electronics

MDPI

Article

Image-Fused-Guided Underwater Object Detection Model Based on Improved YOLOv7

Zhenhua Wang [1,2], Guangshi Zhang [1,2], Kuifeng Luan [1,2], Congqin Yi [1,2] and Mingjie Li [2,3,4,*]

[1] College of Information Science, Shanghai Ocean University, Shanghai 201306, China; zh-wang@shou.edu.cn (Z.W.); m220951662@st.shou.edu.cn (G.Z.); kfluan@shou.edu.cn (K.L.); cqyi@shou.edu.cn (C.Y.)
[2] Key Laboratory of Marine Environmental Survey Technology and Application, Ministry of Natural Resources, Guangzhou 510300, China
[3] South China Sea Institute of Planning and Environmental Research, State Oceanic Administration, Guangzhou 510310, China
[4] Technology Innovation Center for South China Sea Remote Sensing, Surveying and Mapping Collaborative Application, Ministry of Natural Resources, Guangzhou 510310, China
* Correspondence: limingjie@scs.mnr.gov.cn

Abstract: Underwater object detection, as the principal means of underwater environmental sensing, plays a significant part in the marine economic, military, and ecological fields. Due to the degradation problems of underwater images caused by color cast, blurring, and low contrast, we proposed a model for underwater object detection based on YOLO v7. In the presented detection model, an enhanced image branch was constructed to expand the feature extraction branch of YOLOv7, which could mitigate the feature degradation issues existing in the original underwater images. The contextual transfer block was introduced to the enhanced image branch, following the underwater image enhancement module, which could extract the domain features of the enhanced image, and the features of the original images and the enhanced images were fused before being fed into the detector. Focal *EIOU* was adopted as a new model bounding box regression loss, aiming to alleviate the performance degradation caused by mutual occlusion and overlapping of underwater objects. Taking URPC2020 and UTDAC2020 (Underwater Target Detection Algorithm Competition 2020) datasets as experimental datasets, the performance of our proposed model was compared against with other models, including YOLOF, YOLOv6 v3.0, DETR, Swin Transformer, and InternImage. The results show that our proposed model presents a competitive performance, achieving 80.71% and 86.32% in mAP@0.5 on URPC2020 and UTDAC2020, respectively. Comprehensively, the proposed model is capable of effectively mitigating the problems encountered in the task of object detection in underwater images with degraded features and exhibits great advancement.

Keywords: underwater image; deep learning; object detection; YOLOv7

Citation: Wang, Z.; Zhang, G.; Luan, K.; Yi, C.; Li, M. Image-Fused-Guided Underwater Object Detection Model Based on Improved YOLOv7. *Electronics* **2023**, *12*, 4064. https://doi.org/10.3390/electronics12194064

Academic Editor: Stefanos Kollias

Received: 28 August 2023
Revised: 21 September 2023
Accepted: 22 September 2023
Published: 27 September 2023

1. Introduction

Object detection in underwater images is of importance for marine resource exploration, benthic monitoring, seafloor geomorphology observation, and deep-sea archaeology. Differing from generic images accessed in airspace, underwater images universally suffer from severe degradation problems caused by color cast, blurring, and low contrast, which could cause challenges for object detection in underwater images [1]. Therefore, an automatic underwater detection model is of crucial significance and practice, to enhance the accuracy and reduce the time-consuming nature of underwater object observation.

Underwater image enhancement could mitigate the degradation issues of underwater images and attracted extensive attention in past research, especially in the fields of computer vision and digital image processing [2]. Underwater image enhancement method aims to correct the blurred, color-distorted images to enhance their visual quality [3,4], which

can be generally summarized into four classes: optical-principles-based method, physical-degradation-model-based method, non-physical-models-based method, and deep-learning-based method.

An optical-principles-based method utilizes the imaging properties of underwater images. An example of this is through the construction of Stokes vectors containing polarization information to characterize the transmission of light under water. Hu et al. [5] proposed an improved method of correcting the transmittance based on the underwater polarization imaging model to alleviate the incorrectly calculated object irradiance due to the effect of polarization.

A physical-degradation-model-based method needs to estimate unknown parameters based on the constructed physical model with the help of certain a priori knowledge, in order to inversely deduce the attenuation factor of the affected optical component containing information about the object. The dark channel prior (DCP) algorithm was proposed by He et al. [6] as early as 2011, and was based on a priori fact that the intensity of the dark channel tends to be zero in fog-free images, which is based on the foggy sky imaging model and estimates the light transmission map to remove foggy blur in the image. UDCP [7,8] for underwater images, abandons the unreliable red channel, and only considers the use of values in green channel and blue channel to estimate the transmission rate of underwater light accurately.

A non-physical-model-based method resorts to changing the color distribution and radiation intensity of an image in a certain color space by stretching the gray values to achieve the color balance of the image. One such method is based on the Retinex model, which decomposes the image into components containing detailed features of the object itself and the ambient luminance component, and is able to dynamically adjust the edge details, contrast, and color simultaneously due to its nonlinear computation. Fu et al. [9] proposed a Retinex-model-based method that can decompose the image in the illumination and reflectance components, which carries detail information about the object and introduces a new shrinkage factor to assist in the estimation of two unknown components. Zhang et al. [10] proposed a method that maps the image color to the HSV color space after color correcting the original image and gamma-corrects the decomposed components with the Retinex model in order to restore the image's true appearance.

Deep-learning-based methods faces severe challenges in building large-scale datasets, and it is difficult to acquire degraded images and corresponding clear reference images of the same scene [11]. Li proposed twin adversarial contrastive learning [12], which intends to learn a mapping that reflects the relationship between underwater and airborne domain images with both self-supervised and unsupervised processing to alleviate the limitation of the dataset.

The above underwater image enhancement methods enhance the quality of the underwater images, but usually they require a large amount of time due to the complex processing flow, which represents a higher time complexity of processing a single image. Therefore, meeting the requirements of end-to-end detection tasks when combining with deep-learning-based object detection models is challenging.

Deep learning has advanced considerably in the last decade, and some representative generic deep-learning-based object detection models have been developed. Roughly, the deep-learning-based object detection models could be categorized into anchor-based methods and anchor-free methods. Anchor-based methods include Cascade R-CNN [13], YOLOv2 [14], YOLOv3 [15], RetinaNet [16], etc. Such methods assign predefined anchor frames to the detection layer to effectively reduce the search range of the objects and simplify the positive and negative sample matching problem. Anchor-free methods include FCOS [17], CenterNet [18], CornerNet [19], FSAF [20], YOLOX [21], etc., which only need to calculate the center point of the bounding box and position coordinates compared to the pre-set anchor scale and aspect ratio.

On this basis, deep-learning-based object detection models for underwater images face more challenges. In underwater scenes, the water body selectively absorbs different

wavelengths of light with depth, i.e., the less energetic red light is absorbed firstly, which makes the underwater images or video data often present a bluish or greenish color cast. At the same time, the scattering of light by suspended or particulate impurities in water can cause the image to appear foggy and blurred, resulting in a lower contrast. Therefore, existing deep-learning-based object detection models often exhibit enormous limitations when applied directly to object detection in underwater images.

For the object detection in underwater images, some researchers have turned their efforts to construct models that contain both an enhancement part and a detection part, in order to seek the intrinsic connection between the enhanced output results and the input of the detection process. The framework proposed by Liu et al. [22] guides the image features input to the detector by calculating the similarity between the enhanced image feature maps and the original degraded image feature maps. The general underwater object detector (GUOD) [23] incorporates a domain-invariant module in the YOLOv3-based detector with an adversarial training approach to take full advantage of the semantic information of images in different kinds of underwater circumstance, enabling it to exhibit favorable generalizability in different datasets. Joint training of an image enhancement network and a feature extraction network can effectively mitigate the inconsistency of goals between the enhancement tasks and detection tasks. A lightweight network is proposed by Yeh et al. [24], containing a color correction network and an object detection network to simultaneously compute the total loss of both during backpropagation to achieve the unification of the objectives. Fu et al. [25] are devoted to constructing a trainable residual feature transference module to learn mappings between detector-friendly images with heavy degradation and other generic images. Considering the limitation of mining only the features contained in the spectral properties of the underwater images, a 3D convolution was introduced to extract the spatial properties of images as the means of estimating depth [26]. Zhou et al. [27] introduced efficient channel attention (ECA) modules and deep hyperparametric convolution (DO-Conv) as the backbone network to extract semantic information from deep-sea images in their proposed YOLOTrashCan, and designed a convolution module with multi-scale dilation rate in the feature fusion stage to adapt to objects of different sizes.

Most of the existing underwater image object detection models consider the consistency of the image enhancement task and the object detection task [28]. However, the joint training of image enhancement part and feature extraction part leads to enormous parameter computation and resource consumption, which is a great challenge for industrialized applications. At the same time, the intensive distribution and mutual overlapping of underwater objects are unavoidable [29]. This raises challenges for the model's performance.

To resolve the issues above, we proposed an underwater image detection model based on YOLOv7 [30] that fuses enhanced image features.

2. Underwater Object Detection Model

Figure 1 shows the network structure of the proposed underwater object detection model, including an enhanced image branch and a feature extraction branch. The enhanced image branch was constructed to strengthen the feature extraction capability of YOLOv7. In the enhanced image branch, the features of the original underwater images were enhanced by an underwater image enhancement module (as shown in Figure 1 (a)). After that, the features of the enhanced underwater images (output of the underwater image enhancement module) were learned and extracted by the contextual transfer block (CoT) [31] (as shown in Figure 1 (b)). Then all features extracted by the enhanced image branch and the feature extraction branch were fused and fed into the detector of YOLOv7. Aiming to alleviate the reduction in performance due to mutual occlusion and overlapping of underwater objects, the loss function was optimized by focal *EIOU* [32], which was considered as a new model bounding box regression loss (as shown in Figure 1 (c)).

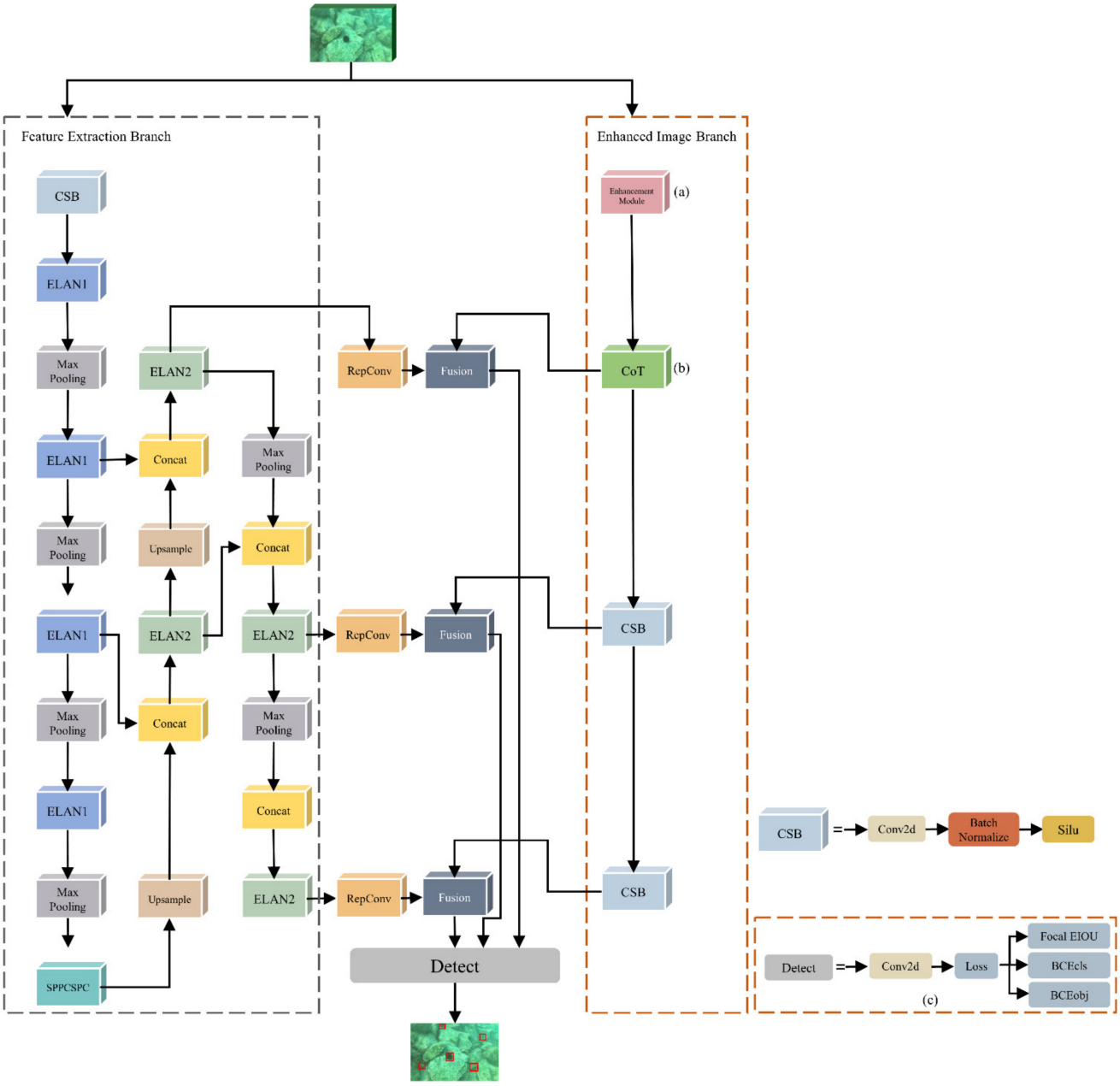

Figure 1. Network structure of the proposed underwater object detection model. (a) Underwater image enhancement module; (b) Contextual transfer block; (c) Focal EIOU.

2.1. Underwater Image Enhancement Module

Due to the submarine physical environment, underwater images undergo severe feature degradation on account of absorption and scattering of light by the water, and exhibit serious color cast, blurring, and low contrast issues. For the solution of the feature degradation issues of underwater images, the underwater image enhancement module was added in the enhanced image branch, which could enhance the degraded image features.

The underwater image enhancement module is designed by deep learning and image formation model [33], represented as

$$J(\chi) = (I(x) - A)e^{\lambda_J d} + Ae^{(\lambda_J - \lambda_A)d}, \quad (1)$$

where $J(x)$ is the radiant intensity of the object itself, which is the energy representation after filtering out the environmental effects. x represents the pixel point at each location of

the image. d represents energy attenuation distance, which is the distance between object and imaging equipment. λ_J and λ_A represent the attenuation coefficients of the object radiant intensity and the ambient radiation, respectively. The image formation model is defined as

$$I(x) = J(x)t(x) + A(1 - t(x)), \tag{2}$$

where $t(x)$ represents the transmission rate of the energy transmission medium, which is expressed as

$$t(x) = e^{-\lambda d(x)}, \tag{3}$$

where λ denotes the attenuation coefficient, and $A(1 - t(x))$ denotes the back scattering effect of the water body.

Since the underwater image enhancement module is not involved in the training process, it could decrease the number of parameters involved in the training process and reduce the time required for the model inference. The output tensor of the underwater image enhancement module is received by the contextual transformer block.

2.2. Contextual Transformer Block

Based on the enhanced images, the contextual transformer block (CoT) is utilized to extract domain features, which could compensate for the detail and texture information.

The output tensor of the underwater image enhancement module, as shallow features, contains more pixel point information, such as edge and arris features. Due to the small receptive field of the shallow feature map, it is hard to fuse with the high-dimensional features extracted by the feature extraction branch, which contains more semantic information. While the CoT is taken as an intermediate stage of enhanced image feature extraction, the feature distribution of the CoT output tensor has a moderate difference with the high-dimensional features, which could compensate for the lost information of the degraded image features effectively.

Figure 2 shows the structure of the contextual transfer block (CoT); an extra convolutional layer is added to this module. The extra convolutional layer could fully utilize the enhanced features for extracting.

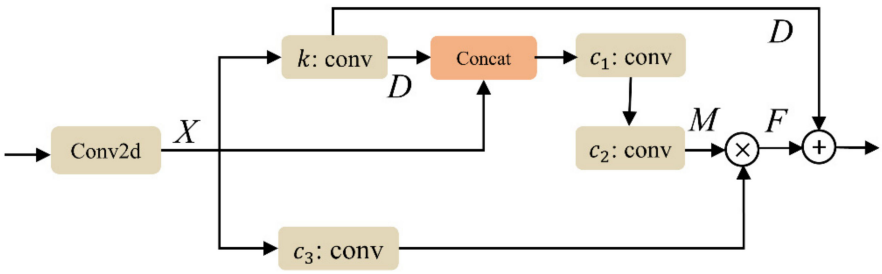

Figure 2. Structure of the contextual transfer block (CoT). × refers to the Hadamard product.

The output tensor of the underwater image enhancement module passes through the extra convolutional layer without activation function, and then is transformed into appropriate dimensions and received by the CoT. The output of the extra convolutional layer is represented as X.

The CoT utilizes a 4×4 group convolution operation to extract the key domain feature mapping at the 4×4 positions of X, while the output of group convolutional operation is represented by D. A concatenation operation is applied to X and D. Then, the result of concatenation operation is fed into the next two consecutive convolutional layers for smoothing of features and the output is represented by M. X is multiplied by M after a convolution layer to obtain F. The final output of the CoT is the result of adding F and D. The final output of the CoT can be expressed by

$$O = F + (X * K), \tag{4}$$

where X denotes the output of the extra convolutional layer, K is the convolutional layers, k and F can be expressed by

$$F = [(X * K, X) * c_1 * c_2] \times (X * c_3), \tag{5}$$

where c_1, c_2, and c_3 are the convolutional layers c_1, c_2, and c_3 respectively, and $\times$ represents the Hadamard product.

2.3. Focal EIOU

CIOU loss, the loss function of YOLOv7, provides a comprehensive review on the model's prediction results in three aspects, including the accuracy of the bounding box regression, the prediction results of the positive samples' categories, and the confidence of the positive samples' objectiveness. *CIOU* is used to calculate bounding box regression loss, describing the difference between the calculated bounding box by the model and the ground truth in the overlap region, coordinates of the center point, and aspect ratio. *CIOU* is represented as

$$L_{CIOU} = 1 - CIOU, \tag{6}$$

where *CIOU* is defined as

$$CIOU = IOU - \left(\frac{\rho^2(b, b_{gt})}{c^2} + \alpha v \right), \tag{7}$$

where v denotes the deviation among the predicted bounding box and the ground truth in terms of aspect ratio. However, v neglects the exact consistency of widths and heights among the predicted bounding box and ground truth.

In underwater images, there is a wide distribution of intensive, mutually obscuring objects in the same category. However, objects in the same category exhibit similar shapes to each other and have similar aspect ratios. Thus, the exact size difference between objects needs to be taken into account rather than the difference in aspect ratio.

To resolve this issue, *CIOU* loss is optimized by *EIOU* loss, which could address the accuracy of the model's predicted bounding box and is able to more effectively and rationally characterize the specific location, scale, and width–height consistency between the regression bounding box and the ground truth. While there are many low-quality boxes in underwater images, the predictions by the low-quality boxes could feed back more gradient information [34], which would introduce some bias to the training results. Thus, the focal loss is utilized and is expected to better balance the positive and negative samples of *EIOU* loss further.

Thus, the loss function of the proposed underwater detection model can be represented as

$$L_{total} = \lambda_1 L_{FocalEIOU} + \lambda_2 L_{BCEcls} + \lambda_3 L_{BCEobj}, \tag{8}$$

where λ_1, λ_2, and λ_3 represent weight coefficients of these three loss functions and are set to 0.05, 0.125, and 0.1, respectively. $L_{FocalEIOU}$ denotes further calculations in terms of focal loss on the basis of *EIOU* loss and is expected to focus more on positive samples with higher accuracy and keep the model sensitive to them, which can be defined as

$$L_{FocalEIOU} = IOU^{\gamma} L_{EIOU}, \tag{9}$$

where γ is set as 0.5 to balance the numbers of positive and negative samples. L_{EIOU} represents the optimized loss function with *EIOU*. The *EIOU* loss is represented as

$$L_{EIOU} = 1 - IOU + \frac{\rho^2(b, b_{gt})}{(w_c)^2 + (h_c)^2} + \frac{\rho^2(w, w_{gt})}{(w_c)^2} + \frac{\rho^2(h, h_{gt})}{(h_c)^2}, \tag{10}$$

where w_c and h_c indicate the width and height of the smallest rectangle enclosing the two boxes, respectively. b and b_{gt} represent the coordinates of the center point of the prediction and the ground truth. w and h denote the width and height of the prediction, while w_{gt} and h_{gt} are the ground truth width and height information.

3. Datasets and Experiment Setup

3.1. Datasets

Experimental datasets employed in the current study included 1200 underwater images, selected from the UCPR2020 dataset [35]. In the experimental images, there were several species of benthos, such as echinus, starfish, holothurian, and scallop. As shown in Figure 3, these images had serious quality problems, including color deviation, blurring, low contrast, and overlapped objects. Another dataset used for experiments is UTDAC2020 from the underwater target detection algorithm competition in 2020, which contains a total of 6461 images in the same four categories as UCPR2020.

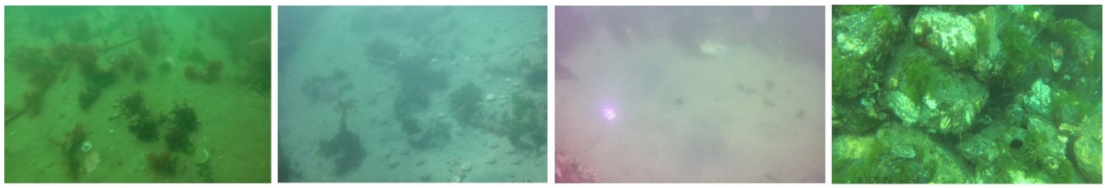

Figure 3. Samples selected from UCPR2020 dataset.

3.2. Experiment Setup

The hardware configuration used for the experiment is shown below: Ubuntu 20.04, a GPU NVIDIA RTX 2080ti, which includes 11 GB memory. Software environment was deep learning framework Pytorch 2.0.0 and Python 3.8. The data used for the ablation experiment and comparison experiment consisted of 1200 randomly selected underwater degradation images from UCPR2020, of which 80% (960 images) were randomly selected as the training set, 10% (120 images) as the validation set, and the remaining 10% (120 images) as the test set. A total of 5168 images from the UTDAC2020 dataset were set as the training set, 693 as the validation set, and 600 as the test set. Comparison experiments were implemented on both the UCPR2020 dataset and the UTDAC2020 dataset in order to obtain a comprehensive set of results.

3.3. Evaluation Metrics

Five metrics were calculated to estimate the performance of the underwater detection model: precision, recall, F1 score, mean average precision (mAP) [36], and GFLPOs [37]. Precision is defined as the ratio of true positive to the sum of true positive and false positive detected objects. Recall measures the proportion of positive detected objects by the model out of all positive objects. F1 score combines precision and recall to evaluate model performance in a more encompassing dimension.

$$Precision = \frac{TP}{TP + FP}, \tag{11}$$

$$Recall = \frac{TP}{TP + FN}, \tag{12}$$

$$F1\_score = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{13}$$

where TP, FP, and FN denote the true positive detected objects, false positive detected objects, and false negative detected objects, respectively.

For multiple categories of detection tasks, the mean average precision is defined as

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i, \tag{14}$$

where N is the total number of categories in the dataset and AP_i refers to the average precision of category i. The metrics mAP@0.5 describes the mAP under the condition that a positive sample is computed when the intersection and concatenation ratio is greater than 0.5.

Giga floating point operations (GFLOPs) were calculated to measure the computational effort of models, denoting the computational complexity of the model.

4. Experiment

To evaluate the performance of the underwater object detection model, two comparative experiments were performed. Experiment one was an ablation experiment. Experiment two was a comparative experiment, where the detection performance of the proposed model was compared against other models, including YOLOF [38], YOLOv6 v3.0 [39], DETR [40], Swin Transformer [41], and InternImage [42].

4.1. Ablation Experiment

The proposed model was compared against baseline with different settings, including baseline with enhanced image fusion, baseline with enhanced image fusion and CoT, baseline with enhanced image fusion, CoT, and focal *EIOU*, where the baseline was set to YOLOv7. Table 1 shows the results of evaluation metrics for the baseline with different settings on the UCPR2020 dataset. For baseline, direct processing of degraded scene images does not fully utilize the model performance. There is a significant improvement in the detection accuracy of the model after incorporating the enhanced image features. mAP @0.5 increases from 62.48% to 75.52%, which is an improvement of 13.04%, and after adding CoT following the enhanced image, the model performance has a slight improvement, with an increase of 2.24% (from 75.52% to 77.76%). Compared to the original bounding box regression loss, after using focal *EIOU* as the optimized bounding box loss function, the mAP@0.5 increases from 77.76% to 80.71%, which is an improvement of 2.95%. With the incorporation of a different module, the computational complexity of the model increases slightly compared to before its use.

Table 1. Evaluation results between of YOLOv7 with different settings on UCPR2020 dataset. √ indicates that the module was used in the experiment.

Model				Precision	Recall	F1 Score	mAP@0.5	GFLOPs
Baseline	Enhanced Image Fusion	CoT	Focal EIOU					
√				73.42%	55.63%	0.63	62.48%	110.35
√	√			80.55%	68.83%	0.74	75.52%	121.47
√	√	√		83.04%	71.64%	0.77	77.76%	180.51
√	√	√	√	84.31%	74.92%	0.79	80.71%	282.05

4.2. Comparison Experiment

To evaluate the performance of underwater object detection, the proposed model was compared with other models, including YOLOF, YOLOv6 v3.0, DETR, Swin Transformer, and InternImage. Tables 2 and 3 show the comparison of underwater object detection evaluation metrics among different models, and Figure 4 shows the detection results by different models.

Table 2. Comparison with other models on UCPR2020 dataset.

Model	Precision	Recall	F1 Score	mAP@0.5	GFLOPs
YOLOF	69.89%	54.34%	0.61	55.72%	98.73
YOLOv6 v3.0	74.33%	71.78%	0.73	62.81%	269.84
DETR	74.28%	60.09%	0.66	60.34%	188.72
Swin Transformer	71.13%	65.50%	0.68	62.45%	351.85
InternImage	73.26%	61.42%	0.67	61.93%	216.37
Ours	84.31%	74.92%	0.79	80.71%	282.05

Table 3. Comparison with other models on UTDAC2020 dataset.

Model	Precision	Recall	F1 Score	mAP@0.5	GFLOPs
YOLOF	71.03%	60.14%	0.65	56.54%	98.73
YOLOv6 v3.0	74.88%	69.37%	0.72	75.94%	269.84
DETR	78.96%	72.70%	0.76	76.13%	188.72
Swin Transformer	81.70%	73.29%	0.77	77.50%	351.85
InternImage	79.44%	70.16%	0.75	76.02%	216.37
Ours	82.71%	80.74%	0.82	86.32%	282.05

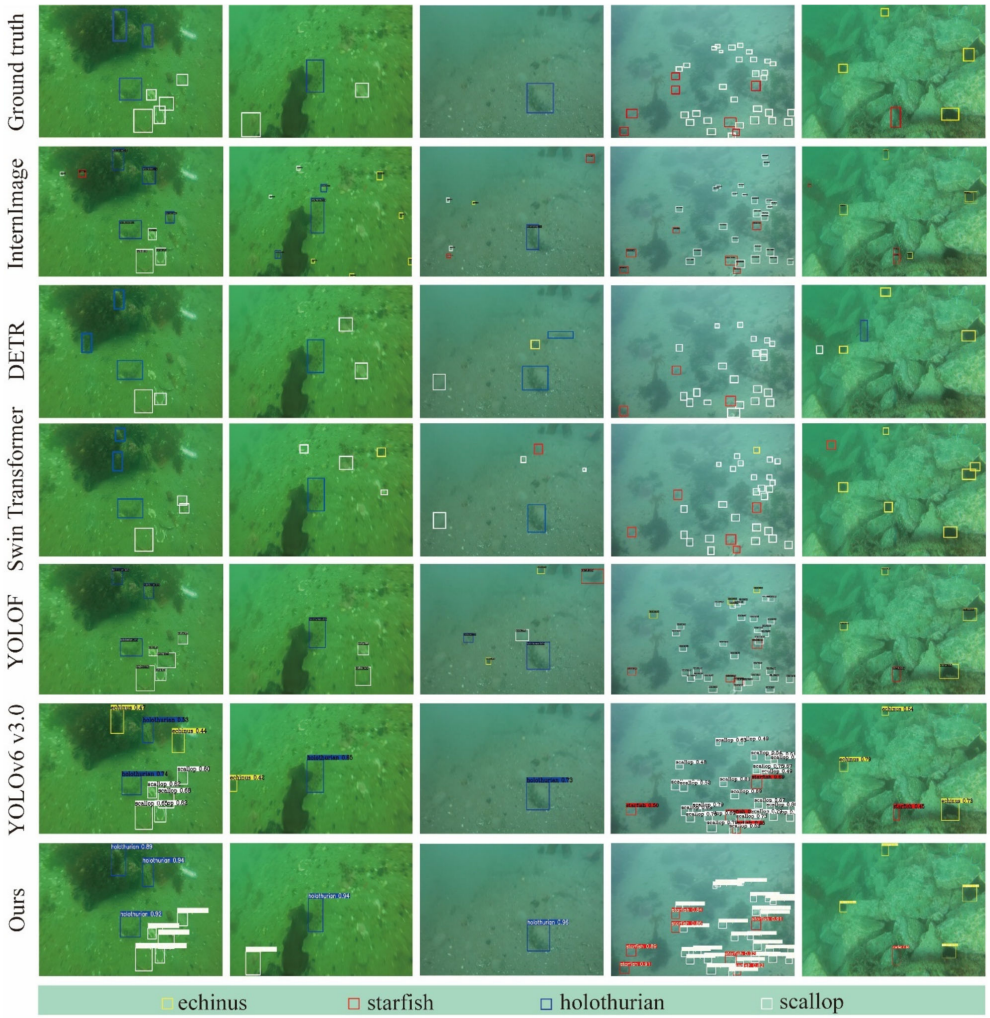

Figure 4. Detection results of different methods on the UCPR2020 dataset. The yellow, red, blue, and white rectangles denote the predicted object categories of echinus, starfish, holothurian, and scallop respectively.

As shown in Figure 4 and Tables 2 and 3, object detection results of our proposed model are close to the ground truth. Some objects that are more similar to the background are more likely to be mis-detected, especially for YOLOv6 v3.0. For Swin Transformer and InternImage, there are more background objects detected falsely as the positive objects. For YOLOF and DETR, there both are missed detections of difficult samples and false detection of background objects. The proposed model achieves the best performance in mAP@0.5. The computational complexity of the proposed model is higher than YOLOF, DETR, InternImage, and YOLOv6 v3.0, while being lower than Swin Transformer.

5. Discussion and Conclusions

Underwater object detection is of great significance for underwater biological resources assessment and ecological environment detection. With the rapid development of monitoring equipment, underwater images have become the main data source for monitoring underwater objects. Considering the limitations of feature degradation in underwater images, this paper proposed an underwater object detection model based on YOLOv7. Firstly, an enhanced image branch was constructed to enhance the feature extract ability of YOLOv7, which consisted of an underwater image enhancement module and a contextual transfer block, while focal *EIOU* was adopted as a new model bounding box regression loss to alleviate the degradation problem.

URPC2020 and UTDAC2020 were taken as experimental datasets, and the proposed underwater object detection model was compared against with other models, including YOLOF, YOLOv6 v3.0, DETR, Swin Transformer, and Intern Image. The results show the proposed model achieves automatic and higher accuracy detection results in underwater images.

However, the proposed model does not guarantee to enhance the original degraded image in a detector-friendly style because the constructed image enhancement module is not involved in the training, which sacrifices the consistency of the enhancement task and the detection task to some extent. In addition, the inference speed of the model still finds it difficult to cope with complex video data, and further optimization of the model's scale and the amount of parameter computation also need to be considered. Besides this, underwater object detection also faces several other challenges. The number of underwater object categories is much larger than that of terrestrial plants and animals, and some of the objects of different categories show similar features, so how to accurately grasp the feature differences between similar objects is a potential problem. Meanwhile, the establishment of a large-scale and sufficient sample set containing different categories to fully exploit the potentialities of deep-learning-based models of underwater object detection needs to be further researched. Moreover, the optimization for underwater object detection models focuses more on solving the problem of the feature extraction process, and the improvement of the inference speed of the model is usually neglected.

Author Contributions: Conceptualization, Z.W. and G.Z.; methodology, G.Z. and K.L.; validation, C.Y. and Z.W.; formal analysis, Z.W., K.L. and M.L.; investigation, G.Z. and K.L.; writing—original draft preparation, G.Z. and Z.W.; writing—review and editing, Z.W., K.L. and M.L.; supervision, Z.W.; funding acquisition, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was generously supported by Key Laboratory of Marine Environmental Survey Technology and Application, Ministry of Natural Resources (grant no. MESTA-2021-B007), and by a grant from the Capacity Development for Local College Project (grant no. 19050502100).

Data Availability Statement: There was no new data created.

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

References

1. Wang, J.; He, X.; Shao, F.; Lu, G.; Jiang, Q.; Hu, R.; Li, J. A novel attention-based lightweight network for multiscale object detection in underwater images. *J. Sens.* **2022**, *2022*, 582687. [CrossRef]
2. Zhou, J.; Sun, J.; Zhang, W.; Lin, Z. Multi-view underwater image enhancement method via embedded fusion mechanism. *Eng. Appl. Artif. Intell.* **2023**, *121*, 105946. [CrossRef]
3. Zhang, W.; Zhuang, P.; Sun, H.-H.; Li, G.; Kwong, S.; Li, C. Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement. *IEEE Trans. Image Process.* **2022**, *31*, 3997–4010. [CrossRef]
4. Jiang, Z.; Li, Z.; Yang, S.; Fan, X.; Liu, R. Target oriented perceptual adversarial fusion network for underwater image enhancement. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6584–6598. [CrossRef]
5. Hu, H.; Zhao, L.; Huang, B.; Li, X.; Wang, H.; Liu, T. Enhancing visibility of polarimetric underwater image by transmittance correction. *IEEE Photonics J.* **2017**, *9*, 6802310. [CrossRef]
6. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 2341–2353.
7. Drews, P.; Nascimento, E.; Moraes, F.; Botelho, S.; Campos, M. Transmission estimation in underwater single images. In Proceedings of the IEEE international Conference on Computer Vision Workshops, Sydney, Australia, 1–8 December 2013; pp. 825–830.
8. Drews, P.L.; Nascimento, E.R.; Botelho, S.S.; Campos, M.F.M. Underwater depth estimation and image restoration based on single images. *IEEE Comput. Graph. Appl.* **2016**, *36*, 24–35. [CrossRef]
9. Fu, X.; Zhuang, P.; Huang, Y.; Liao, Y.; Zhang, X.-P.; Ding, X. A retinex-based enhancing approach for single underwater image. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 4572–4576.
10. Zhang, W.; Li, G.; Ying, Z. A new underwater image enhancing method via color correction and illumination adjustment. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4.
11. 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]
12. Liu, R.; Jiang, Z.; Yang, S.; Fan, X. Twin adversarial contrastive learning for underwater image enhancement and beyond. *IEEE Trans. Image Process.* **2022**, *31*, 4922–4936. [CrossRef]
13. 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.
14. 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.
15. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
16. 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.
17. 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, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
18. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578.
19. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
20. Zhu, C.; He, Y.; Savvides, M. Feature selective anchor-free module for single-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 840–849.
21. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
22. Liu, H.; Jin, F.; Zeng, H.; Pu, H.; Fan, B. Image Enhancement Guided Object Detection in Visually Degraded Scenes. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**. [CrossRef]
23. Liu, H.; Song, P.; Ding, R. Towards domain generalization in underwater object detection. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 1971–1975.
24. 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]
25. Fu, C.; Fan, X.; Xiao, J.; Yuan, W.; Liu, R.; Luo, Z. Learning Heavily-Degraded Prior for Underwater Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2023**. [CrossRef]
26. Li, Q.; Li, J.; Li, T.; Li, Z.; Zhang, P. Spectral-Spatial Depth-Based Framework for Hyperspectral Underwater Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4204615. [CrossRef]
27. Zhou, W.; Zheng, F.; Yin, G.; Pang, Y.; Yi, J. YOLOTrashCan: A Deep Learning Marine Debris Detection Network. *IEEE Trans. Instrum. Meas.* **2022**, *72*, 1–12. [CrossRef]
28. Zhang, J.; Zhu, L.; Xu, L.; Xie, Q. Research on the correlation between image enhancement and underwater object detection. In Proceedings of the 2020 Chinese Automation Congress (CAC), Shanghai, China, 6–8 November 2020; pp. 5928–5933.

29. Liu, Z.; Zhuang, Y.; Jia, P.; Wu, C.; Xu, H.; Liu, Z. A novel underwater image enhancement algorithm and an improved underwater biological detection pipeline. *J. Mar. Sci. Eng.* **2022**, *10*, 1204. [CrossRef]
30. 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, 18–22 June 2023; pp. 7464–7475.
31. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual transformer networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1489–1500. [CrossRef]
32. Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
33. Chen, X.; Zhang, P.; Quan, L.; Yi, C.; Lu, C. Underwater image enhancement based on deep learning and image formation model. *arXiv* **2021**, arXiv:2101.00991.
34. 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.
35. 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 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shenzhen, China, 5–9 July 2021; pp. 1–6.
36. Padilla, R.; Netto, S.L.; Da Silva, E.A. A survey on performance metrics for object-detection algorithms. In Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niteroi, Brazil, 1–3 July 2020; pp. 237–242.
37. Molchanov, P.; Tyree, S.; Karras, T.; Aila, T.; Kautz, J. Pruning convolutional neural networks for resource efficient inference. *arXiv* **2016**, arXiv:1611.06440.
38. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You only look one-level feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13039–13048.
39. Ronkin, M.V.; Akimova, E.N.; Misilov, V.E. Review of deep learning approaches in solving rock fragmentation problems. *AIMS Math.* **2023**, *8*, 23900–23940. [CrossRef]
40. 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.
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.; Dai, J.; Chen, Z.; Huang, Z.; Li, Z.; Zhu, X.; Hu, X.; Lu, T.; Lu, L.; Li, H. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 14408–14419.