Marine Cultural Target Detection for Underwater Robotics Based on Deep Aggregation Network with Deformation Convolution

Yutuo Yang , Wei Liang * , Daoxian Zhou , Yinlong Zhang * , Gaofei Xu

Posted Date: 6 October 2023

doi: 10.20944/preprints202310.0329.v1

Keywords: autonomous underwater vehicle (AUV); underwater target detection; deformable convolution; multi-scale deep aggregation; attention mechanism

Article

Marine Cultural Target Detection for Underwater Robotics Based on Deep Aggregation Network with Deformation Convolution

Yutuo Yang [1,2,3], Wei Liang [1,2,3], Daoxian Zhou [4], Yinlong Zhang [1,2,3,*] and Gaofei Xu [5]

[1] Key Laboratory of Networked Control Systems, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China
[2] State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China
[3] Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China
[4] School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang 110168, China
[5] Institute of Deep-Sea Science and Engineering, Chinese Academy of Sciences, Sanya 572000, China.
* Correspondence: zhangyinlong@sia.cn

Abstract: Underwater robotics equipped with visual detection systems have the function of detecting underwater artifacts, which is of great significance to deep-sea archaeology. Underwater artifacts are located in complex environments with poor imaging conditions, and the targets have challenging problems such as breakage, stacking, and sediment burial, which causes the marine cultural target detection failures. To solve these problems, this paper proposes an underwater cultural target detection algorithm based on the deformable deep aggregation network model for underwater robotics exploration. To fully extract the target feature information of underwater targets in complex environments, this paper designs a multi-scale deep aggregation network with deformable convolutional layers. Besides, the BAM attention module is designed for feature optimization, which enhances the potential feature information of the target while weakening the background interference information. Finally, the target prediction is achieved through feature fusion at different scales. The proposed algorithm has been extensively validated and analyzed on the collected underwater artifact datasets, and the precision, recall, and mAP of the algorithm have reached 93.1%, 91.4%, and 92.8%, respectively. In addition, the present algorithm has been practically deployed on the underwater robotics. In the deep-sea tests, the artifact detection frame rate reaches up to 19 fps, which satisfies the real-time target detection.

Keywords: underwater robotics; marine target detection; deformable convolution; multi-scale deep aggregation; attention mechanism

1. Introduction

In the long history of navigation, maritime accidents have occurred occasionally, and a large number of human artifacts and industrial monuments have been gathered in the underwater seabed environments [1–3], due to the limitations of navigation technology and the influence of extreme weather. These seabed artifacts contain rich historical, cultural, and technological information, which is of vital significance to the in-depth exploration of the progress of human civilization as well as to the promotion of social development. Henceforth, seabed artifacts detection has become a focus of attention at home and abroad [4].

Underwater robotics is a commonly used equipment in seabed exploration, which has the advantages of wide operation range, high detection efficiency, and flexible operation compared with human occupied vehicle (HOV) and remotely operated vehicle (ROV) [5]. Underwater robotics carries two main operational payloads on seafloor exploration missions: side-scan sonar and underwater camera [6]. Side-scan sonar can be used for wide-area rapid search, but the acoustic image obtained is not only of low resolution but also of single content, obtaining limited target

 doi:10.20944/preprints202310.0329.v1

information. Although the detection distance of the underwater camera is small, the optical image obtained can provide rich information such as target shape, color, and texture, which is suitable for close-range fine detection, especially in the detection of underwater small-scale targets in a wide range of applications [7]. In the process of underwater detection operation, if underwater robotics can autonomously detect the targets in the captured video images, it can re-plan the navigation path according to the location of the discovered targets, and carry out more detections around the targets of interest, to facilitate the subsequent analysis and judgment of the seabed artifacts [8]. Therefore, vision-based target detection methods have become a common means of underwater artifact detection for underwater robotics.

Vision-based target detection algorithms can be categorized into two groups: traditional target detection and deep learning-based target detection [9]. In traditional target detection algorithms, the first step involves selecting a region of interest through a sliding window approach [10]. Subsequently, various feature extraction techniques, such as Scale-Invariant Feature Transform (SIFT) [11], Histogram of Oriented Gradients (HOG) [12], are applied to extract features from the selected region. Finally, these extracted features are used for target recognition through the trained classifiers like Support Vector Machine (SVM). Cutter et al. [13] employed Haar-like features and multiple cascaded classifiers to detect fish targets, while Rizzini et al. [14] identified underwater targets based on the uniformity of underwater image color and sharpness information from contours. Qiu et al. [15] proposed an algorithm based on surface feature ripples for detecting underwater moving targets in photopolarimetric imaging mode, which has become a notable example of traditional algorithms in underwater target detection. However, traditional detection methods require the design of various feature extraction models and rely on machine learning techniques for classification. This limits their applicability in real underwater scenarios. Moreover, manually designed feature extraction models primarily capture low and mid-level image features, making it challenging to extract representative semantic information.

With the dramatic improvement of graphic computing hardware such as powerful GPUs and the rapid development of deep neural network models in recent years, target detection algorithms based on deep learning have achieved promising detection performance. Many researchers have applied these methods to underwater target detection scenarios. Chen et al. [16] introduced a novel sample-weighted super network (SWIPENET) to address the blurring problem in underwater images amidst significant noise interference. Lei et al. [17] incorporated the Swin Transformer into the backbone network of YOLOv5, enhancing feature extraction for underwater targets and enabling the network to detect targets in low-quality underwater images. Yan et al. [18] integrated the CBAM attention mechanism into a one-stage target detection model to enable the network focuses more on target feature information, thereby improving detection accuracy. However, the aforementioned methods still suffer the challenges in fully utilizing the characteristics of targets in complex underwater environments. They struggle with detection accuracy when dealing with occlusion and overlapping issues among underwater targets at different scales, as well as problems like leakage and false detection. Song et al. [19] proposed a two-stage underwater target detection algorithm with Boosting R-CNN, which enhances occluded target detection by modeling uncertainty and mining challenging samples. Zeng et al. [20] introduced a Faster R-CNN-AON network based on generative adversarial networks, effectively improving overall detection performance while preventing overfitting. Despite these advancements, it's worth noting that the above-mentioned studies often come with a drawback, i.e., they involve a large number of algorithm parameters, which may not meet the real-time requirements for underwater robotics.

Currently, underwater artifact target detection still faces the following challenges:

(1) Poor imaging quality of underwater target images. Due to the differences in water absorption of light of different wavelengths and the scattering of underwater light, underwater images suffer from color deviation and low visibility [21]. In addition, the imaging quality of underwater images is low due to insufficient underwater illumination conditions of underwater robotics and limited CMOS imaging levels [22].

 doi:10.20944/preprints202310.0329.v1

(2) Target identification failures in the complex underwater environments. Underwater artifacts have different morphologies and tend to accumulate, which makes them easy to be missed or incorrectly detected [23]. In addition, due to the age of underwater artifacts, the artifacts are often covered with sediments, attached by marine organisms, or broken and morphologically mutilated, which leads to difficulties in extracting discriminative features of the artifacts in the process of visual inspection. It causes serious interference in the detection of artifacts.

(3) Difficulty in acquiring underwater target samples. Unlike atmospheric optical images, it is difficult to obtain enough samples with relevant features in the preliminary research of target detection algorithms due to the influence of the complex underwater environment and the limitation of imaging equipment [24].

In order to solve the above problems, we propose an underwater artifact target detection algorithm based on a deformable deep aggregation network model. The main contributions are summarized as follows:

(1) We design a feature extraction network specifically for underwater artifact detection, which enhances the network to extract features from artifact targets in complex scenarios through a deep aggregation structure with deformable convolutional layers and more jump connections.

(2) We introduce the BAM attention mechanism to enhance the features of underwater artifacts and weaken the background redundant information through feature optimization, which improves the model's anti-interference ability, and spares the excessive parameters and computational complexity.

(3) We build a set of underwater robotics underwater visual detection system. By collecting a large number of underwater targets target images, the underwater targets (UCA) dataset is established. The accuracy and real-time performance of the underwater targets target detection algorithm are verified.

The rest of the paper is organized as follows: Section 2 briefly describes our underwater vision inspection system. Section 3 describes the materials and proposed methodology in detail. Section 4 presents the experimental details and system analysis. Section 5 summarizes the entire paper as well as future research directions.

2. Underwater Robotics Visual Detection System

We have constructed an efficient underwater visual inspection system for underwater robotics, the main goal of which is to collect data related to underwater artifacts and verify the effectiveness of our proposed algorithms in real-world environments. As shown in Figure 1, our system is designed to embed the algorithm into an edge computing platform and deploy it on an AUV. The system can utilize images captured by underwater cameras and feed them into our detection algorithm for autonomous target recognition and analysis, and ultimately output detection results. To ensure that we can effectively capture undersea targets, we mounted the underwater camera at the bottom position of the underwater robotics.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2023 doi:10.20944/preprints202310.0329.v1

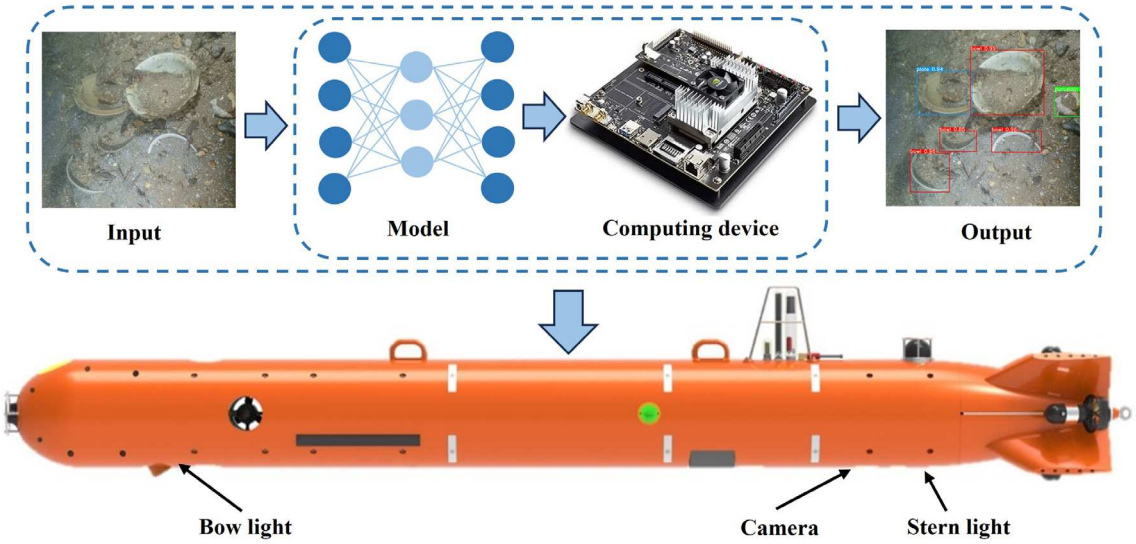


Figure 1. Underwater robotics vision based target detection system.

Therefore, the focus of this study is on the use of an underwater camera to acquire optical images to enable the detection of underwater artifact targets and to test the performance of the algorithm in a real underwater environment.

3. Materials and Methods

We propose an efficient and lightweight underwater targets detection network (UCA-Net) for the task of underwater artifacts target detection in archaeological AUV. UCA-Net combines a deformable convolution module and an attention mechanism to improve the performance of artifact target detection in complex underwater environments. UCA-Net combines deformable convolutional modules and attention mechanisms to improve the performance of artifact target detection in complex underwater environments. As shown in Figure 2, UCA-Net consists of three parts, i.e., feature extraction network, feature optimization network, and feature fusion network. First, the feature extraction network adopts a deep aggregation structure that incorporates deformable convolutional layers and multi-hop connections. The deformable convolutional layer enables the network to better adapt to the complex spatial features of the broken artifacts, and the multi-hop connection helps to capture the multi-scale semantic information of the artifact targets. Secondly, the feature optimization network enhances the key features of underwater artifacts by introducing the BAM attention mechanism by augmenting them in both spatial and channel dimensions while attenuating invalid background information. Finally, the feature fusion network fuses feature from different scales to further enhance the algorithm's representation of the target. With the above design, the UCA-Net algorithm proposed in this paper effectively improves the accuracy and robustness of underwater artifact target detection.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2023 doi:10.20944/preprints202310.0329.v1

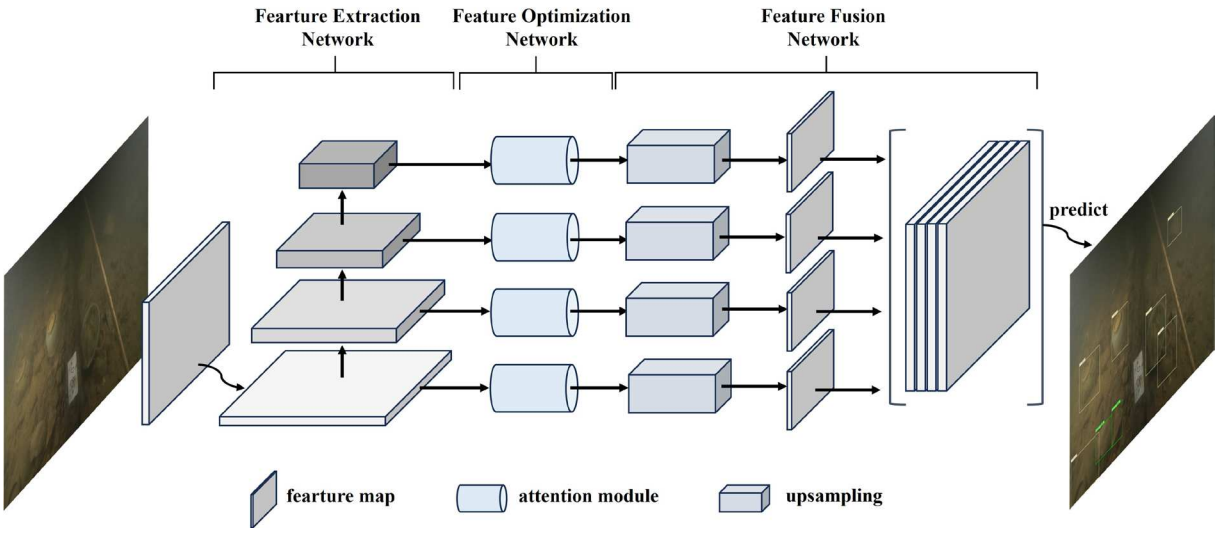


Figure 2. The framework of underwater targets detection network (UCA-Net).

3.1. Feature extraction network

- In the process of target detection for underwater cultural relics, there is a prevalent diversity of target types, sizes, shapes and texture features, which increases the difficulty of detection. In traditional deep learning models, the convolution operation has a fixed structure, which limits the network receptive field, and the network can only capture local information during feature extraction.
- However, due to the issues of breakage and burial of underwater artifacts, they present irregular features. In this case, the traditional convolutional operation makes it difficult to fully extract the features of underwater artifacts leading to detection failure. To enhance the detection ability of convolutional neural networks for underwater artifacts, the long range spatial relationships can be better captured by expanding the receptive field of the network and constructing an implicit spatial model [25]. In complex underwater environments, traditional standard convolution can only perform fixed-size sampling. In contrast, deformable convolution can better learn the features of a target by introducing a learnable offset in the convolution operation, which enables it to dynamically adjust the sampling position and better adapt to the shape of the target such as broken burials [26]. As shown in Figure 3, the deformable convolution module adds a two-dimensional offset to each samples in the convolution kernel based on the traditional standard convolution $\{\Delta p_n \mid n = 1, ..., N\}$, $N = |R|$ mathematically defined in Equation (1).

$$Y(p_0) = \sum_{p_n \in R} w(p_n) \cdot X(p_0 + p_n + \Delta p_n) \tag{1}$$

- where X is the input feature map; R is the 3×3 convolution kernel; p_n is the nth point in the convolution kernel; $w(p_n)$ is the weight corresponding to the p_n point; p_0 is the p_0 point on the input-output feature map; Δp_n is the two-dimensional offset of the deformable convolutional sampling point; Y is the output feature map.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2023 doi:10.20944/preprints202310.0329.v1

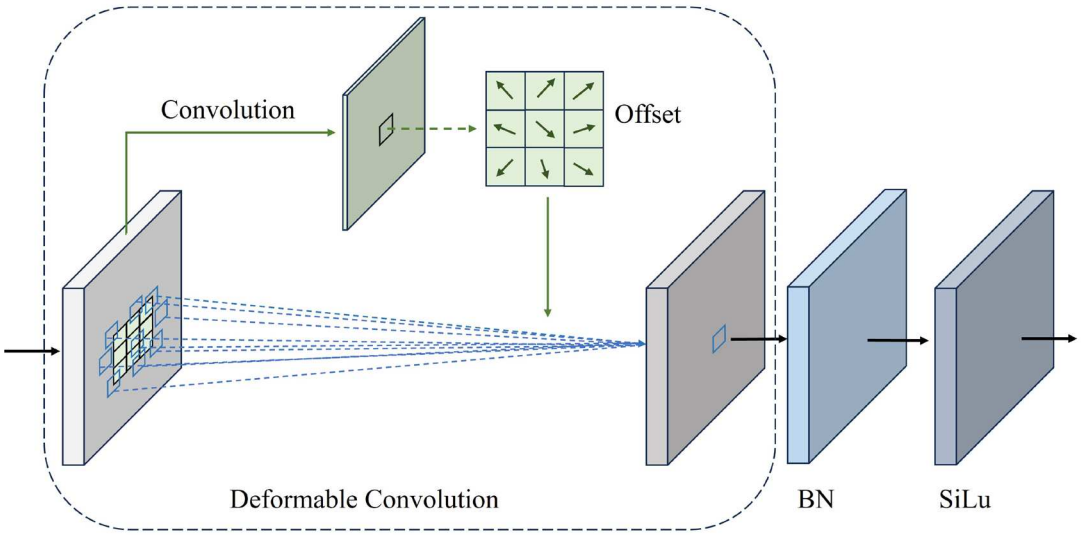


Figure 3. Deformable convolution.

- The deep layer aggregation (DLA) network has been widely used as a compact and efficient feature extraction backbone in computer vision tasks such as target detection and semantic segmentation [27]. DLA network merges the layered feature maps in an iterative manner, which achieves a more accurate representation of the target features while keeping fewer parameters. To adapt to the diverse target sizes and shape features in underwater artifact environments, we design the DLA network structure accordingly so that it can output feature maps with four feature layers of different scales. On this basis, for the irregular artifact morphology present in underwater environments, we introduce deformable convolution to replace the traditional convolution operation to enhance the feature extraction capability of the network for irregular targets. We name the proposed feature extraction network as a multi-scale deep layer aggregation with deformable convolution network (MDLA-DCN). The network shows impressive performances in complex underwater environments and significantly enhances the extraction of features for underwater artifact targets with complex morphology.
- The MDLA-DCN network structure is shown in Figure 4, with four parallel sub-networks with different resolutions. Each sub-network consists of a series of deformable convolutional modules. The same sub-network feature map resolution does not change with the depth of the network, while the feature map resolution of the parallel sub-network decreases sequentially by 1/2. The number of channels increases by a factor of 2. Information exchange across the parallel sub-networks is implemented within the MDLA-DCN network by upsampling so that each sub-network receives the information from the other parallel sub-networks repetitively. Multi-hop connections in the network aggregate features of different resolutions to yield the enhanced underwater artifact features, which are more accurate in terms of spatial and semantic information. In this paper, the 4-, 8-, 16-, and 32-fold downsampled feature maps generated by the parallel sub-networks are used as outputs, in order to fully utilize the multi-scale feature information.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2023 doi:10.20944/preprints202310.0329.v1

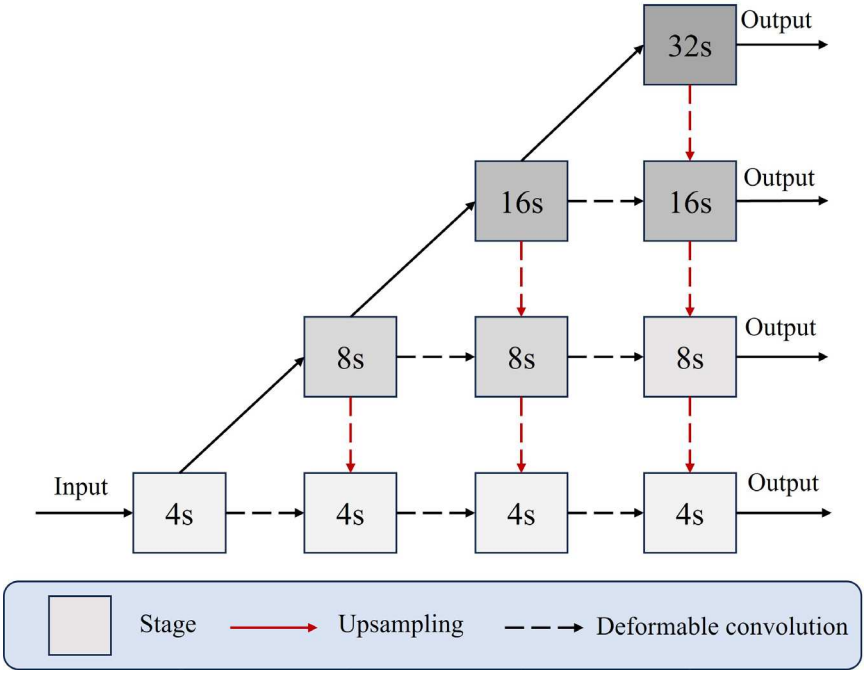


Figure 4. MDLA-DCN network.

3.2. Feature Optimization Network

The feature extraction network generates four different resolutions of feature maps, which contain valid features of the target and also a large number of invalid background features, and there are differences in these four feature maps and their contributions to the final detection results. Therefore, to suppress the invalid features and enhance the target features, and to enable the network to autonomously learn the correlation and importance between feature maps of different resolutions, we introduce BAM attention for feature optimization. Different from the separate channel attention [28] and spatial attention [29], BAM attention enhances features in both spatial and channel dimensions through different branches, the structure of which is shown in Figure 5.

Channel attention branching enables the network focuses on the channel features of interest by modeling the correlation between channels. Firstly, the input feature $F \in \mathbb{R}^{C\times H\times W}$ undergoes global average pooling to encode the global information of each channel and generate a one-dimensional channel vector; then the one-dimensional channel vectors are processed by using the multilayer perceptron (MLP) to estimate the inter-channel attention; finally, the output feature scale is adjusted by using the batch normalization (BN) layer to obtain the channel attention mapping $M_C(F) \in \mathbb{R}^C$. The specific description is shown in Equation (2).

$$M_C(F) = BN\left\{MLP\left[AvgPool(F)\right]\right\} = BN\left\{W_1\left[W_0(AvgPool(F)) + b_0\right] + b_1\right\} \tag{2}$$

where $W_0 \in \mathbb{R}^{\frac{C}{r}\times c}$, $b_0 \in \mathbb{R}^{\frac{C}{r}}$, $W_1 \in \mathbb{R}^{C\times\frac{C}{r}}$, $b_1 \in \mathbb{R}^C$, BN denotes the batch normalization.

Spatial attention branching can effectively capture the spatial location information of features and make the network more concerned about the location information of the target. Firstly, the input $F \in \mathbb{R}^{C\times H\times W}$ is compressed by 1×1 convolution to compress the channel dimension; then two 3×3 null convolutions are used to aggregate the context information with a larger receptive field; finally, the 1×1 convolution is used to map the dimension of the feature map to $\mathbb{R}^{1\times H\times W}$, and a batch normalization layer is used for the scale adjustment to get the spatial attention mapping $M_S(F) \in \mathbb{R}^{H\times W}$. The specific description is shown in Equation (3).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2023 doi:10.20944/preprints202310.0329.v1

$$M_S(F) = BN\left\{f_3^{1\times1}\left\{f_2^{3\times3}\left\{f_1^{3\times3}\left[f_0^{1\times1}(F)\right]\right\}\right\}\right\} \tag{3}$$

where f denotes a convolution operation, the superscripts denote the convolution kernel sizes, and the subscript denotes the order of the convolution operation.

The complete computation of the BAM refinement input feature $F \in \mathbb{R}^{C\times H\times W}$ is shown in Equation (4).

$$F' = F + F \otimes \sigma\left[M_C(F) + M_S(F)\right] \tag{4}$$

where $\otimes$ denotes element-wise multiplication, σ is a sigmoid activation function, $M_C(F)$ and $M_S(F)$ are the channel attention mapping and spatial attention mapping, respectively, which are resized to $\mathbb{R}^{C\times H\times W}$ before being added together.

In general, networks usually overlay the attention mechanism serially, i.e., adding the attention mechanism after most of the convolutional layers. Due to the special characteristics of the feature extraction network structure, the BAM attention mechanism module is only added in parallel to the final output part of the parallel sub-network, which enhances the output features of the sub-network in the spatial and channel dimensions, effectively filters the invalid background features and strengthens the effective target features, and improves the quality of the output features of the sub-network significantly without increasing the parameters of the network too much.

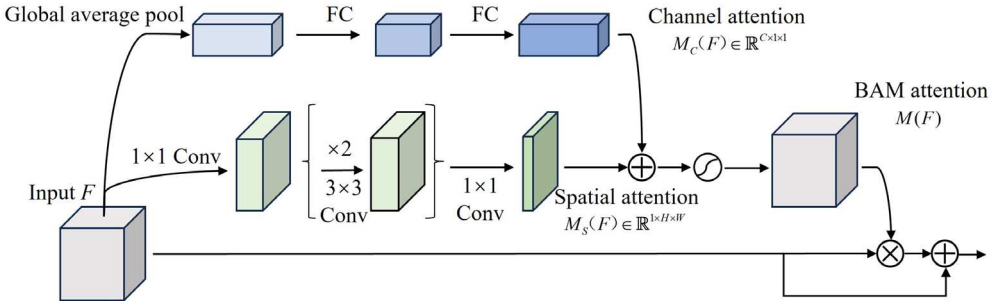


Figure 5. BAM attention mechanism.

3.3. Feature fusion network

After processing by the feature optimization network, feature maps at different scales are obtained, which are used to effectively represent the key features of underwater artifact targets. To realize multi-level feature extraction and fusion for underwater artifacts, we designed a fusion network for combining deep and shallow features.

The feature fusion process is shown in Figure 6, first, channel dimensionality reduction is performed on each source feature map using 3×3 convolution to keep the number of channels consistent while reducing the amount of computation within the network. After that, the low-resolution features are up-sampled using the inverse convolutional layer to keep their resolution consistent with the high-resolution feature maps. Commonly used up-sampling methods include the inverse convolution layer [30] and the bilinear difference method. Since the inverse convolution can provide the network with parameters that can be learned and improve the performance of the network, we choose the inverse convolution for up-sampling. Finally, the four adjusted feature maps are fused by the Concatenation fusion operation for final prediction. Through this multi-scale feature fusion method, the loss of small-scale target features can be effectively reduced and the problem of underestimated utilization of shallow features in spatial locations in the deep network can be solved, thus ensuring the robustness and reliability of target features of underwater artifacts at different scales.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2023 doi:10.20944/preprints202310.0329.v1

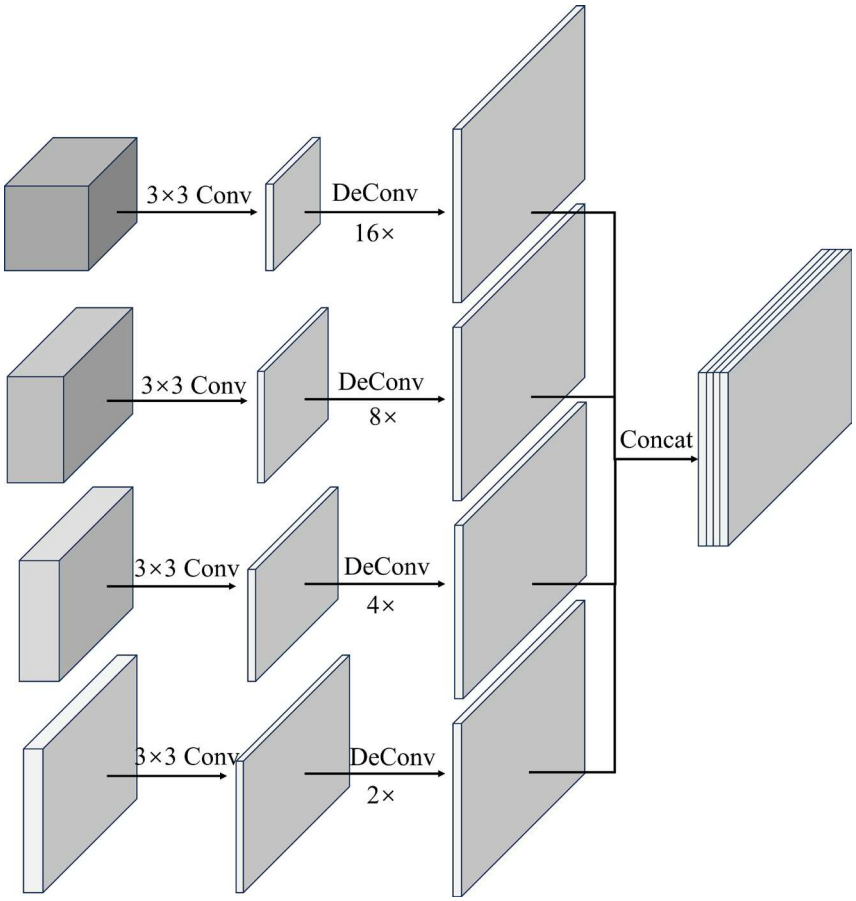


Figure 6. Illustration on the feature fusion network.

4. Experiments

In order to verify the performance of the algorithm proposed in this paper for underwater artifacts target detection, an underwater artifacts dataset is constructed, and the detection model is trained and tested on this dataset. In addition, the algorithm is compared with other mainstream detection algorithms to verify that the algorithm detection performance in complex underwater environments.

4.1. Underwater Target Dataset

The dataset used for the experiments was obtained by photographing real seabed artifact sites with a high-clear-water underwater camera carried by an archaeological underwater robotics. Given the complexity of the underwater environment, the dataset covers a wide range of scenarios, including low light, target stacking, target burial, and target breakage. The UCA underwater artifact dataset was constructed after manual screening, de-duplication, and quality assessment. The dataset contains 10,714 images totally covering five types of targets, namely porcelain plates, bowls, jars, incense burners, and tiles, which are commonly found in underwater archaeological sites. We divide the UCA dataset into the training set, validation set, and test set with the ratio of 6:2:2. Examples of representative images are shown in Figure 7.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2023 doi:10.20944/preprints202310.0329.v1

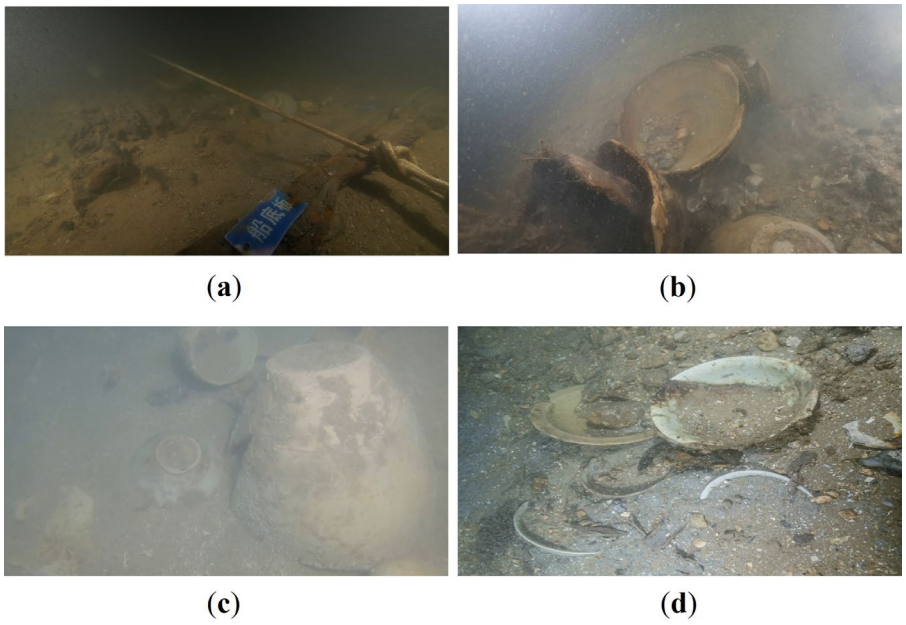


Figure 7. Target images of underwater targets in different scenes: (**a**) Low-light targets; (**b**) Stackable targets; (**c**) Damaged targets; (**d**) Buried targets.

4.2. Experimental Setups

4.2.1. Experimental Environment and Training Parameters

The hardware environment of our experimental platform is a high-performance server, which is configured as follows: Inter Xeon processor with a main frequency of 2.1GHz; 64GB of RAM; and four Nvidia Tesla V100 graphics cards with 32GB of video memory. The software environment is the operating system of Ubuntu18.04, Python 3.7, and CUDA11.0.

The training parameters are as follows: the gradient descent optimizer used to update the parameters of the convolutional kernel is Adam; the optimizer Momentum is 0.937; the learning rate update mode during training is STEP; the maximum learning rate is 0.001; the training batch size is 16; the weight decay coefficient is 0.0005; and the training iteration period Epoch is 300.

4.2.2. Model Evaluation Metrics

We use four main metrics to test the performance of the model in our experiments. Precision(P) denotes the proportion of positive classes that the model considers to be positive and is computed as in Equation (5). Recall(R) denotes the proportion of positive classes classified by the model to the total number of positive classes, and is computed as in Equation (6). *F1* is the harmonic mean of Precision and Recall, and used as a proxy for the model's performance. Average precision (AP) is the area under the curve composed of Precision and Recall taking different thresholds for each class, the larger the value, the better the recognition accuracy of the class, calculated as in Equation (8). The mean Average Precision (mAP) denotes the average *AP* of all the classes, the larger the value, the better the accuracy of the model in recognizing the target, calculated as in Equation (9).

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2023 doi:10.20944/preprints202310.0329.v1

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2}$$

$$AP = \int_0^1 P(R)dR \tag{3}$$

$$mAP = \frac{1}{N}\sum_{n=1}^{N} AP_n \tag{4}$$

where TP denotes the number of positive samples correctly predicted by the mode; FP denotes the number of positive samples predicted by the model that are actually negative samples. FN denotes the number of positive samples predicted by the model to be negative. N denotes the number of all categories, and AP_n denotes the Average Precision of the n_{th} category.

4.3. Comparison with Mainstream Methods

To verify the effectiveness of the underwater target detection algorithm proposed in this paper, we conduct comparison experiments with mainstream target detection algorithms such as Faster-RCNN [31], SSD [32], YOLOv5-l [33] and YOLOv7 [34]. The detection effect and performance comparison is carried out on the real underwater artifact UCA dataset constructed using this paper.

To ensure the comparability of the experiments, we refer to the published code of each comparison algorithm and use the original parameter settings. All comparison algorithms were trained on the same training process for a total of 300 epochs, and the models were analyzed qualitatively and quantitatively to evaluate the performance of each algorithm.

We qualitatively analyze the performance of the algorithms through the detection effects of different models, and the detection effects of Faster-RCNN, SSD, YOLOv5-l, YOLOv7, and the algorithms proposed in this paper are shown in Fig. 7. From the diagram, it can be seen that the SSD algorithm has the worst detection performance, due to the fact that its ability to represent shallow features is not strong enough, which results in more misdetections and omissions. The Faster-RCNN algorithm and the YOLOv5 algorithm have comparable detection effects, and the YOLOv7 algorithm has better effects, but these methods still have omissions when the target appears to be buried or stacked. Compared with the above methods, the algorithm proposed in this paper achieves better detection results, thanks to the optimization of the feature extraction network and the introduction of the BAM attention mechanism, which enables to effectively extract the feature information of the target in the complex environment and improves the algorithm overall robustness.

In order to better verify the superiority of the algorithm proposed in this paper, the UCA data test set is used for comparison with the above algorithm. At the same time, four generalized metrics Precision, Recall, F1, and mAP are introduced to quantitatively evaluate the performance of the algorithms. The comparisons of the algorithms performances are shown in Table 1.

Table 1. Performance comparison of different algorithms.

Method	Precision(%)	Recall(%)	F1(%)	mAP(%)
Faster-RCNN[31]	90.2	88.5	89.3	89.4
SSD[32]	81.9	82.2	82.0	82.8
YOLOv5-l[33]	88.3	87.8	88.1	88.7
YOLOv7[34]	90.1	88.4	89.3	89.9
Ours	**93.1**	**91.4**	**92.2**	**92.8**

Note: Bolded text shows the optimal results for each column.

The experimental results show that our algorithm outperforms others in all metrics, with 93.1%, 91.4%, 92.2%, and 92.8% for Precision, Recall, F1, and mAP, respectively. In this paper, the algorithm

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2023 doi:10.20944/preprints202310.0329.v1

designs a deformable convolution-based multi-scale deep aggregation network for underwater cultural relics targets for feature extraction, which can identify and localize targets in complex environments by better fusing semantic and spatial information. The deformable convolution expands the receptive field of the detection network to effectively extract the broken and irregular artifact features, and the multi-scale deep aggregation network reduces the loss of contextual information of the target features and better captures the global information of the artifact targets. The BAM attention module is introduced for feature optimization, which effectively cuts down the background redundant information and makes the network focus on the target feature information. Finally, progressive feature fusion of different network layers is realized by the multi-scale feature fusion module. Through these targeted network structure and module designs, the inherent features of underwater artifact targets are retained in the deep layer of the network, which enhances the network ability to represent the features of artifact targets in complex environments, thus improving the detection performance.

4.4. Ablation Studies

To demonstrate the effectiveness of the modules, asymptotic performance tests are performed for each improvement modules. The following ablation experiments are performed on the UCA dataset and Table 2 exhibits the ablation results for different variants of the algorithm.

From the experimental results in Table 2, it can be seen that compared with the original DLA network, the multiscale deep aggregation network (MDLA) designed in this paper improves the mAP by 1.7% and the Precision by 1.2%, which effectively enhances the detection ability of different scale targets. The use of DCN deformable convolution instead of ordinary convolution effectively enhances the feature extraction ability of the MDLA network for irregular targets, and mAP is further improved by 0.9%. As DCN expands the receptive field of the detection network, it makes the network enhance the aggregation to capture more comprehensive feature information of the target. With the introduction of the BAM attention module, F1 and mAP are increased by 1.3% and 1.4% respectively, because the attention module enhances the potential information of the target and attenuates the influence of redundant information, which makes the individual indexes further improved and the algorithm has higher detection accuracy. The experiment proves that the addition of deformable convolution and attention module is reasonable in the task of underwater artifact detection in complex environments, which can effectively improve the adaptability and accuracy of the algorithm.

Table 2. Ablation experiments.

Method	Precision(%)	Recall(%)	F1(%)	Map(%)
DLA[27]	88.4	87.3	87.6	88.8
MDLA	89.6	89.2	89.4	90.5
MDLA+DCN	90.9	90.1	90.9	91.4
MDLA+DCN+BAM	**93.1**	**91.4**	**92.2**	**92.8**

Note: Bolded text shows the optimal results for each column. MDLA is a feature extraction network using standard convolution.

4.5. Under Robotics Visual Detection System Performance Test

4.5.1. Underwater robotics Experimental Platform

In order to better test the effectiveness of the visual inspection system in this paper, we embedded the visual detection system into an underwater robotics, and the performance was tested in a real underwater environment.

Underwater robotics is a commonly used equipment in seafloor exploration and plays an important role in underwater related research [35]. The underwater robotics and edge computing device used in the experiment are shown in the Figure 8, and main parameters of underwater robotics are shown in Table 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2023 doi:10.20944/preprints202310.0329.v1

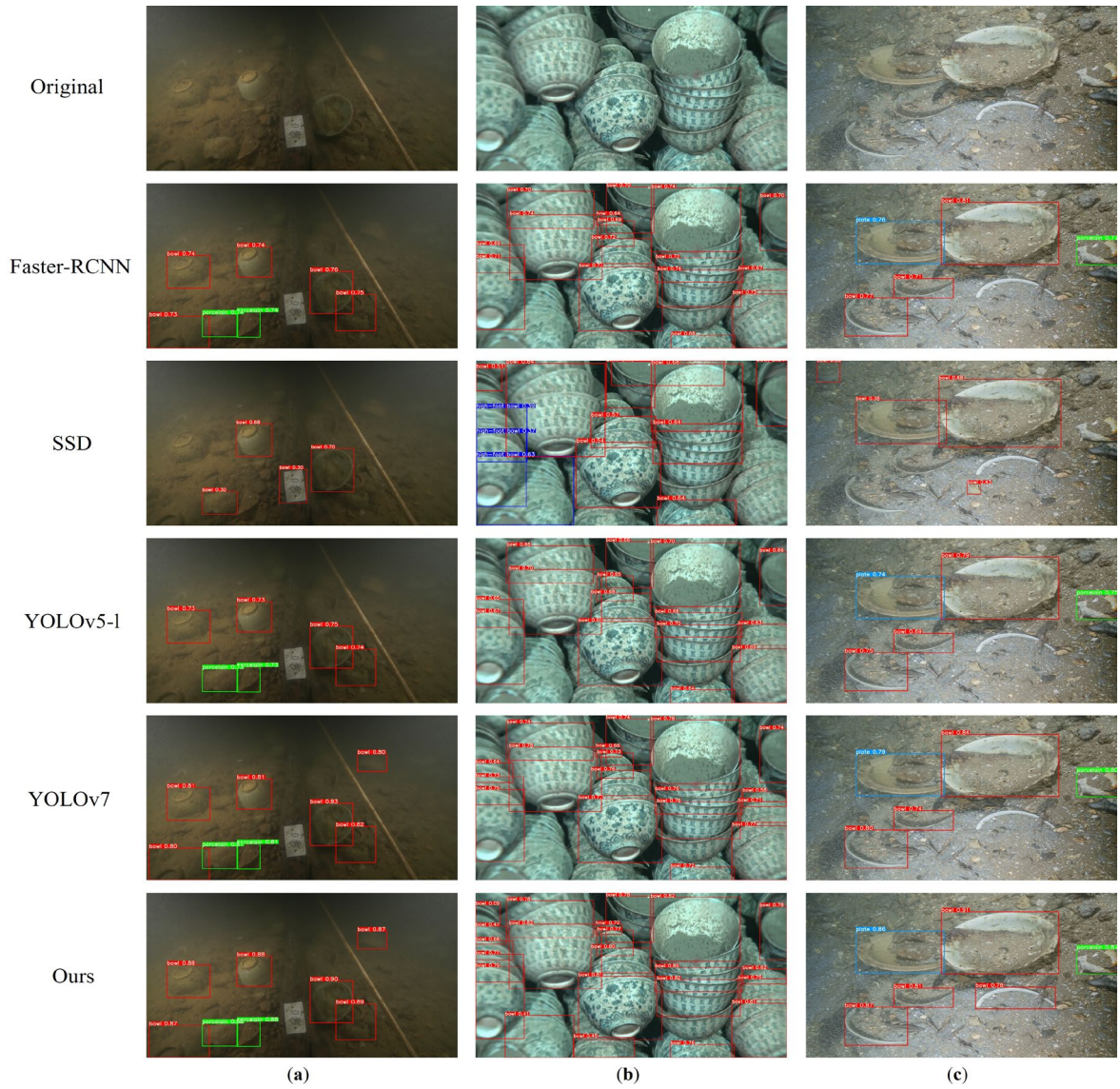


Figure 8. Comparison of the detection results among various algorithms. Different colored squares represent different targets. Red squares represent bowls; green squares represent porcelain items; light blue squares represent plates; and deep blue squares represent high-foot bowls: (**a**) Low-light scene; (**b**) Stacked scene; (**c**) Burial scene.

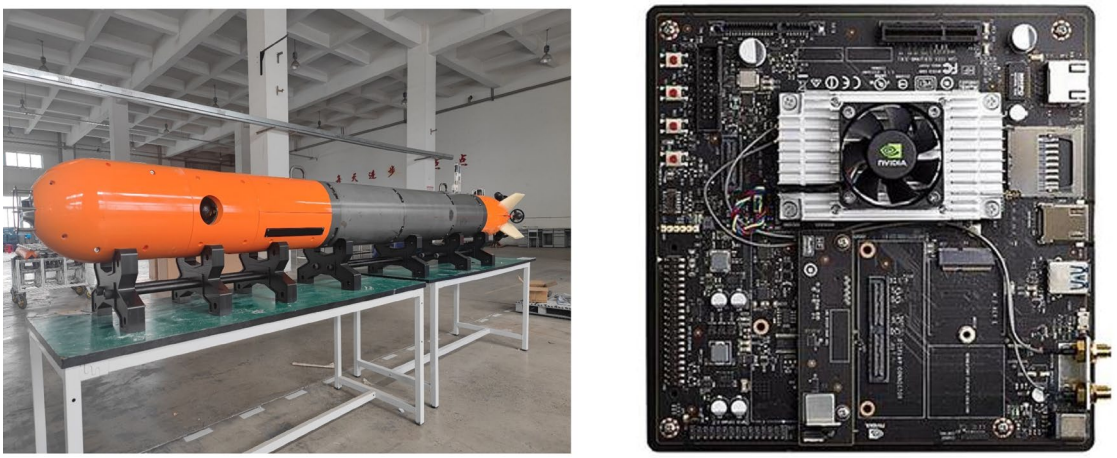
High-power and high-load computing platforms are difficult to be applied in underwater robotics due to space and power constraints. According to the actual demand, the Nvidia Jetson TX2 image edge computing device is selected as the embedded computing platform for underwater robotics. The reasons are as follows: (1) the embedded platform has a size of 50 × 87 mm and a power consumption of only 7.5 W under regular load, which satisfies the power and size requirements of underwater robotics; (2) the CPU is the ARM Cortex-A57 and the GPU is the Nvidia Pascal GPU with 256 CUDA cores, which meets the requirements of the detection algorithm.

4.5.2. Performance Comparison Test

We integrated the visual detection algorithms into the Nvidia Jetson TX2 and deployed it to an underwater robotics for performance testing on images acquired on an underwater archaeological site. The experiments were conducted on a Yuan Dynasty shipwreck site, which is located on the southeast coast with a submerged depth of 30 meters. The length and width of the shipwreck are 13.07 meters and 3.7 meters respectively. The underwater robotics camera examined an area of 48

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2023 doi:10.20944/preprints202310.0329.v1

square meters. The site contained a range of artifacts including porcelain plates, bowls, and incense burners, which were the main targets of this test. The results of this experiment are shown in Figure 9, and effective testing results were achieved.



(a) (b)

Figure 8. underwater robotics platform: (**a**) underwater robotics physical image used in the experiment; (**b**) Embedded Nvidia Jetson TX2 computing devices within underwater robotics.

Table 3. Main parameters of the underwater robotics.

Parameters	Value
Maximum operating depth	1000 m
Cruising speed	2 knots
Maximum speed	5 knots
Diameter	350 mm
Length	3.6 m
Weight in air	250Kg

To evaluate the real-time performance of the proposed target detection algorithm, we select the classical lightweight detection algorithms for comparative analysis. At the same time, two performance metrics - Frames per Second (FPS) and Model Parameters (Params) - are introduced for quantitative evaluation of the algorithms. The system performance metrics are shown in Table 4. The algorithm of this paper detects the frame rate and the number of parameters better than the SSD [32] algorithm, the YOLOv5-l [33] algorithm, and the YOLOv7 [34] algorithm in real tests. The reasons are analyzed as follows: (1) The algorithm in this paper designs MDLA as the basic feature extraction network, which effectively fuses the features of different levels by means of deep aggregation at different scales, thus improving the utilization efficiency of the features. MDLA guarantees detection accuracy while decreasing the number of parameters in the model. (2) The designed attention feature optimization module enhances the target feature information without increasing the number of model parameters. The algorithm in this paper achieves a detection speed of 19 frames per second on an image with a resolution of 640×640, which basically meets the requirements of real-time detection. Because YOLOv5-s and YOLOv7-tiny reduce the depth of the network model more, this paper's algorithm is slightly lower than the two in the detection frame rate, but the mAP is relatively higher which makes up for the disadvantage of the temporal performance.

Table 4. Inference performance of different algorithms.

Method	mAP(%)	Params(M)	Input shape	FPS
SSD[32]	82.8	24.5	640×640	15
YOLOv5-l[33]	88.7	46.5	640×640	11
YOLOv5-s[33]	86.5	14.1	640×640	21
YOLOv7[34]	89.9	74.4	640×640	10

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 October 2023 doi:10.20944/preprints202310.0329.v1

YOLOv7-tiny[34]	86.4	**13.2**	640×640	**22**
Ours	**92.8**	18.9	640×640	18

Note: Bolded text shows the optimal results for each column.

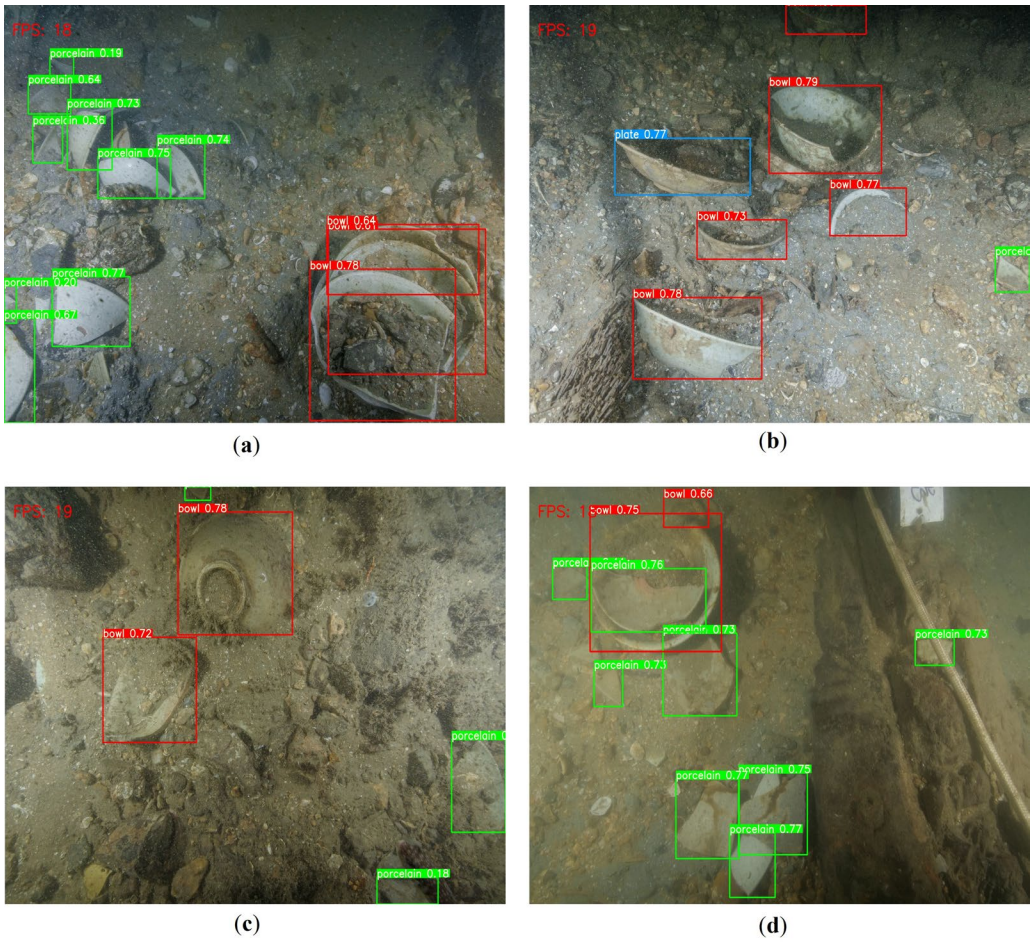


Figure 9. Test results of underwater robotics underwater visual detection system: (**a**) and (**d**) contain a large number of targets, with a detection frame rate of 18 frames; (**b**) and (**c**) contain fewer targets, with a detection frame rate of 19 frames.

5. Conclusions

In this work, we propose an underwater target detection algorithm based on the deformable deep aggregation network model for underwater robotics exploration. In order to fully capture the feature information of the target, we design a MDLA-DCN feature extraction network, in which the deformable convolution is embedded, to ensure the efficient utilization of the feature information of the underwater target in complex scenes. Furthermore, we introduce the BAM attention module for feature optimization to enhance the potential feature of the target while attenuating the background interference information. Finally, we obtain the different scale target predictions by multi-scale feature fusion. The algorithm has lightweight characteristics and is suitable for deployment on image edge computing devices. In order to verify the effectiveness of the proposed algorithm, we constructed a UCA dataset and trained and tested the algorithm. The experimental results show that the algorithm achieves scores of 93.1%, 91.4%, 92.2%, and 92.8% on the precision, recall, F1 value, and mAP metrics, respectively. It should be noted that the algorithm has been deployed to the underwater robotics to achieve a detection speed of 19 frames per second in real scene tests, which meets the real-time detection task requirements.

The algorithm proposed in this paper has high detection accuracy and computational efficiency, which can meet the task requirements of detecting artifact targets in underwater environments. The

innovative ideas of the algorithm can also be applied to other underwater target detection tasks. Although the algorithm in this paper achieves good detection results, there are still some shortcomings, such as the detection failure when marine organisms are attached to the target. In future research, we will focus on solving the problem of detection failure when marine organisms are attached to the target, and further improve the generalization ability of the algorithm model.

References

1. Fanzong, Z.; Xueting, Z.; Jingbiao, L.; Hao L.; Zhengjing, Z.; Shihe, Z. Magnetic gradient tensor positioning method implemented on an autonomous underwater vehicle platform. Journal of Marine Science and Engineering **2023**, 11(10), 1909.
2. Tao, J.; Yize, S.; Hai, H.; Hongde Q.; Xi, C.; Lingyu, L.; Zongyu, Z.; Xinyue, H. Binocular vision based non-singular fast terminal control for the UVMs small target grasp. Journal of Marine Science and Engineering 2023, 11(10), 1905.
3. Jing, Y. Protection of underwater cultural heritage in China: new developments. *International Journal of Cultural Policy* **2019**, *25*(6), 756-764.
4. Pearson, N.; Thompson, B. S. Saving two fish with one wreck: Maximizing synergies in marine biodiversity conservation and underwater cultural heritage protection. *Marine Policy* **2023**, *152*, 105613.
5. Manley, J. E. Unmanned maritime vehicles, 20 years of commercial and technical evolution. *OCEANS 2016 MTS/IEEE Monterey* **2016**, 1-6.
6. An, D.; Mu, Y.; Wang, Y.; Li, B.; Wei, Y. Intelligent Path Planning Technologies of Underwater Vehicles: a Review. *Journal of Intelligent & Robotic Systems* **2023**, *107*(2), 22.
7. Kot, R. Review of Obstacle Detection Systems for Collision Avoidance of Underwater roboticss Tested in a Real Environment. *Electronics* **2022**, *11*(21), 3615.
8. Qin, J.; Yang, K.; Li, M.; Zhong, J.; Zhang, H. Real-Time Positioning and Tracking for Vision-Based Unmanned Underwater Vehicles. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **2022**, *46*, 163-168.
9. Fayaz, S.; Parah, S. A.; Qureshi, G. J. Underwater object detection: architectures and algorithms–a comprehensive review. *Multimedia Tools and Applications* **2022**, *81*(15), 20871-20916.
10. Forsyth, D. Object detection with discriminatively trained part-based models. *Computer* **2014**, *47*(02), 6-7.
11. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* **2004**, *60*, 91-110.
12. Dalal, N.; Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2005; Vol. 1, pp. 886-893.
13. Cutter, G.; Stierhoff, K.; Zeng, J. Automated detection of rockfish in unconstrained underwater videos using haar cascades and a new image dataset: Labeled fishes in the wild. In *2015 IEEE Winter Applications and Computer Vision Workshops* (pp. 57-62). IEEE.
14. Rizzini, D. L.; Kallasi, F.; Oleari, F.; Caselli, S. Investigation of vision-based underwater object detection with multiple datasets. *International Journal of Advanced Robotic Systems* **2015**, *12*(6), 77.
15. Qiu, S.; JIN, W. Radon transform detection method for underwater moving target based on water surface characteristic wave. *Acta Optica Sinica* **2019**, *39*(10), 25-37.
16. Chen, L.; Zhou, F.; Wang, S.; Dong, J.; Li, N.; Ma, H.; Zhou, H. SWIPENET: Object detection in noisy underwater images. *arXiv preprint arXiv:2010.10006*.
17. Lei, F.; Tang, F.; Li, S. Underwater target detection algorithm based on improved YOLOv5. *Journal of Marine Science and Engineering* **2022**, *10*(3), 310.
18. Yan, J.; Zhou, Z.; Zhou, D.; Su, B.; Xuanyuan, Z.; Tang, J.; Liang, W. Underwater object detection algorithm based on attention mechanism and cross-stage partial fast spatial pyramidal pooling. *Frontiers in Marine Science* **2022**, *9*, 1056300.
19. Song, P.; Li, P.; Dai, L.; Wang, T.; Chen, Z. Boosting R-CNN: Reweighting R-CNN samples by RPN's error for underwater object detection. *Neurocomputing* **2023**, *530*, 150-164.
20. Zeng, L.; Sun, B.; Zhu, D. Underwater target detection based on Faster R-CNN and adversarial occlusion network. *Engineering Applications of Artificial Intelligence* **2021**, *100*, 104190.
21. 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 Transactions on Image Processing* **2022**, *31*, 3997-4010.
22. Shortis, M. Camera calibration techniques for accurate measurement underwater. *3D recording and interpretation for maritime archaeology* **2019**, 11-27.
23. Chen, X. Q.; Xia, K.; Hu, W.; Cao, M.; Deng, K.; Fang, S. Extraction of underwater fragile artifacts: research status and prospect. *Heritage Science* **2022**, *10*(1), 9.
24. Hu, K.; Weng, C.; Zhang, Y.; Jin, J.; Xia, Q. An overview of underwater vision enhancement: From traditional methods to recent deep learning. *Journal of Marine Science and Engineering* **2022**, *10*(2), 241.

25. Wei, S. E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2016; pp. 4724-4732.
26. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 2017; pp. 764-773.
27. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018; pp. 2403-2412.
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018; pp. 7132-7141.
29. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A. Gather-excite: Exploiting feature context in convolutional neural networks. *Advances in neural information processing systems* **2018**, *31*.
30. Zeiler, M. D.; Taylor, G. W.; Fergus, R. Adaptive deconvolutional networks for mid and high level feature learning. In *Proceedings of the IEEE conference on computer vision*. 2011; pp. 2018-2025.
31. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **2015**, *28*.
32. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C. Y.; Berg, A. C. Ssd: Single shot multibox detector. In *Proceedings of the European conference on computer vision*. 2016; pp. 21-37.
33. Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; Kwon, Y.; Michael, K.; Mammana, L. ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations. *Zenodo* **2022**.
34. 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*. 2023; pp. 7464-7475.
35. Xu, G.; Liu, K.; Zhao, Y.; Li, S.; Wang, X. Research on the modeling and simulation technology of underwater vehicle. In *OCEANS 2016-Shanghai* (pp. 1-6).