objects provides little visual information, and it is difficult to extract discriminative features for locating and recognizing these objects. In addition, there is a significant imbalance between positive and negative samples in small object detection, which is not conducive to model training, as because background samples dominate training loss and the gradient updates are too deviated to learn the features of foreground positive samples.

Meanwhile, it should be highlighted that existing neural network architectures and datasets are not optimized specifically for small object detection. In the hierarchical features of Convolutional Neural Networks (CNNs), increases in convolution and downsampling operations lead to semantic information being enhanced while details are lost, which is very unfavorable for small object detection. Furthermore, there is a lack of large-scale datasets for small objects, making this problem even more challenging.

2.2.3. Poor Generalization

Unlike scenarios on land, there is a huge discrepancy in environmental conditions between different oceans. Images from the natural seabed at Zhangzidao and Jinshitan, Dalian, China are collected in [20], and the collected images are visualized in Lab Color Space [21]. Experimental result shows that there is a huge discrepancy in image distribution between these two ocean areas.

It is generally accepted that most algorithms based on deep learning strongly rely on the *i.i.d*. assumption between source and target domains [22]. The performance of algorithms often degrades when this assumption is not fulfilled (sometimes known as domain shift), leading to poor generalization. Hence, generic object detectors suffer from seriously degraded detection accuracy when a detector trained for one ocean region is applied to another. The *i.i.d*. assumption is often violated in practical underwater object detection, creating a major generalization challenge.

2.2.4. Real-Time Detection

Real-time detection is a challenge imposed by the limitations of marine robots. Constrained by the limitations of existing technology, marine robots are usually equipped with embedded platforms that only provide extremely limited computing power. However, deep learning models often require high-performance computing hardware that is not easy to deploy. A standard ResNeXt-50 network has about 25.0×10^6 parameters and 4.2×10^9 FLOPS on eight of NVIDIA M40 GPUs [23]. The consumption of computing resources has become a crucial challenge when applying deep learning-based detection models to an underwater scenario. We refer to this research challenge as "real-time detection", as it measures whether a deep learning algorithm can meet the requirements of resource consumption, which is the most important prerequisite for practical applications.

2.2.5. Other Challenges

Another challenge in underwater object detection is the inter-class similarity between different species or background due to similar appearance and camouflage types. This inter-class similarity makes it highly infeasible to distinguish these objects, and has become a huge headache for researchers. This problem has been noted by the community of computer vision researchers [24]; however, due to its extreme intractability the problem of inter-class similarity in underwater object detection does not allow for any clear solution. Here, we treat this challenge as a problem for future work, and do not discuss it further.

3. Underwater Object Detection Based on Deep Learning

Having presented a systematic discussion of the existing research challenges in underwater marine object detection in the previous section, we now move to a review of the related literature on underwater object detection within the framework of the research challenges identified above. As illustrated in Figure 6, image enhancement is first introduced to deal with the image quality degradation problem, then four techniques are proposed to support small object detection in underwater environments. Next, two different strategies that aim to solve the poor generalization problem are systematically analyzed.

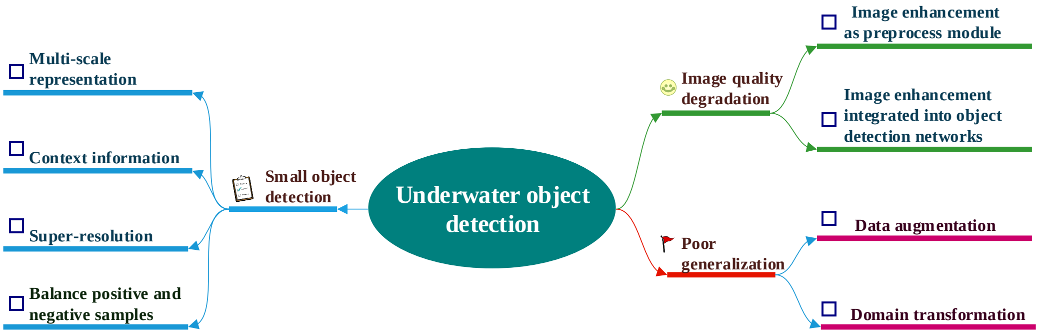

Underwater object detection
Small object detection
Multi-scale representation
Context information
Super-resolution
Balance positive and negative samples
Image quality degradation
Image enhancement as preprocess module
Image enhancement integrated into object detection networks
Poor generalization
Data augmentation
Domain transformation

Figure 6. Framework of underwater object detection according to the main challenges identified in underwater environments.

3.1. Object Detection Coupled with Image Enhancement

Image quality degradation is the main challenge that differentiates underwater object detection from generic object detection on land. To cope with this problem, image enhancement and restoration techniques have been proposed with the aim of correcting color, improving clarity, and addressing blurring and background scattering [25]. In the literature, image enhancement and restoration are used either as a preprocessing module in the underwater object detection pipeline or are integrated into detection networks.

3.1.1. Image Enhancement as Preprocessing Module

In [26], a typical pipeline in which image enhancement works as a preprocessing module in the underwater object detection pipeline is presented for gripping marine products. As illustrated in Figure 7, an effective image enhancement approach based on the Retinex theory is proposed to enhance the images captured by forward and downward views of marine robots. Next, a real-time lightweight object detector based on SSD [27] is proposed for marine object detection. The locations and categories of objects are determined by the detection result. Through enhanced images, underwater robots can examine the seabed environment more precisely.

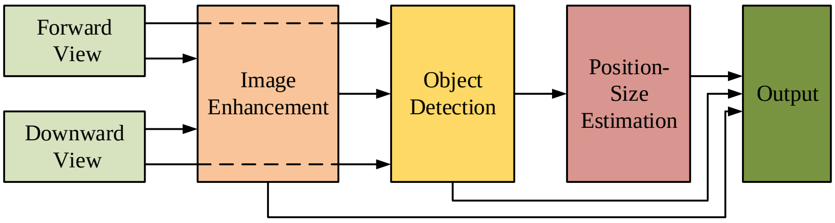


Figure 7. Workflow of the underwater sensing system proposed in [26].

The image enhancement method presented in this pipeline is based on the improved multi-scale Retinex method with color preservation (IMSRCP) [28]. In IMSRCP, color precorrection is first used to balance the color of damaged underwater photos and decrease dominant color. Next, to estimate the reflectance and illumination component, the enhanced multi-scale Retinex is employed in conjunction with intensity computation via channels. Furthermore, the picture is recovered from the logarithmic domain while the compensation dynamics are adjusted. Finally, the color of the original image might be kept depending on the actual application.

Such a pipeline that enhances images before object detection has been a dominant approach in the underwater object detection community for a long time, and research works in this area are ongoing. In [29], the combination of a Max-RGB filter [30] and the Shades of Gray method [31] for improving underwater vision was used to enhance underwater images. Next, a CNN was proposed to handle the weak illumination problem in underwater images. Following image processing, a deep CNN detector was proposed for underwater object detection and classification. Experimental results demonstrated that the improved underwater vision system can assisting a robot in executing underwater missions more efficiently. Similar works on underwater object detection include [32–34].

Underwater image degradation is mainly due to the selective absorption and scattering effect, which can be depicted by a physical image formation model. In [35], a widely used Simplified Model is proposed. As illustrated in Figure 8, natural light penetrates from the air to an underwater scene, and direct transmission reflected off the object can propagate towards the camera. Another portion of the light meets suspended particles, causing much scattering. As such, the radiance perceived by the camera is the sum of two components, that is, the background light formed by multi-scattering and the direct transmission of the reflected light.

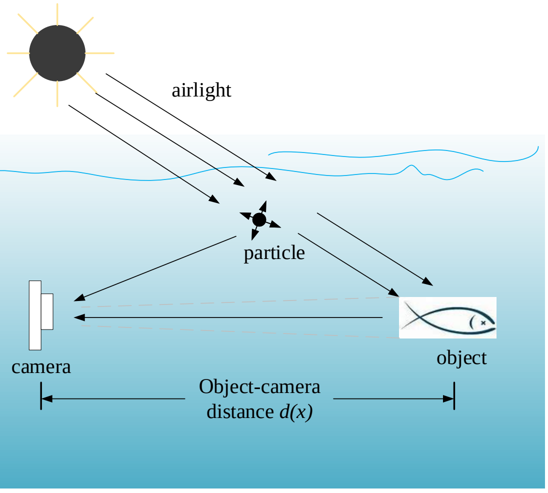


Figure 8. Simplified image formulation model in underwater environment, as proposed in [35].

It is well known that light with different wavelengths are attenuated at different rates in water. Therefore, the image formation model illustrated in Figure 8 can be formulated as follows:

$$\begin{aligned} I_\lambda(x) &= J_\lambda(x) \cdot t_\lambda(x) + (1 - t_\lambda(x)) \cdot B_\lambda, \lambda \in \{r, g, b\}, \\ t_\lambda(x) &= exp(-\beta_\lambda d(x)) \end{aligned} \tag{3}$$

where x denotes a point in the scene, λ denotes the $red/green/blue$ component of the light, $J_\lambda(x)$ is the scene radiance from x, $t_\lambda(x)$ is the transmission of $J_\lambda(x)$ after the light is reflected and has reached the camera, B_λ is the background light, and $I_\lambda(x)$ denotes the captured image. Here, $t_\lambda(x)$ depends on the object–camera distance $d(x)$ and the water attenuation coefficient of each channel β_λ.

Most common scenes of quality degradation in underwater images can be explained based on the image formation model. The different attenuated rates of light describe the selective absorption, which leads to the problem of color distortion. The scattering phenomenon is depicted by the background light component B_λ, which can provide insight for solving this problem.

In [36], Dana et al. proposed an advanced underwater single-image color restoration method based on a physical image formation model. As illustrated in Figure 9, veiling light pixels are first detected by structured edges and used to calculate the background light component B for de-scattering. The object–camera distance is measured by stereo images. Unlike existing works, Dana et al. considered different Jerlov water types in order to estimate transmissions based on a haze-lines model. Finally, color corrections were performed using a physical image formation model and the best result was selected from different water types.

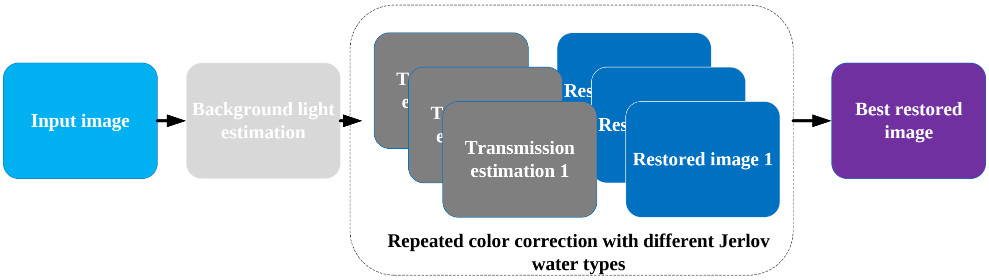

Input image
Background light estimation
T e
T e
Transmission estimation 1
Res
Res
Restored image 1
Repeated color correction with different Jerlov water types
Best restored image

Figure 9. The color restoration and transmission estimation method proposed in [36].

There are other complicated image formation models, such as the revised model proposed in [37] that can depict the process of real-world underwater imaging more precisely by considering the impacts of sensors, ambient illumination, etc. However, due to its complexity, the revised model has received barely any attention, and the simplified model continues to play the vital role in underwater imaging research.

3.1.2. Image Enhancement Integrated into Object Detection Networks

Image enhancement has proven to be helpful for conventional hand-crafted features (e.g., HOG [38], SIFT [39]). How visual enhancement has contributed to deep-learning-based object detection in underwater scenes is discussed in [20]. The authors of [20] showed that enhancement cannot improve within-domain detection accuracy due to the degrading recall efficiency; however, it provides a basic guarantee of generalization in cross-domain scenarios. Many researchers believe that this is caused by the inconsistency resulting from decoupled image enhancement and object detection. As a consequence, another solution for the image quality degradation problem is to integrate image enhancement and object detection into a single model via multi-task learning [40].

In [41], a lightweight deep neural network was proposed for joint learning of color conversion and object detection from underwater images. As illustrated in Figure 10, to alleviate the problem of color distortion, an image color conversion module is first employed to convert color pictures into grayscale images. Then, object detection is performed on the converted images. Color conversion and object detection are learned jointly by a combined loss function:

$$L_{Total} = \lambda_{TV} \times L_{TV} + \lambda_{feature} \times L_{feature} + \lambda_{style} \times L_{style} + \lambda_{detection} \times L_{detection} \quad (4)$$

The color-converted image is measured by three different loss function, namely, the total variation (TV) loss L_{TV} [42], feature reconstruction loss $L_{feature}$ [43], and style reconstruction loss L_{style} [43]. To train the multi-scale object detector, the loss function in YOLO v3 [44] is employed directly, denoted as $L_{detection}$, while λ_{TV}, $\lambda_{feature}$, λ_{style}, and $\lambda_{detection}$ respectively denote the four weight coefficients. With the combined loss function, color conversion is guided in a direction that boosts the performance of object detection.

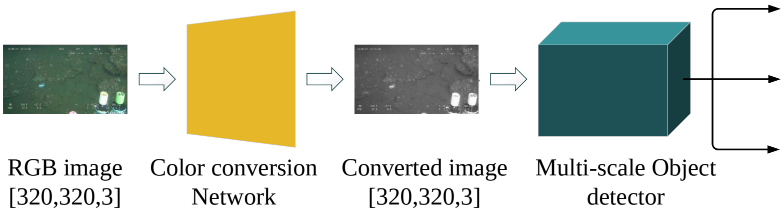


Figure 10. Joint learning of color conversion and object detection in underwater images (figure from [41]).

A comparison is summarized in Table 1. Here, “GFLOPs” denotes gigaflops (floating point operations per second), which reflects the computation complexity of the model, and “Epochs” denotes the number of training epochs. As illustrated in Table 1, experimental results show that the proposed deep model outperforms models without color conversion or with standard color conversion as a preprocessing module when tested on the authors’ custom dataset for object detection. Without interference from color distortion, object detection performance can be improved while maintaining low computational complexity.

Table 1. Comparison of different methods on custom dataset collected in [41]. The numbers in bold indicate the best.

Metric	No Conversion	As Preprocessing	Proposed
mAP	0.878	0.844	**0.896**
GFLOPs	**4.88**	5.02	5.06
Epochs	107	81	**76**

Another similar work is [45], where an end-to-end marine organism detection framework was developed to handle underwater image quality degradation problems caused by noise pollution, color cast, and motion blurring. As illustrated in Figure 11, the framework consists of underwater image enhancement, feature extraction, and back-end detection. Three submodules are employed in the underwater image enhancement module to complete step-by-step denoising, color correction, and deblurring enhancement. Notably, three submodules must be jointly pretrained and optimized with object detection. With image enhancement, the capability of the detector to deal with highly damaged underwater images is significantly promoted. Experimental results show that the proposed framework can improve detection precision by at least 6% compared to existing models for marine organism detection without image enhancement.

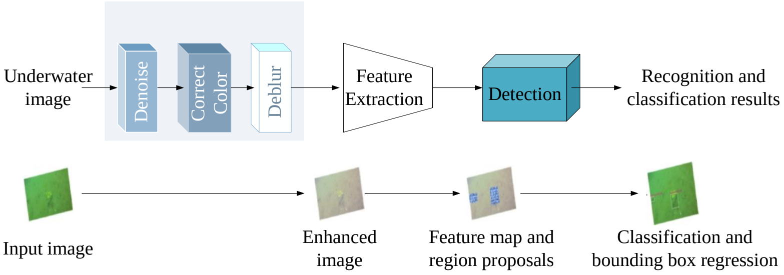

Underwater image
Denoise
Correct Color
Deblur
Feature Extraction
Detection
Recognition and classification results
Input image
Enhanced image
Feature map and region proposals
Classification and bounding box regression

Figure 11. Marine organism detection framework with joint optimization of both image enhancement and object detection (figure from [45]).

Unlike integration of image enhancement using the combined loss function, the Composited FishNet proposed in [46] provides a novel and unified solution for alleviating interference in the underwater environment caused by variance in picture brightness, fish orientation, seabed structure, aquatic plant movement, fish species form, and texture differences.

In Composited FishNet, the authors propose a new composite backbone network, illustrated in Figure 12. The scene change information is encoded by an auxiliary backbone network from the background images without fish; next, the learned background feature is subtracted from the of upper layer feature in the main backbone network. Experimental results show superior performance when the interference is eliminated. Compared with integration through the combined loss function mentioned above, the strategy in Composited FishNet that models the image enhancement process in the neural network is both more elegant and more powerful.

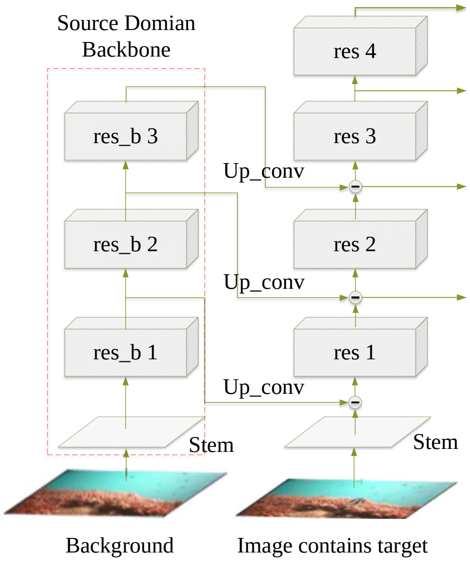


Figure 12. CBresnet backbone network; the interference is encoded and eliminated by the Source Domain Backbone (figure from [46]).

3.1.3. Summary

In this section, we have reviewed two different strategies for solving image quality degradation in underwater object detection. The most prominent characteristic of image enhancement using a preprocessing module is that image enhancement and object detection are decoupled. This decoupling sometimes leads to unexpected results, such that even images with high perceptual quality may not be able to achieve excellent performance in object detection. One possible explanation is that decoupled image enhancement is not desirable for the object detection task. Thus, the second strategy involves joint learning of image enhancement and object detection, which is a more promising technique. In this paradigm, joint learning guides the enhancement module in a direction that can boost detection performance. Research on this topic is in its infant stages, and additional efforts should be directed towards research into joint learning.

3.2. Detection of Small Objects

Small object detection has always been a priority in the marine community due to the characteristics of underwater marine objects, as illustrated in Section 2. Here, we outline the four pillars of small object detection in underwater environments, namely, multi-scale representation, contextual information, super-resolution, and balance of positive and negative examples.

3.2.1. Multi-Scale Representation

In the hierarchical features of CNNs, low-level detailed features are useful for object localization, while semantic information in high-level feature maps is significant for classification of objects. Obviously, most features of a small objects are lost in high-level feature maps, making small object detection extremely difficult. In the literature, a number of researchers have proposed multi-scale feature fusion to generate a more discriminative representation for small object detection.

The first canonical solution for multi-scale representation is the feature pyramid network (FPN) [47]. As illustrated in Figure 13, high-level feature maps are upsampled and added to lower-level ones in order to generate an information-rich representation that is significant for detecting and classifying small objects.

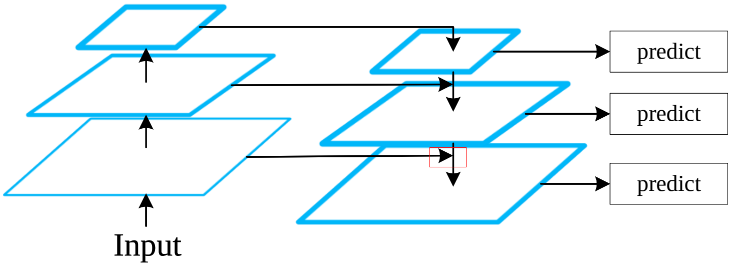


Figure 13. Structure of Feature Pyramid Network (figure from [47]).

The FPN has long been a fundamental block in most deep-learning architectures for underwater object detection. In [9], an efficient detector (termed AquaNet) based on CenterNet [48] was presented, in which multi-scale feature maps from different stages are fused by the FPN to recognize masses of tiny objects from foggy underwater images. In [49], a Multi-scale ResNet (M-ResNet) model based on a modified residual neural network for underwater object detection was proposed. The M-ResNet leverages multi-scale feature fusion by the FPN to allow accurate detection of objects of different sizes, especially small ones. Experimental results show that the proposed approaches always provide improved performance thanks to enhancement of predictive feature representations created by the FPN.

The traditional FPN, on the other hand, is constrained by its one-way top-down information flow. A bottom-up path augmentation can be applied to enrich the entire feature hierarchy in the Path Aggregation Network (PANet) [50] with precise localization signals in lower layers by shortening the information path between the lower layers and the topmost feature. In [50], PANet was integrated with FPN, complementing it in a way that allowed a mixed multi-scale feature structure to be generated, which was then used in an uneaten feed pellet detection model for aquaculture. Experimental results show that the mAP increased by 27.21% compared to the original YOLO-v4 method [51].

To solve the problem of insufficient utilization of semantic information caused by linear upsampling in PANet, an Enhanced Path Aggregation Network (EPANet) was proposed in Composited FishNet [46] for automatic detection and identification of fish from underwater videos. Compared with PANet, the main differences are twofold: (1) a jump connection method is used to merge the backbone network output with the EPANet output information, and (2) the linear interpolation upsampling method is replaced by PixelShuffle [52] to avoid information loss.

Table 2 illustrates a comparison of different feature fusion approaches for underwater object detection on the SeaLife 2017 dataset [46]. Experimental results show that the EPANet exhibits a 1.4% increase in mAP over FPN under the same condition. Compared with PANet, the accuracy is increased by 0.8% due to nonlinear fusion. EPANet outperforms alternatives in terms of the AP_{50} metric, though it suffers from slight degradation in the AP_{75} metric. Comprehensive experiments have demonstrated the effectiveness and efficiency of the EPANet module.

Table 2. Comparison of different feature fusion networks on SeaLife 2017 dataset [46]. The numbers in bold indicate the best.

Methods	AP	AP_{50}	AP_{75}
FPN	0.727	0.921	**0.875**
PANet	0.733	0.917	0.866
EPANet	**0.741**	**0.924**	0.872

Unlike EPANet, the Bidirectional Feature Pyramid Network (BiFPN) [53] is enhanced by removing nodes with only one input edge. These nodes contribute less to the representation that seeks to fuse various features. As illustrated in Figure 14, white nodes represent feature maps from backbone network, arrows represent information flow, and colored nodes represent fused representation. Meanwhile, the BiFPN layers are repeated multiple times to allow additional high-level feature fusion. It is worth mentioning that weighted feature fusion is used in the BiFPN, as feature maps of different resolutions usually contribute to the output representation unequally.

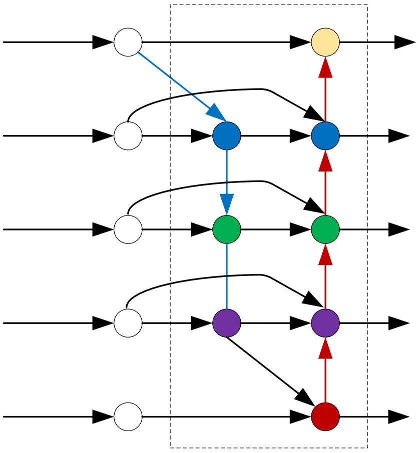


Figure 14. Structure of Bidirectional Feature Pyramid Network (figure from [53]).

Following the improvements in BiFPN, Faster R-CNN was adapted in [54] for underwater organism detection. The authors created a ResNet–BiFPN structure to enhance the capability of feature extraction and multi-scale feature fusion. On the URPC dataset [55], the detection accuracy was 8.26% higher than the original Faster R-CNN. Experimental results showed that BiFPN was able to achieve better accuracy with fewer parameters and FLOPs.

On the other hand, inspired by shortcut connections in residual neural networks [56], a Shortcut Feature Pyramid Network (S-FPN) was proposed in [57] to improve an existing multi-scale feature fusion strategy for holothurian detection. A shortcut connection was attached to the FPN model, with auxiliary information flow merged into conventional multi-scale feature fusion. Shortcut connection enhances feature fusion and reduces information loss in deep networks. Experimental results showed that the mAP of S-FPN reaches 91.5%, outperforming baseline methods on the authors' custom datasets.

Unlike conventional feature fusion, which crosses multiple stages of the backbone, the idea of fusing multi-scale features that cross different channels using different sizes of kernels or complex connections in one block was proposed in [58,59]. This is a method for enhancing the representational learning ability of a neural network by using the idea of multi-scale feature fusion rather than feature fusion across multiple stages of the backbone network.

A Multi-scale Contextual Features Fusion (MFF) block was proposed in AquaNet [9]. In the MFF module, as illustrated in Figure 15, the input is first expanded by 1 × 1 convolution, then split into groups by channel allocation. Each group is convoluted by a kernel, with different kernel sizes used for extraction of multi-scale features. Shortcut connections between different grounds are employed for feature enhancement. For final feature fusion, outputs from all grounds are concatenated and convoluted by 1 × 1 kernels.

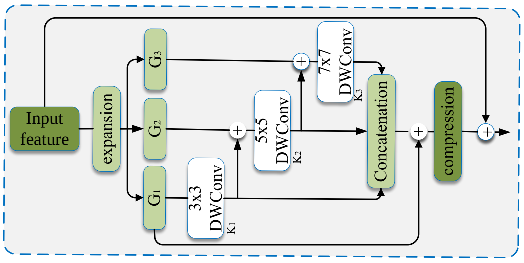


Figure 15. Multi-scale Contextual Features Fusion (MFF) block (figure from [9]).

Most of the aforementioned FPN-based modules and their variants always suffer from a deficiency of context-independent characteristics. There is a substantial risk of information redundancy and negative influence in such feature representations for specific tasks. To alleviate this problem, the attentional mechanism can be integrated into the feature fusion module.

In [60], a Multi-scale Attentional Feature Fusion Module (AFFM) was proposed to fuse semantic and scale-inconsistent features between different convolution layers. As illustrated in Figure 16, feature maps from different stages of the backbone network are first summed and sent into the Multi-Scale Channel Attention Module (MS-CAM). The MS-CAM outputs the attention scores through a *Sigmoid* function. Finally, attention scores are used as references for weighted aggregation of multi-scale features. Experimental results showed that attentional multi-scale feature fusion is superior in underwater object detection.

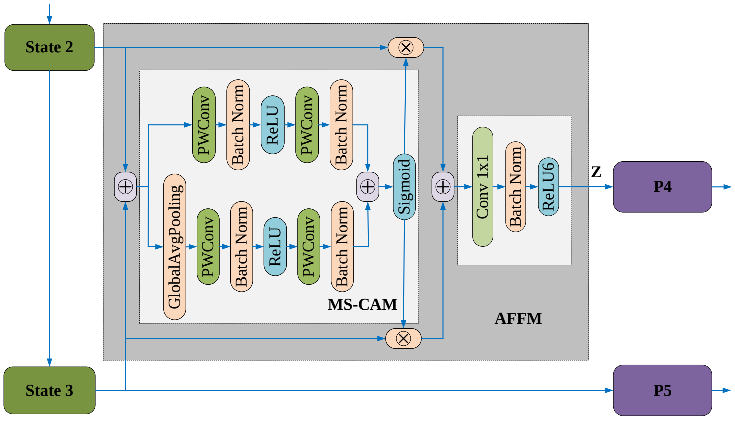


Figure 16. Multi-scale Attentional Feature Fusion Module (figure from [60]).

Similarly, in [61], J. Chen et al. claimed that multi-scale representation should be both spatially aware and scale-aware, and presented a dynamic feature fusion module based on spatial attention and scale attention. This method can fuse different scale feature maps adaptively, thereby enhancing the ability to detect small marine objects by a considerable amount.

From the aforementioned discussion, it is apparent that multi-scale representation for small object detection in underwater environments has been well-studied. However, as an efficient technique of information aggregation, attentional mechanisms have not been studied enough. This creates a need for further investigation of attention-based multi-scale feature fusion. Fully exploiting the potential of attention is expected to effectively promote the development of underwater object detection.

3.2.2. Contextual Information

It is widely acknowledged that lack of visual information severely degrades the performance of small object detection. Many researchers believe contextual information can aid in detecting small objects by leveraging the relationship between objects and their coexisting environment.

(1) Spatial Contextual Information

In the paradigm of anchor-based object detection, the most natural way of extracting contextual information is to enlarge the candidate boxes, which means taking more surrounding environmental information into consideration. In [62], contextual information was extracted from a context region 1.5 times larger than the proposed candidate region. Similarly, in [63], four region with 'foveal' fields of view of $1\times$, $1.5\times$, $2\times$ and $4\times$ of the original proposed box that are all centered on the object of interest were added to the proposed model in order to incorporate contextual information. Experimental results showed that the introduction of spatial contextual information can always result in improved performance.

However, what is noteworthy is that incorporating contextual information for small object detection does not mean taking into account the entire background region, which is redundant and ambiguous, without any restraints. On the contrary, extracting predictive features more precisely is crucial for small object detection.

Deformable convolutional networks are often used to detect marine organisms that are in variable scale forms (e.g., echinus, starfish). In [64], it was employed to improve the capability of the SSD detector in detecting fish, bionics, and other objects in the underwater environment. With precise feature extraction, the detection accuracy and speed in complex underwater environments can be improved significantly.

In [65], two drawbacks that degrade performance of the single-stage detector were identified: first, the detection head regresses coordinates from pre-defined anchors directly, although most anchors are far from matching object regions; second, the information used for classification likely originates from erroneous places, where characteristics may not be precise enough to describe objects. To alleviate these two problems, a novel neural network based on anchor refinement and feature location refinement mechanism was proposed in [65].

As the left image in Figure 17 shows, classic SSD-style detectors rely on hand-crafted anchors that are stiff and invariably inaccurate. The Anchor Refinement Module (ARM) generates refined anchors that provide better initialization for the second-step regression. A location head performs convolution to generate an anchor offset ar using backbone-based ARM features f_{ARM}, i.e.,

$$ar = W_{ar} * f_{ARM}, \tag{5}$$

where the anchor offset ar is the coordinate offset from original anchors, W_{ar} is the learnable convolutional weight, and $*$ denotes convolution. Anchor refinement is urgently needed, because on one hand it greatly relieves the difficulty of localization, while on the other it can guide the refinement of feature location.

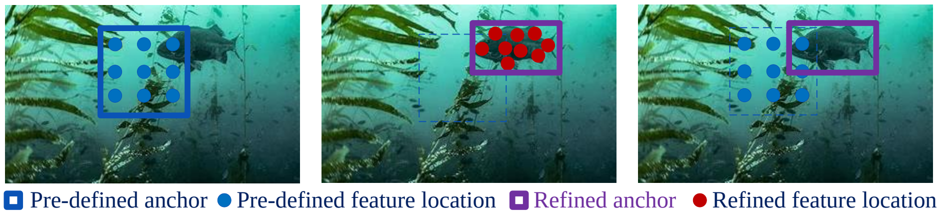


Figure 17. Illustration of Anchor Refinement and Feature Location Refinement mechanisms proposed in [65].

The original *SSD* employs a normal 3×3 grid R (kernel) to estimate the category probability and the coordinates of a feature map cell. Thus, the prediction is provided by

$$P_{p_0} = \sum_{p \in R} w(p) \cdot f_{ODM}(p), \tag{6}$$

where P denotes the prediction of category probability or coordinate offset, w denotes the convolution weight, f_{ODM} represents features from object detection module, p indicates positions in R, and p_0 is the center of the respective field.

It is believed that the respective field of R usually fails to describe the refined anchor region. As such, allowing R to deform to fit various anchor changes, the deformable convolution is developed to capture accurate features with the feature offset δp, as follows:

$$P_{p_0} = \sum_{p \in \mathcal{R}} w(p) \cdot f_{ODM}(p + \delta p) \tag{7}$$

where the feature offsets $\Delta p = \{\delta p\}$ are predicted based on anchor offsets in the Anchor Refinement Module, i.e., feature location refinement is provided by

$$\Delta p = W_{fr} * ar \tag{8}$$

where W_{fr} denotes another learnable convolutional weight and ar is the anchor offset from the ARM module.

Based on the anchor refinement and feature location refinement mechanisms, a Dual Refinement Network (DRNet) has been developed for high-performance detection and applied to online underwater object detection and grasping with an autonomous robotic system.

Spatial context information describes the relationship between objects and their surrounding environment. By leveraging this relationship, objects can usually be recognized more easily, as their camouflage is usually not perfect enough to blend in with the environment. However, the extraction of context information must be carried out cautiously and meticulously, as inappropriate operations are noisy for the detection of underwater marine objects.

(2) Temporal Contextual Information

In addition to spatial contextual information, features from previous frames of underwater video are useful for detecting small objects in the current frame. These features are referred to as temporal context information.

Motion information is an important piece of temporal context information. In [66], motion information was obtained by Gaussian Mixture Modeling (GMM) [67] and optical flow [68] from adjacent frames. As illustrated in Figure 18, by leveraging temporal context information in combination with the original image, a more accurate fish detector based on Faster R-CNN was built for underwater object detection.

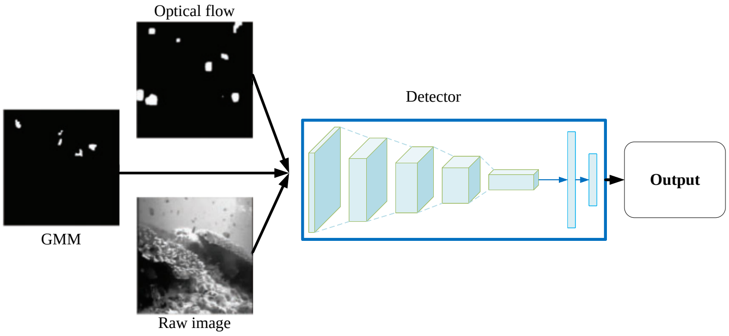


Figure 18. Combining motion information for small underwater object detection (figure from [66]).

The GMM model represents a probability density function $P(x_t)$ of data x at time t as a weighted sum of multiple individual normal distributions $g(x_i)$ for pixel i. The background features can be learned by GMM in a statistical model using the mean and covariance values of the pixel data. Any random and rapid changes in the pixel values of a frame according to the GMM produces a mismatch with the background model of that pixel, in which case motion is presumed to be detected.

The optical flow is represented by a 2D motion vector in the video footage caused by 3D motion of the moving objects. Optical flow is captured by Taylor series approximation with partial derivatives to spatial and temporal coordinates [69]. Based on the optical flow, the proposed method captures a motion between two successive video frames at times t and $t + \delta t$ at every position.

In [70], unlike traditional techniques such as GMM and optical flow, temporal context information is learned by a deep neural network automatically. As illustrated in Figure 19, in order to leverage rich temporal information in videos, each frame of the video images is first processed by a CNN for feature extraction, then these features are fused using the $LH\text{-}LSTM$ module. Through the Conv-LSTM model, temporal contextual information is learned by a neural network automatically and integrated from previous frames for current detection. The ConvLSTM model provides a real-world object detection maneuver for underwater object gripping, providing significant proof of the effectiveness and efficiency of using temporal context information.

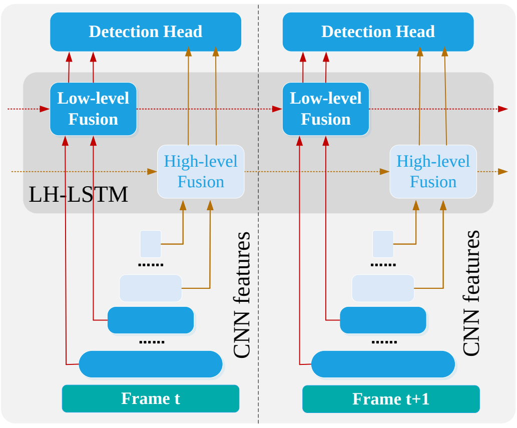


Figure 19. Schematic illustration of TSSD proposed in [70].

In addition to low-level vision features, detection results for the same object in previous frames can provide stronger confidence for detection in future frames. In [71], to preserve rare and endangered objects or eliminate exotic invasive species, a method based on the YOLO framework [72] was proposed for properly classifying objects and counting their numbers in consecutive underwater video images. This approach accumulates object classification results from previous frames to the present frame. The cumulative mean for each object $Avg_i(k)$ can be computed as follows:

$$Avg_i(k) = \frac{(i-1)}{i} Avg_{i-1}(k) + \frac{p_i(k)}{i} \tag{9}$$

where i denotes the number of frames, k denotes a classification object, $p_i(k)$ is the probability of classification for the kth object on the ith frame, and $Avg_i(k)$ denotes the cumulative mean. By applying this heuristic method to YOLO, the proposed method can obtain higher accuracy than the original YOLO detector, which uses only one frame of video images for object detection.

From the aforementioned discussion, it is apparent that both low-level vision information and detection results have been demonstrated to provide effective temporal context information. This is conducive to detection of small underwater objects. However, extraction of temporal context information from previous video frames requires additional computation, resulting in higher computational complexity, which is undesirable in marine robotics.

3.2.3. Super-Resolution

It was demonstrated in [73] that fine details of a small objects for localization can be recovered by Generative Adversarial Networks (GANs). This leads to our discussion of the third pillar of small object detection, i.e., super-resolution, which revolves around converting a low-resolution image with coarse details into a high-resolution image with better visual quality and refined details. In this way, the lack of visual information for small object detection can be circumvented.

In the literature, many deep-learning-based techniques have been employed for super-resolution, including linear networks, residual networks, recursive networks, and more [73]. Linear networks map low-resolution (LR) images to high-resolution (HR) images using a single path for signal flow. The differences between LR and HR images are learned by convolution layers [74]. Super-Resolution Convolutional Neural Network (SRCNN) is one of the outstanding works on super-resolution techniques that rely on pure convolutional layers [75].

In [76], SRCNN is employed to obtain good-quality underwater images in low-light conditions. As illustrated in Figure 20, to obtain LR components, raw data was iteratively processed by Total Variation (TV) regularization [77]. Next, bicubic interpolation, a pre-processing technique for CNNs, was used to increase the LR components. Thereafter, three different convolutional layers were employed to extract features and reconstruct HR patches. All HR patches were merged to generate suitably high HR components, and HR feature vectors were rebuilt. Lastly, the rebuilt pictures were subjected to post-processing, such as noise filtering. Experimental results showed that image reconstruction by the SRCNN method can increase the number of total pixels by about 30%.

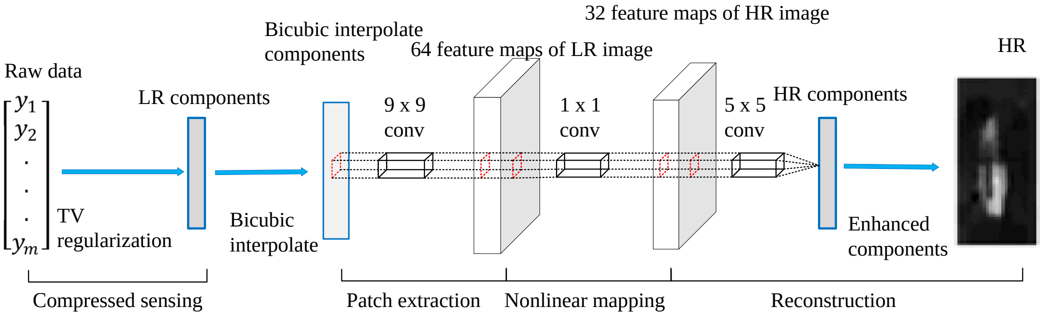


Figure 20. Reconstruction of Underwater Single-Pixel Imaging based on CS-SRCNN (figure from [76]).

Residual networks attach skip connections to the network to learn residues, i.e., high-frequency components between the input and ground truth. Several networks have provided a boost to super-resolution performance using residual learning [78,79]. In [80], a residual-in-residual network-based generative model (Deep SESR) was suggested to handle simultaneous enhancement and super-resolution for underwater robot vision. This model can be trained to restore perceptual picture quality at 2×, 3×, or 4× higher spatial resolution. Residual connection is extensively employed in the feature extraction network of the Deep SESR model, and deconvolution is performed for reconstruction of high-resolution images. Similarly, a deep generative model based on the residual network for single picture super-resolution of underwater photography was presented in [81] that is suitable for object detection in marine robots. Compared with the linear network, the representation capability of residual networks is enhanced significantly by the residual connection, which is desired by super-resolution.

Many other promising techniques have exhibited their effectiveness in super-resolution, though there has been less exploitation in underwater object detection. Recursive networks break down harder super-resolution problems into a set of simpler ones that are easy to solve [82]. Progressive reconstruction design is a promising choice for super-resolution by predicting the output in multiple steps, i.e., 2× followed by 4× and so forth [83]. Super-resolution based on densely connected networks aims at utilizing hierarchical cues available along with the network depth in order to gain more flexibility and richer feature representation [84]. Multi-branch networks in super-resolution aim to obtain a diverse set of features at multiple context scales. Such complementary information can then be fused to obtain a better HR reconstruction, as in [85]. The attention model can boost super-resolution performance by selectively attending to essential features [86].

By recovering fine detail information, super-resolution can help to improve the performance of small object detection in the underwater environment. Due to space constraints, we can only provide a brief introduction to super-resolution techniques here. For more information, please refer to [14].

3.2.4. Balancing of Positive and Negative Examples

There is always an imbalance between positive and negative examples, and small object detection makes this problem more serious. In small object detection, an increase in candidate locations is generated on images, resulting in an extreme class imbalance [19]. In the class imbalance scenario, the easily categorized negative samples overwhelm the training loss; in this case, the direction of gradient updates becomes deviated, resulting in both feature learning of foreground positive samples and the overall training procedure being more difficult. To alleviate this problem, two crucial strategies, sampling and cost-sensitive loss function, have been taken into consideration.

Sampling strategies, accompanied by label assignment procedures, have been developed for a long time to alleviate the class imbalance problem in small object detection. In the training process of Faster R-CNN, the ratio of positive and negative samples in the stage of region proposal is kept at 1:1 by random sampling, and the ratio in the detection head is kept at 1:3 [87]. However, random sampling generally leads to simple samples dominating selected samples. To address this case, an IoU-balanced sampling strategy was proposed in [88] to solve the class imbalance problem more elegantly.

Assume that we need to sample N negative samples from M matching candidates. Under random sampling, the selected probability for each sample is as follows:

$$p = \frac{N}{M} \tag{10}$$

In the IoU-balanced sampling approach of [88], the sample interval is split into K bins according to IoU in order to increase the chosen likelihood of hard negatives, and the N required negative samples are spread evenly throughout the bins. Next, samples are uniformly chosen from the bins. The probability of each sample being selected is computed as follows:

$$p_k = \frac{N}{K} * \frac{1}{M_k}, \quad k \in [0, K) \tag{11}$$

where M_k is the number of sampling candidates in the corresponding interval denoted by k. Experimental results show that IoU-balanced sampling can guide the distribution of training samples to be as near as possible to the distribution of hard negatives, which helps to alleviate the class imbalance challenge. Many other heuristic methods have been employed for sampling [62], although we do not address them further here due to space constraints.

An alternative solution to class imbalance is to design a cost-sensitive loss function with which different penalty factors (weights) for different prediction errors can be used to adjust the balance of classes without increasing computational complexity. The most commonly used method for this solution is Focal Loss (FL) [89], which reshapes cross-entropy loss to a counterpart that downweights the loss assigned to well-classified examples. Specifically, FL is defined as follows:

$$\mathrm{FL}(p_\mathrm{t}) = -\alpha_\mathrm{t}(1 - p_\mathrm{t})^\gamma \log(p_\mathrm{t}) \tag{12}$$

where p_t is the estimated probability, α is a weighting factor to balance the importance of positive/negative examples, and γ indicates a focusing parameter that adjusts the rate at which samples are downweighted smoothly, achieving a balance between hard and easy samples. Due to its superior performance, FL has been used as the de facto loss function in the anchor-free paradigm of object detection [90], and is widely used in the underwater environment [91].

Compared with FL, the Piecewise Focal Loss (PFL) function developed in [57] pays more attention to the difference between the loss weights of hard samples and easy samples. It concentrates more on the training difficulty of hard samples. When the ground truth is 1, those examples with predictive confidence smaller than 0.5 $(p_i < 0.5)$ are referred to as hard samples, as they are hard for the detector to classify. On the contrary, the other samples with predictive confidence greater than 0.5 $(pi \geq 0.5)$ are easy samples. When the ground truth is 0, the interpretations are the other way around. Based on the distinction between hard and easy samples, the PFL is formulated as follows:

$$\begin{aligned} PFL(p_i, p_i^*) &= \begin{cases} -(1-p_i)^{\gamma'} log p_i, if\ p_i^* = 1 \\ -p_i^{\gamma'} log(1-p_i), if\ p_i^* = 0 \end{cases} \\ where\ if\ p_i^* = 1, \gamma' &= \begin{cases} -1/\gamma, if\ p_i < 0.5 \\ \gamma, if\ 0.5 \geq p_i \leq 1 \end{cases} \\ if\ p_i^* = 0, \gamma' &= \begin{cases} \gamma, if\ p_i < 0.5 \\ -1/\gamma, if\ 0.5 \geq p_i \leq 1 \end{cases} \\ p_i^* &= \begin{cases} 0, negative\ label \\ 1, positive\ label \end{cases} \end{aligned} \tag{13}$$

where p_i represents the probability of each candidate box being predicted as an object, p_i^* represents the label of the ground truth, 1 denotes a positive label, and 0 denotes negative label. The term γ' is the adjustment factor that is used to adjust exponential weights of different confidence intervals. This strategy makes the network more inclined to learn from hard samples. Experimental results have demonstrated superior performance of PFL, achieving a mAP of 92.3%, outperforming the baseline cross-entropy of 91.5% and FL of 91.8% on a custom dataset.

There are further studies on cost-sensitive loss functions that have paid attention to different topics, such as hard example mining [92], gradient harmonizing mechanism [93], and long-tail imbalance [94]. However, we mainly focus on sample imbalance between positive and negative examples for small object detection. Due to space constraints, further discussions are omitted.

3.2.5. Summary

In this section, the four pillars of small object detection in underwater environment have been systematically reviewed. Multi-scale representation aims to fuse multi-scale features for discriminative prediction. Attentional multi-scale feature fusion is a research area that is expected to be further explored for the promotion of underwater small object detection. Both spatial contextual information and temporal contextual information are conducive to small object detection by leveraging the relationship between objects and their coexisting environment. However, inappropriate operations on contextual information are be noisy, and result in higher computational complexity. By recovering fine detail information, super-resolution improves the performance of small object detection. Complementing super-resolution with image enhancement is a promising direction for underwater small object detection research. The last pillar, proposed for solving the imbalance between positive and negative samples, involves respective sampling strategies and cost-sensitive loss functions. Both of these can guide the detector to focus on learning features from positive samples, thereby avoiding being overwhelmed by the much larger number of negative samples.

3.3. Generalization in Underwater Object Detection

Generalization is a crucial challenge in underwater object detection and is highly important for real-world applications, as domain shifts can dramatically degrade detection performance. In general, a large-scale dataset with high diversity can address the generalization problem. However, massive and varied underwater scenes make domain-diverse data collection impossible. In this context, two strategies for improving generalization, namely, data augmentation and domain transformation, are discussed.

3.3.1. Data Augmentation

Data augmentation is a basic yet effective technique for improving generalization. It refers to a set of approaches that are capable of increasing the diversity of training datasets [95], such as MixUp [96], Mosaic [51], and PhotoMetricDistortion [97]. These approaches have been fully explored, and have proven to be effective methods of improving generalization in deep learning [98].

In [99], three different data augmentations were employed to train a Faster R-CNN detector for identification and recognition of marine organisms. First, inverse restoration was adopted to mimic various kinds of marine turbulence. Next, perspective transformation was exploited to simulate diverse camera shooting perspectives. Finally, illumination synthesis was employed to simulate varied and uneven lighting situations in underwater environments. Leveraging these three data augmentations, an improvement in both the generalization and robustness of the resulting Faster R-CNN for underwater marine object detection was confirmed by experimental results.

In [100], a detector for detecting marine organisms based only on a tiny underwater dataset with a limited water quality category was proposed. A data augmentation technique based on Water Quality Transfer (WQT) was developed to expand the domain diversity of the initial dataset. WQT is a wavelet-corrected transfer that is based on whitening and coloring transforms. Leveraging WQT, the URPC dataset was transferred to different types of water quality, and promising generalization in underwater object detection was achieved.

In addition to traditional augmentation techniques, further research based on deep learning was conducted in [101,102]. In [101], an approach called Region of Interest Mixed (RoIMix) was proposed to create different training examples by fusing several pictures at the Regions of Interest (RoI) level. RoIMix seeks to mimic overlapping, occlusion, and blurring, allowing the model to implicitly learn to recognize underwater organisms in different conditions.

Let $x \in \mathbf{R}^{H \times W \times C}$ and y represent a proposal and its label. RoIMix seeks to combine two random RoIs, (x_i, y_i) and (x_j, y_j), which when extracted from a batch of images generate virtual proposals $(\tilde{x}, \tilde{y})$:

$$\begin{aligned} \tilde{x} &= \lambda^{'} x_i + (1 - \lambda^{'}) x_j,\ \tilde{y} = y_i, \\ \lambda &= \mathcal{B}(a, a) \\ \lambda^{'} &= max(\lambda, 1 - \lambda) \end{aligned} \tag{14}$$

where $\mathcal{B}(a, a)$ is a Beta distribution with parameter a, as x_i is assigned with a larger coefficient $\lambda^{'}$ and y_i is used to label virtual proposal $\tilde{x}$.

For example, we have x_1 and x_2, which are two RoIs containing a scallop and an echinus, respectively; x_1 is in a blurring environment, while x_3 denotes an occluded sample cropped from a training image, e.g., it represents an echinus lying on a scallop. RoIMix combines x_1 and x_2 to generate a new virtual proposal $\tilde{x}$, similar to x_3, which is used to simulate occlusion and blurring. Experimental results demonstrate that the RoIMix can enhance detection performance on the URPC dataset and Pascal VOC dataset, achieving results 1.18% mAP and 0.8% mAP better, respectively, than the original Faster R-CNN model.

The above proposals can be refined to construct a high-quality training sample for implicitly good generalization. In [102], a proposal-refined object detector was proposed to deal with the detection task in underwater images, which is characterized by insufficient annotation. First, a segmentation network with a poor match for foreground–background segmentation was developed. Next, segmentation was utilized to enhance the region proposals and generate a high-quality training dataset. Leveraging this segmentation, proposal refinement can potentially improve the generalization of object detection.

Traditional augmentation techniques usually work as pre-processing procedures in the detection pipeline, and have become an important cornerstone for superior performance and generalization of underwater object detection. Modern augmentations that leverage deep neural networks represent promising techniques for achieving robust generalization.

3.3.2. Domain Transformation

Domain transformation is an opportunistic strategy for archiving good generalization in underwater object detection by transforming images from distinct domains to middle representations or transforming the trained detector from the source domain to a target domain, as illustrated in Figure 21.

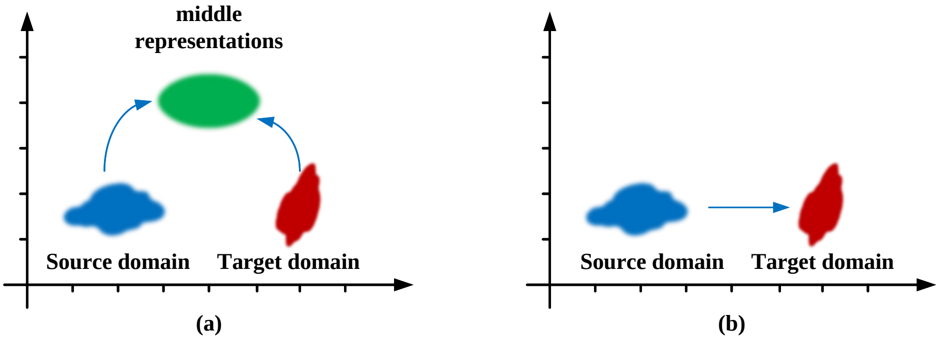


Figure 21. Two kinds of domain transformation for generalization in underwater object detection, where (**a**) transforms distinct domains to middle representations and (**b**) transform source domain to target domain.

In [20], it was shown that distributions of two domains are highly overlapped after the same GAN-based restoration (GAN-RS) is applied to images from different ocean regions. By adopting the GAN-based restoration approach of [21], the problem of domain shift is bypassed subtly. It was shown that detectors developed for one ocean are capable of completing the detection job in another ocean region.

Another experiment using the weak filtering-based restoration approach of [103] was carried out in [20]. Experimental results showed that the generalization problem is alleviated, though the performance does not compare favorably with GAN-based restoration. In [20], it was demonstrated that images from different domains can be transformed into a middle domain representation through enhancement and restoration by the same method; favorable generalization in underwater object detection was achieved. It is apparent that the problem of domain shift can be gradually alleviated by increasing the restoration intensity. As a contribution to domain shift suppression, visual restoration is essential for object detection in underwater environments.

Numerous visual restoration (enhancement) methods have been developed as pre-processing procedures in the automatic pipeline of visually-guided underwater object detection [25]. Visual restoration can provide superior performance for conventional underwater object detection, as richer information can be obtained from enhanced images, and can achieve robust generalization, as it bypasses the problem of domain shift.

Transfer learning is an alternative approach for domain transformation that can achieve good generalization in underwater object detection by storing knowledge gained from the source domain and applying it to a different related target domain. Transfer learning is helpful for handling scenario involving machine learning with a lack of labeled training data.

In [104], transfer learning was employed to alleviate the scarcity of labeled marine data. A marine object detector was pretrained on the ImageNet and MS COCO datasets. In the pre-training technique, the learned knowledge was stored in the parameters of detector and applied to marine object detection, leading to a high-quality detection result with a relatively low number of labeled images. Naturally, transfer learning is a promising technique for addressing the generalization challenge raised by domain shift between ocean regions, as it can transfer the knowledge gained from the source domain to the target domain. It has been proven that transfer learning is extremely effective in underwater object detection [104].

Domain transformation achieves robust generalization in a clever way, helping to achieve significant breakthroughs in underwater marine object detection; as such, further investigation is highly desirable.

3.3.3. Summary

In this section, data augmentation and domain transformation techniques for alleviating the poor generalization challenge raised by domain shift in underwater object detection have been discussed. Both are indispensable for practical applications in real world. In the practical scenario, there is always a lack of labeled data; furthermore, domain shifts occur between the training and test data. Data augmentation can help to increase the diversity of the training data, and modern techniques that leverage deep neural networks have exhibited great potential in such augmentation. Domain transformation helps to bridge the gap in distributions between different domains. Image enhancement and transfer learning are two techniques that can implement domain transformation. Further investigation could help to achieve significant breakthroughs in underwater marine object detection.

3.4. Real-Time Detection

Real-time detection is a crucial indicator that determines whether deep learning models can really work in practical applications. Hence, there is a never-ending quest for excellence in real-time performance of underwater object detection. In this context, we briefly review the evolution of deep leaning-based detectors in underwater object detection and summarize two key techniques for real-time detection. Please see [105] for a more detailed discussion.

3.4.1. Evolution of Deep Learning Detectors for Real-Time Performance

R-CNN (Regions with CNN features) is the first deep-learning-based approach to object detection [106]. It extracts candidate proposals by selective search [107] and employs a CNN to extract features from each proposal. The final classification is performed by an SVM classifier. Fast R-CNN [108] extracts features from the entire image and omits redundant computations caused by selective search. Faster R-CNN proposes a Region Proposal Network (RPN) that predicts candidates directly from the shared feature maps [87]. With these advancements, it can achieve significant increases in detection speed. The R-CNN family has been widely used in underwater object detection for a long time [54,109].

The R-CNN family frames object detection as a "coarse-to-fine" process. First, candidate proposals are extracted, then each proposal is classified. These methods are two-stage algorithms. In the other series of detectors, the YOLO family outputs the extraction of candidate region proposals and predicts detection results from shared feature maps directly [44,72,110]. By discarding the extraction of proposals, the inference time can be reduced to 50 ms, whereas other two-stage competitive models require more than 200 ms. In moving from two stages to one stage, detectors have gained a qualitative leap in real-time performance for underwater object detection [34].

In addition, real-time detection has undergone a paradigm shift from anchor-based to anchor-free. The speed of anchor-free detectors is substantially improved over one-stage detectors while maintaining superior detection accuracy, as the expensive operations associated with anchor mechanisms have been eliminated [9]. The de facto method for real-time detection is now anchor-free. As can be seen from the preceding discussion, removing the complex anchor module with its high degree of computational complexity significantly speeds up detection.

3.4.2. Lightweight Network Design

In addition to the evolution of detectors, there are other amazing efforts devoted to real-time detection. One of these is lightweight network design, which aims to develop effective low-complexity network architectures.

Convolution with a 1×1 kernel is the first significant step, and can minimize computational complexity by reducing the number of feature channels. It is extensively utilized in GoogLeNet [111], SqueezeNet [112], etc. The literature shows that 1×1 convolution is a powerful tool for reducing the parameters of deep learning models, resulting in higher detection speeds while maintaining detection accuracy and avoiding degradation [111,112].

Depth-wise separable convolution factorizes a standard convolution into a depth-wise convolution and a 1×1 point-wise convolution, saving a significant amount of operations and parameters while only slightly diminishing accuracy [113]. ShuffleNet proposes point-wise group convolution and channel shuffling, which can drastically reduce computational costs [114].

Next, 1×1 convolution and depth-wise separable convolution are extensively employed to construct lightweight modules, such as the receptive module in [41]. Low-rank approximation is the underlying theory behind lightweight network design. This theory can be used to develop lightweight network architectures by employing different effective and efficient strategies, and is unquestionably crucial in achieving real-time detection of underwater objects.

3.4.3. Model Compression

Model compression aims to remove redundant parameters (or neurons) in pre-trained neural networks. Parameter redundancy has been proven in the literature, and serves as the theoretical foundation of model compression [115].

Various techniques have been developed for model compression. Network pruning removes neurons (or parameters) based on their importance, which is measured by the number of times it is not zero on a given dataset, over its lifetime, or based on another measure [116]. Knowledge distillation uses a small student model to replicate the actions of a large teacher model. It is possible to preserve a large model's superior performance while shrinking the model's size and resource usage through knowledge distillation [117]. Parameter quantization aims to reorganize the parameters of neural networks with fewer bits. For example, replacing the 32-bit floating parameters with 16-bit ones is the simplest method. The most commonly used technique for quantization is parameter clustering [118].

Model compression is a thriving research area, and in recent years a plethora of compression strategies have emerged in an effort to achieve an acceptable trade-off between processing speed and application accuracy. Model compression has even influenced the design of neural network accelerators, resulting in the achievement of extreme performance improvements [119].

3.4.4. Summary

In this section, we have reviewed the evolution of underwater object detectors, which is a process pursuing real-time performance. Two key techniques dedicated to real-time detection, namely, lightweight network design and model compression, have been discussed. Lightweight network design aims to create elegant lightweight network architectures, while model compression aims to reduce the redundancy in network parameters. Both are important for real-time detection. In order to create a more elegant model, iteratively applying lightweight network design and model compression is recommended, as these approaches are complementary.

4. Popular Datasets

In Section 3, we reviewed the related literature on underwater object detection in light of identified research challenges. In this section, we analyze and discuss the most commonly used datasets for underwater object detection.

4.1. Fish4Knowledge

The Fish4Knowledge database was the first well-known popular dataset [17]. It was developed as part of a project conducted at the University of Edinburgh to study marine ecosystems. This dataset enables comparisons and analysis of fish patterns between different water areas [120]. An underwater dataset for target detection against complex backgrounds is one of the ground truth datasets in the Fish4Knowledge. Figure 22 shows example images from the Fish4Knowledge dataset.

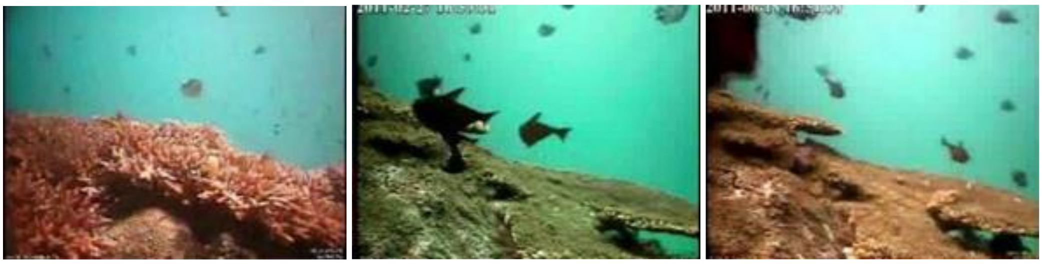


Figure 22. Example images from the Fish4Knowledge dataset (images from [17]).

Annotation: the Fish4Knowledge dataset contains 23 species of fish, and has a total of 27,370 images along with corresponding mask images. All images are annotated manually by marine biologists. Images of different species of fish are stored in different folders. Annotations of bounding boxes for object detection task need to be generated from their mask images.

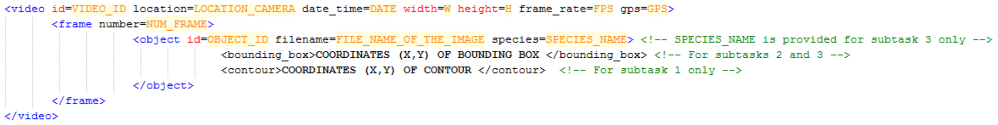
However, the Fish4Knowledge dataset is not well-suited for fish detection, and researchers need to generate annotations from mask images and create custom training/test sets by themselves in order to meet their special research requirements. In addition, the Fish4Knowledge datavase has a serious class imbalance problem, with certain fish categories dominating their classes.

4.2. LifeCLEF 2014

The LifeCLEF2014 dataset was built based on the Fish4Knowledge project, and aims to identify marine organisms in visual data for ecological surveillance and biodiversity monitoring [121]. There are four separate subtasks for fish identification and species recognition in LifeCLEF2014.

For the video-based fish identification subtask, four videos with 21,106 annotations corresponding to 9852 individual fish were fully labeled to form the training dataset, with the aim being to identify moving objects in videos. For the image-based fish identification subtask, the training dataset contains 957 videos with 112,078 labeled fish, with the aim being to detect fish instances in video frames. In the image-based fish identification and species recognition subtask, the training dataset includes 285 videos with 19,868 fishes and the species are annotated, with the aim being to identify species of fish detected in video frames. Image-based fish species recognition is the last challenge, for which the training dataset contains 19,868 annotated fish images and the aim is to identify fish species using only still images containing one fish instance. Matching test datasets are supplied for each subtask.

Annotation: the ground truth annotations of the first three subtasks are provided as XML files which contain species, bounding box coordinates, and contour coordinates, as illustrated in Figure 23.



```
<video id=VIDEO_ID location=LOCATION_CAMERA date_time=DATE width=W height=H frame_rate=FPS gps=GPS>
       <frame number=NUM_FRAME>
                  <object id=OBJECT_ID filename=FILE_NAME_OF_THE_IMAGE species=SPECIES_NAME> <!-- SPECIES_NAME is provided for subtask 3 only -->
                             <bounding_box>COORDINATES (X,Y) OF BOUNDING BOX </bounding_box> <!-- For subtasks 2 and 3 -->
                             <contour>COORDINATES (X,Y) OF CONTOUR </contour>  <!-- For subtask 1 only -->
                  </object>
       </frame>
</video>
```

Figure 23. Ground truth annotations of LifeCLEF2014 dataset (image from ImageCLEF website).

LifeCLEF2014 was significantly extended over the next few years, allowing it to be used with various species of marine organisms as well as in different application tasks. For more information, please refer to [122–124]

4.3. URPC

The National Natural Science Foundation of China's Underwater Robot Professional Contest (URPC) offers a well-known dataset for underwater robot object grabbing [55]. Beginning in 2017, the URPC organizer has published a new dataset each year. The images in the URPC databases depict marine organisms such as holothurians, echinus, scallops, and starfish on an open-sea farm. The URPC datasets are widely used in the community of marine science and engineers working on underwater object detection. Figure 24 provides examples of images from the URPC dataset. Table 3 contains information on diffent URPC datasets over the years. Here, 'Class' means the categories of objects that need to be detected, while 'Train' and 'Test' mean the number of images in the training/test sets, respectively.

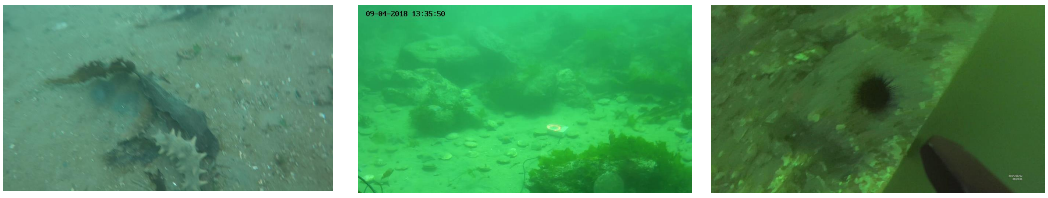


Figure 24. Example images from the URPC dataset.

Table 3. Information on URPC datasets over years.

Dataset	Train	Test	Class	Year
URPC2017	17,655	985	3	2017
URPC2018	2901	800	4	2018
URPC2019	4757	1029	4	2019
URPC2020_ZJ	5543	2000	4	2020
URPC2020_DL	6575	2400	4	2020

Annotation: the annotations of URPC dataset are provided in XML files, with one file for each image that contains the label and coordinates of the bounding box for each object in the image.

The URPC dataset does not provide annotation files for testing, and cannot be downloaded after the contest ends. Researchers must split the original training data to create a custom training set and test set. This custom split makes the comparison with state-of-the-art methods impossible. In this context, the URPC datasets are unsuitable for use as benchmark datasets.

4.4. UDD and DUO

To overcome the shortcomings of the URPC dataset, Dalian University of Technology (DUT) has presented a new dataset (UDD) comprising 2227 photos in three categories (holothurian, echinus, and scallop) to improve the object grasping capabilities of underwater robots for open-sea farming [9]. This dataset comprises 1827 training images and

400 testing images with a maximum resolution of 3840 × 2160 pixels. The UDD is a high-quality and small-scale dataset; hence, it is not a good choice as a benchmark for deep learning-based underwater object detection.

The DUT further gathers and re-annotates the URPC datasets as well as the UDD dataset, yielding 7782 images. Finally, [6] constructed a new Detecting Underwater Objects (DUO) dataset comprising a diversity of underwater scenes and more reasonable annotations. In the DUO dataset, 6671 photos are used for training and 1111 for testing.

Annotation: the annotations of the UDD and DUO datasets are prepared in MS COCO format and provided in a JSON file containing the category label, bounding box, segmentation, and other information.

However, the UDD and DUO datasets, similar to the URPC dataset, were generated for the task of grabbing marine organisms using robots. These datasets only have three/four class objects, respectively, and are therefore unsuitable for generic underwater object detection tasks.

4.5. Summary

In this section, we have reviewed several popular datasets for underwater marine object detection. At present, there is no universally accepted benchmark dataset, as each of the aforementioned datasets has its own deficiencies. Several of them, such as Fish4Knowledge, are poorly structured, and researchers need to build their own training and testing sets by custom split, making comparison with state-of-the-art methods infeasible [99]. The others are small in scale, which is undesirable for deep learning. These problems have impeded the progress of underwater object detection research. In fact, the lack of annotated data is a major challenge in underwater object detection, prompting the need for further investigation into the construction of a universally accepted benchmark dataset.

5. Discussions and Analysis

In this section, we primarily compare our work with previous reviews and discuss which directions in underwater object detection research are important in the future.

5.1. Comparison with Previous Reviews

Notable surveys on underwater object detection are summarized in Table 4. The first survey is focused on monitoring of underwater ecosystems [125]. In ecosystem protection research, massive images are frequently collected, prompting a need for automatic object detection and classification. In [125], related works that employ deep learning for underwater imagery analysis are systematically described and categorized according to the category of objects, including fish, plankton, coral, and sea grass. However, a taxonomy based on categorization of objects is not suitable, making high-level understanding of this research topic nearly impossible. On the other hand, the challenges around underwater object detection were barely mentioned in [125].

The second survey reviewed underwater target recognition methods based on deep learning [126]. In [126], the focus was on dangerous underwater target recognition, with methods divided into three categories, namely, target recognition based on shape features, unsupervised recognition techniques, and deep learning methods based on CNNs. The pros and cons of various methods, as well as the challenges in few-shot target recognition and target recognition under environmental interference, were discussed in [126]. Finally, different algorithms were compared on the UDD dataset. The authors concluded that the poor generalization performance of algorithms is always a major problem in this field. However, although a variety of topics in underwater object detection were covered in [126], there was no mention of marine organisms, and a reasonable taxonomy was lacking. The authors briefly identified two challenges in underwater object detection; however, they did not provide a comprehensive analysis.

Table 4. Summary of related surveys on underwater object detection.

No.	Survey title	Content	Deficiency	Year	References
1	Deep Learning in Underwater Marine Object Detection: A Survey	Summarized deep learning object detection and classification methods with respect to the categories of objects, including fish, plankton, coral, and sea grass.	Unsuitable taxonomy; the main challenges of underwater object detection were not discussed.	2017	[125]
2	Underwater Target Recognition Methods Based on the Framework of Deep Learning: A Survey	Described the application of deep learning in underwater dangerous target and man-made object recognition. Discussed two main challenges around few-shot learning and environmental interference.	Does not contain marine organisms; used unsuitable taxonomy; analysis of challenges not comprehensive.	2020	[126]
3	Robust Underwater Object Detection with Autonomous Underwater Vehicles: A Comprehensive Study	Briefly discussed conventional methods and deep learning methods for underwater object detection with AUVs.	Challenges in underwater object detection were not discussed; taxonomy simple and crude; similarities and differences between two categories of methods not described.	2020	[13]

In another survey of underwater object detection [13], several deep learning approaches and conventional object detection methods that can be deployed on autonomous underwater vehicles were reviewed and a comparison between these methods was carried out. Unfortunately, this survey used a coarse taxonomy, and different research works were only elaborated in the framework of conventional methods and deep learning methods. Moreover, the outstanding challenges in underwater object detection were not presented, and the pros and cons of different methods were not discussed.

The main differences between our paper and existing surveys are highlighted below:

1. **Systematic analysis of research challenges in underwater object detection:** this paper extensively reviews the related literature on underwater object detection. The main challenges around object detection in underwater environments are identified and analyzed systematically. Four challenges, namely, image quality degradation, small object detection, poor generalization, and real-time detection, are discussed at length.
2. **Comprehensive review in light of identified challenges:** an excellent literature review should provide information on the research topic, the challenges researchers are confronted with, and the different solutions that have been proposed. An in-depth exploration of crucial technologies and state-of-the-art methods under the framework of four research challenges makes this survey paper valuable and useful in understanding of underwater marine object detection.
3. **Deeper insight into futuristic trends:** a discussion of how the research topic in question is expected to evolve is an important part of any survey. Based on the state-of-the-art methods in underwater object detection and consideration of next-generation artificial intelligence, this paper provides deeper insight into future trends in underwater object detection.

5.2. Future Trends in Underwater Object Detection

Based on the aforementioned discussion, we believe that future directions in underwater marine object detection must include the following aspects:

(1) A well-constructed benchmark dataset. The success of deep-learning-based underwater object detection is dependent on the use of a well-constructed benchmark dataset. For any research topic, a good benchmark dataset should be large-scale, diversified, class-balanced, and well-constructed [12,127,128]. The Fish4Knowledge is a large-scale dataset that includes 23 different types of fish species. However, the number of images in each category varies greatly, resulting in a serious class imbalance problem. Although the Life-

CLEf datasets are relatively good, they only include fish and ignore other important marine organisms. To the best of our knowledge, no widely accepted benchmark is available at present. Creating a large-scale and well-constructed benchmark dataset is an important topic for future research.

(2) Inter-class similarities in underwater object detection. In order to avoid being harmed by predators and improve their survival ability, many marine organisms are very good at camouflage. This presents a challenge in terms of high inter-class similarity, as marine organisms have great similarity with the background or other species. This phenomenon makes it difficult even for humans to recognize marine organisms from their environment, and poses a great challenge for machine learning algorithms. There is no doubt that accounting for subtle differences between objects and their environments can improve detection performance significantly. While it is known that deep learning is relatively robust against inter-class similarities, research works in the area of inter-class similarity are rare [24]. Addressing this problem is an important challenge in moving underwater marine object detection forward.

(3) Underwater object detection based on transformers. One of the biggest surprises in artificial intelligence research in recent years is the development of transformers. As a promising fundamental infrastructure, transformers have dominated research on object detection during the 2020s thanks to competitive performance and tremendous potential [129]. However, the most famous vanilla transformers are used with large models with high computational complexity, and require very long training times to converge. Due to these intrinsic limitations, transformers are much slower than traditional convolution with the same FLOPs (at least 3× slower) [130]. In addition, transformer suffer from a relatively low performance on small object detection tasks due to their non-local sensitivity [131]. Considering the constraints in underwater environments, an efficient transformer designed specifically for small objects could be of great significance for underwater marine object detection.

(4) Multi-modal data fusion. Most research studies on underwater object detection rely solely on optical or acoustic sensors. However, it is well-known that each of these sensors has its own distinct characteristics and limitations. Optical imaging provides high imaging resolution and rich information, as well as being more intuitive for humans. However, the sensing range and image quality are severely influenced by absorption and scattering effects. Sonar imaging based on acoustic waves, on the other hand, has the advantages of low loss and long propagation distance. The active [132,133] and passive [134] acoustic approaches continue to play an important role in underwater exploration. Hence, complementary employment of these two kinds of methods represents a way to greatly improve object detection performance. Multi-modal data fusion is highly desirable in underwater exploration.

(5) Multi-task learning. Rather than extending existing models to new tasks, today's AI models are typically trained to perform only one task. However, the human brain does not work in this manner. A variety of capabilities that can be accessed as needed and combined to perform new and/or more complex tasks is highly desirable [135]. Due to the challenges of image quality degradation, new image enhancement and/or restoration techniques are always beneficial. Therefore, engaging both underwater image enhancement and object detection in a well-designed multi-task learning framework is a promising direction. Multi-task learning can improve data efficiency as well, as this paradigm does not require learning every task from scratch.

(6) Sparse model with efficient computing. It is widely accepted that the success of deep learning is due to the unprecedented availability of big data and concurrent advancements in computing power [136]. In the traditional dense model, the entire neural network is activated to complete a task regardless of how simple or complex it is [137], which is inefficient and goes against the way the human brain works. Creating a sparse model in which only small segments of the network are activated as needed can, in theory, result in a higher capability for different tasks, and is more energy efficient as well. Sparse models with efficient computing can undoubtedly make multi-task learning more feasible, and represent a key technological infrastructure for the next generation of ocean exploration techniques.

6. Conclusions

Deep learning-based underwater object detection has received significant attention in the community of marine science and engineering researchers thanks to its superior performance. It has tremendous potential to support a wide range of marine activities. This paper presents a comprehensive and critical review of deep learning-based underwater object detection techniques. Four research challenges in the underwater environment are identified in this paper, namely, image quality degradation, small object detection, poor generalization, and real-time detection. In light of these identified challenges, a comprehensive analysis is presented to provide a thorough understanding of the subject matter. Finally, popular datasets and future directions in underwater object detection are discussed. We hope that readers find this survey informative and useful in helping them to understand the recent advances in underwater object detection, and that it can guide future research in this exciting area.

Author Contributions: Writing—original draft preparation, J.C.; writing—review and editing, M.J.E., J.C., Y.Z. and W.G.; supervision, M.J.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fundamental Research Funds for the Central Universities under Grant 3132019344 and the Leading Scholar Grant, Dalian Maritime University, under Grant 00253007.

Acknowledgments: The authors would like to acknowledge the support of the Fundamental Research Funds for the Central Universities under Grant 3132019344 and the Leading Scholar Grant, Dalian Maritime University, under Grant 00253007.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Costello, C.; Cao, L.; Gelcich, S.; Cisneros Mata, M.Á.; Free, C.M.; Froehlich, H.E.; Golden, C.D.; Ishimura, G.; Maier, J.; Macadam Somer, I.; et al. The future of food from the sea. *Nature* **2020**, *588*, 95–100. [CrossRef] [PubMed]
2. Santos, J.D.; Vitorino, I.; Reyes, F.; Vicente, F.; Lage, O.M. From ocean to medicine: Pharmaceutical applications of metabolites from marine bacteria. *Antibiotics* **2020**, *9*, 455. [CrossRef] [PubMed]
3. Norman, T.; Pedro, R.; Ricardo, I.J. Seabed mineral resources, an alternative for the future of renewable energy: A critical review. *Ore Geol. Rev.* **2020**, *126*, 103699.
4. Sahoo, A.; Dwivedy, S.K.; Robi, P. Advancements in the field of autonomous underwater vehicle. *Ocean Eng.* **2019**, *181*, 145–160. [CrossRef]
5. Zou, Z.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *arXiv* **2019**, arXiv:1905.05055.
6. Liu, C.; Li, H.; Wang, S.; Zhu, M.; Wang, D.; Fan, X.; Wang, Z. A Dataset And Benchmark Of Underwater Object Detection For Robot Picking. In Proceedings of the IEEE International Conference on Multimedia & Expo Workshops, Shenzhen, China, 5–9 July 2021; pp. 1–6.
7. Ditria, E.M.; Lopez Marcano, S.; Sievers, M.; Jinks, E.L.; Brown, C.J.; Connolly, R.M. Automating the analysis of fish abundance using object detection: Optimizing animal ecology with deep learning. *Front. Mar. Sci.* **2020**, *7*, 429. [CrossRef]
8. Li, J.; Xu, C.; Jiang, L.; Xiao, Y.; Deng, L.; Han, Z. Detection and analysis of behavior trajectory for sea cucumbers based on deep learning. *IEEE Access* **2019**, *8*, 18832–18840. [CrossRef]
9. Liu, C.; Wang, Z.; Wang, S.; Tang, T.; Tao, Y.; Yang, C.; Li, H.; Liu, X.; Fan, X. A New Dataset, Poisson GAN and AquaNet for Underwater Object Grabbing. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 2831–2844. [CrossRef]
10. Hodges, R.P. *Underwater Acoustics: Analysis, Design and Performance of Sonar*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
11. Yu, Y.; Zhao, J.; Gong, Q.; Huang, C.; Zheng, G.; Ma, J. Real-Time Underwater Maritime Object Detection in Side-Scan Sonar Images Based on Transformer-YOLOv5. *Remote Sens.* **2021**, *13*, 3555. [CrossRef]
12. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
13. Gomes, D.; Saif, A.S.; Nandi, D. Robust Underwater Object Detection with Autonomous Underwater Vehicle: A Comprehensive Study. In Proceedings of the International Conference on Computing Advancements, Dhaka, Bangladesh, 10–12 January 2020; pp. 1–10.

14. Anwar, S.; Li, C. Diving deeper into underwater image enhancement: A survey. *Signal Process. Image Commun.* **2020**, *89*, 115978. [CrossRef]
15. Yang, L.; Liu, Y.; Yu, H.; Fang, X.; Song, L.; Li, D.; Chen, Y. Computer vision models in intelligent aquaculture with emphasis on fish detection and behavior analysis: A review. *Arch. Comput. Methods Eng.* **2021**, *28*, 2785–2816. [CrossRef]
16. Han, F.; Yao, J.; Zhu, H.; Wang, C. Marine organism detection and classification from underwater vision based on the deep CNN method. *Math. Probl. Eng.* **2020**, *2020*, 3937580. [CrossRef]
17. Fisher, R.B.; Chen Burger, Y.H.; Giordano, D.; Hardman, L.; Lin, F.P. *Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 104.
18. Tong, K.; Wu, Y.; Zhou, F. Recent advances in small object detection based on deep learning: A review. *Image Vis. Comput.* **2020**, *97*, 103910. [CrossRef]
19. Liu, Y.; Sun, P.; Wergeles, N.; Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert Syst. Appl.* **2021**, *172*, 114602. [CrossRef]
20. Chen, X.; Lu, Y.; Wu, Z.; Yu, J.; Wen, L. Reveal of domain effect: How visual restoration contributes to object detection in aquatic scenes. *arXiv* **2020**, arXiv:2003.01913.
21. Chen, X.; Yu, J.; Kong, S.; Wu, Z.; Fang, X.; Wen, L. Towards real-time advancement of underwater visual quality with GAN. *IEEE Trans. Ind. Electron.* **2019**, *66*, 9350–9359. [CrossRef]
22. Li, H.; Pan, S.J.; Wang, S.; Kot, A.C. Domain generalization with adversarial feature learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5400–5409.
23. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
24. Sun, G.; Cholakkal, H.; Khan, S.; Khan, F.; Shao, L. Fine-Grained Recognition: Accounting for Subtle Differences between Similar Classes. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12047–12054.
25. Wang, Y.; Song, W.; Fortino, G.; Qi, L.; Zhang, W.; Liotta, A. An experimental-based review of image enhancement and image restoration methods for underwater imaging. *IEEE Access* **2019**, *7*, 140233–140251. [CrossRef]
26. Wang, Y.; Tang, C.; Cai, M.; Yin, J.; Wang, S.; Cheng, L.; Wang, R.; Tan, M. Real-time underwater onboard vision sensing system for robotic gripping. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 1–11. [CrossRef]
27. 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, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
28. Tang, C.; von Lukas, U.F.; Vahl, M.; Wang, S.; Wang, Y.; Tan, M. Efficient underwater image and video enhancement based on Retinex. *Signal Image Video Process.* **2019**, *13*, 1011–1018. [CrossRef]
29. Han, F.; Yao, J.; Zhu, H.; Wang, C. Underwater image processing and object detection based on deep CNN method. *J. Sens.* **2020**, *2020*, 6707328. [CrossRef]
30. Land, E.H. The retinex theory of color vision. *Sci. Am.* **1977**, *237*, 108–129. [CrossRef] [PubMed]
31. Finlayson, G.D.; Trezzi, E. Shades of gray and colour constancy. In Proceedings of the Color and Imaging Conference. Society for Imaging Science and Technology, Scottsdale, AZ, USA, 9–12 November 2004; Volume 2004, pp. 37–41.
32. Madhan, E.; Kannan, K.; Rani, P.S.; Rani, J.V.; Anguraj, D.K. A distributed submerged object detection and classification enhancement with deep learning. *Distrib. Parallel Databases* **2021**, 1–17.
33. Hu, K.; Lu, F.; Lu, M.; Deng, Z.; Liu, Y. A Marine Object Detection Algorithm Based on SSD and Feature Enhancement. *Complexity* **2020**, *2020*, 5476142. [CrossRef]
34. Yang, H.; Liu, P.; Hu, Y.; Fu, J. Research on underwater object recognition based on YOLOv3. *Microsyst. Technol.* **2021**, *27*, 1837–1844. [CrossRef]
35. Chiang, J.Y.; Chen, Y.C. Underwater image enhancement by wavelength compensation and dehazing. *IEEE Trans. Image Process.* **2011**, *21*, 1756–1769. [CrossRef] [PubMed]
36. 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.* **2021**, *43*, 2822–2837. [CrossRef]
37. Akkaynak, D.; Treibitz, T. A revised underwater image formation model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6723–6732.
38. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
39. Ng, P.C.; Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **2003**, *31*, 3812–3814. [CrossRef]
40. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* **2017**, arXiv:1706.05098.
41. Yeh, C.H.; Lin, C.H.; Kang, L.W.; Huang, C.H.; Lin, M.H.; Chang, C.Y.; Wang, C.C. Lightweight Deep Neural Network for Joint Learning of Underwater Object Detection and Color Conversion. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6129–6143. [CrossRef]
42. Mahendran, A.; Vedaldi, A. Understanding Deep Image Representations by Inverting Them. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5188–5196.

43. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-time Style Transfer and Super-resolution. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 694–711.
44. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
45. Zhang, X.; Fang, X.; Pan, M.; Yuan, L.; Zhang, Y.; Yuan, M.; Lv, S.; Yu, H. A Marine Organism Detection Framework Based on the Joint Optimization of Image Enhancement and Object Detection. *Sensors* **2021**, *21*, 7205. [CrossRef] [PubMed]
46. Zhao, Z.; Liu, Y.; Sun, X.; Liu, J.; Yang, X.; Zhou, C. Composited FishNet: Fish Detection and Species Recognition From Low-Quality Underwater Videos. *IEEE Trans. Image Process.* **2021**, *30*, 4719–4734. [CrossRef]
47. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
48. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
49. Pan, T.S.; Huang, H.C.; Lee, J.C.; Chen, C.H. Multi-scale ResNet for real-time underwater object detection. *Signal Image Video Process.* **2020**, *15*, 941–949. [CrossRef]
50. Hu, X.; Liu, Y.; Zhao, Z.; Liu, J.; Yang, X.; Sun, C.; Chen, S.; Li, B.; Zhou, C. Real-time detection of uneaten feed pellets in underwater images for aquaculture using an improved YOLO-V4 network. *Comput. Electron. Agric.* **2021**, *185*, 106135. [CrossRef]
51. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
52. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
53. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10781–10790.
54. Shi, P.; Xu, X.; Ni, J.; Xin, Y.; Huang, W.; Han, S. Underwater Biological Detection Algorithm Based on Improved Faster-RCNN. *Water* **2021**, *13*, 2420. [CrossRef]
55. Chen, L.; Liu, Z.; Tong, L.; Jiang, Z.; Wang, S.; Dong, J.; Zhou, H. Underwater object detection using Invert Multi-Class Adaboost with deep learning. In Proceedings of the IEEE International Joint Conference on Neural Networks, Glasgow, UK, 19–24 July 2020; pp. 1–8.
56. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
57. Peng, F.; Miao, Z.; Li, F.; Li, Z. S-FPN: A shortcut feature pyramid network for sea cucumber detection in underwater images. *Expert Syst. Appl.* **2021**, *182*, 115306. [CrossRef]
58. Gao, S.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P.H. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–662. [CrossRef]
59. Tan, M.; Le, Q.V. Mixconv: Mixed depthwise convolutional kernels. *arXiv* **2019**, arXiv:1907.09595.
60. 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]
61. Chen, J.; Er, M.J.; Zhang, Y.; Gao, W.; Wu, J. Novel Dynamic Feature Fusion Stragegy for Detection of Small Underwater Marine Object. In Proceedings of the 5th International Conference on Intelligent Autonomous Systems (ICoIAS), Dalian, China, 23–25 September 2022; pp. 24–30.
62. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A unified multi-scale deep convolutional neural network for fast object detection. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 354–370.
63. Zagoruyko, S.; Lerer, A.; Lin, T.Y.; Pinheiro, P.O.; Gross, S.; Chintala, S.; Dollár, P. A multipath network for object detection. *arXiv* **2016**, arXiv:1604.02135.
64. Qiang, W.; He, Y.; Guo, Y.; Li, B.; He, L. Exploring underwater target detection algorithm based on improved SSD. *Xibei Gongye Daxue Xuebao/J. Northwestern Polytech. Univ.* **2020**, *38*, 747–754. [CrossRef]
65. Chen, X.; Yu, J.; Kong, S.; Wu, Z.; Wen, L. Joint Anchor-Feature Refinement for Real-Time Accurate Object Detection in Images and Videos. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 594–607. [CrossRef]
66. Salman, A.; Siddiqui, S.A.; Shafait, F.; Mian, A.; Shortis, M.R.; Khurshid, K.; Ulges, A.; Schwanecke, U. Automatic fish detection in underwater videos by a deep neural network-based hybrid motion learning system. *ICES J. Mar. Sci.* **2020**, *77*, 1295–1307. [CrossRef]
67. Stauffer, C.; Grimson, W.E.L. Adaptive background mixture models for real-time tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Fort Collins, CO, USA, 23–25 June 1999; pp. 246–252.
68. Warren, D.H.; Strelow, E.R. *Electronic Spatial Sensing for the Blind: Contributions from Perception, Rehabilitation, and Computer Vision*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; Volume 99.
69. Beauchemin, S.S.; Barron, J.L. The computation of optical flow. *ACM Comput. Surv.* **1995**, *27*, 433–466. [CrossRef]
70. Chen, X.; Yu, J.; Wu, Z. Temporally identity-aware SSD with attentional LSTM. *IEEE Trans. Cybern.* **2019**, *50*, 2674–2686. [CrossRef]
71. Park, J.H.; Kang, C. A Study on Enhancement of Fish Recognition Using Cumulative Mean of YOLO Network in Underwater Video Images. *J. Mar. Sci. Eng.* **2020**, *8*, 952. [CrossRef]
72. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.

73. Yang, W.; Zhang, X.; Tian, Y.; Wang, W.; Xue, J.H.; Liao, Q. Deep learning for single image super-resolution: A brief review. *IEEE Trans. Multimed.* **2019**, *21*, 3106–3121. [CrossRef]
74. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
75. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [CrossRef]
76. Li, M.; Mathai, A.; Lau, S.L.; Yam, J.W.; Xu, X.; Wang, X. Underwater Object Detection and Reconstruction Based on Active Single-Pixel Imaging and Super-Resolution Convolutional Neural Network. *Sensors* **2021**, *21*, 313. [CrossRef]
77. Strong, D.; Chan, T. Edge-preserving and scale-dependent properties of total variation regularization. *Inverse Probl.* **2003**, *19*, S165–S187. [CrossRef]
78. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
79. Ahn, N.; Kang, B.; Sohn, K.A. Fast, accurate, and lightweight super-resolution with cascading residual network. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 252–268.
80. Islam, M.J.; Luo, P.; Sattar, J. Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception. *arXiv* **2020**, arXiv:2002.01155.
81. Islam, M.J.; Enan, S.S.; Luo, P.; Sattar, J. Underwater image super-resolution using deep residual multipliers. In Proceedings of the IEEE International Conference on Robotics and Automation, Paris, France, 31 May–31 August 2020; pp. 900–906.
82. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1637–1645.
83. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 624–632.
84. Tong, T.; Li, G.; Liu, X.; Gao, Q. Image super-resolution using dense skip connections. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4799–4807.
85. Ren, H.; El Khamy, M.; Lee, J. Image super resolution based on fusing multiple convolution neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 54–61.
86. Choi, J.S.; Kim, M. A deep convolutional neural network with selection units for super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 154–160.
87. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
88. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra r-cnn: Towards balanced learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 821–830.
89. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
90. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9627–9636.
91. Cao, S.; Zhao, D.; Sun, Y.; Liu, X.; Ruan, C. Automatic coarse-to-fine joint detection and segmentation of underwater non-structural live crabs for precise feeding. *Comput. Electron. Agric.* **2021**, *180*, 105905. [CrossRef]
92. Wang, X.; Zhang, S.; Wang, S.; Fu, T.; Shi, H.; Mei, T. Mis-classified vector guided softmax loss for face recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12241–12248.
93. Li, B.; Liu, Y.; Wang, X. Gradient harmonized single-stage detector. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8577–8584.
94. Wang, J.; Zhang, W.; Zang, Y.; Cao, Y.; Pang, J.; Gong, T.; Chen, K.; Liu, Z.; Loy, C.C.; Lin, D. Seesaw loss for long-tailed instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 9695–9704.
95. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
96. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
97. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.
98. Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X.D. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef]
99. Huang, H.; Zhou, H.; Yang, X.; Zhang, L.; Qi, L.; Zang, A.Y. Faster R-CNN for marine organisms detection and recognition using data augmentation. *Neurocomputing* **2019**, *337*, 372–384. [CrossRef]
100. Liu, H.; Song, P.; Ding, R. Towards domain generalization in underwater object detection. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 1971–1975.

101. Lin, W.H.; Zhong, J.X.; Liu, S.; Li, T.; Li, G. Roimix: Proposal-fusion among multiple images for underwater object detection. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 2588–2592.
102. Lv, X.; Wang, A.; Liu, Q.; Sun, J.; Zhang, S. Proposal-refined weakly supervised object detection in underwater images. In Proceedings of the International Conference on Image and Graphics, Beijing, China, 23–25 August 2019; pp. 418–428.
103. Chen, X.; Wu, Z.; Yu, J.; Wen, L. A real-time and unsupervised advancement scheme for underwater machine vision. In Proceedings of the IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems, Sheraton Princess Kaiulani, HI, USA, 31 July–4 August 2017; pp. 271–276.
104. Levy, D.; 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/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1466–14668.
105. Er, M.J.; Chen, J.; Zhang, Y. Marine Robotics 4.0: Present and Future of Real-Time Detection Techniques for Underwater Objects. In *Industry 4.0—Perspectives and Applications*; Gordan, D.M., Ghaedi, D.K., Saleh, D.V., Eds.; IntechOpen: Rijeka, Croatia, 2022; Chapter 8.
106. 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, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
107. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
108. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
109. 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 IEEE Conference on OCEANS, Washington, DC, USA, 19–22 October 2015; pp. 1–5.
110. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
111. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
112. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
113. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
114. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
115. Cheng, Y.; Yu, F.X.; Feris, R.S.; Kumar, S.; Choudhary, A.; Chang, S. An exploration of parameter redundancy in deep networks with circulant projections. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2857–2865.
116. LeCun, Y.; Denker, J.; Solla, S. Optimal brain damage. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 27–30 November 1989; Volume 2.
117. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
118. Wu, J.; Leng, C.; Wang, Y.; Hu, Q.; Cheng, J. Quantized convolutional neural networks for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4820–4828.
119. Deng, L.; Li, G.; Han, S.; Shi, L.; Xie, Y. Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey. *Proc. IEEE* **2020**, *108*, 485–532. [CrossRef]
120. Boom, B.J.; He, J.; Palazzo, S.; Huang, P.X.; Beyan, C.; Chou, H.M.; Lin, F.P.; Spampinato, C.; Fisher, R.B. A research tool for long-term and continuous analysis of fish assemblage in coral-reefs using underwater camera footage. *Ecol. Inform.* **2014**, *23*, 83–97. [CrossRef]
121. Joly, A.; Goëau, H.; Glotin, H.; Spampinato, C.; Bonnet, P.; Vellinga, W.P.; Planque, R.; Rauber, A.; Fisher, R.; Müller, H. Lifeclef 2014: Multimedia life species identification challenges. In Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Sheffield, UK, 15–18 September 2014; pp. 229–249.
122. Joly, A.; Goëau, H.; Glotin, H.; Spampinato, C.; Bonnet, P.; Vellinga, W.P.; Planqué, R.; Rauber, A.; Palazzo, S.; Fisher, B.; et al. LifeCLEF 2015: Multimedia life species identification challenges. In Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Toulouse, France, 8–11 September 2015; pp. 462–483.
123. Joly, A.; Goëau, H.; Glotin, H.; Spampinato, C.; Bonnet, P.; Vellinga, W.P.; Champ, J.; Planqué, R.; Palazzo, S.; Müller, H. LifeCLEF 2016: Multimedia life species identification challenges. In Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Évora, Portugal, 5–8 September 2016; pp. 286–310.

124. Joly, A.; Goëau, H.; Glotin, H.; Spampinato, C.; Bonnet, P.; Vellinga, W.P.; Lombardo, J.C.; Planqué, R.; Palazzo, S.; Müller, H. Lifeclef 2017 lab overview: Multimedia species identification challenges. In Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Dublin, Ireland, 11–14 September 2017; pp. 255–274.
125. 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, Antwerp, Belgium, 18–21 September 2017; pp. 150–160.
126. Teng, B.; Zhao, H. Underwater target recognition methods based on the framework of deep learning: A survey. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 1729881420976307. [CrossRef]
127. Vicente, S.; Carreira, J.; Agapito, L.; Batista, J. Reconstructing pascal voc. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 41–48.
128. Pang, Y.; Cao, J.; Li, Y.; Xie, J.; Sun, H.; Gong, J. TJU-DHD: A diverse high-resolution dataset for object detection. *IEEE Trans. Image Process.* **2020**, *30*, 207–219. [CrossRef]
129. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef]
130. Hu, H.; Zhang, Z.; Xie, Z.; Lin, S. Local Relation Networks for Image Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 3463–3472.
131. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
132. Testolin, A.; Kipnis, D.; Diamant, R. Detecting Submerged Objects Using Active Acoustics and Deep Neural Networks: A Test Case for Pelagic Fish. *IEEE Trans. Mob. Comput.* **2022**, *21*, 2776–2788. [CrossRef]
133. Testolin, A.; Diamant, R. Combining Denoising Autoencoders and Dynamic Programming for Acoustic Detection and Tracking of Underwater Moving Targets. *Sensors* **2020**, *20*, 2945. [CrossRef] [PubMed]
134. Haesang, Y.; Keunhwa, L.; Youngmin, C.; Kookhyun, K. Underwater Acoustic Research Trends with Machine Learning: Passive SONAR Applications. *J. Ocean. Eng. Technol.* **2020**, *34*, 227–236.
135. Zhang, Y.; Yang, Q. An overview of multi-task learning. *Natl. Sci. Rev.* **2018**, *5*, 30–43. [CrossRef]
136. Shrestha, A.; Mahmood, A. Review of deep learning algorithms and architectures. *IEEE Access* **2019**, *7*, 53040–53065. [CrossRef]
137. Lennie, P. The cost of cortical computation. *Curr. Biol.* **2003**, *13*, 493–497. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.