electronics

Article

Underwater Fish Object Detection with Degraded Prior Knowledge

Shijian Zheng [1,2,*], Rujing Wang [1,3] and Liusan Wang [1,*]

[1] Intelligent Agriculture Engineering Laboratory of Anhui Province, Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230000, China; rjwang@iim.ac.cn
[2] Department of Electronic Engineering, School of Information Science and Engineering, Southwest University of Science and Technology, Mianyang 621010, China
[3] Department of Automation, University of Science and Technology of China, Hefei 230026, China
* Correspondence: zhengshijian_swust@126.com (S.Z.); lswang@iim.ac.cn (L.W.)

Abstract: Understanding fish distribution, behavior, and abundance is crucial for marine ecological research, fishery management, and environmental monitoring. However, the distinctive features of the underwater environment, including low visibility, light attenuation, water turbidity, and strong currents, significantly impact the quality of data gathered by underwater imaging systems, posing considerable challenges in accurately detecting fish objects. To address this challenge, our study proposes an innovative fish detection network based on prior knowledge of image degradation. In our research process, we first delved into the intrinsic relationship between visual image quality restoration and detection outcomes, elucidating the obstacles the underwater environment poses to object detection. Subsequently, we constructed a dataset optimized for object detection using image quality evaluation metrics. Building upon this foundation, we designed a fish object detection network that integrates a prompt-based degradation feature learning module and a two-stage training scheme, effectively incorporating prior knowledge of image degradation. To validate the efficacy of our approach, we develop a multi-scene Underwater Fish image Dataset (UFD2022). The experimental results demonstrate significant improvements of 2.4% and 2.5%, respectively, in the mAP index compared to the baseline methods ResNet50 and ResNetXT101. This outcome robustly confirms the effectiveness and superiority of our process in addressing the challenge of fish object detection in underwater environments.

Keywords: underwater fish image; object detection; degraded prior knowledge; convolutional neural network; image enhancement

Citation: Zheng, S.; Wang, R.; Wang, L. Underwater Fish Object Detection with Degraded Prior Knowledge. *Electronics* **2024**, *13*, 2346. https://doi.org/10.3390/electronics13122346

Academic Editor: George A. Tsihrintzis

Received: 23 April 2024
Revised: 10 June 2024
Accepted: 10 June 2024
Published: 15 June 2024

1. Introduction

In recent years, underwater environments have become increasingly important areas of study due to their ecological significance, commercial potential, and strategic importance. Among the various tasks associated with underwater exploration, the detection and tracking of marine organisms, mainly fish, have garnered significant attention. Understanding fish species' distribution, behavior, and abundance is essential for various applications such as aquatic ecology research, fisheries management, environmental monitoring, and underwater robotics. However, detecting fish objects in underwater environments poses significant challenges compared to similar tasks in terrestrial or aerial settings. The underwater domain introduces unique complexities, including poor visibility, light attenuation, turbidity, and distortion caused by waves and water currents [1]. These factors severely degrade the quality of sensory data collected by underwater imaging systems, making it difficult to detect and classify fish objects [2] accurately.

Traditional approaches to underwater fish detection often rely on manually engineered features and shallow learning algorithms, which need help to cope with underwater imagery's inherent variability and complexity. Moreover, these methods typically assume that

the input data are of high quality, ignoring the pervasive effects of degradation and noise commonly encountered in real-world underwater scenarios. Recent research has focused on leveraging advanced deep-learning techniques to address these challenges and improve underwater fish detection performance. Deep neural networks have shown remarkable capabilities in learning complex patterns and representations directly from raw data, enabling them to adapt to underwater imagery's degraded and noisy nature. One approach involves the development of complex depth detectors [3], wherein researchers endeavor to create intricate, deep architectures to enhance image quality or extract features beneficial for underwater object detection (UOD) tasks. An intuitive strategy involves cascading an underwater image enhancement (UIE) network with a general depth detector [4,5]. However, despite producing visually appealing outputs, the UIE module's results do not consistently translate to high accuracy for the depth detector. Another strategy entails amassing a large corpus of training examples [6]. Nonetheless, such efforts are hindered by poor detection performance due to environmental degradation obscuring many crucial scene features. Additionally, some studies leverage Generative Adversarial Networks (GANs) [7] to synthesize underwater images from natural counterparts. While these comprehensive methods augment the training sample size and introduce guidance from natural images, their efficacy for real-world UOD remains questionable, given the substantial disparities between underwater and natural objects/scenes.

While the role of visual restoration in enhancing traditional features is established [8], the relationship between image quality and convolutional representation remains elusive. Take underwater scenes as an example (depicted in Figure 1): they exhibit various degradation issues and diverse styles. By employing two restoration techniques, Learnable Full-frequency Transformer Dual Generative Adversarial Network (LFT-DGAN) [9], and Divide-and-Conquer network (DC-net) [10], higher-quality images can be generated. Each column represents the same scene but with varying detection outcomes using the cascaded RCNN detector, suggesting a potential link between underwater image restoration and object detection. It prompts the question: how does visual restoration contribute to object detection in underwater environments? To address this inquiry, this paper initially delves into the interplay between visually restored image quality and detection efficacy, shedding light on the domain's influence on object detection and constructing a dataset favorable for object detection by utilizing image quality assessment metrics. Subsequently, it introduces an underwater fish object detection method leveraging prior knowledge of degradation. Specifically, the algorithm comprises two core components: a prompt-based degradation feature learning module and a two-stage training regimen. The prompt-based degradation feature learning module is built upon the amplitude-frequency feature relationship among images, facilitating comprehensive learning and transfer of image feature representations conducive to object detection. The two-stage training regimen enhances the detector's suitability for underwater fish detection tasks by adapting the training protocol. Finally, the construction of the UFD2022 multi-category underwater fish dataset validates the proposed method's efficacy. Overall, the main contributions of this study are as follows:

- It investigated the relationship between visually restored image quality and detection performance, highlighting the impact of a domain on object detection efficacy. It also constructed a dataset favorable for object detection by utilizing image quality assessment metrics.
- It introduces a novel underwater fish object detection method that utilizes prior knowledge of degradation, consisting of a prompt-based degradation feature learning module and a two-stage training scheme.
- A multi-scene Underwater Fish Image Dataset (UFD2022) was compiled, comprising 2800 images. This dataset offers a more comprehensive and rich resource for related research endeavors.

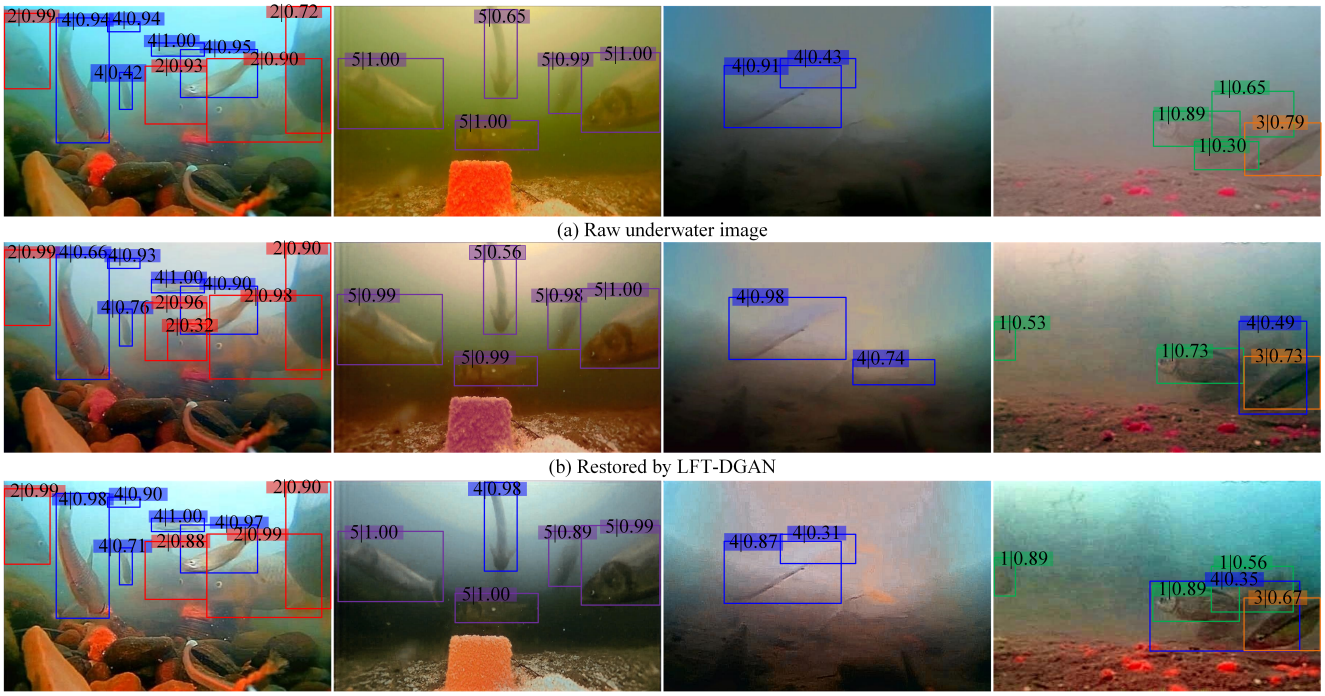

Figure 1. Visualization of object detection with different enhancement algorithms. The numerical marks in the specific figures correspond to the numerical marks in Table 1 and indicate different types of fish.

Table 1. UFD2022 dataset fish categories and sample attributes.

Fish Type	Fish Name	Train Number	Test Number
1	Crucian	1204	848
2	Carp	2603	429
3	Pseudorasbora parva	1240	357
4	Hemiculter leucisculus	1588	145
5	Silver bighead carp	466	374

2. Related Work

Underwater image target detection presents more intricate characteristics than ground images, mainly when gathering uncontrolled fish data. It encounters many challenges, such as weather fluctuations, variations in lighting (including day and night), and changes in water quality. These factors significantly elevate the complexity of crafting underwater target detection models. The evolution of fish target detection models is intricately tied to conventional image target detection technologies, transitioning gradually from manual feature extraction and traditional machine learning methods towards the realm of deep learning [11]. In the domain of fish detection, the extraction of image features assumes paramount importance. Numerous researchers have introduced algorithms like SIFT, SURF, and ORB based on crucial point information and HOG, LBP, and HAAR algorithms based on surface information for fish image target feature extraction. While these methods excel in detecting fish based on appearance attributes such as size, shape, and color in stationary scenes with minimal fish movement, they heavily rely on extensive prior knowledge. However, relying solely on color and texture features for distinguishing fish from the background poses challenges, particularly in camouflage scenarios, where fish and background pixels may exhibit textural ambiguity and color similarities. Furthermore, utilizing color information or other features becomes difficult in uncontrolled underwater environments with light attenuation. In contrast, while focusing on fish boundaries, shape features often overlook the internal information within the shape, potentially simplifying the fish

detection process. In response to these challenges, researchers have proposed innovative methodologies. For instance, M. Ravanbakhsh et al. [12] devised a shape-based level set framework for automated fish detection, effectively tackling issues such as background interference, contrast boundary disparities, and occlusion in underwater imagery. Additionally, M.C. Chuang et al. [13] introduced deformable multi-core technology to address camera movement concerns. While these methods demonstrate commendable performance in specific contexts, their efficacy may vary with changing environmental conditions.

The studies [14] above showcase numerous underwater fish image detection methods, primarily emphasizing the extraction of traditional low-level features. These features often involve small details in the images, such as critical points, colors, textures, shapes, and regions of interest. Additionally, they require a significant amount of manual feature engineering, which introduces additional uncertainties. In practical applications, the effectiveness of methods based on these traditional features often must catch up to expectations. On the other hand, deep learning involves the representation of data at multiple levels, from a low level to a high level, with higher-level features built upon lower-level ones, carrying rich semantic information that can be utilized for object detection in images. In 2016, qin et al. [15] proposed an object detection framework for underwater fish detection. By training the model on publicly available datasets with authentic images, the network model achieved a 15% and 10% improvement in detection accuracy compared to SVM and Softmax, respectively, making automatic detection more precise. However, the CNN architecture used in this model was resource-intensive, particularly with its sliding window detection approach. In 2020, Salman et al. [16] built upon RCNN, utilizing fish motion information and raw images to generate candidate regions. This method achieved at least a 16% improvement in accuracy over Gaussian mixture models on the FCS dataset. Due to the time and computational resources required for CNN training and limitations in sample size and processing speed of pre-trained models, conventional CNN-based methods have spurred the development of approaches like GANs, SSD, YOLO, etc., in the general image domain to address their shortcomings. Some scholars have integrated these methods with underwater-specific image features to create new solutions. In 2018, Zhao et al. [17] proposed a practical and effective semi-supervised learning model based on an improved deep convolutional generative adversarial network for live fish detection in aquaculture. It addressed the issue of CNN-based object detection algorithms heavily relying on training samples and their annotation quality. In 2020, Fan et al. [18] addressed problems such as underwater image blurring, scale and color variations, color shifts, and texture distortions by proposing an underwater detection framework with feature enhancement and anchor constraints. Extensive experiments on the UWD dataset demonstrated the proposed framework's excellent performance in accuracy and robustness. While single-network structures have shown promising results, some scholars are considering parallel processing using multiple-network architectures. For example, in 2022, Kristian, M. et al. [19] proposed a two-step deep learning approach to detecting and classifying temperate fish without pre-training. The first step involved using YOLO object detection to detect each fish in the image. The second step utilized a convolutional neural network with a Squeeze-and-Excitation (SE) structure to classify the fish species in the images. Transfer learning was employed to overcome the limited training samples of temperate fish and improve classification accuracy. In 2021, Y. Wageeh et al. [20] utilized a combination of algorithms to detect fish quantity and trajectories. They first enhanced underwater turbid images using the multiscale Retinex algorithm, followed by the YOLO model for training. Finally, they combined YOLO with optical flow algorithms to track the movement of fish in each frame of the video, thus obtaining fish trajectories.

Fish frequently utilize different habitats for activities such as foraging and shelter, and the differences between habitats may mean that a deep learning model trained on one habitat may be unreliable in another [21,22]. For instance, the complexity of the structure may affect the model's performance, as background clutter and foreground camouflage may affect accuracy. To maximize the effectiveness and accuracy of monitoring and analysis,

deep learning models must be able to adapt to changes in image backgrounds (e.g., across habitats). This issue can be addressed from two perspectives. The first approach is the most straightforward, involving data augmentation and object detection in two-stage processing. While image augmentation methods can help centralize cross-domain data processing, they may not necessarily improve object detection accuracy. There has been extensive research into how degraded images affect object detection accuracy. For example, in 2018, Pei et al. [23] investigated whether dehazing algorithms could improve image classification results and found that there was not always a strict positive correlation between classification accuracy and visual quality; sometimes, it could even decrease image classification performance. In 2021, Pei et al. [24] studied the effects of image degradation and restoration on classification algorithms based on CNN networks and found that it might be because many important low-level features are not well utilized in the first few layers. However, their study was limited to image classification based on the CNN framework and did not consider emerging technologies such as Transformer structures and other advanced image processing methods. Additionally, the division of degraded images was based solely on personal experience or generated data parameters, needing a more scientific basis. Scholars have proposed relevant algorithms based on this research. In 2021, Kazuki, E. et al. [25] proposed a convolutional network that utilizes restoration networks and ensemble learning methods to classify degraded images. This network automatically infers ensemble weights by estimating the degradation level of degraded images and restoring image features, allowing for the accurate classification of degraded images at different degradation levels. In 2019, Prasun, R. et al. [26] explored the effects of degraded images on CNNs and capsule networks and obtained results similar to those described earlier. To address the issues of increased training time and incomplete degraded image introduction associated with training networks on mixed degraded samples, they proposed a new network step to improve the robustness of CNN architectures to some degradation scenarios and introduced a new capsule network structure. The literature described above shows that enhancement and high-level image processing have different optimization goals; enhanced images may lack some structural information, affecting high-level image processing. It may also be because data augmentation and high-level processing are treated as two independent processes cascaded in separate frameworks. In 2020, Chen et al. [8] utilized deep learning techniques and physical prior knowledge for underwater image enhancement. By learning features from underwater images using deep learning networks and combining them with physical prior knowledge, they achieved better reconstruction and enhancement of images. The research results showed that utilizing deep learning and prior knowledge for underwater image enhancement can significantly improve the perception effect of object detection. The other approach involves improving model structures. In 2020, Ellen M. et al. [27] demonstrated the ability of deep learning algorithms to transcend habitats across spatial scales, showing that deep learning models can accurately and consistently extract helpful information from video clips and perform across habitat types when trained on various habitat type clips. In 2021, Zhao et al. [28] proposed a novel composite fish detection framework to solve fish detection and positioning problems in complex underwater environments. They redesigned the composite backbone network to reduce the interference of underwater environments on object features. They introduced an enhanced path aggregation network to address the problem of insufficient utilization of semantic information.

In summary, based on the low-entropy characteristic of underwater images and the diverse features of fish species, designing corresponding modules to improve conventional image object detection algorithms and thus develop new underwater fish detection techniques will become one of the critical directions for future development. Utilizing a two-stage processing approach involving data augmentation and target detection and optimizing objectives through an end-to-end model, a new end-to-end model tightly integrates image enhancement features with target detection, achieving more precise underwater fish detection. This method better

adapts to the complexity of underwater environments and the diversity of fish targets, bringing breakthroughs and advancements to underwater target detection.

3. Materials and Methods

3.1. Fish Image Collection

Evaluating detection algorithm performance is crucial to advancing the automatic monitoring of underwater fish targets. Although there are public underwater image datasets such as Fish4-Knowledge and WildFish, they are mainly collected based on marine environments, which differs from freshwater fish's research background. To meet this practical need, this study carefully constructed a multi-scene underwater fish image dataset called UFD2022, which contains 2800 images. In Table 1, we detail the names, instance images, and training and test sample sizes of various types of fish in the UFD2022 dataset. All images in this dataset are from professional underwater equipment and online video resources. As shown in Figure 2, these underwater equipment were equipped with a front-mounted Sony CCD 1200-line industrial-grade camera to ensure high-quality underwater shooting effects. In addition, the equipment also integrates multiple fish-attracting, white, and infrared lights to optimize shooting conditions. When the equipment captures underwater fish information, it is displayed in real time on the 7 inch display screen at the back end through a transmission line, and the data are stored in the memory card synchronously for subsequent data analysis and processing. When constructing the UFD2022 dataset, we paid particular attention to fish samples' diversity and image quality. To this end, we excluded images that lacked fish or were severely impeded by fish to ensure the accuracy and reliability of the dataset. In addition, we invited several experts in the field of fish research to participate in the guidance and used the LabelMe 4.5.6 software to annotate the fish in the pictures accurately. We used rectangular boxes to annotate the fish to ensure the accuracy and consistency of the annotations. In addition, we also analyzed the distribution relationship between the UFD2022 dataset and the conventional COCO and VOC datasets in Appendix A.1.

Figure 2. Underwater fish image collection equipment. (**a**) Equipment appearance, (**b**) underwater fish image.

3.2. Proposed Method

In this section, we elaborate on the proposed method in-depth. First, we propose a data partitioning scheme using quality evaluation metrics. We observed significant distribution differences between severe image degradation caused by underwater environments and standard datasets beneficial to target detection. We carefully designed a suggestive degradation feature learning module to overcome this challenge of learning the mapping relationship between friendly data and severely degraded data regions. Next, we propose a two-stage training scheme to optimize the object detection network further. As shown in

Figure 3, combined with the detailed steps of Algorithm 1, the core of this framework lies in two key designs: the hint-based degradation feature learning module and the two-stage training strategy. The hint-based degradation feature learning module can capture and learn critical features in the image degradation process, thereby helping the network better adapt to the particularities of underwater images. The two-stage training scheme gradually guides the network to adjust to the degradation characteristics of underwater photos. It improves the accuracy and robustness of target detection by first pre-training on high-quality datasets and then fine-tuning on degraded datasets. In addition, to further verify the effectiveness of the image degradation feature transfer method, we also explored the application of low-quality image feature transfer in target detection tasks. Experimental results show that through reasonable feature transfer strategies, we can effectively utilize useful information in low-quality images to improve target detection performance. Please refer to Appendix A.3 for specific experimental results and a detailed analysis.

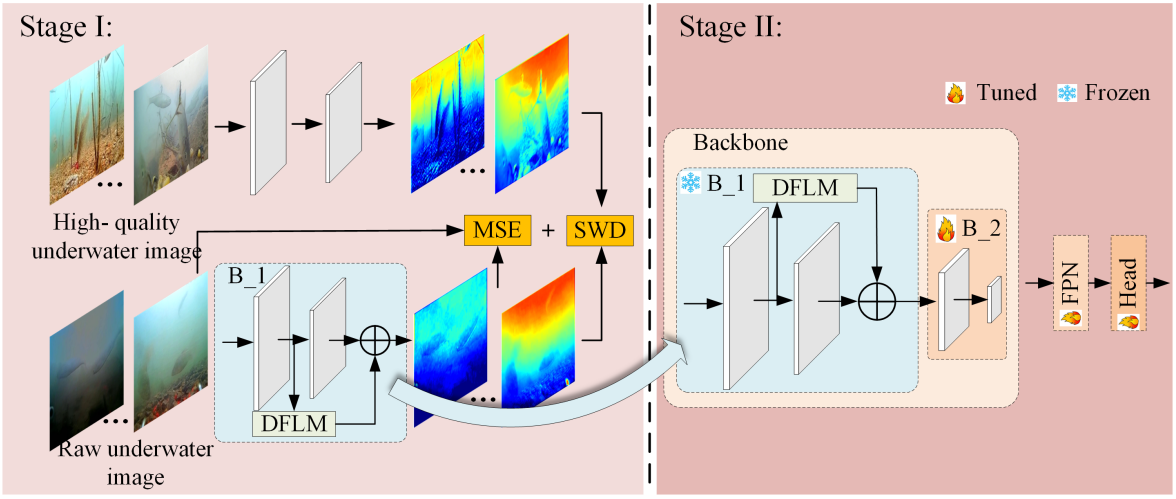

Figure 3. A framework for underwater image object detection based on prior knowledge.

3.2.1. Underwater Image Data Partitioning Scheme

Unlike methods requiring high-quality images or features, our objective is to explore the prior knowledge learnable from pictures in the original underwater dataset that is conducive to object detectors. This prior knowledge guides the detector to eliminate unfriendly features, avoid unstable steps caused by environmental degradation, and reconcile inconsistencies between detection and visual quality. We conducted a series of statistical experiments on the UFD2022 underwater dataset, effectively dividing the dataset into datasets that are beneficial to target detection and datasets that are harmful to target detection. Notably, the mapping relationship between these two datasets embodies the prior knowledge we seek. It enables the use of high-quality image data as instructive guidance for mitigating degradation interference. The description of the entire exploration process is outlined below.

For the UFD2022 training dataset, we explored the impact of different domain images (obtained by different enhancement methods) on the target detection network, and we found an interesting phenomenon: the data domain quality significantly affects the target detection performance. For example, the results in Section 4.3.1 show that under the framework of image domain hybrid learning, lower-quality domains contribute less to the overall performance, resulting in insufficient learning from such samples. Consequently, it becomes imperative to implement effective strategies for appropriately partitioning the original dataset, thereby mitigating the adverse effects of low-quality data on model training.

Algorithm 1 Underwater object detection network based on learning prior knowledge of degraded features.

(1) Data partitioning scheme:

Require: underwater image dataset $D \in \mathbb{R}^{B\times C\times H\times W}$, Object Detection Model M.

Ensure: Suitable for object detection image dataset $x \in \mathbb{R}^{B\times C\times H\times W}$, Harmful to object detection datasets $xh \in \mathbb{R}^{B\times C\times H\times W}$.

Exploring the impact of different domain quality images on object detection performance.

Use $UIQM$ and $MUSIQ$ to divide the dataset D into degradation levels and obtain different subsets D_1, D_2, D_3, D_4.

for $D_i \in D_1, D_2, D_3, D_4$ **do**

Run the object detection model M on D_i.

Evaluate and record the performance of the model on D_i.

end for

Analyze the performance evaluation results and divide the datasets that are beneficial and harmful to target detection.

(2) Unsupervised training:

Require: Suitable for object detection image dataset $x \in \mathbb{R}^{B\times C\times H\times W}$, underwater image dataset $D \in \mathbb{R}^{B\times C\times H\times W}$

Ensure: DFLM module parameters θ

for $i = 1$ to B **do**

Randomly select B images from the valuable dataset for object detection, calculate the magnitude of the Fast Fourier Transform, and calculate channel-level statistics (mean and variance).

Calculate the mean and variance of the amplitude values of the underwater image data D.

The mean and variance of the amplitude values of the two images are exchanged, and the enhanced image is calculated using the AdaIN method.

Calculate the loss function of the enhanced image and D, and update the gradient.

end for

(3) Fine-tuning training:

Require: Underwater image dataset, DFLM module parameters θ

Ensure: Detection output

for $i = 1$ to B **do**

The input image is first processed by a priori degradation knowledge module with fixed parameters θ.

The processed results are processed by the back-end object detection module.

Update gradients for object detection.

end for

return Detection output

Next, we used image quality evaluation indicators to build a data partitioning model. Specifically, we used the normalized UIQM and MUSIQ indicators (see Appendix A.2 for an analysis of selected UIQM and MUSIQ) to establish coordinates for each image, forming a spatial coordinate system based on the center to evaluate image quality more accurately. We divided the image dataset into four levels based on this coordinate system. Images

in the third quadrant are classified as level IV, indicating that these images are the most severely degraded. Given the strong monotonicity of the MUSIQ indicator, we assigned it a higher weight, so images in the fourth quadrant were classified as level III. Images in the first quadrant were rated as level II, while images in the second quadrant were rated as level I, indicating that the images were the least degraded. In this way, we successfully classified the image data according to quality.

Finally, we constructed a specific evaluation framework that quantifies the prediction of target detection models for each degradation degree dataset (as shown in Algorithm 2) and further divides the original underwater image dataset into a target detection data subset and data that are detrimental to the target detection datasets.In detail, we first removed data subsets with different degradation levels from the source degraded dataset to construct a series of carefully designed training datasets. Then, based on these diverse training sets, we trained multiple models and carefully fine-tuned them to optimize their performance on object detection tasks. We then used a precise mathematical formula (see Formula (1)) to measure the contribution of data with different levels of degradation to object detection performance. This approach allows us to accurately identify those datasets with specific levels of degradation that have a significant positive or negative impact on overall object detector performance.

$$Infl[m \rightarrow t] = \Gamma_S[f(t;S)|m \in S] - \Gamma_S[f(t;S)|m \notin S] \tag{1}$$

where $f(t;S)$ is the output of an object detection model trained on a subset S of the source dataset. Positive influence values indicate that including the source degradation level m helps the model correctly predict the object example t. On the other hand, a negative impact value indicates that source degradation level actually hurts the model's performance on object example t.

Algorithm 2 Estimating the impact of dataset degradation on object detection performance.

Require: Source training dataset $S = U_{m=1}^{m} D_m$ (degradation level m), object dataset $T = (t_1, t_2, \ldots, t_n)$, training algorithm A (Cascade RCNN), and the number of m models

Data Partition: Divide source data into equal number of subsets $S_1, S_2, \ldots, S_m \in S$

for each $i \in [1, m]$ **do**

Training model f_i is obtained by training algorithm A on dataset S_i.

end for

for each $m \in [1, m]$ **do**

for each $j \in [1, n]$ **do**

$Infl[D_m \rightarrow t_j] = \Gamma_S[f(t_j;S)|D_m \in S] - \Gamma_S[f(t_j;S)|D_m \notin S]$

end for

end for

return $Infl[D_m \rightarrow t_j]$, j, m

3.2.2. Prompt-Based Degradation Feature Learning Module (DFLM)

The Fourier amplitude spectrum information contains rich characteristic details. Through simple exchange operations, the frequency domain information of the image can be fully utilized to retain its details and structural features, improve its quality and usability, and suppress noise [29–31]. While considerable research has explored related concepts, investigations into high-level visual tasks still need to be conducted. Therefore, this section describe relevant experiments conducted to address this gap. Two underwater images exhibiting distinct conditions (i.e., low quality and high quality) were initially pro-

vided. These images underwent an amplitude component exchange while retaining their original phase components, creating new data. Subsequently, object objects were detected using the object detection algorithm. As depicted in Figure 4, this experiment illustrates that amplitude components influence image quality and impact high-level tasks such as object detection. Consequently, it is inferred that enhancing the amplitude components in Fourier space can potentially improve the detection accuracy of low-quality images.

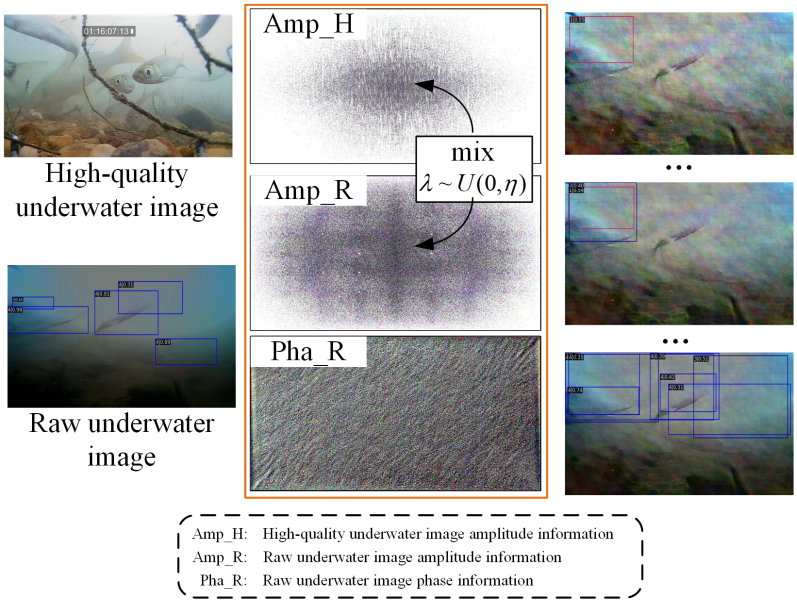

Figure 4. Example of amplitude information exchange between two images.

A prompt-based feature transfer learning module was developed to address the issue of significant background interference caused by directly swapping the phase between high- and low-quality images, which is detrimental to the task of object detection. This module aims to model a learnable function that captures the differences in knowledge between the two datasets. Building upon the concept above and the AdaIN method [32], the prompt-based module endeavors to transform the feature distribution from one dataset, which is unfavorable for object detection, to that of another dataset, which is favorable for object detection. Here, the prompt encapsulates knowledge extracted from the dataset conducive to object detection, facilitating the transfer of relevant feature information. The workflow of the Degradation Feature Prior Knowledge Module is illustrated in Figure 5.

Given a mini-batch of input underwater image features $\{x_i\}_{i=1}^{B} \in \mathrm{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$, the phase value $\{P^u(x_i)\}_{i=1}^{B}$ and amplitude $\{A^u(x_i)\}_{i=1}^{B}$ of these features are obtained using Equations (3) and (4). Similarly, mini-batch image features $\{y_i\}_{i=1}^{B}$ are randomly selected from the dataset that facilitates object detection. The corresponding image amplitude values $\{A^h(x_i)\}_{i=1}^{B}$ are obtained through the Fourier transform equation. The specific formula is as follows:

$$F(x_i)(u,v,c) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x_i(h,w,c) e^{-j2\pi(\frac{h}{H}u + \frac{w}{W}v)} \tag{2}$$

$$A(x_i) = abs(F(x_i)(u,v,c)) \tag{3}$$

$$P(x_i) = angle(F(x_i)(u,v,c)) \tag{4}$$

Among them, H, W, and C represent the height, width, and number of image feature channels, respectively. Here, h and w denote the spatial domain coordinates, while u and v represent the Fourier space coordinates. The *abs* function denotes the absolute value, and *angle* represents the corresponding angle function. The inverse Fourier transform is defined as F^{-1}.

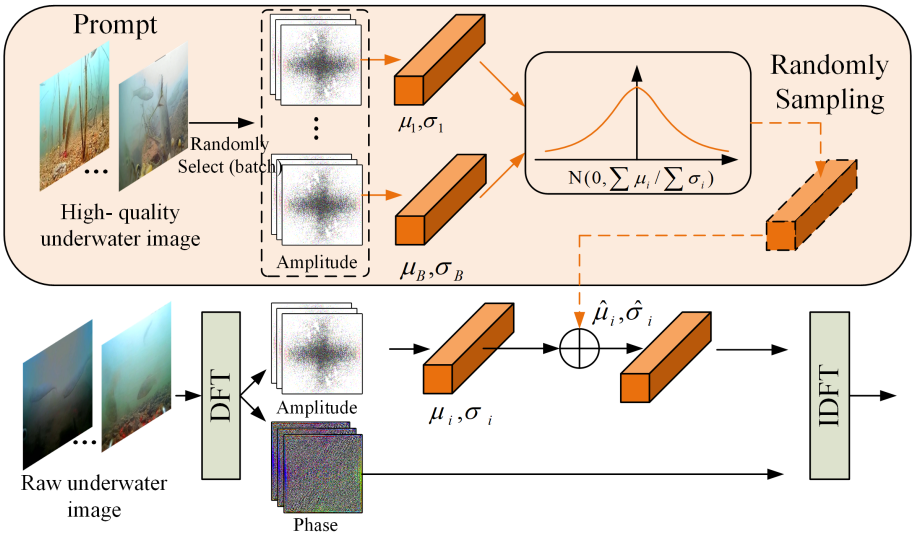

Figure 5. Prompt-based degradation feature learning Module.

Next, for the amplitude values of the object detection dataset, channel-level statistics (mean and standard deviation) are calculated as follows:

$$\mu(A_i^h) = \frac{1}{\sqrt{HW}} \sum_{u=1}^{H} \sum_{v=1}^{H} A_i^h(h,w,c) \tag{5}$$

$$\delta(A_i^h) = \frac{1}{\sqrt{HW}} \sum_{u=1}^{H} \sum_{v=1}^{H} [A_i^h(h,w,c) - \mu(A_i^h)]^2 \tag{6}$$

Assuming that the distribution of each statistic follows a Gaussian distribution, the standard deviation of the statistic is calculated as follows:

$$\sum_{\mu}(A_i^h) = \frac{1}{B} \sum_{i=1}^{B} [\mu(A_i^h) - E[\mu(A_i^h)]]^2 \tag{7}$$

$$\sum_{\delta}(A_i^h) = \frac{1}{B} \sum_{i=1}^{B} [\delta(A_i^h) - E[\delta(A_i^h)]]^2 \tag{8}$$

Thus, a Gaussian distribution of the probability statistics of the amplitude spectrum of the friendly object detection dataset is established, from which a new mean $\overline{\mu}$ and standard deviation $\overline{\delta}$ are randomly drawn:

$$\overline{\mu}(A_i^h) = \mu(A_i^h) + \xi_{\mu} \sum_{\mu}(A_i^h), \xi_{\mu} \sim N(0,\alpha) \tag{9}$$

$$\overline{\delta}(A_i^h) = \delta(A_i^h) + \xi_{\delta} \sum_{\delta}(A_i^h), \xi_{\delta} \sim N(0,\alpha) \tag{10}$$

where $\alpha \in (0,1]$ represents the intensity of the disturbance.

Finally, the phase spectrum of the underwater image is reconstructed:

$$\overline{A_i^u} = \overline{\mu}(A_i^h)\left(\frac{A_i^u - \mu(A_i^u)}{\delta(A_i^u)}\right) + \overline{\delta}(A_i^h) \tag{11}$$

$$output = F^{-1}(\overline{A_i^u}, P_i^u) \tag{12}$$

3.2.3. Two-Stage Training Scheme

Unsupervised transfer training phase. As depicted in Figure 3, we utilized a pre-trained network (trained on the UDFD2022 dataset) to extract low-level and mid-level features from both underwater images and detector-friendly underwater images. Here, the terms "low-level features" and "mid-level features" refer to the outputs of the shallow and mid-level stages of the network, respectively. For instance, in ResNet50/101, "stage 0" denotes the shallow stage, while "stage 1" represents the mid-level stage. During the unsupervised training phase, we aim to transfer the features from underwater images to those of a detector-friendly detector. We employ the DFLM module to learn the transfer of severely degraded features to achieve this. Throughout the training process, we keep the shallow and middle layer stages of the feature extraction network fixed while continuously updating the DFLM module. To ensure the conversion of underwater features into detector-beneficial features, we apply the SGD function [33] and the MSE function to relax the constraints on the features. The MSE function primarily aims to preserve the primary underwater feature information of the processed image. In contrast, the SGD loss function is focused on transferring the underwater image feature distribution of the friendly detector to that of the underwater image feature distribution. The overall loss function can be defined as follows:

$$LOSS = \underset{f}{min}\ E_{Y\sim PY}(\|Y - f(Y)\|^{\beta}) + \lambda d(P_X, P_{\overline{X}}) \tag{13}$$

The fine-tuning phase. As illustrated in Figure 3, we incorporate the trained degradation transfer module into the existing feature extraction network, specifically between the shallow and mid-level stages. Subsequently, we fine-tune the subsequent detection components, including the high-level stages of the feature extraction network, RPN, Neck, and Head, using expected detection losses. The detection losses can be referenced as follows:

$$L_{det} = L_{bbox} + L_{cls} \tag{14}$$

Among these, L_{bbox} represents the bounding box regression loss, and L_{cls} denotes the classification loss. The second stage involves fine-tuning the advanced stages of the feature extraction network of the object detector, including RPN, Neck, and Head, on the training set of the underwater dataset. Following the second stage, our proposed method can be evaluated on underwater datasets.

4. Results and Discussion

4.1. Experimental Settings

All experiments were conducted on a Dell Precision T3630 workstation equipped with an NVIDIA RTX 2080Ti GPU with 24GB of memory. The software environment consisted of Ubuntu 18.04.5 and Python 3.8. The code was developed using the MMdetection toolbox. During the network training process, we employed the SGD optimizer with a momentum of 0.9. The models were trained for 12 epochs with a learning rate of 0.0025. Starting from the 9th epoch, the learning rate was reduced by one-tenth sequentially. Unless otherwise stated, the parameters for each comparative experimental algorithm were set to default values.

The research presented in this article is of significant importance as it primarily focuses on the UFD2022 dataset. This dataset was meticulously partitioned into training and test sets at an 8:2 ratio, ensuring the fairness and validity of our research. We utilized commonly adopted evaluation metrics in object detection to assess model performance, namely, average precision (AP) and mean average precision (mAP). Average precision (AP) is defined as the area under the Precision–Recall (P-R) curve, calculated as follows:

$$AP = \int_0^1 P(R)d(R) \tag{15}$$

Mean average precision (mAP) represents the average of AP values across multiple categories and is defined as follows:

$$mAP = \frac{1}{classes} \sum_{0}^{classes} \int_{0}^{1} P(R)d(R) \tag{16}$$

Note: AP_{50} and AP_{75} in the text represent the average accuracy under different Intersection over Union (IoU) thresholds.

4.2. Experimental Result

4.2.1. Research on the Relationship between Image Enhancement and Object Detection

To explore various data domains, let us consider the UFD2022 dataset as homogeneous data, disregarding domain discrepancies among the data. Based on this dataset, three distinct data domains were generated: Domain-O: This comprises the original dataset containing the training set (train-O) and the test set (test-O). Domain-L: Here, all original data samples undergo enhancement and processing through the LFT-DGAN algorithm, resulting in the creation of the training set (train-L) and the test set (test-L) for training purposes. Domain-D: Similar to Domain-L, all original data samples are subjected to enhancement and processing, this time through the DC-Net algorithm, to generate the training set (train-D) and the test set (test-D) for training.Mixed Domain: In this domain, with a fixed total number of samples, an equal number of data samples is randomly selected from the Domain-O, Domain-L, and Domain-D datasets and combined to form the training set (train-all) and the test set (test-all) for training purposes.

As illustrated in Figure 1, Domain-O exhibits pronounced color distortion, blurriness, and low contrast, whereas the degraded visual samples in Domain-L and Domain-D are effectively restored. Image quality is assessed through various evaluation metrics, with the hierarchy being Domain-O $<$ Domain-L $<$ Domain-D. Specific data results can be found in Appendix A.1. This section delves into a comprehensive examination of Cascade RCNN, Faster RCNN, RetinaNet, Sparse RCNN, and Swin methods within the detection framework. These methods are trained on four distinct training datasets: train-O, train-L, train-D, and train-all. The objective is to explore the influence of various architectural networks and domain datasets on model accuracy.

(1) In-domain dataset analysis

Quantitative numerical analysis: Training and evaluating Cascade RCNN detectors with different backbones (Resnet50, Resnet101 (depth-expanded), Resnetx101 (width-expanded)) and different detection methods. Specific results are shown in Table 2. Across domain-O, -L, and -D, as the image restoration intensity increases, object detection accuracy decreases. For instance, in the Cascade RCNN-R50 method, the object detection accuracy in domain-D $<$ domain-L $<$ domain-O. Similar trends are observed in the other two different backbone object detection methods. As the complexity of the backbone network increases, the object detection accuracy also improves accordingly. Using the Resnetx101 backbone network generally yields the best results because a larger receptive field facilitates the collection of more image feature information. In addition to exploring object detection methods with different backbone networks, exploration is also conducted on different detection methods. Different object detectors achieve the highest mAP in domain-O and the lowest mAP in domain-D.

Table 2. Fish object detection results based on different methods.

Method	Train	Test	mAP	1	2	3	4	5
Cascade RCNN-R50	Train-O	Test-O	58.1	58.9	63.5	49.0	54.9	65.8
	Train-L	Test-L	58.0	58.6	63.3	47.7	53.3	67.1
	Train-D	Test-D	56.4	58.5	62.6	47.3	51.6	62.3
Cascade RCNN-R101	Train-O	Test-O	60.4	60.4	66.2	51.2	55.8	68.2
	Train-L	Test-L	60.0	60.4	65.1	52.5	56.7	69.1
	Train-D	Test-D	59.6	59.7	64.9	50.0	54.8	68.4
Cascade RCNN-X101	Train-O	Test-O	61.0	60.0	66.3	51.3	58.1	69.4
	Train-L	Test-L	60.8	60.2	65.5	51.2	57.0	70.1
	Train-D	Test-D	59.5	59.6	65.0	49.2	56.4	67.4
Faster RCNN-R50	Train-O	Test-O	55.4	56.4	60.5	46.6	52.0	61.4
	Train-L	Test-L	54.6	55.5	59.9	45.4	50.6	61.7
	Train-D	Test-D	54.4	55.6	60.4	44.5	50.7	60.9
RetinaNet	Train-O	Test-O	46.2	51.0	53.0	33.1	42.4	51.5
	Train-L	Test-L	45.4	50.7	53.0	30.1	41.7	51.2
	Train-D	Test-D	44.1	50.5	52.4	31.0	40.6	46.1
Sparse RCNN-R50	Train-O	Test-O	43.8	45.3	47.6	33.7	39.7	52.5
	Train-L	Test-L	42.9	44.4	48.7	30.5	39.5	51.6
	Train-D	Test-D	39.6	42.8	47.5	27.4	37.2	42.9
Swin	Train-O	Test-O	54.2	56.3	59.8	41.8	48.2	64.8
	Train-L	Test-L	50.8	52.7	58.3	38.3	45.9	59.1
	Train-D	Test-D	46.9	51.3	53.1	33.2	42.4	54.4

Convolutional visualization: Compared to low-quality domains, high-quality image samples contain pronounced object representations, making it easier for humans to detect objects in high-quality domain images. This study investigates object saliency in CNN-based detectors. Figure 6 illustrates the multi-scale features of Cascade RCNN, serving as inputs to the detection head and the final convolutional features utilized for detection. Despite domain diversity, the variations in object saliency across multi-scale feature maps are relatively minor. Consequently, the influence of domain quality on convolutional representations is considered negligible concerning object saliency.

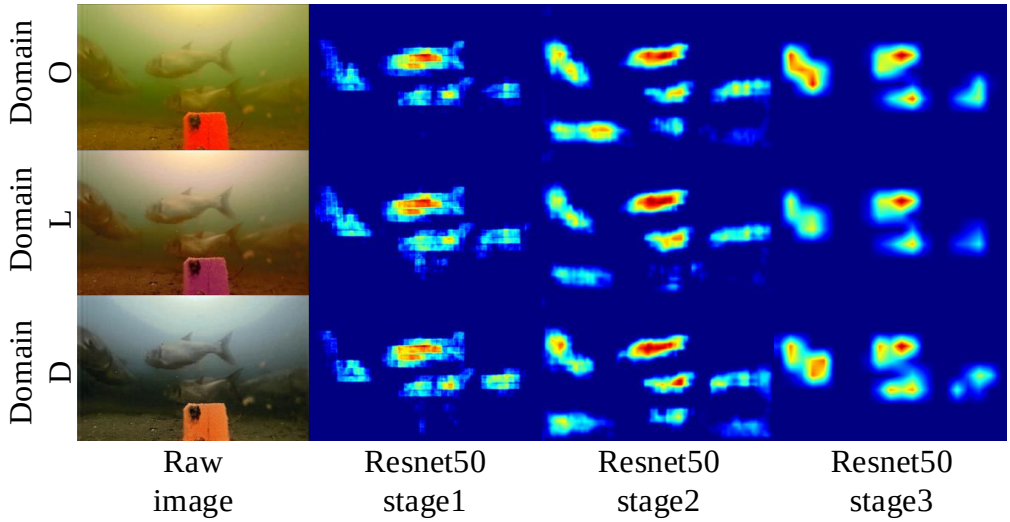

Figure 6. Visualization of convolutional features at different feature scales for detectors based on resnet50 backbone network.

Precision–Recall Analysis: Experimentation was conducted using two representative methods: the two-stage Cascade RCNN-R50 and the single-stage RetinaNet. Different

object precision–recall curves were visualized, as shown in Figure 7, demonstrating the superiority of the two-stage object detection method over the single-stage one, which aligns with conventional wisdom. The following conclusions can be drawn from the two-stage object detection results: (1) the high-precision portion comprises detections with high confidence, with highly overlapping curves across different domains. For instance, for the "1" class objects detected by Cascade RCNN-R50, the curves for domains-O, -L, and -D are almost indistinguishable when the recall is less than 0.8. It suggests that domain differences can be mostly disregarded when detecting highly confident objects. (2) The low-precision portion exhibits separated curves. Specifically, the domain-D < domain-L < domain-O curve indicates that false positives increase with the improvement of domain quality when detecting challenging objects (low-confidence detection results). For example, when the recall for detecting "4" class objects exceeds 0.8, the recall efficiency gradually decreases with increasing image restoration intensity, particularly pronounced in the single-stage method. In addition to the above patterns observed in the two-stage process, the following conclusions were drawn from the single-stage object detection results: (1) different domain data significantly impact the single-stage method, with distinct separations observed in the precision–recall curves for different data domains. (2) The experimental results are consistent with the characteristics of two-stage object detection, where the precision of single-stage object detection is lower than that of two-stage object detection. Moreover, single-stage object detection is significantly influenced by the detection features of the object. For instance, the processed precision is relatively high since there are more samples of the "1" class fish in the data. In contrast, the "4" class fish objects are relatively minor, resulting in more significant fluctuations in detection precision.

Based on the analysis above, the training and evaluation of the object detector rely on the same data domain, highlighting two main points: (1) the quality of the data domain profoundly influences the performance of object detection. (2) To address low recall rates, image enhancement may yield little improvements in in-domain detection performance. It is worth noting that a low recall efficiency corresponds to a low precision under the same recall rate.

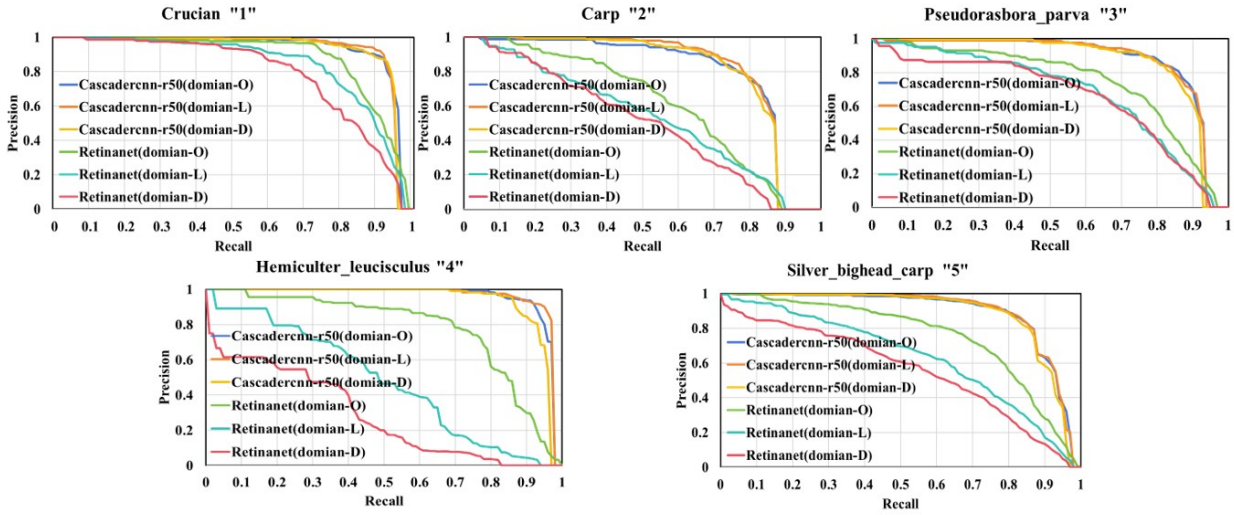

Figure 7. Precision–recall curves for different detectors.

(2) Cross-domain dataset analysis

Cross-Domain Evaluation: Domain-O and Domain-L are used to assess domain shift. Object detectors were trained on train-O and evaluated on test-L, and vice versa. The cross-domain results were computed by subtracting the corresponding in-domain results to highlight the experimental results. Specifically, as shown in Table 3, for the Test-L dataset, the mAP metrics of Cascade RCNN-R50 and RetinaNet models trained on Train-O

decreased by 4.2% and 2.7%, respectively, compared to models trained on Train-L. Similarly, for the Test-O dataset, the map metrics of Cascade RCNN-R50 and RetinaNet models trained on Train-L decreased by 6.1% and 5.9%, respectively, compared to models trained on Train-O. Different degrees of decrease in mAP metrics across different categories indicate varied cross-domain performance degradation. Furthermore, models trained on Train-O achieve higher metrics on the Test-L dataset than those trained on Train-L, suggesting better cross-domain generalization capability for lower-quality domains.

Table 3. Cross-domain test evaluation results. "↓" and "↑" indicate that the performance index of the cross-domain model is lower and higher, respectively, than that of the intra-domain model.

Method	Train	Test	mAP	1	2	3	4	5
Cascade RCNN-R101	Train-O	Test-L	55.8 4.2 ↓	53.5 6.9 ↓	61.8 3.3 ↓	45.7 6.8 ↓	54.3 2.4 ↓	63.8 5.3 ↓
	Train-L	Test-O	54.3 6.1 ↓	54.4 6 ↓	55.5 9.6 ↓	45.8 5.4 ↓	54.8 1.0 ↓	61.0 7.2 ↓
RetinaNet	Train-O	Test-L	42.7 2.7 ↓	40.4 10.3 ↓	52.1 0.9 ↓	31.8 1.7 ↑	40.0 1.7 ↓	50.2 1 ↓
	Train-L	Test-O	40.3 5.9 ↓	51.6 0.6 ↑	43.3 9.7 ↓	32.5 0.6 ↓	40.0 2.4 ↓	45.5 6 ↓

Cross-Domain Training: To explore the detection performance of domain-mixed learning, object detectors were trained on the Train-all dataset and evaluated on the Test-O, Test-L, and Test-D datasets. As shown in Table 4, for the same training set, the testing accuracy in different domains decreased to a certain extent, with more considerable disparities in the testing object detection accuracy observed in higher-quality images. That is, when using Train-all to train models, the high-quality data (Train-D), to some extent, lose their effectiveness. Therefore, cross-domain training is not suitable for improving detection performance. Additionally, data domains of different qualities contribute differently to the learning process. Mixing low-quality samples with high-quality ones hinders effective learning of the low-quality samples.

Table 4. Cross-domain training evaluation results. "↓" and "↑" indicate that the performance index of the cross-domain model is lower and higher, respectively, than that of the intra-domain model.

Method	Train	Test	mAP	1	2	3	4	5
Cascade RCNN-R101	Train-all	Test-O	58 2.4 ↓	57.7 2.7 ↓	64.1 2.1 ↓	50.7 0.5 ↓	55.5 0.3 ↓	62.1 6.1 ↓
		Test-L	57.7 2.3 ↓	57.7 2.7 ↓	63.6 1.5 ↓	44.6 7.9 ↓	54.5 2.2 ↓	68 1.1 ↓
		Test-D	52.1 7.5 ↓	58.4 1.3 ↓	63 1.9 ↓	47.2 2.8 ↓	53.7 1.1 ↓	38.1 30.3 ↓
RetinaNet	Train-all	Test-O	44.8 1.4 ↓	49.7 1.3 ↓	52.5 0.5 ↓	33.4 0.3 ↑	42.2 0.2 ↓	46.5 5 ↓
		Test-L	43.6 1.8 ↓	49.5 1.2 ↓	53 0	29 1.1 ↓	39.8 1.9 ↓	46.4 4.8 ↓
		Test-D	40 4.1 ↓	49.5 1 ↓	52.3 0.1 ↓	30.8 0.2 ↓	38.6 2 ↓	28.7 17.4 ↓

In this section of the study, Cascade RCNN-R101 and RetinaNet detectors were trained and evaluated on different data domains. It primarily reveals three key observations: (1) domain shift leads to decreased accuracy. (2) Cross-domain inference shows that learning from low-quality domains facilitates better generalization to high-quality domains. (3) In domain-mixed learning, the contribution of low-quality domains is relatively minor, which hinders effective learning of low-quality samples.

4.2.2. Research Results of High-Quality Underwater Image Dataset

Relevant experiments were conducted on an underwater fish image dataset to assess how training Cascade RCNN with image datasets augmented using UIQM and MUSIQ performs under various learnable parameters. The model was trained for the original training dataset F_{train} by partitioning it based on different parameters and excluding data from high-quality underwater images. Specifically, the new training dataset can be described as follows:

$$\{F_i^f\}_{i=1} = F_{train} \ominus \{F_{(F_{mi}^h, F_{ui}^h)}\}_{i=1}, i = 1, 2, 3, 4 \tag{17}$$

where $\ominus$ denotes the removal of a portion of the dataset, and $F_{(F_{mi}^h, F_{ui}^h)}$ signifies the division of the dataset based on learnable parameters (MUSIQ and UIQM). We are focusing on data falling within the third quadrant. Each learnable parameter is sorted from small to large according to the image dataset metric and divided at 1/8, 1/4, 3/8, and 1/2 of the dataset quantity. The MUSIQ axis is represented by F_{m1}^h, F_{m2}^h, F_{m3}^h, and F_{m4}^h while the UIQM axis is represented by F_{u1}^h, F_{u2}^h, F_{u3}^h, and F_{u4}^h.

The specific experimental results are illustrated in Figure 8. The following observations can be made: (1) the further away from the center of the dataset, the more the lower quality data tend to affect object detection accuracy adversely. For instance, object detection accuracy in the upper right corner surpasses that of the model trained on the source dataset by 1%. Removing detrimental object detection data can effectively enhance the accuracy of the object detection model. (2) When fixing the learnable parameters of MUSIQ and UIQM separately, as indicated by the two directions of the arrow, the variability range of MUSIQ is greater than that of UIQM. It implies that the object detection accuracy model relies more heavily on MUSIQ parameters. It also confirms that MUSIQ is better suited for evaluating the extent of image degradation. Following this experimental guideline, the dataset is partitioned with fixed parameters (F_{mi}^h, F_{ui}^h), where data falling within the third quadrant are deemed as harmful object detection data, and the remaining data are considered as friendly object detection data.

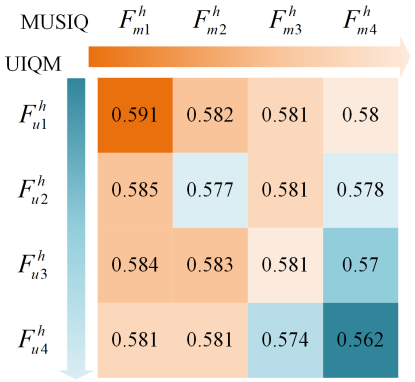

Figure 8. Experimental results of learnable parameters UIQM and MUSIQ.

4.2.3. Comparative Results with State-of-the-Art Methods

This paper utilizes the UFD2022 dataset to validate the efficacy of the proposed method and compare it with baseline image object detection techniques. It encompasses diverse deep learning-based object detection methods employing various architectures, including CNN, Transformer, Diffusion, and UIE + UOD. In the UIE + UOD architecture detector, cutting-edge image enhancement methods such as PUIE [34], CWR [35], DC-Net, and LFTDGAN are chosen for preprocessing. The preprocessed images are then fed into the detector (Cascade RCNN [36] and Faster RCNN [37]) for subsequent training and testing phases. For CNN-based architecture detectors, both single-stage models (e.g., FCOS [38] and YOLO series [39]) and two-stage models (e.g., Faster RCNN and Dynamic RCNN [40]) are selected for training and testing purposes. Regarding Transformer-based

architecture detectors, the latest research models from representative series like Swin [41] and PVT [42,43], along with DINO [44] and DDQ [45], are used for training and testing. Finally, for detectors based on the Diffusion architecture, the DiffusionDet [46] model is the selection for the training and testing phases. We also added the FERNet detector [18] designed specifically for underwater object detection methods.

Table 5 shows in detail the evaluation results of different architectural methods in the object detection task. Our method significantly outperformed existing state-of-the-art algorithms on various key metrics. Specifically, compared with the baseline detector, our method significantly improved the mapped metric of Cascade RCNN (ResNet50) by 2.4%, while the AP50 and AP75 metrics also increased by 0.9% and 2.6%, respectively. Compared with Cascade RCNN (ResNetXT101), our method achieved a 2.5% improvement in mAP, and the improvements in AP50 and AP75 reached 3.8% and 2%, respectively. However, it is worth noting that the UIE + UOD type detector did not perform as well as we expected. It is mainly attributed to the inconsistency in their respective goals, with UIE aiming to optimize image quality and UOD focusing on improving detection accuracy. Currently, CNN-based detectors dominate the object detection field, but our method shows clear advantages. Although YOLOV7 performed well on multiple metrics, it still lagged behind our ResNetXT101 model by 1.5%, 1.3%, and 1.5% in mAP, AP50, and AP75, respectively. In addition, the Transformer class detector, as an emerging detector type, also failed to surpass our proposed method. For example, our ResNetXT101 model achieved 63.5% in mAP, 1.9% higher than the best-performing DDQ model in its class. Although diffusion-based detectors are the latest object detection technology that introduces diffusion models, they could still not compete with our proposed method regarding actual detection effects. Despite performance improvements, the network designed for underwater object detection still lagged behind our ResNetXT101 model by 4.3% in mAP. It fully proves the advancement and effectiveness of our proposed method in the field of object detection.

Table 5. Accuracy comparison experiment on UFD2022 dataset.

Methods	Backbone	mAP	AP50	AP75
Baseline:				
	ResNet50	0.581	0.886	0.671
	ResNetXT101	0.61	0.903	0.708
UIE + UOD:				
DC-Net + Cascade RCNN	ResNet50	0.564	0.886	0.659
LFTDGAN + Cascade RCNN	ResNet50	0.58	0.884	0.662
PUIE + Faster RCNN	ResNet50	0.558	0.889	0.631
CWR + Faster RCNN	ResNet50	0.537	0.872	0.603
CNN:				
FCOS	ResNet50	0.418	0.72	0.452
YOLOF	ResNet50	0.465	0.824	0.475
YOLOX	DarkNet53	0.602	0.917	0.694
YOLOV7	CSPDarkNet	0.62	0.928	0.713
Dynamic RCNN	ResNet50	0.559	0.868	0.647
Faster RCNN	ResNet50	0.56	0.889	0.638
Transformer:				
Swin	Swin-T-P4	0.542	0.876	0.617
PVTv1	PVT-Medium	0.562	0.897	0.621
PVTv2	PVTv2-B4	0.607	0.922	0.7
DINO	Swin-L	0.585	0.878	0.67
DDQ	DETR	0.616	0.914	0.716
Diffusion:				
DiffusionDet	ResNet50	0.6	0.895	0.697
Underwater:				
FERNet	—	0.592	0.915	0.712
Ours:				
	ResNet50	0.605	0.914	0.697
	ResNetXT101	0.635	0.941	0.728

Figure 9 illustrates several methods' detection results, further substantiating our proposed approach's efficacy across diverse environments. We showcase the baseline method Cascade RCNN (ResNet50, ResNetXT101) alongside the most representative methods from each category, such as LFTDGAN+Cascade RCNN in the UIE+UOD class, YOLOV7 in the CNN class, DDQ in the Transformer class, DiffusionDet in the Diffusion class, and our proposed method. For images with color distortion (first row), our proposed method and most others successfully executed the detection task. However, in scenarios with low visibility (second and third rows) and fog effects (fourth row), most methods struggled to complete the detection, resulting in false positives and missed detections. In contrast, our proposed method accurately performed the detection task. Notably, methods like DDQ and Cascade RCNN (ResNet50) exhibited errors and missed detections. These qualitative findings underscore the robust performance of our proposed method across diverse environmental conditions.

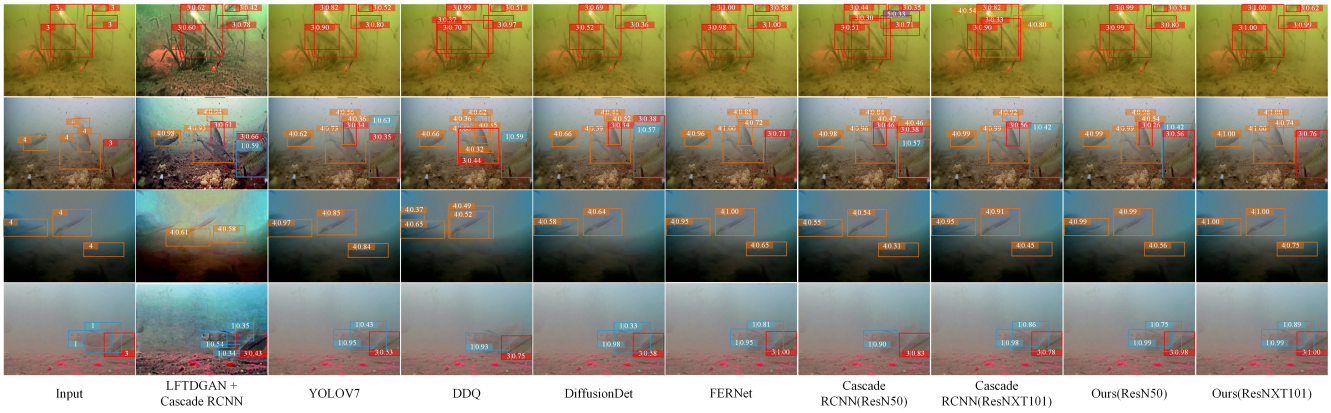

Figure 9. Some qualitative examples on the UFD2022 dataset. Examples from top to bottom are color distortion, low contrast, and fog effects. The numerical marks in the specific figures correspond to the numerical marks in Table 1 and indicate different types of fish.

To further demonstrate the effectiveness of our proposed method, we visualize the backbone feature maps of some methods in Figure 10. Our method excels at distinguishing between object and background features, indicating its effectiveness.

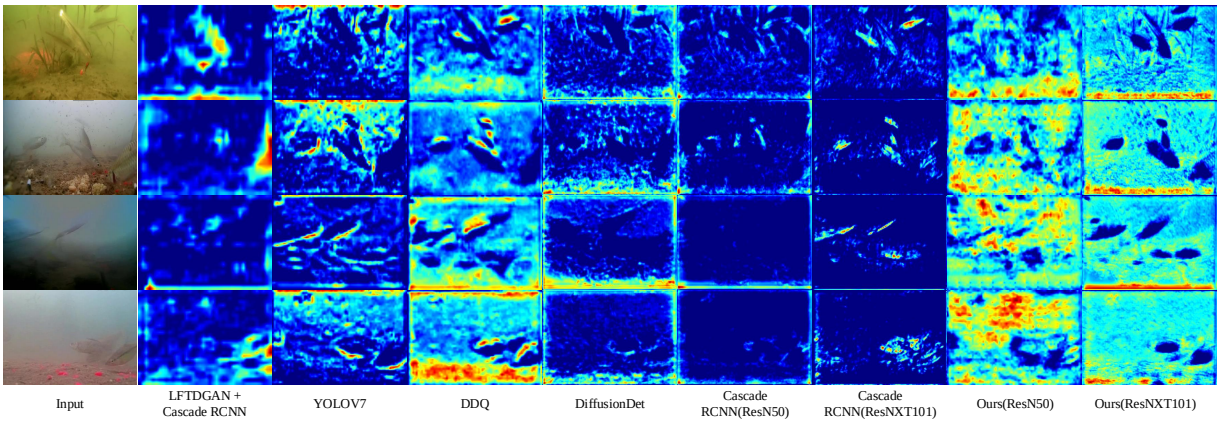

Figure 10. Feature map visualization results of different methods on the UFD2022 dataset. Examples from top to bottom are color distortion, low contrast, and fog effects.

4.3. Ablation Experiment

4.3.1. Performance of Prompt-Based Degradation Feature Learning Module

The degradation prior knowledge learning module aims to transfer highly degraded regions in underwater images. To evaluate the transferability of the DFLM's feature distribution, we visualize the feature distributions of three subsets: images processed by DFLM, the original image set, and a set of low-quality underwater photos. Each group of images contains 30 randomly selected images for visualization. As shown in Figure 11, the distribution of the original image set is closer to that of high-quality underwater image datasets, indicating that the degradation prior knowledge successfully learned the previous understanding of underwater images.

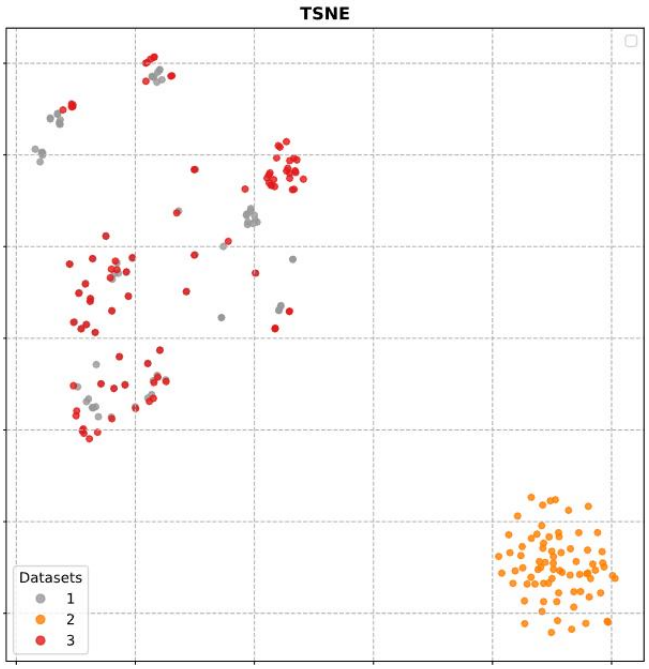

Figure 11. Visualization results of TSNE performance of degraded prior knowledge module. Here, "1" represents a dataset that is not conducive to object detection, "2" represents a dataset that is conducive to object detection, and "3" represents DFLM processed data.

4.3.2. Structure of Prompt-Based Degradation Feature Learning Module

An ablation study was performed to analyze the effectiveness of different parameter settings in the degraded prior knowledge learning module. The specific settings are as follows:

- Pha-Amp: Directly exchange image amplitude and frequency features.
- Pos0: Place the DFLM module before stage0 of resnet50.
- Pos1: Place the DFLM module after stage1 of resnet50.
- Pos2: Place the DFLM module after stage2 of resnet50.

The results of the ablation experiment are presented in Table 6. The proposed method outperforms the Pha-Amp method by 9.3% in the mAP index, 6% in the AP50 index, and 12.1% in the AP75 index. The proposed degradation module exhibits its most effective functionality during the intermediate stage, resulting in the highest accuracy index achieved.

Table 6. Ablation experiment's accuracy on UFD2022 dataset.

Method	Pha-Amp	Pos0	Pos1	Pos2	Ours (ResNet50)
mAP	0.512	0.591	0.596	0.589	0.605
AP50	0.854	0.892	0.904	0.891	0.914
AP75	0.576	0.676	0.684	0.664	0.697

4.3.3. The Impact of Training Strategies

In this subsection, we describe ablation experiments conducted to analyze the effectiveness of various training strategies in the proposed method. Table 7 presents the results, showcasing the performance improvements achieved with the incremental addition of each step.

- CU: Baseline experiments were conducted by training Cascade RCNN (ResNet50) on the UFD2022 training dataset.
- CU + FOD: We trained Cascade RCNN on a subset of the UFD2022 dataset that is a high-quality dataset for friendly object detection (FOD).
- CU + TF: The DFLM module does not perform advance training and was directly added to the backbone to train Cascade RCNN on a subset of data in the UFD2022 dataset that is conducive to object detection.
- No FT: Without fine-tuning subsequent components of the detector, Cascade RCNN was trained on a subset of the UFD2022 dataset that is beneficial to object detection.

As depicted in Table 7, the CU + FOD, CU + TF, and No FT training strategies demonstrate varying degrees of improvement compared to the baseline CU. The proposed method attained optimal results, suggesting that degradation feature transfer plays a crucial role. Moreover, fine-tuning further enhances perception capabilities, underscoring its significance in the training process.

Table 7. Ablation experimental accuracy on UFD2022 dataset.

Training Strategies	mAP	AP50	AP75
CU	0.581	0.897	0.657
CU + FOD	0.591	0.903	0.681
CU + TF	0.596	0.912	0.689
No FT	0.585	0.907	0.675
Ours(ResNet50)	0.605	0.914	0.697

5. Conclusions

In this paper, we present a novel fish target detection network leveraging prior knowledge of image degradation to mitigate the feature mismatch between underwater image degradation and target detection. Firstly, we investigated the correlation between visually restored image quality and detection efficiency, elucidating the impact of this relationship on target detection. We constructed a dataset conducive to target detection using image quality evaluation metrics. Secondly, we propose a simple yet effective prompt-based degradation feature learning module (DFLM) to learn features conducive to target detection. The DFLM can be learned unsupervised and utilized as a randomly inserted module. Moreover, we introduce a novel two-stage training scheme that enhances the detector's adaptability to underwater fish detection tasks by dynamically adjusting the training scheme. Finally, we curated a multi-scene underwater fish dataset (UFD2022), laying a robust foundation for underwater fish target detection endeavors. On the UFD2022 dataset, our proposed method demonstrated a 2.4% and 2.5% improvement in the mean average precision (mAP) metric over the baseline ResNet50 and ResNetXT101, respectively. Similarly, our proposed method, ResNetXT101, achieved a 1.5% enhancement over the state-of-the-art method YOLOV7. Ablation studies further validated the efficacy of each component in our model for high-density counting tasks.

Author Contributions: Methodology, writing—original draft preparation, writing—review and editing, S.Z.; visualization, formal analysis, L.W.; funding acquisition, project administration, R.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China of Funder grant number 31671586.

Data Availability Statement: The data can be shared up on request.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

Appendix A.1. UFD2022 Dataset Analysis

Due to the presence of underwater fish images from various scenes within the UFD2022 dataset, this factor significantly influences the accuracy of fish object algorithms. To further analyze the disparities between the UFD2022 dataset and two commonly used general-purpose object detection datasets, PASCAL VOC [47] and MS COCO [48], an equivalent number of samples were randomly sampled from the two general datasets. Utilizing clustering TSNE algorithms, the dataset was visually clustered for comparison, as depicted in Figure A1. It can be observed that the overall distribution of samples in the UFD2022 dataset is relatively dispersed, whereas the clustering of samples in the general-purpose object detection datasets is more concentrated. This indicates that the UFD2022 dataset encompasses a larger cross-domain variety of images from different underwater scenes.

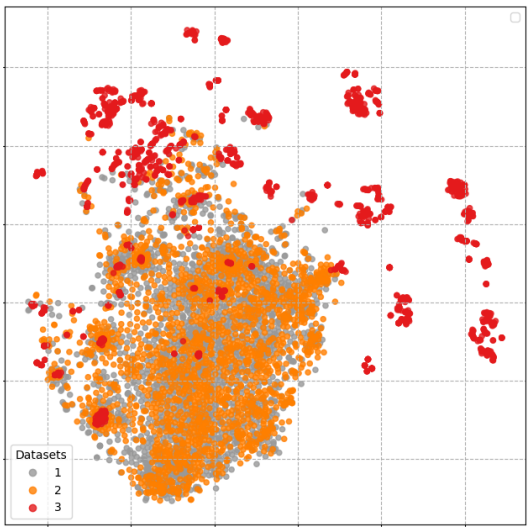

Figure A1. TSNE visualization of the UFD2022 dataset; "1" represents the PASCAL VOC dataset, "2" represents the MS COCO dataset, and "3" represents the UFD2022 dataset.

Appendix A.2. Data Analysis in Different Domains

It is difficult to quantitatively distinguish image quality from the visual enhancement effect of underwater images. Therefore, a quantitative evaluation of underwater images is carried out through different image quality evaluation indicators. The Underwater Color Image Quality Evaluation Metric (UCIQE) quantifies image quality through hue, saturation, and contrast; Underwater Image Quality Metric (UIQM) is a comprehensive quality representation of underwater images; MUSIQ [49] is a Transformer-based multi-scale image quality assessment model; Brisque [50] evaluates image quality by calculating image naturalness; entropy is the information entropy of the image; NIQE is based on the statistical characteristics of natural images for modeling and comparison. The evaluation indicators of different domains are shown in Table A1. The evaluation indicators of different domains show the relative order of underwater image quality, where Domain-O $<$ Domain-L $<$ Domain-D.

Table A1. Results of different evaluation indicators. The optimal results are marked in red.

Method	UIQM	UCIQE	MUSIQ	Entropy	Brisque	NIQE
Domain-O	0.6074	4.2723	45.0560	7.3642	38.8077	6.1737
Domain-L	0.7297	5.0423	45.6851	7.6452	37.1034	5.8911
Domain-D	0.8136	4.5908	45.9735	7.6370	33.0612	5.0216

The aforementioned evaluation indicators primarily focus on the overall scope of the dataset and may not effectively highlight the domain range, distribution, and other characteristics of local datasets. To address this limitation, it is recommended to visualize the dataset distribution by using two different indicators as the x and y axes. However, it should be noted that some metrics may not accurately reflect the quality of recovered underwater images, as discussed in the literature [51]. Therefore, establishing reliable degradation levels solely based on these indicators might be questionable. To identify the two best non-reference image quality assessment (NR-IQA) methods for underwater images, an analysis of six NR-IQA indicators is conducted. The analysis involves a degraded underwater image x and a pair of clean images y, which are linearly combined according to the mixing theory to generate a set of images with varying qualities. The specific formula for this process can be expressed as follows:

$$S = \{\alpha_i x + (1 - \alpha_i) y\}_{i=1}^{10} \qquad s.t. \quad \alpha_i = 0.1 \times i, i = 1, 2, \ldots, 10 \tag{A1}$$

If the score of an NR-IQA method on the underwater fish dataset decreases with an increase in α_i, it signifies the reliability of that evaluation method. Adhering to this criterion, experiments were performed on the underwater fish dataset using six different NR-IQA methods, and the results are illustrated in Figure A2. Notably, MUSIQ and UIQM indicators exhibited the highest conformity to the monotonicity requirement. Hence, the UIQM and MUSIQ indicators were selected for visualizing the distribution of the dataset. By leveraging the UIQM and MUSIQ indicators, a more comprehensive assessment of underwater image quality can be achieved. This enables the identification of suitable image enhancement techniques or algorithms to enhance the visual quality of underwater images.

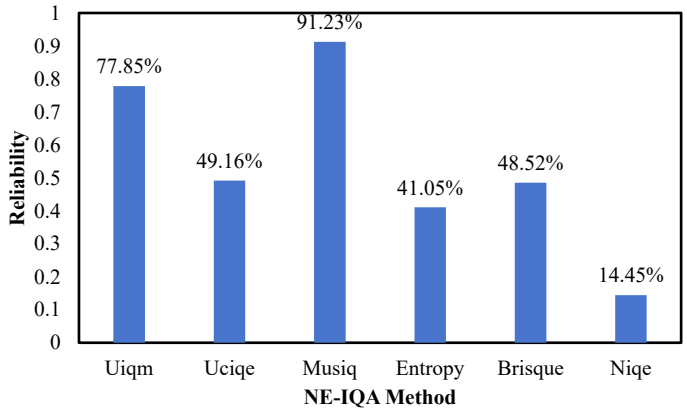

Figure A2. Results of different NR-IQA metrics on the UFD2022 dataset.

We chose the UIQM and MUSIQ indicators (normalized) to compose the coordinates for each image and visualized the results, as shown in Figure A3. To provide a better description, a spatial coordinate system was established with the center position, leading to the following observations: (1) despite the superior visual and quantitative evaluation results achieved by the Domain-D and Domain-L data, they exhibit a more uniform and widespread distribution compared to the Domain-O data. (2) The effectiveness of image enhancement correlates with the capability to transform lower-quality images into higher-quality ones. For instance, in the bottom-left quadrant of Figure A3, where Domain-O contains lower-quality images, algorithmic processing leads to quality improvements, with Domain-D demonstrating superior processing performance. (3) It is worth noting that image enhancement algorithms may sometimes yield lower quantitative scores for some initially high-quality images, despite an overall improvement in quality. For example, while most of the original Domain-O data reside in the second quadrant, after enhancement, some data points disperse into other lower-quality quadrants (such as Domain-D and Domain-L images).

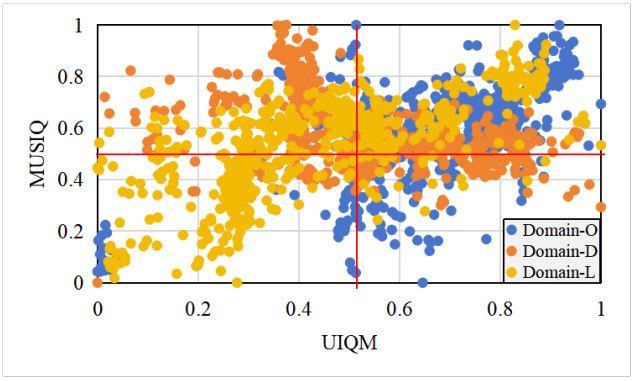

Figure A3. Visualization of local dataset metrics.

Appendix A.3. Research on Distribution of Prior Knowledge of Degradation of Underwater Images

Based on the illustration in Figure A4, we conducted a study on the effects of transferring feature distributions from severely degraded datasets to other domains on the accuracy of object detection in underwater images. We utilized various image enhancement methods to improve low-quality images (i.e., performing feature distribution transfer) and assessed fish object detection accuracy using Cascade RCNN. Drawing from the analysis of data domain distribution transfer, we arrived at the following two key conclusions: (1) low-quality datasets hinder the improvement of accuracy in underwater image object detection. This indicates that low-quality data may lead to a decline in the performance of object detection algorithms, thereby impacting overall detection accuracy. (2) Transferring the feature distribution from low-quality datasets to high-quality datasets can maximize object detection accuracy. This suggests that by effectively transferring features from low-quality samples to high-quality ones, the performance of object detection algorithms can be significantly enhanced, consequently improving the accuracy of object detection in underwater images. Hence, it can be inferred that the contributions of low-quality and high-quality samples to object detection accuracy differ. Therefore, effective methods need to be explored to learn from and process low-quality samples, aiming to enhance the performance of object detection algorithms and consequently improve object detection accuracy in underwater images.

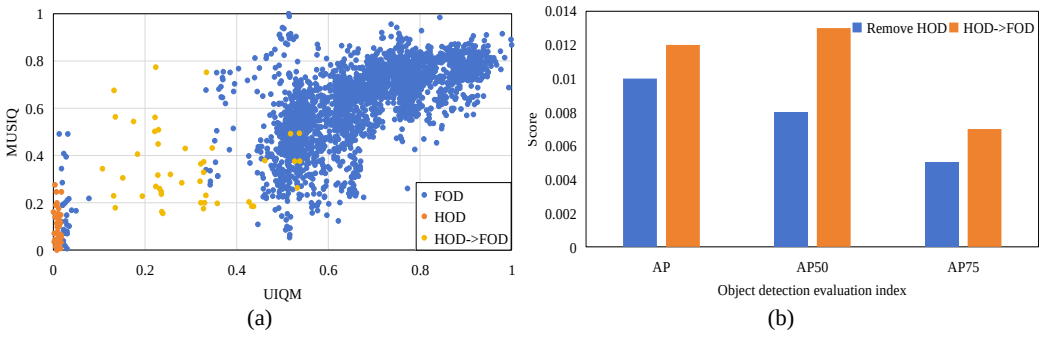

Figure A4. Object detection results of transformed distribution on UFD2022 degraded dataset. (**a**) The UFD2022 image dataset is divided into a high-quality dataset for friendly object detection (FOD) and a low-quality dataset for harmful object detection (HOD). (**b**) The low-quality dataset that is not conducive to object detection is removed. The distribution characteristics of low-quality datasets that are not conducive to object detection are transferred to the object detection results of high-quality datasets that are conducive to object detection. The score is represented by the difference between the object detection accuracy of the model trained on the processed data and the object test accuracy of the model trained on the original data.

References

1. Zheng, S.; Wang, R.; Zheng, S.; Wang, F.; Wang, L.; Liu, Z. A Multi-scale feature modulation network for efficient underwater image enhancement. *J. King Saud-Univ.-Comput. Inf. Sci.* **2024**, *36*, 101888. [CrossRef]
2. Wang, H.; Sun, S.; Bai, X.; Wang, J.; Ren, P. A reinforcement learning paradigm of configuring visual enhancement for object detection in underwater scenes. *IEEE J. Ocean. Eng.* **2023**, *48*, 443–461. [CrossRef]
3. Xu, S.; Zhang, M.; Song, W.; Mei, H.; He, Q.; Liotta, A. A systematic review and analysis of deep learning-based underwater object detection. *Neurocomputing* **2023**, *527*, 204–232. [CrossRef]
4. Fayaz, S.; Parah, S.A.; Qureshi, G.; Lloret, J.; Del Ser, J.; Muhammad, K. Intelligent Underwater Object Detection and Image Restoration for Autonomous Underwater Vehicles. *IEEE Trans. Veh. Technol.* **2023**, *73*, 1726–1735. [CrossRef]
5. 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] [PubMed]
6. 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]
7. Jiang, L.; Wang, Y.; Jia, Q.; Xu, S.; Liu, Y.; Fan, X.; Li, H.; Liu, R.; Xue, X.; Wang, R. Underwater species detection using channel sharpening attention. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, China, 20–24 October 2021; pp. 4259–4267.
8. Chen, L.; Jiang, Z.; Tong, L.; Liu, Z.; Zhao, A.; Zhang, Q.; Dong, J.; Zhou, H. Perceptual underwater image enhancement with deep learning and physical priors. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 3078–3092. [CrossRef]
9. Zheng, S.; Wang, R.; Zheng, S.; Wang, L.; Liu, Z. A Learnable Full-frequency Transformer Dual Generative Adversarial Network for underwater image enhancement. *Front. Mar. Sci.* **2024**, *11*, 1321549. [CrossRef]
10. Zheng, S.; Wang, R.; Chen, G.; Huang, Z.; Teng, Y.; Wang, L.; Liu, Z. Underwater image enhancement using Divide-and-Conquer network. *PLoS ONE* **2024**, *19*, e0294609. [CrossRef]
11. 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. [CrossRef]
12. Ravanbakhsh, M.; Shortis, M.R.; Shafait, F.; Mian, A.; Harvey, E.S.; Seager, J.W. Automated Fish Detection in Underwater Images Using Shape-Based Level Sets. *Photogramm. Rec.* **2015**, *30*, 46–62. [CrossRef]
13. Chuang, M.C.; Hwang, J.N.; Ye, J.H.; Huang, S.C.; Williams, K. Underwater fish tracking for moving cameras based on deformable multiple kernels. *IEEE Trans. Syst. Man Cybern. Syst.* **2016**, *47*, 2467–2477. [CrossRef]
14. Liu, H.; Ma, X.; Yu, Y.; Wang, L.; Hao, L. Application of deep learning-based object detection techniques in fish aquaculture: A review. *J. Mar. Sci. Eng.* **2023**, *11*, 867. [CrossRef]
15. Qin, H.; Li, X.; Liang, J.; Peng, Y.; Zhang, C. DeepFish: Accurate underwater live fish recognition with a deep architecture. *Neurocomputing* **2016**, *187*, 49–58. [CrossRef]
16. 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]
17. Zhao, J.; Li, Y.; Zhang, F.; Zhu, S.; Liu, Y.; Lu, H.; Ye, Z. Semi-supervised learning-based live fish identification in aquaculture using modified deep convolutional generative adversarial networks. *Trans. ASABE* **2018**, *61*, 699–710. [CrossRef]
18. Fan, B.; Chen, W.; Cong, Y.; Tian, J. Dual refinement underwater object detection network. In Proceedings of the European Conference Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 275–291.
19. Knausgård, K.M.; Wiklund, A.; Sørdalen, T.K.; Halvorsen, K.T.; Kleiven, A.R.; Jiao, L.; Goodwin, M. Temperate fish detection and classification: A deep learning based approach. *Appl. Intell.* **2022**, *52*, 6988–7001. [CrossRef]
20. Wageeh, Y.; Mohamed, H.E.D.; Fadl, A.; Anas, O.; ElMasry, N.; Nabil, A.; Atia, A. YOLO fish detection with Euclidean tracking in fish farms. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *12*, 5–12. [CrossRef]
21. Zhao, T.; Zhang, G.; Zhong, P.; Shen, Z. DMDnet: A decoupled multi-scale discriminant model for cross-domain fish detection. *Biosyst. Eng.* **2023**, *234*, 32–45. [CrossRef]
22. Gong, B.; Dai, K.; Shao, J.; Jing, L.; Chen, Y. Fish-TViT: A novel fish species classification method in multi water areas based on transfer learning and vision transformer. *Heliyon* **2023**, *9*, e16761. [CrossRef]
23. Pei, Y.; Huang, Y.; Zou, Q.; Lu, Y.; Wang, S. Does haze removal help cnn-based image classification? In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 682–697.
24. Pei, Y.; Huang, Y.; Zou, Q.; Zhang, X.; Wang, S. Effects of image degradation and degradation removal to CNN-based image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1239–1253. [CrossRef] [PubMed]
25. Endo, K.; Tanaka, M.; Okutomi, M. CNN-based classification of degraded images with awareness of degradation levels. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 4046–4057. [CrossRef]
26. Roy, P.; Ghosh, S.; Bhattacharya, S.; Pal, U. Effects of degradations on deep neural network architectures. *arXiv* **2018**, arXiv:1807.10108.
27. Ditria, E.M.; Sievers, M.; Lopez-Marcano, S.; Jinks, E.L.; Connolly, R.M. Deep learning for automated analysis of fish abundance: The benefits of training across multiple habitats. *Environ. Monit. Assess.* **2020**, *192*, 698. [CrossRef] [PubMed]
28. 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] [PubMed]
29. Wang, H.; Wu, X.; Huang, Z.; Xing, E.P. High-frequency component helps explain the generalization of convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8684–8694.

30. Chen, G.; Peng, P.; Ma, L.; Li, J.; Du, L.; Tian, Y. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 458–467.
31. Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; Madry, A. Adversarial examples are not bugs, they are features. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
32. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.
33. Lee, C.Y.; Batra, T.; Baig, M.H.; Ulbricht, D. Sliced wasserstein discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10285–10295.
34. Fu, Z.; Wang, W.; Huang, Y.; Ding, X.; Ma, K.K. Uncertainty inspired underwater image enhancement. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 465–482.
35. Han, J.; Shoeiby, M.; Malthus, T.; Botha, E.; Anstee, J.; Anwar, S.; Wei, R.; Petersson, L.; Armin, M.A. Single underwater image restoration by contrastive learning. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 2385–2388.
36. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
37. 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]
38. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
39. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
40. Zhang, H.; Chang, H.; Ma, B.; Wang, N.; Chen, X. Dynamic R-CNN: Towards high quality object detection via dynamic training. In Proceedings of the European Conference Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 260–275.
41. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
42. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pvt v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media* **2022**, *8*, 415–424. [CrossRef]
43. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF international Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578.
44. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9650–9660.
45. Zhang, S.; Wang, X.; Wang, J.; Pang, J.; Lyu, C.; Zhang, W.; Luo, P.; Chen, K. Dense distinct query for end-to-end object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7329–7338.
46. Chen, S.; Sun, P.; Song, Y.; Luo, P. Diffusiondet: Diffusion model for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 19830–19843.
47. Everingham, M.; Eslami, S.A.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
48. Chen, X.; Fang, H.; Lin, T.Y.; Vedantam, R.; Gupta, S.; Dollár, P.; Zitnick, C.L. Microsoft coco captions: Data collection and evaluation server. *arXiv* **2015**, arXiv:1504.00325.
49. Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; Yang, F. Musiq: Multi-scale image quality transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 5148–5157.
50. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. [CrossRef] [PubMed]
51. Huang, S.; Wang, K.; Liu, H.; Chen, J.; Li, Y. Contrastive semi-supervised learning for underwater image restoration via reliable bank. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18145–18155.