Journal of *Marine Science and Engineering*

MDPI

Article

Object Detection for Underwater Cultural Artifacts 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; yangyutuo@sia.cn (Y.Y.); weiliang@sia.cn (W.L.)
[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; zdx20110908@163.com
[5] Institute of Deep-Sea Science and Engineering, Chinese Academy of Sciences, Sanya 572000, China; xugf@idsse.ac.cn
* Correspondence: zhangyinlong@sia.cn

Abstract: Cultural artifacts found underwater are located in complex environments with poor imaging conditions. In addition, the artifacts themselves present challenges for automated object detection owing to variations in their shape and texture caused by breakage, stacking, and burial. To solve these problems, this paper proposes an underwater cultural object detection algorithm based on the deformable deep aggregation network model for autonomous underwater vehicle (AUV) exploration. To fully extract the object feature information of underwater objects in complex environments, this paper designs a multi-scale deep aggregation network with deformable convolutional layers. In addition, the approach also incorporates a BAM module for feature optimization, which enhances the potential feature information of the object while weakening the background interference. Finally, the object 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, our method has been practically deployed on an AUV. In the field testing over a shipwreck site, the artifact detection frame rate reached up to 18 fps, which satisfies the real-time object detection requirement.

Keywords: autonomous underwater vehicle; cultural artifact object detection; deformable convolution; multi-scale deep aggregation; attention mechanism

Citation: Yang, Y.; Liang, W.; Zhou, D.; Zhang, Y.; Xu, G. Object Detection for Underwater Cultural Artifacts Based on Deep Aggregation Network with Deformation Convolution. *J. Mar. Sci. Eng.* **2023**, *11*, 2228. https://doi.org/10.3390/jmse11122228

Academic Editor: Marco Cococcioni

Received: 3 October 2023
Revised: 10 November 2023
Accepted: 19 November 2023
Published: 24 November 2023

1. Introduction

During the long history of navigation, maritime losses have occurred from time to time, and a large number of shipwrecks and the objects and cargoes they contain have accumulated on the seabed [1]. Reasons for these losses are varied and include the limitations of navigation technology, the influence of extreme weather, human errors, and wars. These seabed artifacts contain rich historical, cultural, and technological information, which is of great help to the in-depth exploration of human civilization.

Detection of underwater artifacts is essential for underwater archaeological research and heritage management. It is important to understand the location of underwater archaeological sites and their conditions and contents. This is important to facilitate research and effective heritage management. Underwater heritage management is particularly important given threats such as illegal salvage or looting and the increasing expansion of offshore and seabed industries. Large-scale cataloging of shipwreck archaeological sites has typically been done by manually identifying sites from seafloor mapping data generated

from marine geophysical data such as MBES (Multibeam Echo Sounders) or SSS (Side-Scan Sonar) [2,3], although there have been some recent attempts to increase automation [4,5]. Given the increasing numbers of seafloor surveys implemented for scientific purposes, it can be argued that there is a need for increasing automation in underwater cultural artifact detection.

The increasing use of autonomous underwater vehicles (AUVs) for seafloor exploration offers great potential for realizing this goal. In seafloor exploration operations, AUVs have the advantages of a wide operating range, high detection efficiency, and flexible operation [6]. AUVs usually carry side-scan sonars and underwater cameras [7]. The side-scan sonar can be used for rapid searching of a wide range of underwater sites over large areas. The underwater camera can obtain optical images of artifacts containing rich information such as object shape, color, texture, etc., which is suitable for close-range fine detection [8]. In the process of underwater detection operation, if the AUV can autonomously recognize the object in the captured video image, it can re-plan the navigation path according to the location of the discovered object and carry out more detections around the object of interest, so as to facilitate the subsequent analysis and judgment of the seafloor artifacts [9].

Over the past two decades, researchers have made significant strides in the application of machine vision technology for underwater site detection [10,11]. Jaklic et al. [12] modeled ancient Roman shipwreck cargo sites using 3D point cloud technology. The mapping of underwater sites has been achieved through three-dimensional image processing techniques by Menna et al. [13]. Character et al. [14] employed a deep-learning algorithm model for sonar-based shipwreck image detection. However, there is a noticeable scarcity of object detection algorithms designed for optical images of underwater artifacts in the current literature, particularly in conjunction with AUVs. Therefore, the study of vision-based object detection methods for underwater artifacts holds great significance.

Vision-based object detection algorithms can be categorized into two groups: traditional object detection and deep learning-based object detection [15]. In traditional object detection algorithms, the first step involves selecting a region of interest through a sliding window approach [16]. Subsequently, various feature extraction techniques, such as Scale-Invariant Feature Transform (SIFT) [17] and Histogram of Oriented Gradients (HOG) [18], are applied to extract features from the selected region. Finally, these extracted features are used for object recognition through trained classifiers like Support Vector Machine (SVM). Cutter et al. [19] employed Haar-like features and multiple cascaded classifiers to detect fish objects, while Rizzini et al. [20] identified underwater objects based on the uniformity of underwater image color and sharpness information from contours. Qiu et al. [21] proposed an algorithm based on surface feature ripples for detecting underwater moving objects in photopolarimetric imaging mode, which has become a notable example of traditional algorithms in underwater object 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, object detection algorithms based on deep learning have achieved promising detection performance. Many researchers have applied these methods to underwater object detection scenarios. Chen et al. [22] introduced a novel sample-weighted super network (SWIPENET) to address the blurring problem in underwater images amidst significant noise interference. Lei et al. [23] incorporated the Swin Transformer into the backbone network of YOLOv5, enhancing feature extraction for underwater objects and enabling the network to detect objects in low-quality underwater images. Yan et al. [24] integrated the CBAM attention mechanism into a one-stage object detection model to enable the network to focus more on object feature information, thereby improving detection accuracy. However, the afore-

mentioned methods still suffer challenges in fully utilizing the characteristics of objects in complex underwater environments. They struggle with detection accuracy when dealing with occlusion and overlapping issues among underwater objects at different scales, as well as problems like leakage and false detection. Song et al. [25] proposed a two-stage underwater object detection algorithm with Boosting R-CNN, which enhances occluded object detection by modeling uncertainty and mining challenging samples. Zeng et al. [26] introduced a Faster R-CNN-AON network based on generative adversarial networks, effectively improving overall detection performance while preventing overfitting. Despite these advancements, it is 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 AUVs.

Currently, object detection applied to underwater cultural artifacts still faces the following challenges:

(1) Poor quality of underwater 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 [27]. In addition, the imaging quality of underwater images is low due to the insufficient underwater illumination conditions of AUVs and limited CMOS imaging levels [28].

(2) Object identification failures in complex underwater environments. Underwater artifacts have different morphologies and tend to accumulate, which makes them easily missed or incorrectly detected [29]. In addition, due to the age of underwater artifacts, the artifacts are often covered with sediments, encrusted by marine organisms, or broken, 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 samples of underwater artifacts. Unlike atmospheric optical images, it is difficult to obtain enough samples with relevant features in the preliminary research of object detection algorithms due to the influence of the complex underwater environment and the limitation of imaging equipment [30].

In order to solve the above problems, we propose an object detection algorithm specifically designed for archaeological artifacts located underwater 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 objects in complex scenarios through a deep aggregation structure with deformable convolutional layers and more jump connections.

(2) We introduce the bottleneck attention module (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 redundant parameters and computational complexity.

(3) We build a database of underwater archaeological artifacts. By collecting a large number of underwater object images, the underwater cultural artifacts (UCAs) dataset is established. The accuracy and real-time performance of the underwater objects object 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. AUV Visual Detection System

We have constructed an efficient underwater visual inspection system for AUVs, 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, in our preliminary work, we employed a robot equipped with an underwater camera to capture images at the shipwreck site and create a custom dataset for our algorithmic study.

In the subsequent phase, we integrated the algorithmic model proposed in this paper into an edge computing platform and deployed it on an autonomous underwater vehicle. During real-world testing, the system utilizes the images captured by the underwater camera, processes them using our detection algorithm for autonomous object recognition and analysis, and ultimately produces the detection results.

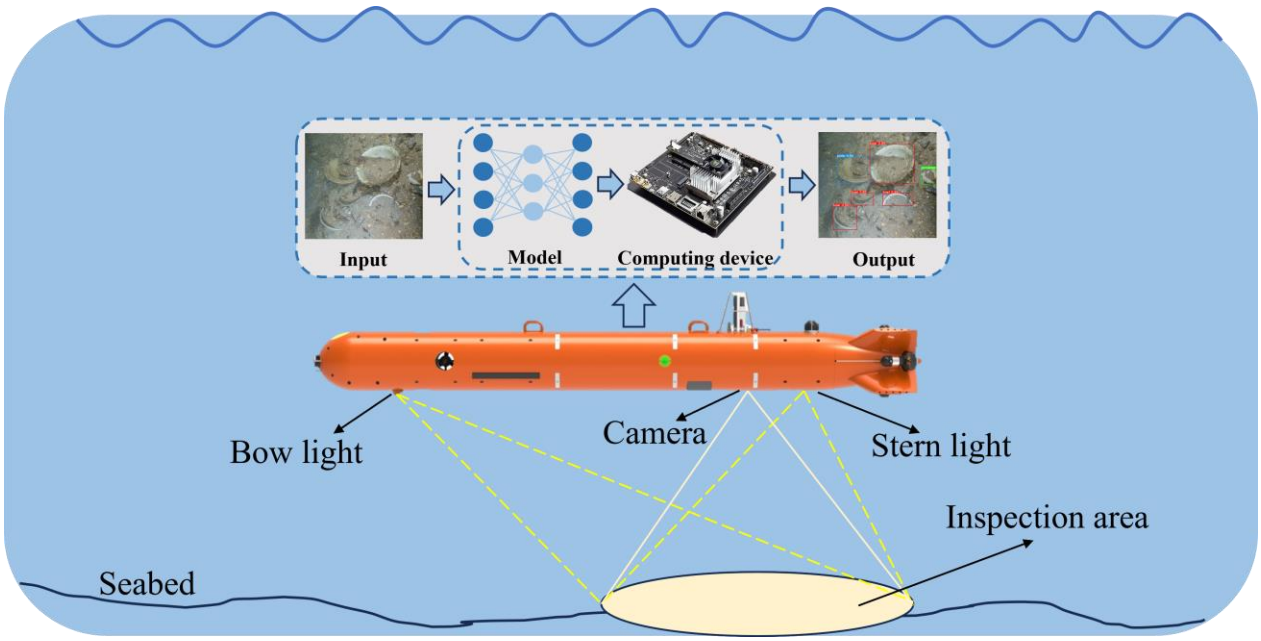


Figure 1. AUV vision-based object detection system.

Therefore, the focus of this study was to develop an artifact object detection algorithm for AUVs for shipwreck sites and to test the performance of the algorithm in real underwater environments.

3. Materials and Methods

Our underwater cultural artifacts object detection network (UCA-Net) combines a deformable convolution module and an attention mechanism to improve the performance of artifact object detection in complex underwater environments. As shown in Figure 2, UCA-Net consists of three parts: a feature extraction network, a feature optimization network, and a 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 objects. Secondly, the feature optimization network enhances the key features of underwater artifacts by introducing the BAM attention mechanism, augmenting them in both the spatial and channel dimensions while attenuating invalid background information. Finally, the feature fusion network fuses features from different scales to further enhance the algorithm's representation of the object. With the above design, the UCA-Net algorithm proposed in this paper effectively improves the accuracy and robustness of underwater artifact object detection.

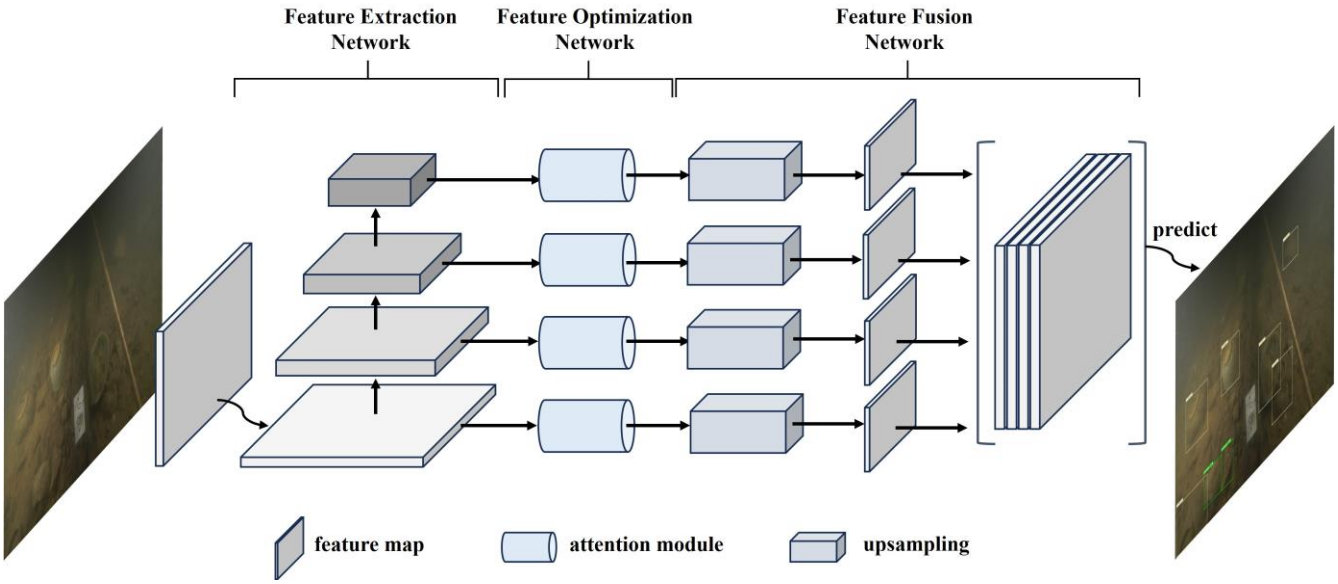


Figure 2. The framework of underwater objects detection network (UCA-Net).

3.1. Feature Extraction Network

The process of object detection for underwater archaeological artifacts is made difficult by the diversity of object types, shapes, and textures. 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 [31]. In complex underwater environments, traditional standard convolution can only perform fixed-size sampling. In contrast, deformable convolution can better learn the features of an object 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 objects such as broken burials [32]. As shown in Figure 3, the deformable convolution module adds a two-dimensional offset to each sample in the convolution kernel based on the traditional standard convolution $\{\Delta p_n | n = 1, \ldots, N\}$, $N = |R|$, 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; and Y is the output feature map.

The deep layer aggregation (DLA) network has been widely used as a compact and efficient feature extraction backbone in computer vision tasks such as object detection and semantic segmentation [33]. A DLA network merges the layered feature maps in an iterative manner, which achieves a more accurate representation of the object features while keeping fewer parameters. To adapt to the diverse object sizes and shape of archaeological artifacts found in underwater environments, we designed the DLA network structure accordingly so that it could output feature maps with four feature layers of different scales. On this basis, we introduced deformable convolution to replace the traditional convolution operation to enhance the feature extraction capability of the network for irregular objects.

We named the proposed feature extraction network as a multi-scale deep layer aggregation with a deformable convolution network (MDLA-DCN). The network shows impressive performances in complex underwater environments and significantly enhances the extraction of features for underwater artifact objects with complex morphology.

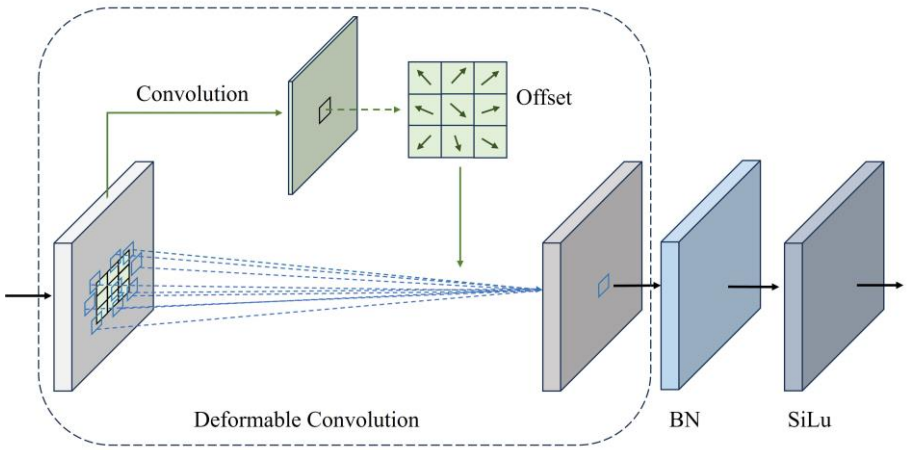


Figure 3. Deformable convolution.

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 via 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 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.

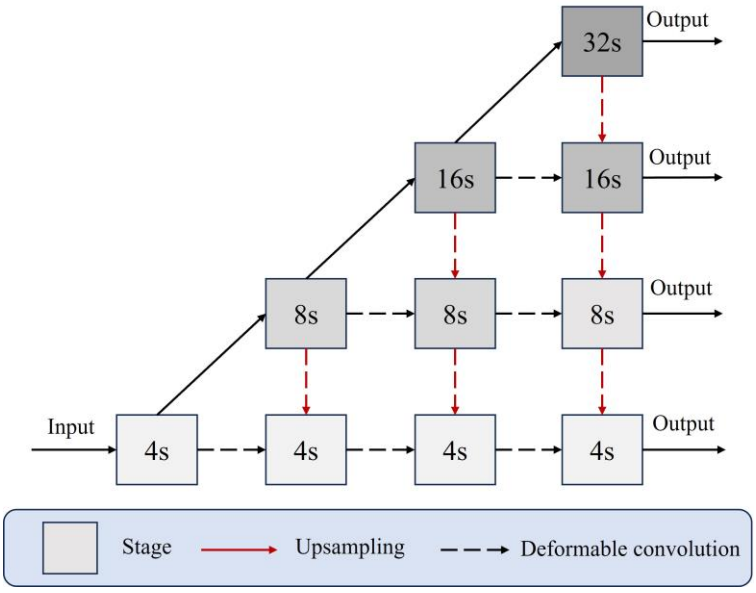


Figure 4. MDLA-DCN network.

3.2. Feature Optimization Network

The feature extraction network generates four different resolutions of feature maps, which contain valid features of the object 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 object features, as well as to enable the network to autonomously learn the correlation and importance between feature maps of different resolutions, we introduced BAM attention for feature optimization. Different from the separate channel attention [34] and spatial attention [35], BAM attention enhances features in both the spatial and channel dimensions through different branches, the structure of which is shown in Figure 5.

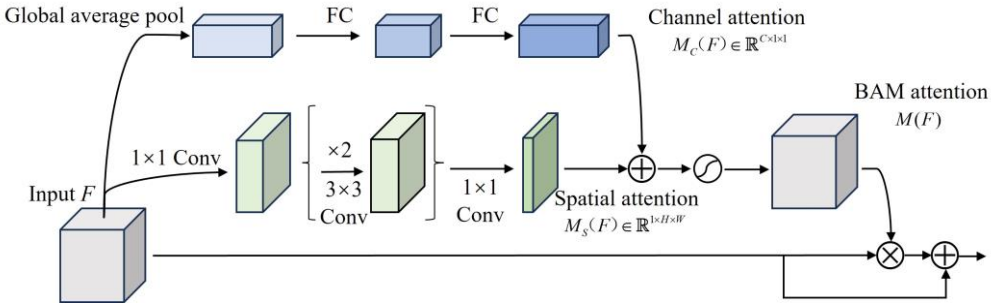


Figure 5. BAM attention mechanism.

Channel attention branching enables the network to focus 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).

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

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 object. 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 obtain the spatial attention mapping $M_S(F) \in \mathbb{R}^{H\times W}$. The specific description is shown in Equation (3).

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

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).

$$F' = F + F \otimes \sigma[M_C(F) + M_S(F)] \quad (4)$$

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 object features, and improves the quality of the output features of the sub-network significantly without increasing the parameters of the network too much.

3.3. Feature Fusion Network

After processing with the feature optimization network, feature maps at different scales were obtained, which were used to effectively represent the key features of underwater artifact objects. 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. Afterwards, 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 [36] 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 chose the inverse convolution for up-sampling. Finally, four adjusted feature maps are fused with the Concatenation fusion operation for final prediction. Through the multi-scale feature fusion, the loss of small-scale object 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 object features of underwater artifacts at different scales.

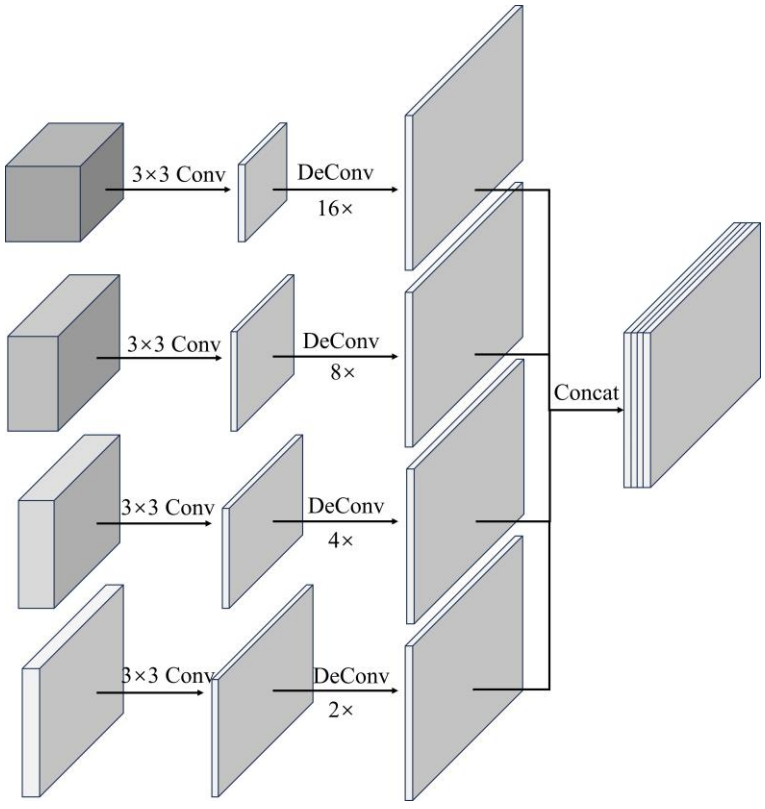


Figure 6. Illustration of the feature fusion network.

4. Experiments

In order to verify the performance of the algorithm proposed in this paper, a database of underwater archaeological artifacts was built and used for training and testing the detection model. In addition, the algorithm was compared with other state-of-the-art detection algorithms to verify the detection performance in complex underwater environments.

4.1. Underwater Object Dataset

The images of underwater artifacts were captured from two different underwater archaeological sites. Both sites are located in the sea off Guangdong Province, China, where the water depth ranges from 23 to 30 m. One site is a Southern Song Dynasty (12th century A.D.) shipwreck and the other is a Ming Dynasty (16th century A.D.) shipwreck. Both sites have large cargoes of porcelain artifacts which are scattered over the seabed. All photographs in this dataset were taken with an underwater camera carried by an AUV. Given the complexity of the underwater environment, the dataset covers a wide range of scenarios, including low light, object stacking, object burial, and object breakage. The underwater cultural artifacts (UCAs) dataset was constructed after manual screening, de-duplication, and quality assessment. The dataset consists of 10,714 images and fully covers five types of objects, namely porcelain plates, bowls, jars, incense burners, and tiles, which are commonly found in Chinese maritime trade shipwrecks from the 11th to the 17th centuries A.D. We divided the UCA dataset into training, validation, and test sets according to the ratio of 6:2:2. Examples of representative images are shown in Figure 7.

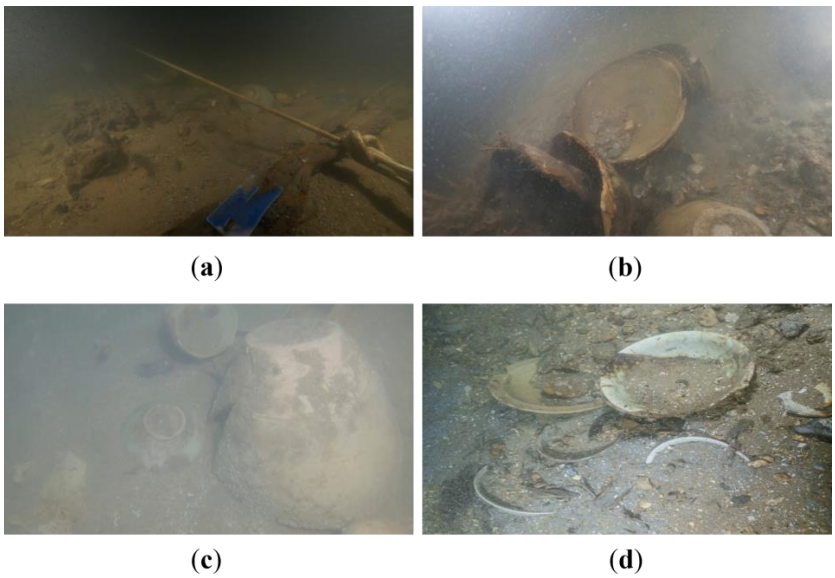


Figure 7. Illustration of the underwater cultural artifacts in different scenes: (**a**) low-light artifact; (**b**) stackable artifact; (**c**) damaged objects; (**d**) buried artifact.

4.2. Experimental Setups

4.2.1. Experimental Environment and Training Parameters

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

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

4.2.2. Model Evaluation Metrics

We used four main metrics to test the performance of the model. 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. It is used as a proxy for the model's performance and is calculated as in Equation (7). 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 object, calculated as in Equation (9).

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

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

$$AP = \int_0^1 P(R)dR \tag{8}$$

$$mAP = \frac{1}{N}\sum_{n=1}^{N} AP_n \tag{9}$$

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. Ablation Studies

To demonstrate each individual module's contribution to overall effectiveness, performance tests were conducted by successively adding or modifying the modules. Furthermore, four generalized metrics, precision, recall, F1, and mAP, were introduced to quantitatively evaluate the performance of the algorithms. The initial test used the original DLA network, which was then replaced with the MDLA network and then successively enhanced by addition of, firstly, the DCN, and secondly, the BAM module. Results for each of the model variants are shown in Table 1.

Table 1. Ablation experiments.

Method	Precision (%)	Recall (%)	F1 (%)	Map (%)
DLA [33]	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.

From the experimental results in Table 1, 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 objects. The use of DCN deformable convolution instead of ordinary convolution effectively enhances the feature extraction ability of the MDLA

network for irregular objects, 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 object. 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 object and attenuates the influence of redundant information, which further improves the individual indexes and causes the algorithm to have higher detection accuracy. The experiment proves that the addition of the 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.

4.4. Comparison with Mainstream Methods

To verify the effectiveness of the underwater object detection algorithm proposed in this paper, we conducted comparisons with mainstream object detection algorithms such as Faster-RCNN [37], SSD [38], YOLOv5-l [39], and YOLOv7 [40]. All comparisons were carried out using the aforementioned UCA dataset.

To ensure the comparability, 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 analyzed 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 Figure 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, while the YOLOv7 algorithm has better effects, but these methods still have omissions when the object 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 effective extraction of the feature information of the object in complex environments and improves the algorithm's 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. The comparisons of the algorithms' performances are shown in Table 2.

Table 2. Performance comparison of different algorithms.

Method	Precision (%)	Recall (%)	F1 (%)	mAP (%)
Faster-RCNN [37]	90.2	88.5	89.3	89.4
SSD [38]	81.9	82.2	82.0	82.8
YOLOv5-l [39]	88.3	87.8	88.1	88.7
YOLOv7 [40]	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.

Comparing the metrics of different algorithms in the table, our algorithm outperforms the others in all metrics. Compared with the SSD, which uses predictors directly based on multi-scale feature maps, the map of ours is improved by 10%. The results of Faster-RCNN and YOLOv7 are close to each other, and the map of ours is higher than both of them by 3.4% and 2.9%, respectively. Obviously, for underwater artifact objects, ours shows better detection performance. It can be seen that through the proposed network structure, the inherent features of underwater artifact objects are retained in the deep layer of the

network, which enhances the network's ability to represent the features of artifact objects in complex environments, thus improving the detection performance.

4.5. Real World Testing

4.5.1. Experimental Platform

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

The AUV and edge computing device used in the experiment are shown in Figure 8, and the main parameters of the AUV are shown in Table 3.

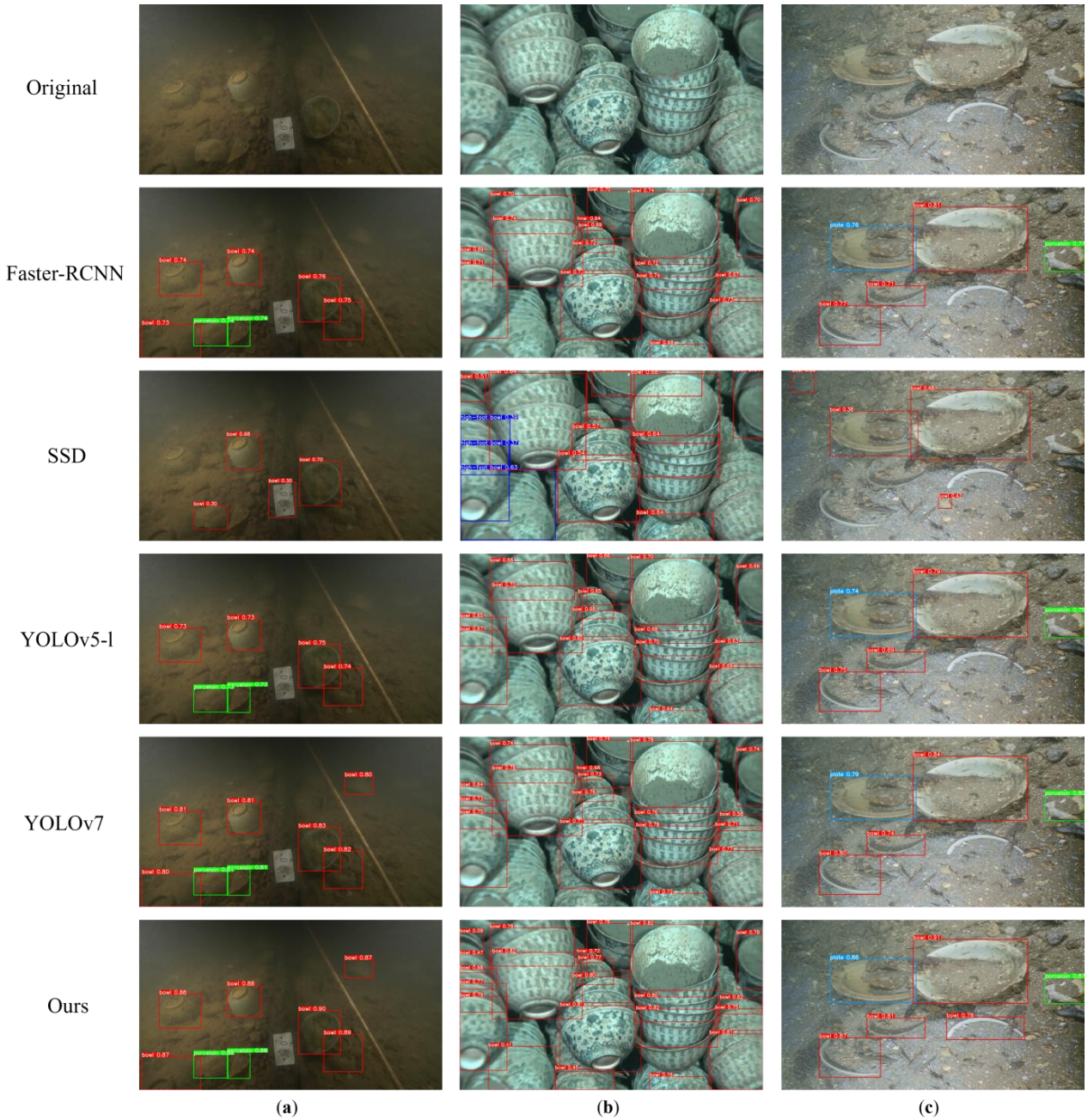


Figure 8. Comparison of the detection results among various algorithms. Different colored squares represent different objects. Red squares represent bowls; green squares represent porcelain items; light blue squares represent plates; and deep blue squares represent high-foot bowls. (**a**) Low-light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

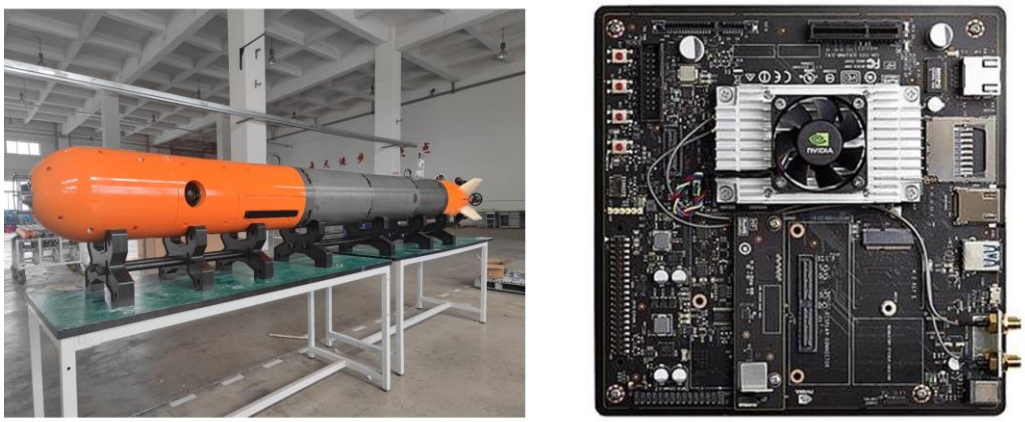
Table 3. Main parameters of the AUV.

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	250 kg

High-power and high-load computing platforms are difficult to install in underwater vehicles due to space and power constraints. While considering the actual demand, the Nvidia Jetson TX2 image edge computing device was selected as the embedded computing platform for the AUV. The reasons are as follows: (1) the embedded platform is of small size (50 × 87 mm) and low power consumption (7.5 W under regular load); (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 the AUV for performance evaluation on the images collected on an underwater archaeological site. The field experiments were conducted on a Yuan Dynasty (13th century A.D.) shipwreck site located in the southeastern waters of Fujian Province, China, at a submerged depth of 30 m. The shipwreck was chosen as the test object for this experiment because it contains a range of artifact types similar to those in the UCA dataset. The length and width of the shipwreck are 13.07 m and 3.7 m, respectively. We surveyed an area of 48 square meters, which covered the cargo hold portion of the shipwreck. The site contained a range of artifacts, including porcelain plates, bowls, and incense burners, which were the main objects of this test. The results are shown in Figure 9.



(**a**) (**b**)

Figure 9. (**a**) AUV platform; (**b**) the embedded Nvidia Jetson TX2 computing devices.

To evaluate the real-time performance of the proposed object detection algorithm, we have selected 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. The system performance metrics are shown in Table 4. The algorithm of this paper detected the frame rate and the number of parameters better than the SSD [38] algorithm, the YOLOv5-l [39] algorithm, and the YOLOv7 [40] 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 object 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 (see Figure 10). Because YOLOv5-s and YOLOv7-tiny reduce the depth of the network model more, this paper's algorithm is slightly lower than these two in the detection frame rate, but the mAP is relatively higher, which makes up for the disadvantage of the temporal performance.

Table 4. Comparisons on the inference performance between ours and state-of-the-art methods.

Method	mAP (%)	Params (M)	Input Shape	FPS
SSD [38]	82.8	24.5	640 × 640	15
YOLOv5-l [39]	88.7	46.5	640 × 640	11
YOLOv5-s [39]	86.5	14.1	640 × 640	21
YOLOv7 [40]	89.9	74.4	640 × 640	10
YOLOv7-tiny [40]	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.

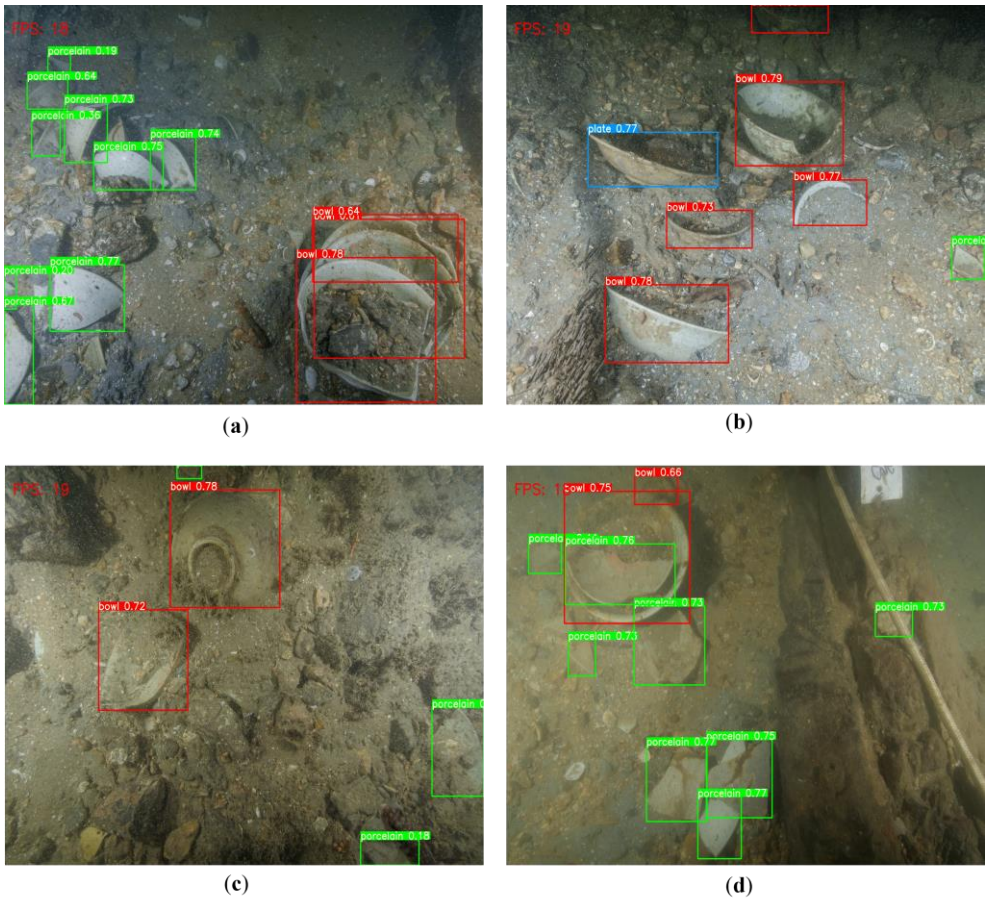


Figure 10. Typical results of cultural artifacts detection: (**a**,**d**) contain a large number of objects, and the visual detection system achieves the detection rate of 18 frames per second; (**b**,**c**) contain fewer objects, and the visual detection system achieves the detection rate of 19 frames per second.

5. Discussion

Underwater artifacts are affected by the complex environment in which they are located, as well as changes in their own shape and texture, and these problems hinder

the effective detection of underwater cultural artifacts. We have designed the proposed algorithm components to effectively improve the detection of underwater artifacts.

In this paper, we designed a deformable convolution-based multi-scale deep aggregation network for underwater cultural relics object feature extraction, which can identify and localize objects in complex environments by 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 object features and better captures the global information of the artifact objects. The BAM attention module is introduced for feature optimization, which effectively cuts down the background redundant information and makes the network focus on the object feature information. Finally, progressive feature fusion of different network layers is realized by the multi-scale feature fusion module.

From the experimental results, the algorithm in this paper has achieved better detection results. However, the seabed environment where the underwater artifacts are located is complex, and the algorithm may fail to detect them in special cases, such as the appearance of marine organisms attached to the object.

6. Conclusions

In this work, we propose an underwater cultural artifact detection algorithm based on the deformable deep aggregation network model for AUV exploration. In order to fully capture the feature information, we designed an MDLA-DCN feature extraction network in which the deformable convolution is embedded to ensure the efficient utilization of the feature information of the underwater object in complex scenes. Furthermore, we introduce the BAM attention module for feature optimization to enhance the potential feature of the object while attenuating the background interference information. Finally, we obtain the different scale object predictions using multi-scale feature fusion. The algorithm has lightweight characteristics and is suitable for deployment on edge computing devices. In order to verify the effectiveness of the proposed algorithm, we have built a UCA dataset. The experimental results show that the algorithm achieves 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 on the AUV to achieve a detection rate of 18 frames per second in real scene tests, which meets the real-time detection requirements.

The algorithm proposed in this paper has high detection accuracy and computational efficiency which can meet the task requirements of detecting artifact objects in underwater environments. The innovative ideas of the algorithm can also be applied to other underwater object detection tasks. Although the algorithm in this paper achieves good detection results, there are still some shortcomings. In future research, we will focus on solving the problem of detection failure when marine organisms are attached to the object and further improve the generalization ability of the algorithm model.

Author Contributions: Conceptualization, Y.Y. and W.L.; methodology, D.Z. and Y.Z.; software, D.Z.; validation, D.Z.; formal analysis, Y.Z.; investigation, G.X.; resources, G.X.; writing—original draft preparation, D.Z. and Y.Y.; writing—review and editing, D.Z. and Y.Z.; visualization, W.L.; supervision, W.L.; project administration, W.L. and G.X.; funding acquisition, Y.Z. and G.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China grant number 62273332, Youth Innovation Promotion Association of the Chinese Academy of Sciences grant number 2023386 and 2022201, the National Key Research and Development Program of China grant number 2020YFC1521704, Guangdong Basic and Applied Basic Research Foundation grant number 2023A1515011363.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

1. Jing, Y. Protection of underwater cultural heritage in China: New developments. *Int. J. Cult. Policy* **2019**, *25*, 756–764. [CrossRef]
2. Geraga, M.; Christodoulou, D.; Eleftherakis, D.; Papatheodorou, G.; Fakiris, E.; Dimas, X.; Georgiou, N.; Kordella, S.; Prevenios, M.; Iatrou, M.; et al. Atlas of Shipwrecks in Inner Ionian Sea (Greece): A Remote Sensing Approach. *Heritage* **2020**, *3*, 1210–1236. [CrossRef]
3. McCartney, I. Scuttled in the Morning: The discoveries and surveys of HMS Warrior and HMS Sparrowhawk, the Battle of Jutland's last missing shipwrecks. *Int. J. Naut. Archaeol.* **2018**, *47*, 253–266. [CrossRef]
4. Davis, D.S.; Buffa, D.C.; Wrobleski, A.C. Assessing the Utility of Open-Access Bathymetric Data for Shipwreck Detection in the United States. *Heritage* **2020**, *3*, 364–383. [CrossRef]
5. Bingham, B.; Foley, B.; Singh, H.; Camilli, R.; Delaporta, K.; Eustice, R.; Mallios, A.; Mindell, D.; Roman, C.; Sakellariou, D. Robotic tools for deep water archaeology: Surveying an ancient shipwreck with an autonomous underwater vehicle. *J. Field Robot.* **2010**, *27*, 702–717. [CrossRef]
6. Manley, J.E. Unmanned maritime vehicles, 20 years of commercial and technical evolution. In Proceedings of the OCEANS 2016 MTS/IEEE Monterey, Monterey, CA, USA, 19–23 September 2016; pp. 1–6.
7. An, D.; Mu, Y.; Wang, Y.; Li, B.; Wei, Y. Intelligent Path Planning Technologies of Underwater Vehicles: A Review. *J. Intell. Robot. Syst.* **2023**, *107*, 22. [CrossRef]
8. Kot, R. Review of Obstacle Detection Systems for Collision Avoidance of AUVs Tested in a Real Environment. *Electronics* **2022**, *11*, 3615. [CrossRef]
9. Qin, J.; Yang, K.; Li, M.; Zhong, J.; Zhang, H. Real-Time Positioning and Tracking for Vision-Based Unmanned Underwater Vehicles. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *46*, 163–168. [CrossRef]
10. Drap, P.; Seinturier, J.; Long, L. A photogrammetric process driven by an Expert System: A new approach for underwater archaeological surveying applied to the 'Grand Ribaud F' Etruscan wreck. In Proceedings of the 2003 Conference on Computer Vision and Pattern Recognition Workshop, Madison, WI, USA, 16–22 June 2003; Volume 1, p. 16.
11. Drap, P.; Long, L. Towards a digital excavation data management system: The "grand ribaud f" estruscan deep-water wreck. In Proceedings of the 2001 Conference on Virtual Reality, Archeology, and Cultural Heritage, Glyfada, Greece, 28–30 November 2001; pp. 17–26.
12. Jaklič, A.; Erič, M.; Mihajlović, I.; Stopinšek, Ž.; Solina, F. Volumetric models from 3D point clouds: The case study of sarcophagi cargo from a 2nd/3rd century AD Roman shipwreck near Sutivan on island Brač, Croatia. *J. Archaeol. Sci.* **2015**, *62*, 143–152. [CrossRef]
13. Menna, F.; Agrafiotis, P.; Georgopoulos, A. State of the art and applications in archaeological underwater 3D recording and mapping. *J. Cult. Herit.* **2018**, *33*, 231–248. [CrossRef]
14. Character, L.; Ortiz, A., Jr.; Beach, T.; Luzzadder-Beach, S. Archaeologic Machine Learning for Shipwreck Detection Using Lidar and Sonar. *Remote Sens.* **2021**, *13*, 1759. [CrossRef]
15. Fayaz, S.; Parah, S.A.; Qureshi, G.J. Underwater object detection: Architectures and algorithms—A comprehensive review. *Multimed. Tools Appl.* **2022**, *81*, 20871–20916. [CrossRef]
16. Forsyth, D. Object detection with discriminatively trained part-based models. *Computer* **2014**, *47*, 6–7. [CrossRef]
17. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
18. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
19. 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 Proceedings of the 2015 IEEE Winter Applications and Computer Vision Workshops, Waikoloa, HI, USA, 6–9 January 2015; pp. 57–62.
20. Rizzini, D.L.; Kallasi, F.; Oleari, F.; Caselli, S. Investigation of vision-based underwater object detection with multiple datasets. *Int. J. Adv. Robot. Syst.* **2015**, *12*, 77. [CrossRef]
21. Qiu, S.; Jin, W. Radon transform detection method for underwater moving object based on water surface characteristic wave. *Acta Opt. Sin.* **2019**, *39*, 25–37.
22. Chen, L.; Zhou, F.; Wang, S.; Dong, J.; Li, N.; Ma, H.; Zhou, H. SWIPENET: Object detection in noisy underwater images. *arXiv* **2020**, arXiv:2010.10006.
23. Lei, F.; Tang, F.; Li, S. Underwater object detection algorithm based on improved YOLOv5. *J. Mar. Sci. Eng.* **2022**, *10*, 310. [CrossRef]
24. 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. *Front. Mar. Sci.* **2022**, *9*, 1056300. [CrossRef]
25. 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. [CrossRef]

26. Zeng, L.; Sun, B.; Zhu, D. Underwater object detection based on Faster R-CNN and adversarial occlusion network. *Eng. Appl. Artif. Intell.* **2021**, *100*, 104190. [CrossRef]
27. 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] [PubMed]
28. Shortis, M. Camera calibration techniques for accurate measurement underwater. In *3D Recording and Interpretation for Maritime Archaeology*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 11–27.
29. Chen, X.Q.; Xia, K.; Hu, W.; Cao, M.; Deng, K.; Fang, S. Extraction of underwater fragile artifacts: Research status and prospect. *Herit. Sci.* **2022**, *10*, 9. [CrossRef]
30. Hu, K.; Weng, C.; Zhang, Y.; Jin, J.; Xia, Q. An overview of underwater vision enhancement: From traditional methods to recent deep learning. *J. Mar. Sci. Eng.* **2022**, *10*, 241. [CrossRef]
31. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4724–4732.
32. 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, Venice, Italy, 22–29 October 2017; pp. 764–773.
33. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep layer aggregation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2403–2412.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
35. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A. Gather-excite: Exploiting feature context in convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, Montréal, QC, Canada, 3–8 December 2018.
36. 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, Washington, DC, USA, 20–25 June 2011; pp. 2018–2025.
37. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
38. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
39. 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. Available online: https://ui.adsabs.harvard.edu/abs/2022zndo...7002879J/abstract (accessed on 2 October 2023).
40. 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.