frontiers | Frontiers in Marine Science

TYPE Methods
PUBLISHED 24 May 2023
DOI 10.3389/fmars.2023.1112310

OPEN ACCESS

EDITED BY
Zhiqun Daniel Deng,
Pacific Northwest National Laboratory (DOE), United States

REVIEWED BY
Shaoqiong Yang,
Tianjin University, China
Nicola Secciani,
University of Florence, Italy

*CORRESPONDENCE
Wei Guo
✉ guow@idsse.ac.cn
Yinlong Zhang
✉ zhangyinlong@sia.cn

RECEIVED 30 November 2022
ACCEPTED 08 May 2023
PUBLISHED 24 May 2023

CITATION
Xu G, Zhou D, Yuan L, Guo W, Huang Z and Zhang Y (2023) Vision-based underwater target real-time detection for autonomous underwater vehicle subsea exploration.
Front. Mar. Sci. 10:1112310.
doi: 10.3389/fmars.2023.1112310

Vision-based underwater target real-time detection for autonomous underwater vehicle subsea exploration

Gaofei Xu[1], Daoxian Zhou[2,3,4], Libiao Yuan[2,3,4], Wei Guo[1]*, Zepeng Huang[5] and Yinlong Zhang[2,3,4]*

[1]Institute of Deep-Sea Science and Engineering, Chinese Academy of Sciences, Sanya, China, [2]State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China, [3]Key Laboratory of Networked Control Systems, Chinese Academy of Sciences, Shenyang, China, [4]Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, China, [5]Underwater Archaeology Department, National Center for Archaeology, Beijing, China

Autonomous underwater vehicles (AUVs) equipped with online visual inspection systems can detect underwater targets during underwater operations, which is of great significance to subsea exploration. However, the undersea scene has some instinctive challenging problems, such as poor lighting conditions, sediment burial, and marine biofouling mimicry, which makes it difficult for traditional target detection algorithms to achieve online, reliable, and accurate detection of underwater targets. To solve the above issues, this paper proposes a real-time object detection algorithm for underwater targets based on a lightweight convolutional neural network model. To improve the imaging quality of underwater images, contrast limited adaptive histogram equalization with the fused multicolor space (FCLAHE) model is designed to enhance the image quality of underwater targets. Afterwards, a spindle-shaped backbone network is designed. The inverted residual block and group convolutions are used to extract depth features to ensure the target detection accuracy on one hand and to reduce the model parameter volume on the other hand under complex scenarios. Through extensive experiments, the precision, recall, and mAP of the proposed algorithm reached 91.2%, 90.1%, and 88.3%, respectively. It is also noticeable that the proposed method has been integrated into the embedded GPU platform and deployed in the AUV system in the practical scenarios. The average computational time is 0.053s, which satisfies the requirements of real-time object detection.

KEYWORDS

autonomous underwater vehicle, subsea exploration, real-time target detection, lightweight convolutional neural network, underwater image enhancement

1 Introduction

The ocean occupies the most extensive space on Earth and is an important place for the spread, exchange, and development of human civilization. During their long history, influenced by navigation technology and unpredictable marine weather conditions, a large number of ships, unfortunately sank at sea, and the ships themselves, together with the cargo they carried, were scattered on the seabed and remained dormant for thousands of years. These cultural relics are scattered on the seabed and contain rich information on history, culture, and the level of technological development and are of great significance to the study of human civilization and economic and social development. At the same time, the ocean features in a large amount of minerals and a place for military and economic activities. In addition to shipwrecks and historical cultural relics, researchers are also concerned about subsea targets such as rare biological resources, mineral resources such as manganese nodules, special geological formations such as hydrothermal and cold springs, underwater lost objects and military targets on the seafloor. The abovementioned targets, together with underwater artifacts, are the focus of attention for seabed exploration.

Autonomous underwater vehicles (AUVs) are commonly used equipment in subsea exploration and have advantages in operational range, detection efficiency, and operational flexibility compared with human-occupied vehicles (HOVs), remotely operated vehicles (ROVs) and other underwater exploration equipment (Manley, 2016). During subsea exploration missions, AUVs usually carry acoustic and optical loads, such as forward-looking sonar (FLS), 3D multibeam echosounder, side-scan sonar, camera, and so on. For the AUV studied in this paper, a forward-looking sonar, a side-scan sonar, and a camera were installed, taking into account the price, size, and weight of the load, as well as the operational objectives of the AUV. Among them, the forward-looking sonar is used to detect obstacles ahead, and the side-scan sonar and the camera are used to detect targets on the seafloor. For subsea target detection applications, the side scan sonar has a large detection range, but its resolution is low, and the target information that can be obtained is limited. For the AUV mentioned above, the side scan sonar is mainly used for rapid searches over a large range. The camera has a small detection range, but it can obtain rich information such as target shape, color, and texture, which is convenient for underwater target recognition. For the AUV mentioned above, the camera is mainly used for close range fine detection on the deep seafloor where the water quality is relatively good.

In the process of traditional underwater exploration based on AUVs, AUVs usually perform comb searches (lawn mower mode) on the seafloor according to the preplanned navigation path and take video image information of the seafloor. After the AUV is recovered to the research vessel, it is then manually judged whether the target to be searched exists in the captured video. If the target is found in the video, the location of the found target is inferred from the navigation information recorded by the AUV. When operating in this mode, the AUV cannot determine whether it has photographed the target, so even if the target is encountered in the underwater search process, it can only follow a preplanned path and cannot conduct further detailed exploration of the target and its surroundings. During subsea explorations, if the AUV can autonomously identify the targets in the captured video images, it can replan the navigation path based on the location of the discovered targets and take more shots around the targets of interest for subsequent analysis and judgment (Lin and Zhao, 2020). Therefore, it is necessary to carry out research on vision-based underwater target recognition methods so that AUVs can autonomously analyze the videos captured during operations online to improve the operational efficiency and intelligence of AUVs.

There are still many challenges for underwater target recognition, especially online underwater target recognition based on AUV platforms.

(1) The poor image quality of underwater target images. Underwater images typically suffer from color deviations and low visibility due to wavelength-dependent light absorption and scattering (Zhang et al., 2022). At the same time, insufficient lighting and low-end underwater imaging devices on board AUVs further degrade the quality of underwater images (Qiang et al., 2020; Lei et al., 2022).

(2) Underwater target samples are difficult to obtain. Due to their long history, underwater targets usually have problems such as sediment cover, marine organism attachment, damage, and incomplete shape. However, it is difficult to obtain enough samples with relevant characteristics in the early research on target recognition algorithms.

(3) Low arithmetic resources for embedded computers. Due to the limitations of equipment size and power supply capacity, high-performance computing equipment such as mainframe computer workstations commonly used in the laboratory cannot be used on AUV, and only embedded computers, for example, the NVIDIA series, can be selected for online target detection (Lin and Zhao, 2020).

In response to the above problems, this paper proposes a vision-based algorithm for real-time underwater target detection, with the following main contributions:

(1) Design of the underwater image enhancement network component. The contrast limited adaptive histogram equalization with the fused multicolor space (FCLAHE) algorithm is designed to improve the quality of underwater target images. It achieves this by performing histogram equalization in multiple color spaces.

(2) An underwater target detection algorithm based on a convolutional neural network is designed. Through group convolution and inverted residual blocks, the lightweight and efficient feature extraction network design is completed, which can further reduce the model calculation while ensuring the accuracy of the algorithm.

(3) An underwater visual inspection system that can meet the needs of the AUV online application has been built. By collecting a large number of underwater target images, the

establishment of the UCR dataset is completed, and the system proves the robustness of the algorithm through a large number of tests.

Section 2 of this paper presents related research work in the area of underwater target detection algorithms. Section 3 details the algorithmic framework and specific design elements. Section 4 describes the acquisition and production of the datasets used for the experiments. Section 5 is the experimental results section, where the advancement of this algorithm is analyzed through test comparisons. Section 6 summarizes the whole paper, illustrating the advantages of our approach in high-precision and real-time underwater target detection scenarios, as well as future research directions.

2 Related work

There are two important branches of underwater target detection algorithm research as follows. One is based on the traditional image processing method, which enhances the target detail features and then accomplishes the target detection task. The other is a deep learning-based approach, which analyzes image features and designs a feature extraction network to accomplish target detection (Yeh et al., 2021).

Traditional underwater target detection algorithms first digitize the images and then analyze them by modeling them with statistical learning theory to finally complete the detection task. One of the representatives of traditional underwater target detection is the surface feature ripple extraction underwater target detection algorithm proposed by Xu et al. (2019), which models the photoelectric polarization image and then performs underwater target detection. However, this algorithm presents different shapes when imaging at different angles, resulting in a complex mathematical model that cannot be applied in real underwater scenarios. The traditional underwater target detection algorithm has problems such as a low detection rate and poor real-time performance when detecting multicategory targets in low-light underwater environments.

The rapid expansion of underwater image data has spurred research into deep learning-based detection algorithms for various underwater marine targets. Compared with traditional algorithms, underwater target detection algorithms using deep learning have significantly improved accuracy and robustness (Moniruzzaman et al., 2017). Valdenegro-Toro (2016) introduced a CNN-based approach to build an end-to-end system, designing shared convolutional layers for object detection and recognition in sonar images. Zacchini et al. (2020) designed a deep learning-based underwater automatic target recognition (ATR) system for identifying and locating potential targets in FLS images. Song et al. (2023) proposed a two-stage underwater target detection algorithm with boosting R-CNN, which improves the detection of buried and obscured targets by modeling the uncertainty of underwater targets and difficult sample mining. Zeng et al. (2021) proposed the Faster R-CNN-AON network for the case of limited underwater samples, which effectively improved the overall detection performance by introducing an AON adversarial network to prevent the detection network from overfitting. Although the above studies have significantly improved the performance, the algorithms are poor in real-time and cannot be applied to online detection systems. To address these problems, Lei et al. (2022) introduced the Swin Transformer into the backbone network of YOLOv5 to enhance feature extraction from underwater blurred images, making the network suitable for underwater detection tasks with blurred targets. Yan et al. (2022) added the CBAM attention mechanism to the one-stage target detection model to make the network more focused on target feature information, improving detection accuracy and reducing the model. Guo et al. (2021) combined target keyframe extraction with network channel pruning to reduce the model complexity. Deep learning-based underwater target detection algorithms show better performance in complex underwater environments, but the above algorithms still cannot adapt to low computing power resources and thus cannot achieve online target detection.

3 Algorithm framework

The vision-based lightweight underwater target detection algorithm proposed in this paper and the overall framework of the algorithm are shown in Figure 1. This algorithm framework contains the following main components: (1) This paper designs an underwater image enhancement model of FCLAHE with fused multicolor space to enhance the quality of underwater target images. (2) A spindle-shaped backbone network is designed by introducing inverted residual blocks to improve the extraction of target feature information. (3) Use group convolution instead of the original standard convolution to reduce model parameters and improve inference speed.

3.1 Image enhancement algorithm of FCLAHE

Contrast limited adaptive histogram equalization (CLAHE) is a classical image enhancement algorithm that sets a threshold for each region of the histogram and spreads the number of pixels above that threshold evenly to other regions of the histogram, avoiding over-enhancement (Aggarwal and Ryoo, 2011). The specific steps of the CLAHE algorithm are as follows:

(1) Image subregion division: the original image is divided into several subregions of equal size, each subregion is nonoverlapping and contiguous, and the number of pixels in each subregion is C;

(2) A histogram of a subregion denoted by $H_{ij}(\mathrm{k})$;

(3) Calculate the threshold value: calculate the truncation limit value according to equation (1)

$$\beta = \frac{c}{L}\left(1 + \frac{\alpha}{100}(S_{\max} - 1)\right) \tag{1}$$

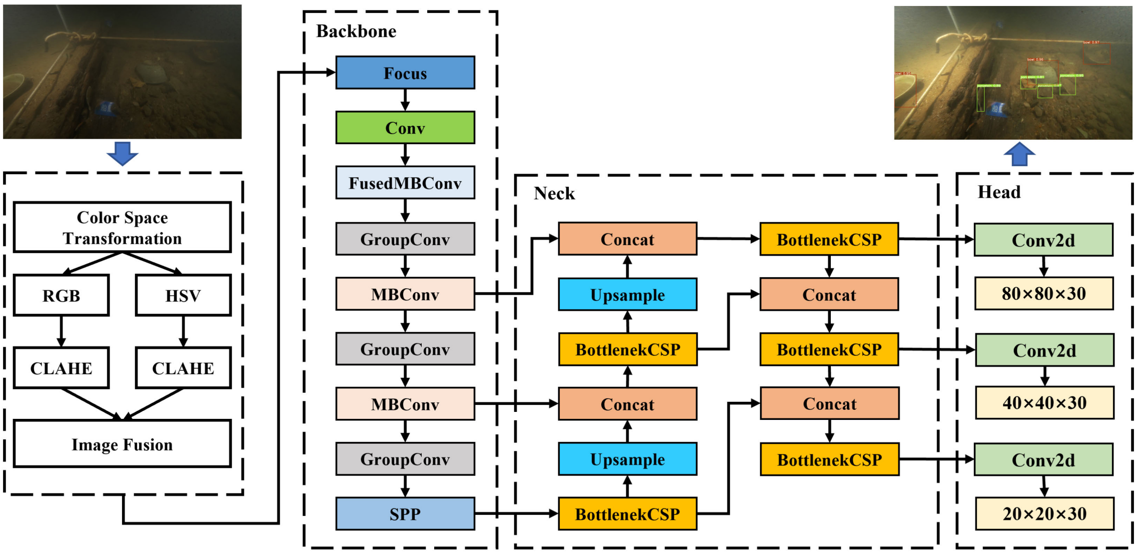

FIGURE 1
Algorithm framework.

where β is the calculated limit value; α is the truncation factor, whose value ranges from [0, 100]; and Smax is the maximum slope.

(4) Reallocation of pixel points: each subregion, H_{ij}(k) is cropped using the corresponding β value. The cropped pixels are reassigned to each gray level of the histogram in a loop until all the cropped pixels are assigned;

(5) Histogram equalization is performed separately for the cropped grayscale histogram of each subregion;

(6) Reconstructing pixel point grayscale values: the center point of each subregion is used as a reference point to obtain its gray value, and the gray value of each pixel in the output image is calculated by linear interpolation using a bilinear interpolation method.

Although the CLAHE algorithm achieves better results in image enhancement, its application in underwater environments results in color deviation and contrast reduction in underwater images due to light scattering, and the single use of the CLAHE algorithm leads to poor contrast enhancement (Ancuti et al., 2012). Therefore, this paper proposes the FCLAHE algorithm with fused multicolor space to enhance the image by pulling the gray dynamic range of the image, thus enhancing the image contrast.

To solve the problems in the CLAHE algorithm, this paper designs the algorithm of FCLAHE with fused multicolor space, which is shown in Figure 2, where the specific steps are as follows:

(1) the input images are converted into RGB and HSV color spaces and input into the CLAHE algorithm module to obtain the enhanced images I_{rgb_c} and I_{hsv_c};

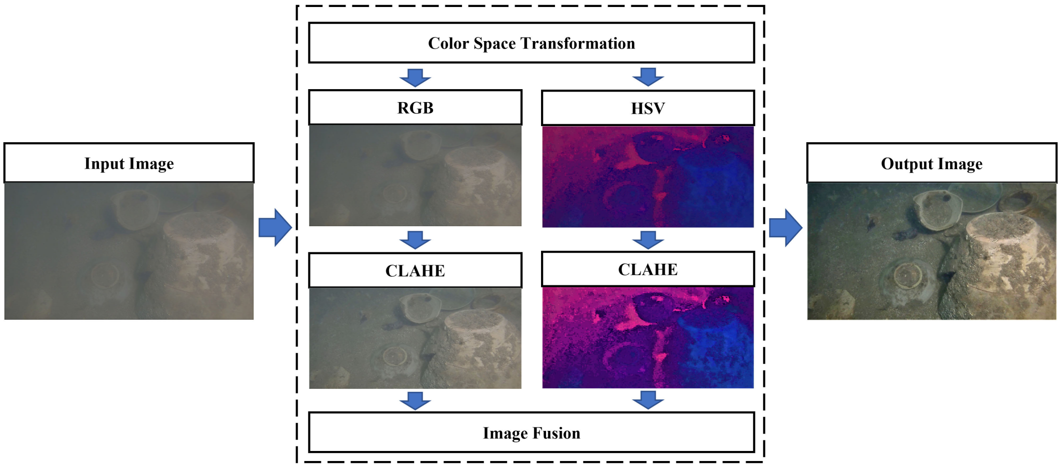

FIGURE 2
The flow of the FCLAHE algorithm.

(2) Euclidean parametric calculation of the enhanced image I_{rgb_c}, I_{hsv_c}, according to equation (2), completes the image fusion.

$$I_{merge}(i,j) = \sqrt{I^2_{rgb_c}(i,j) + I^2_{hsv_c}(i,j)} \quad (i = 0, 1\ldots, M-1; j = 0, 1\ldots, N-1) \tag{2}$$

In this paper, the algorithm is applied to underwater artifact images, and a partial comparison of the enhancement effect is shown in Figure 3.

3.2 Inverted residual block

The inverted residual block enhances the gradient propagation of the feature extraction network and significantly reduces the memory footprint required for inference. Under the same amount of computation, the network consisting of an inverted residual block contains more model parameters and is more efficient in feature extraction. The method uses the crush-and-excite attention (Hu et al., 2018) mechanism in the channel dimension to make the feature extraction network focus more on the information-rich channel features and remove the unimportant channel features, which makes it easier to distinguish between the sensing target and background information and further improves the model accuracy. The inverted residual block is mainly divided into two structures, the MBConv block (Howard et al., 2019) and Fused-MBConv (Xiong et al., 2021), as shown in Figures 4A, B.

Since most existing GPU gas pedals are optimized for standard 3×3 convolution, the MBConv block, while having fewer parameters and smaller computation, cannot take advantage of existing gas pedals, resulting in computational inefficiencies. However, the Fused-MBConv block can take advantage of GPU gas pedals to achieve a more ideal state of computational efficiency.

Therefore, the accuracy of the underwater target detection algorithm is ensured while further reducing the computational effort of the model. In this paper, we design the strategy of using the Fused-MBConv block and MBConv block together, placing the Fused-MBConv block in the shallow layer of the network and the MBConv block in the deep layer of the network. This strategy greatly improves the training and prediction speed of the model by making full use of CPU and GPU. The description of the backbone feature extraction network is shown in Table 1. The input of the backbone network is the image with a size of 640×640×3, and the feature is extracted through the Focus and Conv layers. The output is the feature map (map size: 320×320×64) which is extracted through the Fused-MBConv to obtain the shallow features of the target. The output 160×160×64 feature map goes through two MBConv layers of different scales to extract deeper feature information about the target. With the Fused-MBConv block and MBConv block, the backbone network can fully extract the features of the input image.

3.3 Group convolution

With the strategy of using the Fused-MBConv block and MBConv block together, the detection accuracy is guaranteed, and part of the computation of the model is reduced. To further satisfy the low arithmetic resource scenario of AUVs, this paper adopts group convolution instead of the original standard convolution. By grouping the input feature maps by channel, each group of feature maps is convolved with the corresponding convolution kernel in the group. Each convolution kernel is not involved in the convolution operation of the rest of the group, reducing the dimensionality of the convolution kernels and thus the computational effort.

The calculation of standard convolution and group convolution is shown in (Figures 5A, B), where C is the number of channels of the input feature map, H and W represent the height and width of the feature map, respectively, and N is the number of channels of the output feature map. When the number of convolution kernels is and the size is , the operations of the two convolutions are analyzed as follows: the size of the convolution kernel of standard

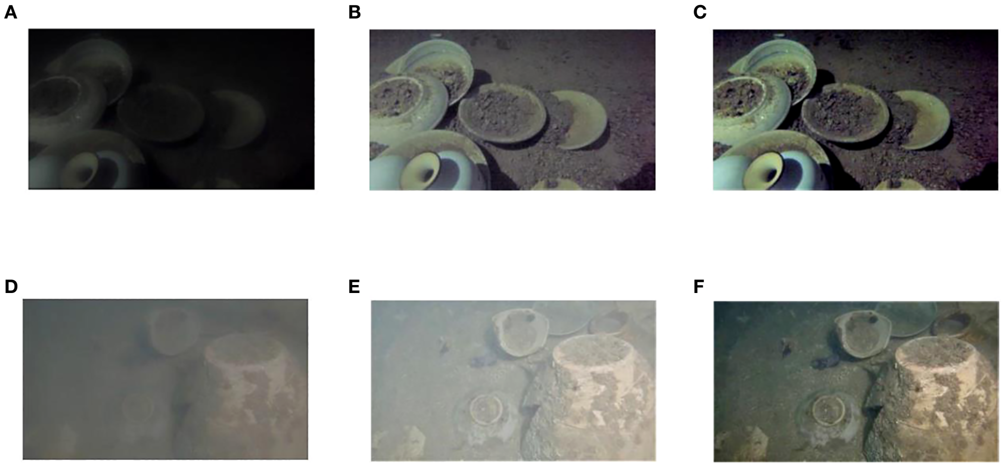

FIGURE 3
Comparison of enhancement algorithms: **(A, D)** original image, **(B, E)** CLAHE, **(C, F)** FCLAHE.

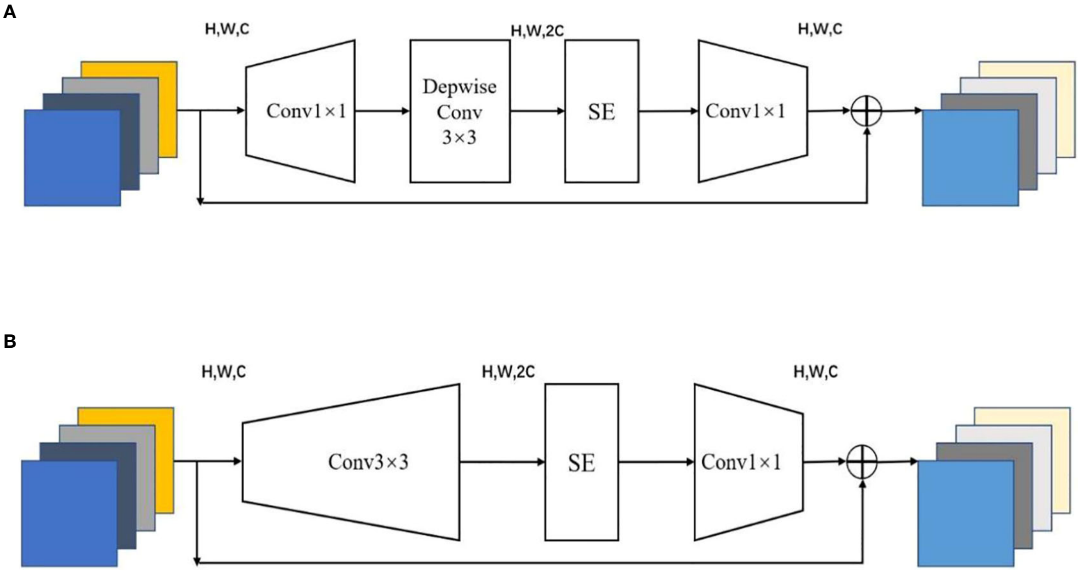

FIGURE 4
Inverted residual block: **(A)** MBConv, **(B)** Fused-MBConv.

convolution is $C \times K \times K$, and the total operation of the network is shown in equation (3).

$$Q_1 = N \times C \times K \times K \tag{3}$$

Let the number of convolutional groups be G, the number of channels per group of feature maps is $\frac{C}{G}$, the number of channels per group of output feature maps is $\frac{C}{G}$, the size of each convolutional kernel be $\frac{C}{G} \times K \times K$, the number of convolutional kernels per group is $\frac{N}{G}$, and the total number of network operations be as shown in equation (4).

$$Q_2 = N \times \frac{C}{G} \times K \times K \tag{4}$$

From equations (3)(4), it can be seen that the group convolution is $1/G$ of the total number of operations of the standard convolution under the same input conditions, and the use of group convolution significantly reduces the number of model operations.

4 Dataset

4.1 Image acquisition

The underwater target dataset constructed in this paper is obtained from two sources, one from real seafloor photography and the other from experimental simulations. The real underwater target data are images of various types of porcelain taken at an underwater archaeological site, totaling 10,000 images (resolution 1920×1080). In the experimental simulation scenario, porcelain plates, bowls, jars, and other types of porcelain targets are dropped into the water, showing different scattered states, and then filmed from different angles using underwater cameras. A total of 10,000 (resolution 1920×1080) simulated underwater target data were collected to simulate the scenes when the AUV was conducting subsea exploration, including four cases of blurring, burial, stacking, and low light. Some of the acquired images are shown in Figures 6A–D.

4.2 Production of dataset

Using Labelimg, the locations of underwater targets were manually labeled with rectangular boxes in each image, and real box labeling files in text format were obtained. The images in the dataset were enhanced using the FCLAHE image enhancement algorithm based on fused multicolor space to improve the generalization of the model to underwater blurred images. The final UCR underwater dataset was created. The dataset comprises five categories of objects, including porcelain plates, bowls, jars, censers, and porcelain fragments. These five types of porcelain are commonly found as underwater cultural relics. The completed

TABLE 1 Specification of the designed network.

Input	Operator	Channels	Activation Function	Output
640×640×3	Focus			320×320×12
320×320×12	Conv, k3×3	64	SiLu	320×320×64
320×320×64	Fused-MBConv, k3×3	64	ReLu6	160×160×64
160×160×64	MBConv, k5×5	128	H-swish	80×80×128
80×80×128	MBConv, k3×3	256	ReLu6	40×40×256
40×40×256	SPP, k5×5, 9×9, 13×13	512		20×20×512

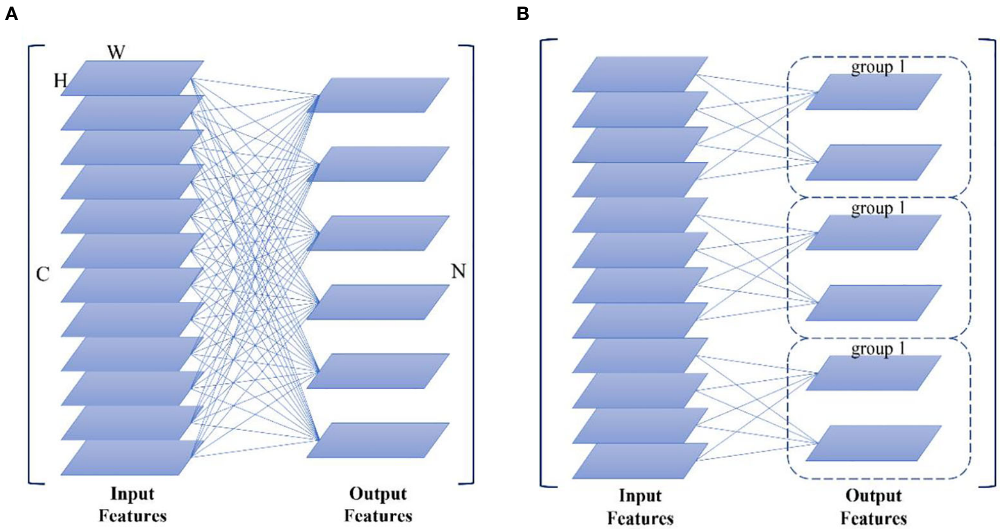

FIGURE 5
Convolution blocks: **(A)** standard convolution, **(B)** group convolution.

annotated dataset was randomly divided into a training set, a validation set and a test set at a ratio of 8:1:1. The training set was used to train the model, the validation set was used for iterative optimization during model training, and the test set was used to test the accuracy of the optimal model.

5 Algorithm comparison and experimental analysis

The hardware environment of the experimental platform in this paper is a high-performance server, which is configured as follows: Intel Core i7 processor, CPU main frequency is 3.6 GHZ, 16 GB RAM, and equipped with four Nvidia Geforce GTX 1080Ti graphics cards with 11 GB of video memory. The software environment is the Ubuntu 18.04 operating system, Python 3.7, and CUDA 11.0.

The relevant training parameters in the experiments are shown below. The gradient descent optimizer used to update the convolution kernel parameters is Adam, the optimizer momentum is 0.937, the maximum learning rate is 0.001, the batch size of training is 16, the weight attenuation coefficient is 0.0004, and the epoch (training iteration cycle) is 200.

To verify the superiority of the proposed algorithm, the following experiments are conducted: (1) comparative performance analysis of underwater vision detection algorithms. (2) The AUV deploys a visual inspection system and tests the relevant performance in an underwater environment.

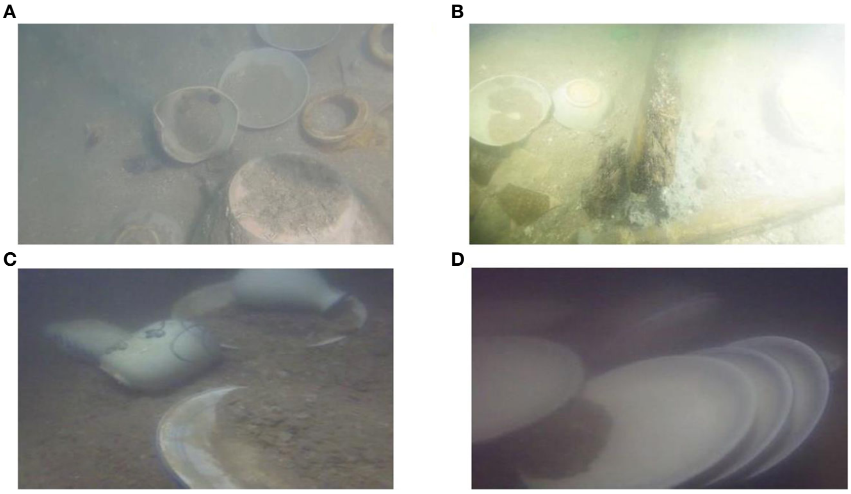

FIGURE 6
Example datasets: **(A, B)** real underwater target data, **(C, D)** simulated underwater target data.

5.1 Underwater vision inspection algorithm performance comparison analysis

To better validate the detection performance of the proposed underwater vision detection algorithms in this paper, three typical strategies are selected: DPM+SVM (Felzenszwalb et al., 2009), YOLO algorithm (Bochkovskiy et al., 2020) (Jocher, 2020), and Mask R-CNN (He et al., 2017) are compared with the algorithm in this paper.

The detection results of each strategy are shown in Figures 7A–D. Our method benefits from the optimization of the backbone feature extraction network, which can detect all targets even when they appear to be buried and overlapped. And the algorithm is robust. The DPM +SVM algorithm is the least effective, and the model generalization becomes poor when the target presents multiple angles or occlusion, and cannot be effectively detected, and the algorithm has a high rate of missed detection. The YOLO algorithm appears to miss detection when the target buried part is too large, and the robustness of the algorithm is poor. Mask R-CNN detection is obviously better than DPM+the SVM algorithm and YOLO algorithm, and can detect all targets even when they are buried and overlapped, but the confidence of the algorithm is lower than this method. The algorithm in this paper has better recognition than DPM+SVM, YOLO, and Mask R-CNN.

To better observe the superiority of each module of the algorithm proposed in this paper, the YOLO and Mask R-CNN algorithms with better detection effects were selected and three sets of comparison experiments were conducted using the UCR dataset as follows: the first group shows the performance comparison of the detection network; the second group shows the performance comparison of the algorithm with the addition of the CLAHE image enhancement module; the third group shows the performance comparison of the algorithm with the addition of the FCLAHE image enhancement module. To evaluate the algorithm performance, five general performance metrics are introduced: Precision, Recall, F1, mean Average Precision (mAP), and Parameters for evaluating the algorithm performance. The performance metrics of the three groups of algorithms are shown in Table 2.

This method outperformed other algorithms in all indicators, with 91.2%, 90.1%, 87.9%, and 88.3% for precision, recall, F1, and mAP, respectively, for the following analytical reasons. (1) The backbone network of the underwater target detection algorithm proposed in this paper uses a combination of inverse residual blocks and SE attention mechanism with feature correction capability in the channel direction, which enables the network to enhance the effective feature channels and achieve adaptive calibration of the feature channels. The algorithm helps to distinguish the foreground and background of the image more clearly, and the detection results of the model are more accurate. In the UCR dataset test, the single underwater target detection network detects better than the original YOLO algorithm and Mask R-CNN algorithm. In the second and third groups of comparison experiments, the detection network in this paper has higher detection performance improvement and better compatibility with image enhancement modules compared to the two comparison algorithm networks. (2) The FCLAHE image enhancement model designed in this paper reduces the probability of color bias in the CLAHE algorithm by fusing multiple color spaces and improves the underwater image enhancement performance. Compared with the CLAHE model, the image enhancement model designed in this paper improves the network detection performance, with 4.8%, 4.7%, 2.6%, and 2.9% improvement in precision, recall, F1, and mAP, respectively. The experimental data show that the FCLAHE image enhancement model has a more positive effect on underwater target recognition.

5.2 AUV vision inspection system performance testing

To better test the performance of the visual detection system in this paper, the system was deployed to the actual embedded computer on board the AUV, and the algorithm performance was tested using real underwater images taken at the underwater archaeological site.

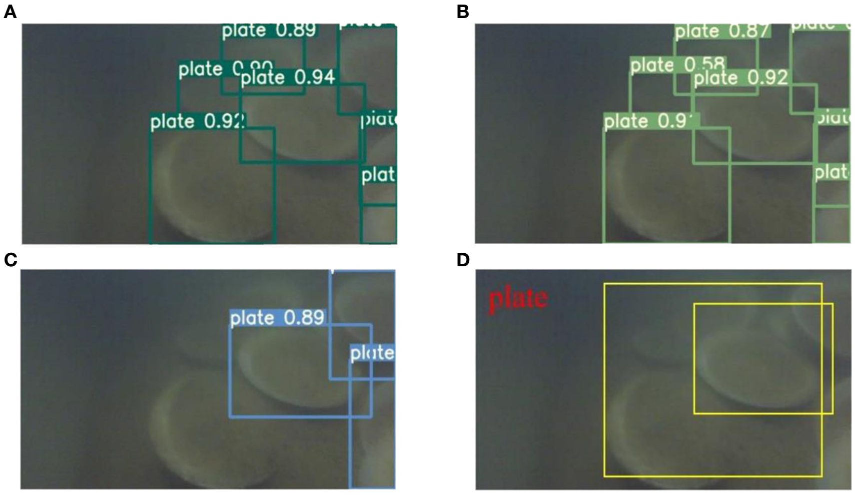

FIGURE 7
Detection results: **(A)** Ours, **(B)** DPM+SVM, **(C)** YOLO, **(D)** Mask R-CNN.

TABLE 2 Algorithm performance metrics.

Algorithm	Precision(%)	Recall(%)	F1(%)	mAP(%)	Parameters(M)
YOLO (Jocher, 2020) Mask R-CNN(He et al., 2017) SS-net	81.8 86.4 86.1	80.4 85.3 85.3	81.5 85.5 84.7	81.4 85.3 84.9	45.5 65.1 10.8
CLAHE+YOLO(Jocher, 2020) CLAHE+Mask R-CNN(He et al., 2017) CLAHE+SS-net	84.5 89.2 89.2	82.8 88.2 88.3	83.2 87.2 87.1	82.6 87.4 87.5	47.0 66.2 12.3
FCLAHE+YOLO(Jocher, 2020) FCLAHE+Mask R-CNN(He et al., 2017) FCLAHE+SS-net	86.4 90.3 **91.2**	83.9 89.7 **90.1**	85.1 87.8 **87.9**	84.2 88.0 **88.3**	48.2 67.5 **13.3**

The detection network of this paper in the experiment is called SS-net.
The bold value is the best value in the comparison.

5.2.1 Introduction of the AUV experimental platform

AUVs are commonly used equipment in seafloor exploration and have an important role in underwater-related research (Xu et al., 2016). An AUV with an online target detection function was developed for the search of underwater artifacts, shipwrecks, and other undersea targets, the main parameters of the AUV are shown in Table 3. The conceptual design of this AUV is shown in Figure 8.

The camera is arranged on the bottom of the AUV to facilitate the filming of subsea targets. To improve the underwater lighting effect, one lighting lamp is arranged at the bow and one at the stern of the AUV. The bow light was designed to be tilted backward to better illuminate the camera, as shown in Figure 9.

5.2.2 Performance test

High-power, high-load computing platforms are difficult to apply in AUVs due to space and power constraints. The Nvidia Jetson TX2 image edge computing device was selected as the AUV embedded platform based on actual requirements. The reasons are as follows: (1) The embedded platform is 50×87 mm in size and consumes only 7.5 W under regular load, meeting the power consumption and size requirements of AUVs; (2) The CPU adopts ARM Cortex-A57, and the GPU adopts Nvidia Pascal GPU with 256 CUDA cores to meet the requirements of algorithm operation.

The vision inspection system in this paper was deployed to the actual Nvidia Jetson TX2 on board the AUV, and performance tests were conducted using images obtained from the underwater archaeological site. The experiment was conducted at the shipwreck site of the Yuan Dynasty, situated on the southeast coast, at a depth of 30 meters underwater. The wreck measures 13.07 meters in length and 3.7 meters in width. The employed AUV camera examined an area spanning 48 square meters. The site contains an array of cultural artifacts, including porcelain plates, bowls, and incense burners, which constituted the primary targets of this test. The empirical results of this study are illustrated in Figure 10. This method has achieved effective detection results. The implementation of the channel attention mechanism within the inverse residual block has enabled the proposed algorithm to exhibit a high degree of robustness, particularly in circumstances where the cultural relics are obscured or partially buried.

To assess the real-time performance of the proposed algorithm in target detection, two superior strategies from Section 5.1, namely FCLAHE+YOLO and FCLAHE +Mask R-CNN, were selected for comparative analysis. Concurrently, two performance metrics—Frame Per Second (FPS) and Parameters—were introduced to facilitate a quantitative evaluation of the algorithm. The system performance metrics are presented in Table 4. This algorithm detects the frame rate FPS higher than the FCLAHE+YOLO algorithm and FCLAHE +Mask R-CNN algorithm in the actual test. The reasons for the analysis are as follows: (1) This algorithm uses group convolution instead of the original standard convolution, making the number of model parameters significantly reduced; (2) By introducing the inverse residual block and use strategy, the memory space occupation is reduced during inference, and the GPU inference acceleration is improved. This algorithm achieves a detection speed of 20 frames per second in images with a resolution of 1280×720 and 16 frames per second in images with a resolution of 1920×1080, which basically meets the requirements of real-time detection.

6 Conclusions and discussions

To achieve real-time online detection of underwater targets that meet the requirements of AUV applications with limited image quality and processor computing performance, in this paper, a vision-based lightweight underwater target detection algorithm is designed. To improve underwater imaging, an image enhancement

TABLE 3 Main parameters of the AUV.

Parameters	Value
Maximum operating depth	1000 m
Cruising speed	2 knots
Maximum speed	5 knots
Diameter	φ350 mm
Length	3.6 m
Weight in air	250Kg

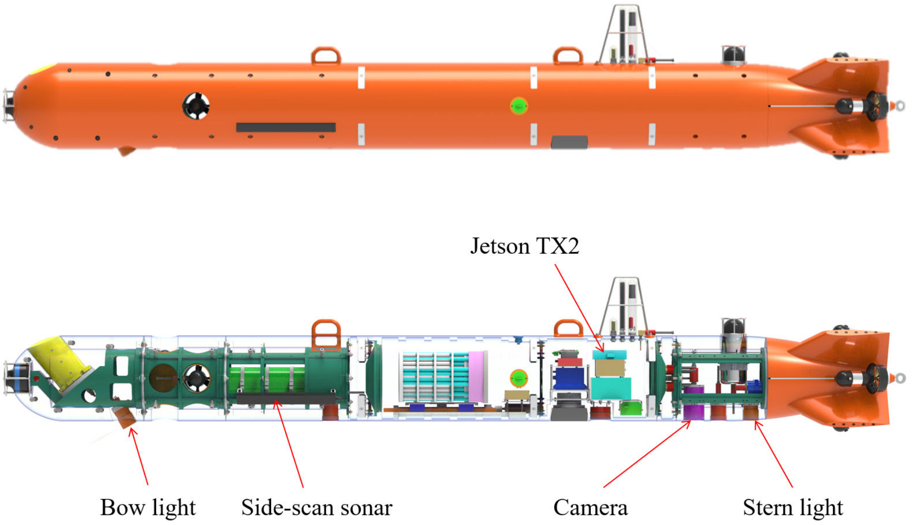

FIGURE 8
Conceptual design of the AUV.

module is added in front of the detection network, and an FCLAHE enhancement algorithm with fused multicolor space is designed for underwater image enhancement. In the detection algorithm backbone network, a new lightweight and efficient feature extraction network is designed using group convolution and inverse residual blocks to ensure the feature extraction depth while reducing the number of model parameters. This algorithm is lightweight and suitable for deployment on image edge computing devices. To verify the effectiveness of this algorithm, the UCR dataset is built to train and test the algorithm. The experimental results show that the algorithm achieves scores of 91.2%, 90.1%, 87.9%, and 88.3% for precision, recall, F1, and mAP, respectively. It is also noticeable that this method has been integrated into the embedded GPU platform and deployed to the AUV system in the real test scenario. The average computational time is 0.053 s, which satisfies the requirements of real-time object detection.

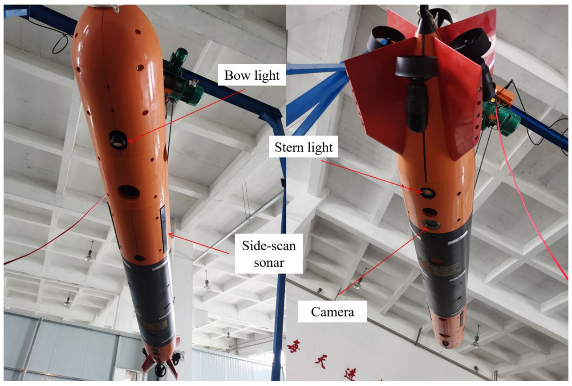

FIGURE 9
Camera and light on the bottom of the AUV.

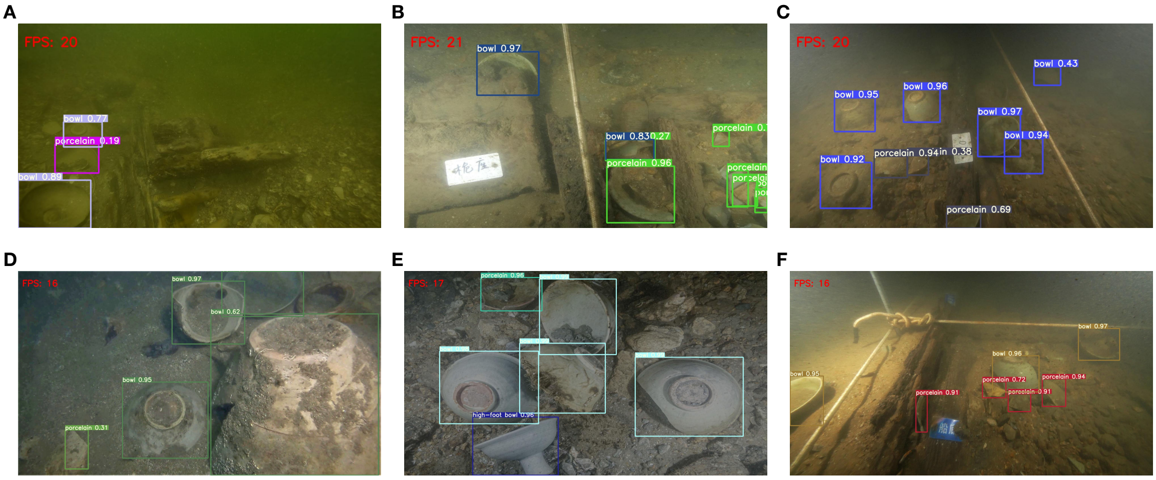

FIGURE 10
Comparison of the proposed method at different image resolutions: **(A–C)** 1280×720, **(D–F)** 1920×1080.

TABLE 4 Algorithm performance metrics.

Algorithm	mAP (%)	Parameters (M)	Input shape	FPS
FCLAHE+YOLO(Jocher, 2020)	84.2	48.2	1280×720	12
			1920×1080	9
FCLAHE+Mask R-CNN(He et al., 2017)	88.0	67.5	1280×720	7
			1920×1080	5
FCLAHE+SS-net	**88.3**	**13.3**	1280×720	**20**
			1920×1080	**16**

The bold value is the best value in the comparison.

The algorithm proposed in this paper has high detection accuracy and computational efficiency, which can satisfy the requirements of detecting artifact targets in underwater environments. The lightweight idea of this algorithm can also be applied to other underwater target detection tasks. However, there are still some problems, such as detection failures when marine organisms are attached to the target. In future research, we will expand richer datasets to further improve the generalization ability of the algorithm model. It should also be noted that instead of choosing the NAS-like architecture search strategy to obtain the optimized models in the designed hardware platform, we design a new strategy to seamlessly integrate the MBConv and Fused-MBConv together based on their structural characteristics, which can take full advantage of the hardware performance in the prediction process. Concretely, the authors put the Fused-MBConv block in the shallow layer of the network and the MBConv block in the deep layer of the network.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

Author contributions

GX designed the study, analyzed the data and wrote the initial draft of the paper. DZ and LY performed the research, analyzed the data and wrote the initial draft of the paper. WG designed the study and revised the initial draft of the paper. ZH collected the data and revised the initial draft of the paper. YZ conceived of the research, contributed the central idea and revised the initial draft of the paper. All authors contributed to the article and approved the submitted version.

Funding

This work is supported by the National Key Research and Development Program of China (2020YFC1521704), the National Natural Science Foundation of China (62273332), the Youth Innovation Promotion Association of the Chinese Academy of Sciences (2023386).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aggarwal, J. K., and Ryoo, M. S. (2011). Human activity analysis: a review. *ACM Computing Surveys (Csur)* 43 (3), 1–43. doi: 10.1145/1922649.1922653

Ancuti, C., Ancuti, C. O., Haber, T., and Bekaert, P. (2012). "Enhancing underwater images and videos by fusion," in *Proceedings of the IEEE conference on computer vision and pattern recognition.* (Providence, RI, USA: IEEE), 81–88.

Bochkovskiy, A., Wang, C. Y., and Liao, H. Y. M. (2020). Yolov4: optimal speed and accuracy of object detection. *arXiv preprint arXiv* 2004, 10934. doi: 10.48550/arXiv.2004.10934

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2009). "Object detection with discriminatively trained part-based models," in *IEEE Transactions on Pattern Analysis and Machine Intelligence* (IEEE), Vol. 32, 1627–1645.

Guo, W., Zhang, Y., Zhou, Y., Xu, G., and Li, G. (2021). Underwater real-time target detection based on key frame and model compression. *J. Physics: Conf. Ser.* 1800 (1), 012001. doi: 10.1088/1742-6596/1800/1/012001

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. *Proc. IEEE Int. Conf. Comput. Vision*, 2961–2969. doi: 10.1109/ICCV.2017.322

Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., et al. (2019). Searching for mobilenetv3. *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 1314–1324. doi: 10.1109/ICCV.2019.00140

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition.* (Salt Lake City, UT, USA: IEEE), 7132–7141.

Jocher, G. (2020) *Yolov5*. Available at: https://github.com/ultralytics/yolov5.

Lei, F., Tang, F., and Li, S. (2022). Underwater target detection algorithm based on improved YOLOv5. *J. Mar. Sci. Eng.* 10 (3), 310. doi: 10.3390/jmse10030310

Lin, S., and Zhao, Y. (2020). Review on key technologies of target exploration in underwater optical images. *Laser Optoelectronics Prog.* 57 (6), 060002. doi: 10.3788/LOP57.060002

Manley, J. E. (2016). Unmanned maritime vehicles, 20 years of commercial and technical evolution. *OCEANS 2016 MTS/IEEE Monterey*, 1–6. doi: 10.1109/OCEANS.2016.7761377

Moniruzzaman, M., Islam, S. M. S., Bennamoun, M., and Lavery, P. (2017). "Deep learning on underwater marine object detection: a survey," in *Proceedings of the 18th International Conference on Advanced Concepts for Intelligent Vision Systems.* (Springer Cham), 150–160.

Qiang, W., He, Y., Guo, Y., Li, B., and He, L. (2020). Exploring underwater target detection algorithm based on improved SSD. *Xibei Gongye Daxue Xuebao/Journal Northwestern Polytechnical Univ.* 38 (4), 747–754. doi: 10.1051/jnwpu/20203840747

Song, P., Li, P., Dai, L., Wang, T., and Chen, Z. (2023). Boosting r-CNN: reweighting r-CNN samples by RPN's error for underwater object detection. *Neurocomputing* 530, 150–164. doi: 10.1016/j.neucom.2023.01.088

Valdenegro-Toro, M. (2016). "End-to-end object detection and recognition in forward-looking sonar images with convolutional neural networks," in *2016 IEEE/OES Autonomous Underwater Vehicles (AUV)*. (Tokyo, Japan: IEEE), 144–150.

Xiong, Y., Liu, H., Gupta, S., Akin, B., Bender, G., Wang, Y., et al. (2021). "Mobiledets: searching for object detection architectures for mobile accelerators," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (Nashville, TN, USA: IEEE), 3825–3834.

Xu, G., Liu, K., Zhao, Y., Li, S., and Wang, X. (2016). Research on the modeling and simulation technology of underwater vehicle. *OCEANS 2016-Shanghai*, 1–6. doi: 10.1109/OCEANSAP.2016.7485607

Xu, M., Qiu, S., Jin, W., Yang, J., and Guo, H. (2019). Radon transform detection method for underwater moving target based on water surface characteristic wave. *Acta Optica Sin.* 39 (10), 25–37. doi: 10.3788/AOS201939.1001003

Yan, J., Zhou, Z., Su, B., and Xuanyuan, Z. (2022). Underwater object detection algorithm based on attention mechanism and cross-stage partial fast spatial pyramidal pooling. *Front. Mar. Sci.* 2299. doi: 10.3389/fmars.2022.1056300

Yeh, C. H., Lin, C. H., Kang, L. W., Huang, C. H., Lin, M. H., Chang, C. Y., et al. (2021). "Lightweight deep neural network for joint learning of underwater object detection and color conversion," in *IEEE Transactions on Neural Networks and Learning Systems* (IEEE), Vol. 33, 6129–6143.

Zacchini, L., Franchi, M., Manzari, V., Pagliai, M., Secciani, N., Topini, A., et al. (2020). "Forward-looking sonar CNN-based automatic target recognition: an experimental campaign with FeelHippo AUV," in *2020 IEEE/OES Autonomous Underwater Vehicles Symposium (AUV)*. (St. Johns, NL, Canada: IEEE), 1–6.

Zeng, L., Sun, B., and Zhu, D. (2021). Underwater target detection based on faster r-CNN and adversarial occlusion network. *Eng. Appl. Artif. Intell.* 100, 104190. doi: 10.1016/j.engappai.2021.104190

Zhang, W., Zhuang, P., Sun, H. H., Li, G., Kwong, S., and Li, C. (2022). Underwater image enhancement *via* minimal color loss and locally adaptive contrast enhancement. *IEEE Trans. Image Process.* 31, 3997–4010. doi: 10.1109/TIP.2022.3177129