Underwater Biological Target Detection Algorithm and Research Based on YOLOv7 Algorithm

Hongwei Zhuang, Weisheng Liu

***Abstract*—Underwater target detection is an important method for detecting marine organisms. However, due to the image occlusion of underwater targets, blurred water quality, poor lighting conditions, small targets, and complex backgrounds, the detection of underwater biological targets has posed significant challenges. In the intricate underwater environment, the conventional feature extraction method has a few drawbacks, including imprecise feature extraction, sluggish detection speed, and inadequate robustness. Consequently, an underwater target detection method based on the enhanced You Only Look Once 7 (YOLOv7) is proposed in this study. The network architecture is reconstructed, and the Deformable Convolutional Network (DCN) modules replace some 3×3 convolutional blocks in the ELAN structure to offset sampling points and reduce background interference. Skip connections and 1× 1 convolutional architecture are added to the DCN module to improve the model's perception of image details. In addition, Contextual Transformer 3 (COT3) is also incorporated to improve visual performance. Finally, to improve the detection efficiency of small objects, the CIoU loss function is finally replaced by the Normalized Wasserstein Distance (NWD) algorithm. The mAP of DCCN-YOLOv7 on the URPC dataset is 80.4%, according to the experimental results, 2.8% higher than the YOLOv7 network model that is used as a baseline. Furthermore, in contrast to the original YOLOv7 algorithm, the detection speed and accuracy are higher, making it more appropriate for target recognition underwater.**

***Index Terms*—COT3, DCN, Loss function, Underwater target detection, YOLOv7**

I. Introduction

UNDERWATER biological monitoring is a crucial topic within the field of underwater target detection. Its primary objective is to locate and identify targets in underwater scenes, thereby offering valuable insights into the abundance of marine organisms and their response to environmental changes [1]. This field of study has garnered considerable interest owing to its broad range of applications in oceanography, underwater navigation and aquaculture [2]. For example, it enables autonomous and intelligent identification and analysis of seafood, such as sea cucumbers, scallops and other marine organisms, which has traditionally been done manually. Autonomous detection and accurate identification of seafood can help farmers manage growth and habitat changes. It can also free humans from labour-intensive and dangerous tasks. However, underwater target detection poses many challenges due to the complex environment and lighting conditions. While deep learning-based target detection systems have shown promising results in various applications, there are still significant technical hurdles within the specific domain of underwater target detection. Firstly, physical phenomena like light scattering and absorption in the underwater environment lead to distorted underwater images and incomplete extraction of submerged objects. This in turn reduces the accuracy of target detection [3]. Secondly, accurately extracting small objects from complex underwater environments poses a challenge, as underwater targets predominantly consist of small objects that frequently blend into intricate backgrounds.

Traditional target detection methods typically involve three main stages: region selection, feature extraction, and feature classification. Since the target can appear at any position in the image, with uncertain size and aspect ratio, traditional target detection methods face two main challenges: (1) the sliding window selection strategy lacks focus, resulting in high time complexity and redundant windows; and (2) the hand-designed features lack robustness.

The field of computer vision has been transformed by the emergence of deep learning, leading to significant progress in target detection. In particular, algorithms such as Faster RCNN [4] and SSD [5] have been widely applied. Among these methods, target detection using the YOLO (You Only Look Once) algorithm [6] is representative because of its efficiency and accuracy. It transforms target detection from a classification problem into a regression problem, thus enabling end-to-end detection. Efficient and accurate target detection capabilities are crucial for underwater robots to be commercially viable. Researchers have made significant contributions in this area. To tackle the issue of inconsistent scale features in multi-scale detection using the YOLOv3 algorithm, Li Chong [7] proposed an adaptive spatial feature fusion method. This method addresses the problem of inconsistent feature scaling in the YOLOv3 algorithm. This method spatially filters incoming information to suppress inconsistencies. Furthermore, the IOU loss function used is

Manuscript received June 10, 2023; revised April 11, 2024.

This work was supported by the Special Fund for Scientific Research Construction of University of Science and Technology Liaoning.

Hongwei Zhuang is a postgraduate student of the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: zhw17615148162@163.com).

Weisheng Liu is a professor of the College of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, CO 114051, China (corresponding author to provide fax: 0412-5929809; e-mail: succman@163.com).

the transgressive regression loss, which improves the localization accuracy of the YOLOv3 prediction frames. Li et al. [8] employed the densely connected YOLOv3 model to detect zooplankton in their natural environment. The aim was to minimize feature loss during network transmission and to address the issue of species distribution imbalance by using samples generated by CycleGAN. Chen et al. [9] presented SWIPENet, a network designed to detect underwater targets with small sample sizes. The proposed network incorporated an Inverted Multi-Class Adaboost (IMA) sample weighting algorithm, resulting in improved detection accuracy. Shi et al. [10] used the YOLOv4 algorithm to recognize sea cucumbers, sea urchins, scallops, and starfish. Their research demonstrated the effectiveness of the convolutional attention mechanism and data augmentation in improving detection accuracy. Cao et al. [11] proposed a detector that can detect live crabs underwater in real-time and in a robust manner. To achieve this, the detector utilized a single-shot multi-box detector (SSD) that had MobileNetV2 as its backbone. To improve performance, the traditional convolution operation was substituted with depth-separable convolution.

Although the previously mentioned detection algorithms improve detection accuracy, they are still not sufficiently high for small-scale targets. Underwater biological detection is challenging due to significant interference during image acquisition. In addition, mobile acquisition can lead to image distortion and reduced visibility. Therefore, improving the accuracy of underwater target detection is necessary to enhance portability. To tackle these challenges, we propose an experimentally validated model, DCCN-YOLOv7. When faced with complex stimuli, the human visual system quickly focuses on a target or region of interest. Inspired by this, we introduce the COT3 [12] module into the feature fusion model. This module includes a static context that receives a 3×3 convolution and a dynamic context based on situational self-attention. This promotes the learning of visual representations, thereby increasing the efficiency of the attention mechanism in deep learning tasks. Furthermore, the deformation convolution has been enhanced to shift the sampling points towards the foreground target. This adjustment minimizes background interference, enhancing feature extraction accuracy. To tackle the issue of detecting small targets, the Normalized Wasserstein Distance (NWD) regression loss [13] is used instead of the CIoU loss function [14] to more accurately measure the similarity between two bounding boxes (BBoxes), resulting in an algorithm that is more robust [15] and stable. The third and fourth sections discuss the algorithm structure and present the experimental results.

II. Related Work

A. YOLOv7 Model

YOLOv7 [16] represents a significant development within the YOLO family of algorithms. This algorithm is designed to detect targets in a single stage. Its main contribution is the compilation of several existing techniques, including the re-parameterisation of the modules and the strategies for the dynamic allocation of labels. Ultimately, this compilation outperforms all other known target detectors in both speed and accuracy. It achieves performance ranging from 5 to 160 Frames Per Second (FPS). During testing on the GPU V100, the model achieved an accuracy of 56.8% AP and a detection rate of over 30 FPS (batch=1). This is evidence that even today is high accuracy detectors are capable of over 30 FPS. Compared to other network models within the YOLO series, YOLOv7's detection approach is similar to that of YOLOv5 [17]. Both models use a deep convolutional neural network (CNN) for feature extraction. They share similarities in terms of multi-scale prediction and high efficiency.

B. Deformable Convolution

A fixed convolutional kernel is typically used in the convolutional operation that extracts features from the input data. However, when dealing with tasks involving deformed objects, the use of a fixed kernel becomes challenging because it cannot adapt well to things with varying degrees of deformation. This limitation leads to suboptimal feature extraction. While this approach effectively captures local feature correlations, it also results in equal weighting of foreground and background features. This limitation hinders the effectiveness of convolution in scenarios with complex objects.

Compared to traditional convolution, a two-dimensional offset is added to the original convolution kernel and sampled. This offset is applied in both the X and Y directions. This allows the sampling point to be moved to any position within the neighborhood. This effectively overcomes the limitations of insufficient sampling of a fixed rectangular structure. The feature map's receptive field is increased, improving the detection of complex objects.

C. Loss Function

In target detection networks, target localization heavily relies on a module that performs bounding box regression. The intersection over IoU loss function is commonly employed to assess the similarity between predicted and actual bounding boxes. It performs well when the bounding boxes have an overlapping area [18]. However, IoU has limitations, especially when one bounding box contains another. To address this issue, Generalized IoU [19] introduces the concept of the minimum bounding box, which includes both the predicted and actual boxes. This approach is employed to ascertain the weights within the bounding box. Nonetheless, GIoU degrades to IoU when one bounding box completely encloses another. Additional metrics such as DIoU [20] and CIoU [21] have been suggested to address the shortcomings of bounding box regression and improve its accuracy. DIoU and CIoU consider three geometric properties: overlap area, centroid distance, and aspect ratio. They aim to provide a more comprehensive evaluation of the bounding boxes. However, the CIoU loss function may not give satisfactory results for small targets. This is because the proximity of small targets often results in a short distance, which leads to a significant CIoU value, even with small differences in centroid and area ratios. This makes it difficult to effectively discriminate between them. This paper presents a more detailed representation of the loss function to address this issue. It allows for smoother convergence and improves

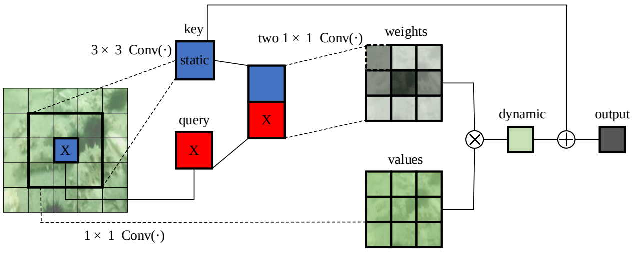

Fig. 2. Block COT3

the localization accuracy of target images. By effectively taking into account the characteristics of small target objects, the proposed approach aims to improve their detection.

III. Experimental Methods

This section details the improved design of the YOLOv7 model for detecting biological targets underwater. It introduces the DCN-ELAN structure, the COT3 module, and the NWD loss function. These improvements aim to enhance mutual occlusion feature extraction of organisms and small targets.

A. Deformable Convolution Improvement Module

In natural underwater environments, blurred background features can pose a challenge to accurate foreground detection. Standard convolutional operations may not effectively address this problem. In contrast, deformable convolution introduces a two-dimensional offset to the original convolution kernel, allowing adaptive sampling. These offsets are optimized by backpropagation. This approach improves the extraction of underwater biometric features while reducing the influence of unclear background features.

In underwater target detection, the shape and posture of underwater organisms can be affected by water flow, swimming and other factors, and the use of deformable convolution can better capture these dynamic changes. However, due to the limitations of a single DCN with fixed receptive field, when the boundary details and small local features of the target require more accurate positioning, a single DCN may not provide sufficient accuracy, and its learning ability is limited, which is unable to cope with complex target deformation and attitude changes. To solve this problem, we have made improvements to the DCN module structure, which is illustrated in Fig. 1. By integrating the residual structure into the DCN, adding skip connections and 1×1 convolutional branches, this design facilitate the network not only to process signals through the deformable convolutional layer, the batch normalization layer and the activation function layer, but also to add the final output to the original input. This mechanism enables the network to learn residual information between the input and output, thereby enhancing the transfer and extraction of features and improving the network’s capability to model spatial structure.

B. CoT3 Model

Convolutional neural networks are powerful models capable of learning discriminative visual representations. They have demonstrated remarkable performance across a spectrum of tasks, including image recognition, object detection, and mental state analysis. However, existing attention mechanisms often rely solely on independent query keys to compute the attention matrix, ignoring the valuable contextual information between neighboring keys. In contrast, the CoT block introduces a novel approach to enhance the capability of visual representation. The CoT block first uses a 3×3 convolutional operation to capture the static contextual information between keys. This mining process helps incorporate relevant information from neighboring keys. Then, based on the query and contextualized keys, two successive 1×1 convolutional operations are applied to compute self-attention and generate dynamic contexts. These dynamic contexts can adaptively capture critical relationships and generate informative representations. Finally, the static and dynamic contexts are combined to produce the output, improving the model's overall performance. An advantage of the CoT block is its compatibility with existing ResNet [22] architectures.

By adding a CoT3 block in the feature fusion stage, performance can be improved while keeping the parameter budget within a reasonable range. Fig. 2 provides a visual representation of the fusion process and illustrates how the CoT block improves the visual representation.

Although traditional self-attentive mechanisms can facilitate feature interactions across different spatial locations, they predominantly rely on single queries and key pairs for learning. This approach overlooks the rich contextual relationships between queries and other keys, resulting in a limited understanding of visual representations on 2D feature maps. In contrast, our context converter block [23] adopts a unified learning approach that encompasses contextual relationships between keys and self-attention on the feature graph, all while staying within a limited parameter budget. By incorporating contextual relationships into the learning process, our model can capture the dependencies and interactions among different variables, resulting in improved performance in the learning task. The context converter block enables the model to learn more comprehensive and informative visual representations [24]. The specific structure of the context converter block is illustrated in Fig. 3.

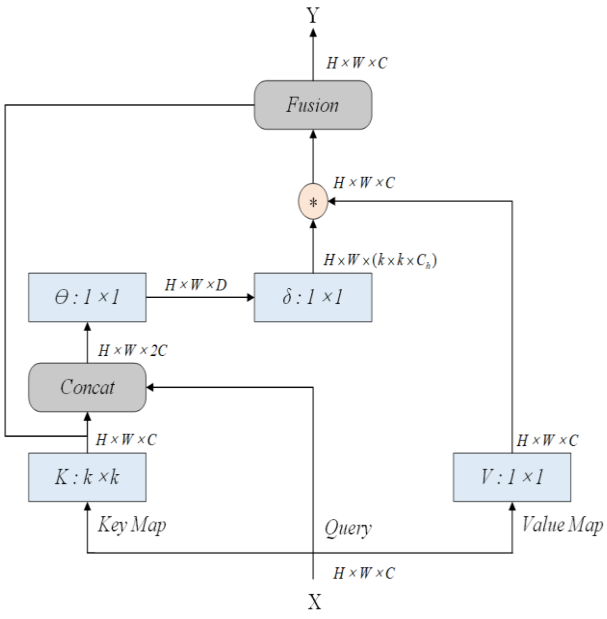

Fig. 3. Context Converter (CoT) block

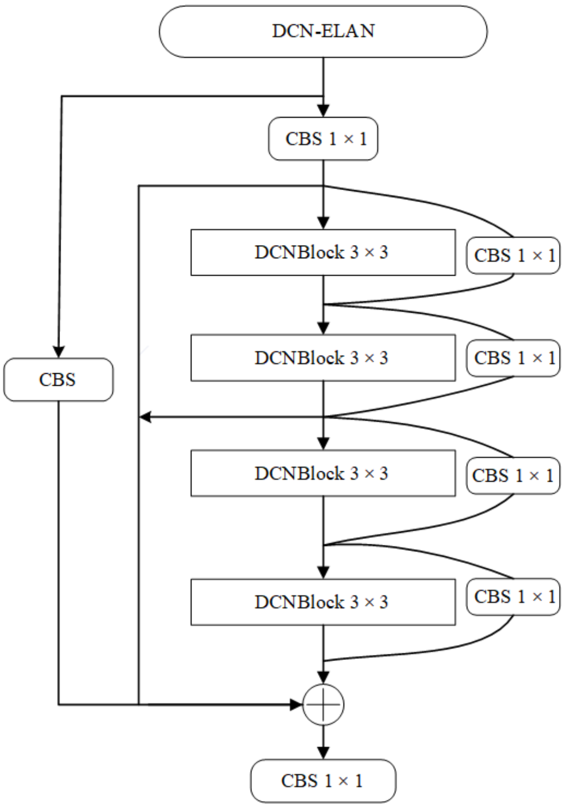

Fig. 1. Improved DCN module structure diagram

C. Loss Function

The loss function in the YOLOv7 network model is shown in Equation (1):

$$L_{obgect,Loss} = L_{loc,Loss} + L_{conf,Loss} + L_{class,Loss} \tag{1}$$

$L_{loc,Loss}$ indicates location loss. $L_{conf,Loss}$ indicates confidence loss. $L_{class,Loss}$ indicates classification loss.

Both confidence loss and classification loss were calculated using the BCEWithLogits Loss function. The coordinate loss was calculated using CIoU with the following formula:

$$L_{CIoU} = 1 - I_{IoU} + \frac{\rho^2(b, b_{gt})}{c^2} + av \tag{2}$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w_{gt}}{h_{gt}} - \arctan\frac{w}{h})^2 \tag{3}$$

$$a = \frac{v}{(1 - I_{IoU}) + v} \tag{4}$$

b denotes the prediction frame. b_{gt} means the true frame; c indicates the diagonal distance of the smallest closure region that can encompass both the predicted frame and the true frame; α is a balancing parameter; and ν is used to quantify the consistency of the aspect ratio. From equation (3), it can be seen that ν equals 0 when the aspect ratio of the predicted frame and the actual frame are equal. At this point, the penalty term for the aspect ratio has no effect, and the CIoU loss function lacks a stable expression.

Therefore, NWD replaces CIoU in the original network. The NWD index calculates the Wasserstein distance between two bounding boxes to measure their similarity. The NWD measure is shown in Equation (5):

$$NWD(N_a, N_b) = \exp-\left(\frac{\sqrt{W_2^2(N_a, N_b)}}{C}\right) \tag{5}$$

$W_2^2(N_a, N_b)$ is a distance variable, and C is a constant closely related to the dataset and robust within a certain range. The NWD measurement is designed as a loss function, as shown in Equation (6):

$$L_{NWD} = 1 - NWD(N_p, N_g) \tag{6}$$

Where N_p is the Gaussian distribution model of prediction box P, and N_g is the Gaussian distribution model of gt box G. Even in both cases, the NWD-based losses can provide gradient |P∩G|=0 and |P∩G|=P/G.

The regression loss based on NWD ensures smoother detection of small underwater objects, benefiting from the scale invariance property, and the capability to evaluate the similarity among disjoint or mutually encompassing bounding boxes. This allows the robustness and stability of the algorithm to be assessed.

D. The Proposed DCCN-YOLOv7 Model

In the proposed DCCN-YOLOv7 model, the original ELAN network of YOLOv7 is improved through the design of the DCN structure. The 3×3 DCN convolutional block is replaced by the original 3×3 convolutional block, and jump joins and 1×1 convolution are added to DCN to improve the model's perception of image details. In addition, by introducing the CoT3 module in the feature fusion, the model can encode and model the context information before the detection result is formed, providing a more accurate and effective feature representation for target localisation and classification. By replacing the CIoU loss function with the NWD loss function, the detection capability of small underwater biological targets is effectively improved. The improved model diagram is shown in Fig. 4.

IV. Experimental Results And Analysis

A. Dataset Source and Processing

The experiments performed in this paper use an underwater dataset from the Underwater Robot Target Grab Competition (URPC), which consists mainly of five categories of underwater objects: starfish, sea urchins, scallops, sea cucumbers and seagrass. First, the images are pre-processed by removing a subset of the seagrass data and their associated labels, while preserving the remaining relevant target information. Next, the images are normalized and the corresponding target labels are stored in XML format. Finally, consistent with the scope of this study, an adaptive brightness data augmentation method is employed to mitigate the impact of light changes and other factors on the algorithm performance, and the total number of samples is expanded from 2326 to 4642. In this paper, the data were divided into three sets based on the ratio of 6:2:2 for training, testing, and

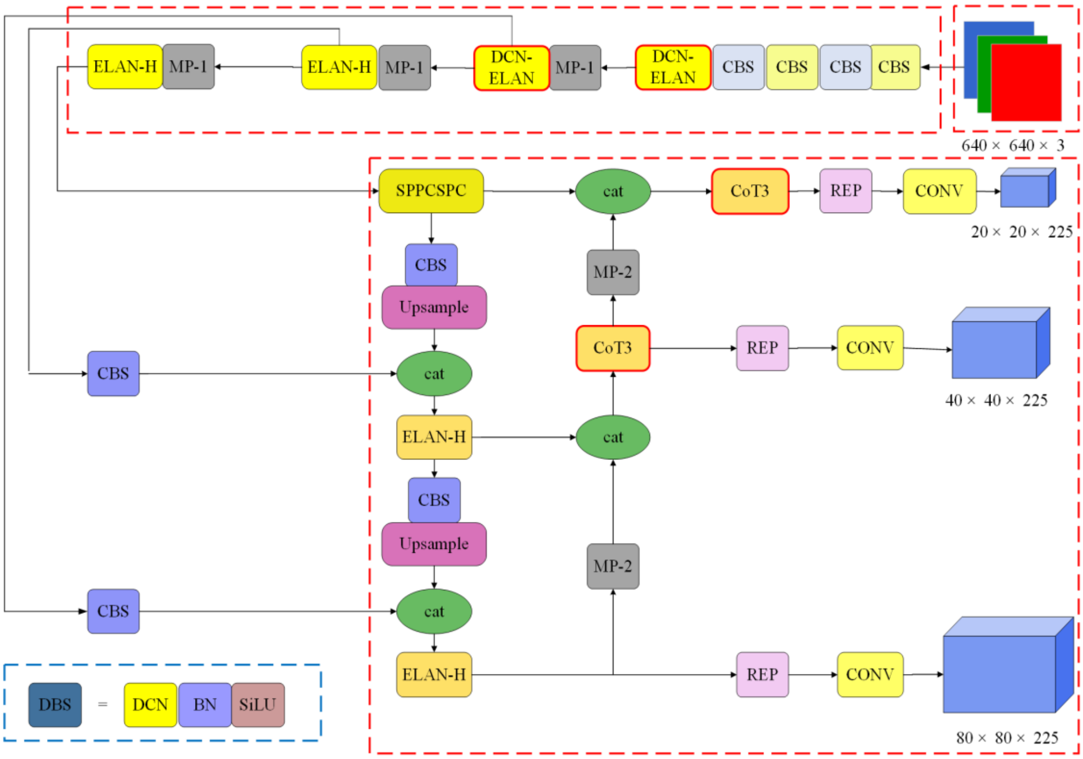

Fig. 4. Structure diagram of the DCCN-YOLOv7 model

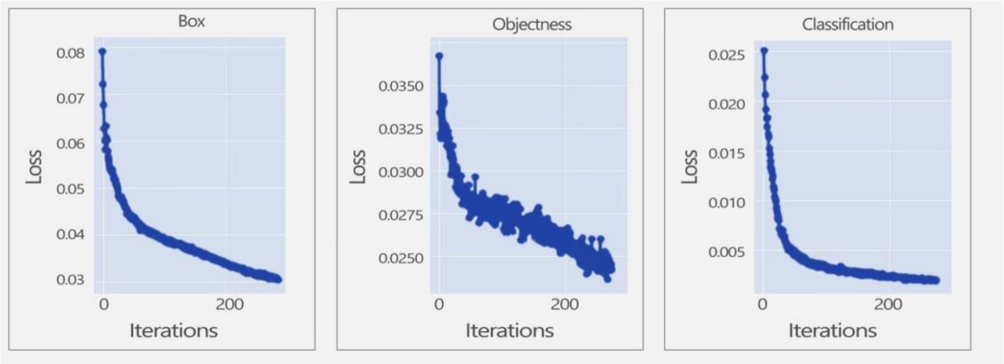

Fig. 5. Variation curves of CIoU loss values

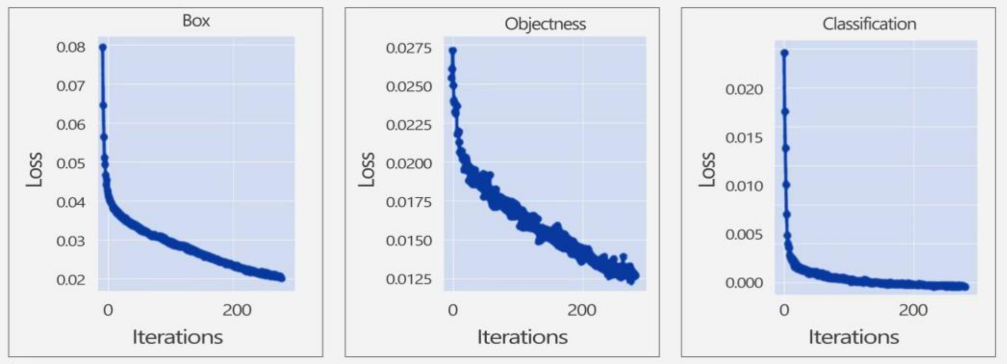

Fig. 6. Variation curves of NWD loss values

validation respectively. This means that 2785 samples in the training set, 928 samples in the testing set, and 928 samples in the validation set.

B. Experimental Environment and Evaluation Index

The performance strengths and weaknesses of the underwater target detection model are evaluated based on both detection accuracy and image processing speed, and the model performance is evaluated qualitatively by comparing whether there are false positives and false negatives. In the experiment, the Precision-Recall (P-R) curve, Average Precision (AP) and Mean Average Precision (mAP) are primarily chosen as the evaluation metrics to calculate the detection precision. They are calculated as follows:

$$R = \frac{TP}{TP + FP} \tag{7}$$

$$P = \frac{TP}{TP + FN} \tag{8}$$

$$AP = \int_0^1 P(R)dR \quad (9)$$

$$mAP = \frac{1}{N}\sum AP \quad (10)$$

In the formula, TP, FP and FN are the number of objects accurately detected, incorrectly detected and not detected respectively. P denotes the accuracy rate; R denotes the recall rate. The area between the P-R curve and the axes is known as the Average Precision (AP) value. The mean average precision (mAP) evaluates detection performance across all categories in the target detection network model by calculating the average AP values. As the model should take into account not only the accuracy but also the frequency of detection (FPS) of the model, the FPS is also an important evaluation index.

C. Experimental Results

a) Convergence Comparison of Loss Function

The convergence of the YOLOv7 loss function was verified in an identical experimental environment using the same network model. The loss curves for the two different loss functions, CIoU and NWD, are shown in Fig. 5 and Fig. 6 respectively. These curves depict the progress of the edge loss over the iterations.

By observing Fig. 5 and Fig. 6, it becomes evident that the CIoU and NWD tend to converge with an increase in the number of iterations. the NWD loss value is comparatively smaller and more stable than the CIoU loss value. Therefore, using NWD as the bounding box loss function in this paper's dataset is more effective in enhancing the network model's performance.

This observation suggests that by using NWD as the bounding box loss function, we can more accurately measure the similarity between small objects, particularly those of varying scales. The NWD loss function is not affected by object overlap and is particularly suitable for evaluating the similarity of small objects. During testing, we found that using NWD as the loss function can significantly enhance the accuracy of underwater object detection, which is of practical value in real-world applications.

b) DCN position analysis

A series of comparative experiments were designed to add DCN modules at different locations in the network in order to investigate the optimal placement of DCN modules. One of the modules is placed in a backbone network that preserves the original information and is used to extract features from the input images and reduce dimensionality, where the feature maps generated by each convolutional layer represent an abstract representation of the original information. The other position is placed in the feature fusion (Neck), which can enhance the comprehensive perception and reasoning ability of the model through feature fusion. In the YOLO model, multiple convolutional layers with different step sizes to gradually acquire the feature maps of the sensory field at different scales. As depicted in Table I, the DCN module placed in the backbone network outperforms feature fusion, which is better than that of the original model. Because the DCN module is placed in the feature fusion, the DCN module requires additional parameters to control the shape and size of the deformable convolution kernel, which may increase the model parameters and decrease the model efficiency, thus affecting the detection performance.

c) Ablation Experiment

In this paper, ablation experiments were carried out to assess the detection performance of underwater organisms. The effectiveness of each optimization module was verified through these experiments, with the YOLOv7 model being chosen as the original architecture to establish a performance baseline. The model's overall performance was assessed and presented in Table II.

Based on the experimental findings presented in Table I, when performing ablation experiments on the YOLOv7 model and applying different enhancement methods, each individual enhancement significantly improves the model performance. Incorporating the DCN-ELAN module into the model improved the mean average precision (mAP) by 0.7%. In addition, the experiments validated the impact on performance of introducing the CoT3 module and replacing the CIOU loss function with NWD, which increased mAP by 1.5% and 1.1% respectively. Taken together, the overall mAP increased by 2.8% and the model improved its detection performance at almost the same FPS. Therefore, it can be concluded that the DCCN-YOLOv7 model is an effective method of enhancement.

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT DCN LOCATIONS

Method	YOLOv7	In Backbone	In Neck
mAP/%	77.6	78.3	77.9
Recall/%	74.5	75.2	74.9

TABLE II
COMPARISON OF ABLATION EXPERIMENTS

Model	DCN-ELAN	COT3	NWD	mAP@0.5/%	Precision/%	Recall/%	FPS
YOLOv7	-	-	-	77.6	79.3	74.5	64
YOLOv7	√			78.3	82.7	75.2	66
YOLOv7		√		79.1	79.8	76.5	67
YOLOv7			√	78.7	80.1	77.5	65
YOLOv7	√	√	√	80.4	82.7	76.7	67

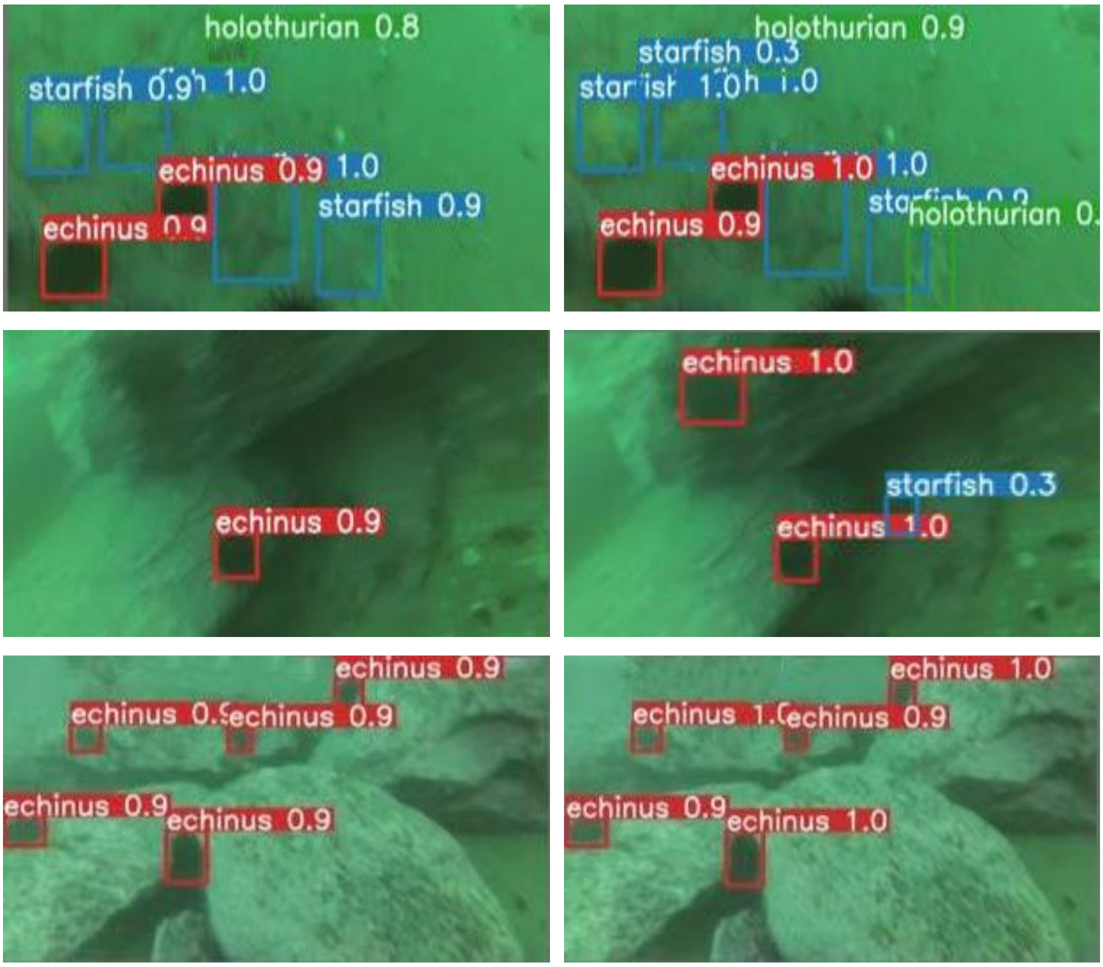

Fig. 7. Detection results of YOLOv7 (left) and DCCN-YOLOv7 (right) in underwater scenes

TABLE III
COMPARISON EXPERIMENT

Network model	mAP /%	scallop AP/%	echinus AP /%	starfish AP /%	holothurian AP /%
SSD	63.5	44.6	76.4	69.1	63.9
YOLOv4	64.5	42.9	84.1	78.0	52.8
YOLOv5-m	73.4	51.3	89.7	84.3	68.1
RetinaNet	75.5	53.6	90.5	86.2	71.7
YOLOv7	77.6	58.8	91.8	88.1	71.8
YOLOv8	79.1	62.8	92.1	88.7	72.8
DCCN-YOLOv7	80.4	65.3	93.0	89.6	73.7

d) Comparison Between DCCN-YOLOv7 Network Model and other Network Models

To assess the effectiveness of the algorithm proposed in this paper compared to other algorithms, a comparative evaluation was conducted, six different algorithms, namely SSD, YOLOv4, YOLOv5-m, RetinaNet, YOLOv7 and YOLOv8, were selected for comparison. The comparison was made using the same experimental equipment and data set. The outcomes are depicted in Table III. It was observed that the DCCN-YOLOv7 network model outperformed other classical network models in terms of mAP values when using images of the same input size. This suggests that the DCCN-YOLOv7 model is more suitable for the detection of underwater organisms.

The test results diagram in Fig. 7 shows a significant enhancement in the detection performance of underwater organisms using the proposed DCCN-YOLOv7 model.

V. CONCLUSION

A new detection model called DCCN-YOLOv7 is proposed in this paper, which incorporates jump connection and 1 × 1 convolutional structure in the deformable convolutional module to facilitate the transfer of information, which can retain the features of the previous layer and enhance the feature representation ability. Consequently, it leads to improved computational efficiency during network training and inference. In addition, CoT3 is integrated into the backbone network to improve the model's visual representation of underwater targets. On this basis, the model loss function is optimized by introducing the NWD regression loss function, which can measure the similarity of distribution without considering any overlap between small targets. This is particularly useful for accurately measuring similarity among small targets since NWD is independent of targets of different scales. Experimental evaluations using the URPC dataset were conducted to compare the proposed DCCN-YOLOv7 model with other popular target detection algorithms in complex underwater environments. The results demonstrate that the proposed model exhibits superior accuracy in terms of robustness compared to other models, making it highly valuable for applications. The next step would be to collect a large number of different samples as targets and use image enhancement techniques to improve the underwater dataset to enhance the detection performance of the model in practical applications.

REFERENCES

[1] R. Schettini, and S. Corchs, “Underwater Image Processing: State of the Art of Restoration and Image Enhancement Methods,” *EURASIP Journal on Advances in Signal Processing*, 2010:746052.

[2] M. R. Heithaus, and L. M. Dill, “Food Availability and Tiger Shark Predation Risk Influence Bottlenose Dolphin Habitat Use,” *Ecology*, 83(2), pp. 480–491, 2002.

[3] A. Rova, G. Mori, and L. M. Dill, “One Fish, Two Fish, Butterfish, Trumpeter: Recognizing Fish in Underwater Video,” *DBLP* (2007).

[4] K. M. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *Advances in Neural Information Processing Systems*, pp. 91–99, 2015.

[5] A. C. Berg, W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed et al., “SSD: Single Shot Multi-Box Detector,” *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer International Publishing, 2016.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” *IEEE Computer Vision and Pattern Recognition*, pp. 779-788, 2016.

[7] C. C. Li, “Detection and Recognition of Underwater Fish Targets Based on YOLOv3,” *Xianyang: Northwest Agriculture and Forestry University of Science and Technology A&F University*, 2020.

[8] Y. Li, J. Guo, X. Guo, and J. Zhao, “Toward in Situ Zooplankton Detection with a Densely Connected YOLOv3 Model,” *Applied Ocean Research*, pp. 114, 2021.

[9] L. Chen, Z. H. Liu, Z. H. Jiang, L. Tong, S. K. Wang, and J. Y. Dong et al., “Underwater Object Detection Using Invert Multi-Class Adaboost with Deep Learning,” *International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, 2020.

[10] P. F. Shi, S. Han, and J. J. Ni, “Underwater Target Detection Algorithm Enhanced and Improved by YOLOv4 Combined with Data,” *Chinese Journal of Electronic Measurement and Instrumentation*, 36(3), pp. 113-121, 2022.

[11] S. Cao, D. Zhao, and X. Liu, “Real-Time Robust Detector for Underwater Live Crabs Based on Deep Learning,” *Computer Electron Agric*, pp. 172, 2020.

[12] Y. H. Li, Y. Ting, Y. W. Pan, and M. Tao, “Contextual Transformer Networks for Visual Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2), pp. 1489-1500, 2022.

[13] J. Wang, C. Xu, W. Yang, and L. Yu, “A Normalized Gaussian Wasserstein Distance for Tiny Object Detection,” *ArXiv Preprint ArXiv*: 2110.13389 (2021).

[14] Z. H. Zheng, P. Wang, and D. W. Ren, “Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance 8:2 Segmentation,” *IEEE Transactions on Cybernetics*, 52(8), pp. 8574-8586, 2022.

[15] C. Chen, M. Y. Liu, and O. Tuzel, “R-CNN for Small Object Detection,” *Proceedings of IEEE International Conference on Computer Vision*, pp. 214-230, 2016.

[16] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, “YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7464-7475, 2023.

[17] Q. Song, S. Li, and Q. Bai, “Object Detection Method for Grasping Robot Based on Improved YOLOv5,” *Micromachines*, 12(11), pp. 1273, 2021.

[18] G. Brauwers, and F. Frasincar, “A General Survey on Attention Mechanisms in Deep Learning,” *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[19] D. F. Zhou, J. Fang, and X. B. Song, “IoU Loss for 2D/3D Object Detection,” *Proceedings of 2019 International Conference on 3D Vision. Washington D. C, USA: IEEE Press*, pp. 85-94, 2019.

[20] Z. H. Zheng, P. Wang, and W. Liu, “Distance IoU Loss: Faster and Better Learning for Bounding Box Regression,” *Artificial Intelligence*, 34(7), pp. 12993-13000, 2020.

[21] Z. H. Zheng, P. Wang, and W. Liu, “Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation,” *IEEE Transactions on Cybernetics*, 52(8), pp. 8574-8586, 2022.

[22] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[23] Y. Li, T. Yao, Y. Pan, and T. Mei, “Contextual Transformer Networks for Visual Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2), pp. 1489-1500, 2022.

[24] Z. Yu, J. Yu, and J. Fan, “Multi-Modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering,” *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1821-1830, 2017.