Instant Deep Sea Debris Detection for Maneuverable Underwater Machines to Build Sustainable Ocean using Deep Neural Network

Baoxiang Huang[a,b], Hongfeng Zhang[a], Guojia Hou[a], Linyao Ge[c], Milena Radenkovic[d]

[a] *School of Computer Science and Technology, Qingdao University , China*
[b] *Laboratory for Regional Oceanography and Numerical Modeling, Qingdao National Laboratory for Marine Science and Technology, china*
[c] *School of Marine Technology, Institute for Advanced Ocean Study, Ocean University of China, china*
[d] *School of Computer Science and Information Technology, The University of Nottingham, United Kingdom*

Abstract

Deep sea debris is any persistent man-made material that ends up in the deep sea. The scale and rapidly increasing amount of sea debris are endangering the health of the ocean. So, many marine communities are struggling for the objective of a clean, healthy, resilient, safe, and sustainably harvested ocean. That includes deep sea debris removal with maneuverable underwater machines. Previous studies have demonstrated that deep learning can successfully extract features from seabed images or videos, and are capable of identifying and detecting debris to facilitate debris collection. In this paper, the lightweight neural network (termed DSDebrisNet), which can leverage the detection speed and identification performance to achieve instant detection with high accuracy, is proposed to implement compound-scaled deep sea debris detection. In DSDebrisNet, a hybrid loss function considering the illumination and detection problem was also introduced to improve performance. In addition, the DSDebris dataset is constructed by extracting images and video frames from the JAMSTEC dataset and labeled using a graphical image annotation tool. The experiments are implemented on the deep sea debris dataset, and the results indicate that the proposed methodology can achieve promising detection accuracy in real time. The in-depth study also provides significant evidence for the successful extension branch of artificial intelligence to the deep sea research domain.

Keywords:
Deep sea debris, deep learning-based detection method, Marine pollution, Seafloor.

1. Introduction

The history of marine debris can be traced back to entanglement reports and plastics ingestion and edible reports from the 1960s. In the early 1970s, the presence of plastic on the ocean floor was discovered from some comments and had a strong effect on marine animalsSoliño et al. [2022]. In the early 1980s, growing concern about the impact of marine litter led to a series of conferences on marine debris. By the end of 1980, the problems related to marine debris were well understood and observations were transferred to research on effective solutions to the problem Galgani et al. [1996]; Zhang et al. [2021]. Plastic production has grown dramatically over the past few decades Topouzelis et al. [2021]. In 2021, the report 'From Pollution to Solution: a global assessment of marine litter and plastic pollution' reveals that marine debris impacts the health of ecosystems, wildlife, and humans Schmaltz et al. [2020]; UNEP [2021]. As typical, global cumulative production of plastics is estimated to grow from 9.2 million tons in 2017 to 34 million tons by 2050 Geyer [2020]. Consequently, it is urgent to reduce the amount of uncontrolled or poorly managed waste flowing into the ocean Schlining et al. [2013]; Woodall et al. [2014].

The deep sea is the ultimate concentration of land debris and environmental pollutants Zhang and Peng [2022]; Agostini et al. [2021]. Debris Tracker is an open data citizen science movement that aims to track and prevent ocean plastic pollution, thereby addressing the global challenge through widespread awareness and action Jones et al. [2022]. Unquestionably, deep sea debris can definitely break, smother, and even destroy sensitive ecosystems Cui et al. [2020]; Cau et al. [2019], so the sustainable development of the ocean requires an active treatment of marine litter Botero et al. [2020]; Amon et al. [2020].

Therefore, the intellectualization of debris detection is a critical problem for the sustainable development of the marine environment and marine ecology all over the world Nurlatifah et al. [2021]. Effectively tackling the problem of deep sea debris requires a wide range of actions Peng et al. [2018]. With the help of submersibles, which are underwater robots, the seafloor environments can be visualized, sampled, and surveyed for scientific analysis Fulton et al. [2019]. The deep sea debris database provided by the Global Oceanographic Data Center (GODAC) of the Japan Agency for Marine Earth Science and Technology (JAMSTEC) is available for Marine Earth Science and Technology [2018].

To effectively remove deep sea debris, detection algorithms must be able to operate in near real-time on remotely operated vehicle underwater platforms Xue et al. [2021a]. In terms of detection methods, image processing

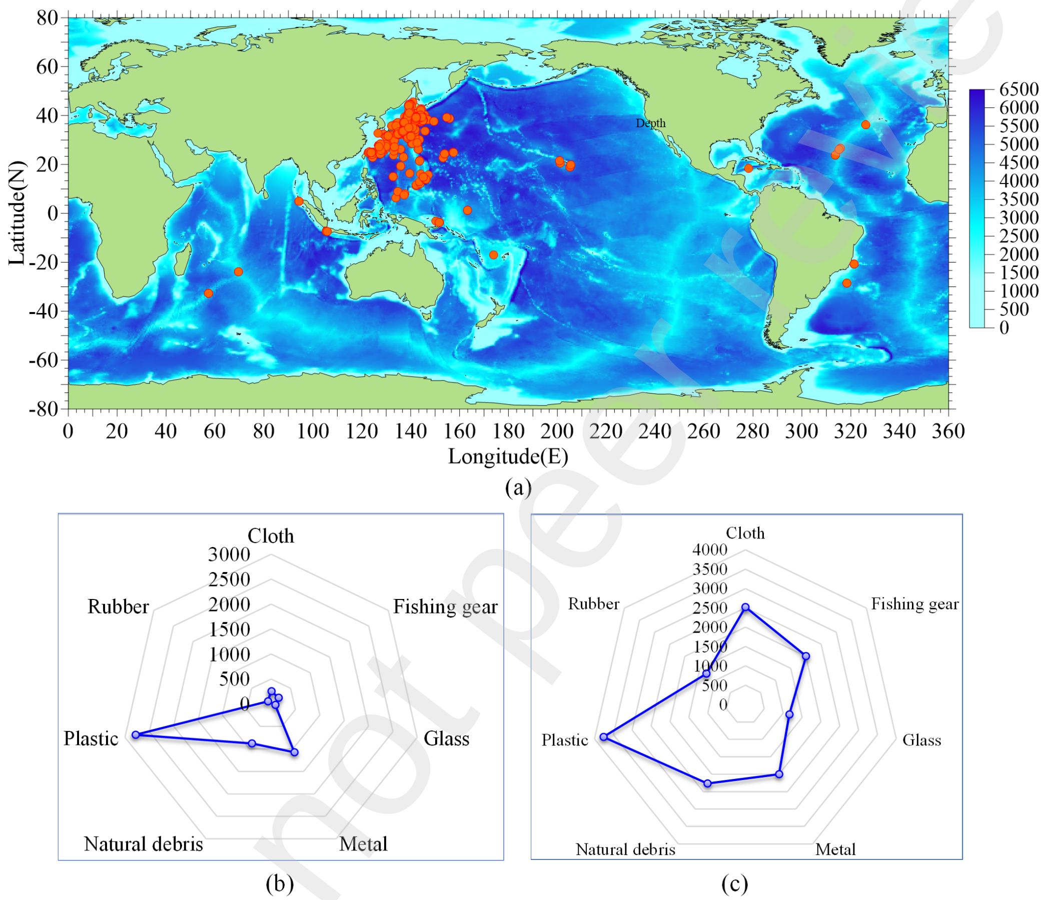

Figure 1: The overall description of the DSDebris dataset. (a) General bathymetry and location of the study area, jointly with the detailed location of observation points marked with red points, (b) Distribution of different classes in the JAMSTEC dataset, including videos and images, (c) Distribution of different classes in the DSDebris dataset, which is more balanced.

techniques and deep learning methods LeCun et al. [2015] are both available. Image processing techniques Huang et al. [2021] typically do not require a large amount of training data and are essentially unsupervised Politikos et al. [2021]. However, these technologies are limited by various factors, such as lighting effects, complex scenarios, and so on. The upper issues can be better tackled with deep learning methods. To identify the floating marine plastics, deep learning VGG16 Simonyan and Zisserman [2015] architecture is adopted to distinguish bottles, buckets, and straw debris. The performance of six well-known deep convolutional neural networks (CNNs), namely VGG19 Simonyan and Zisserman [2015], InceptionV3 Szegedy et al. [2016], ResNet50He et al. [2016], Inception-ResNetV2 Szegedy et al. [2017], DenseNet121 Huang et al. [2017], and MobileNetV2 Howard et al. [2017], are utilized as feature extractors according to three different extraction schemes for the identification of underwater marine debris. Faster-RCNN Ren et al. [2017] with transfer learning of ResNet-50 architecture was employed to detect the debris. With dataset JAMSTEC, four deep learning architectures, including YOLO (You only look once) Redmon et al. [2016], Faster RCNN, Tiny-YOLO, and Single Shot MultiBox Detector (SSD) Liu et al. [2016] with MobileNetv2 were trained using standard fine-tuning procedures to demonstrate the effectiveness of deep learning for the deep sea debris detection problem. Meanwhile, the one-stage detection network ResNet50-YOLOV3 was constructed to improve the detection performance of deep sea debrisXue et al. [2021b]. Previous studies have confirmed that the development of deep learning methods Ren et al. [2022] can facilitate the oceanographic research Han et al. [2022]; Chen et al. [2022], including marine debris removal, while there is little research to improve detection speed and real-time detection to ensure efficient debris collection procedures.

The objective of this paper is to solve the problem of instant detecting debris with the deep neural network. The underwater videos and images are degraded by uneven absorption of light due to the particles in the water. A lightweight deep learning model named DSDebrisNet is proposed in this paper. The main contributions include three aspects: (1) The number of parameters, operation cost, and weight of the proposed DSDebrisNet are decreased, thus greatly improving the operation speed; (2) The proposed DSDebrisNet is easier to deploy to maneuverable underwater machines because of the lightweight neural network; (3) Experiments of deep sea debris detection indicate that proposed DSDebrisNet methodology can achieve better index results and higher speed than the competing detection methods.

The remaining parts of the paper proceed as follows. Section 2 illustrates the dataset and problem formulation. Next, the deep sea debris detection methodology is addressed systematically in Section 3. Then, section 4 illuminate the experimental results to demonstrate the strength of the proposed DSDebrisNet. Finally, some remarks are concluded in Section 5.

2. Preliminaries

2.1. Data description

The JAMSTEC launched deep sea debris database in March 2017. This dataset provides type-specific marine debris data collected from the deep sea in the form of photos and videos, which have been taken dating back from 1983 with the help of ROVs and submersible, "SHINKAI6500", "HYPER-

DOLPHIN", etc. Only a few cases have been investigated at the deepest depths of the oceans and this dataset also consists of marine litter present at depths greater than 6000m as well. The general bathymetry and location of the study area are illustrated in Fig.1(a).

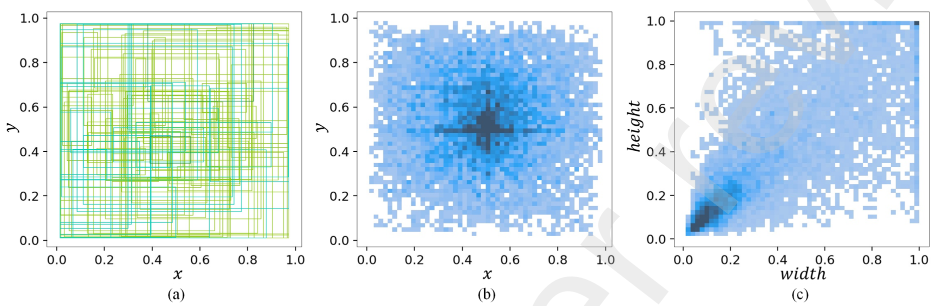

Figure 2: The location analysis of the objects to be detected with the bounding boxes. (a) Distribution of 100 random bounding boxes, (b) and (c) are the position and size normalization plots of random bounding boxes, respectively.

The videos and photos of the JAMSTEC database are categorized by shapes and materials, the categories including fishing gear, cloth, glass, metal, plastic, rubber, and natural debris. In addition, corresponding information on debris sunken to the deep sea by the locations is also available. The number of glass and rubber is at least 82 and 84, respectively, while the number of plastic is at most 2787. There is a highly unequal distribution of classes, the imbalance of the dataset can also be seen intuitively in Fig.1(b).

In this paper, the debris detection dataset, defined as the Deep Sea Debris dataset, or DSDebris dataset, is built by extracting video frames and combining them with the original image. It is noted that the imbalance of the original dataset is reduced by controlling the number of images extracted from the video and image enhancement, the distribution of different classes in DSDebris dataset is shown in Fig.1(c). Then, the bounding boxes are labeled with a graphical image annotation tool. Meanwhile, the annotations are saved as XML files to support the specific format.

Finally, the DSDebris dataset contains about 11,600 images divided into seven categories: fishing gear, cloth, glass, metal, natural debris, plastic, and rubber. To facilitate image processing, we adjusted the resolution of all images to 480*320. At the same time, due to the discrete distribution characteristics of deep sea garbage individuals, the images in the DSDebris dataset may contain more than one debris individual, meaning that some images contain multiple debris categories. The detailed information is shown in Table.1.

Table 1: The detailed number of different objects in the JAMSTEC and DSDebris datasets.

Categories \ Dataset	JAMSTEC	DSDebris
Fishing net Rope	194	2003
Cloth	253	2518
Glass	82	1161
Metal	1075	1999
Natural debris	877	2268
Plastic	2787	3768
Rubber	84	1285
Total	5352	15002

Further, the size of the target object is crucial for the performance of the detection network. To represent the distribution of location and size, the ground truth of 100 objects is randomly selected, as shown in Fig.2(a). At the same time, the bounding boxes are normalized to obtain the detailed position coordinates of the object, namely x, y, $width$ and $height$, x and y are the coordinates of the left top point of the bounding box, *width* and *height* are the width and height of the object. By carefully observing the x and y comparison graph in Fig.2(b), it can be indicated that most objects are in the central region of the image. Further, the width and height map in Fig.2(c) verifies the rationality of the aspect ratio. So the DSDebris dataset can reflect the complexity of deep sea environment as well as meet the distribution pattern required by the detection network.

2.2. Problem formulation

To better describe the detection process of deep sea debris, let $x \in \mathbb{R}^{3\times N}$ denote the input image or the frame of video with $N = m \times\ n$ pixels, where m and n are defined as the number of rows and columns in the input image, respectively, $k \equiv \{1, 2, \cdots, K\}$ denotes a vector of K class labels. The purpose of detection is to achieve multiple instances of object location with the bounding box and the corresponding object class. The task associated with object localization is to discern whether a position $x_{:,i}$ belongs to the object of a certain class k. For this purpose, a feature extractor $f(\cdot)$ and s score estimator $e(\cdot)$ are learned to extract the pixel-level features $Z = f(x) \in \mathbb{R}^{C\times N}$ and estimate the localization score $Y^* = e(Z) \in \mathbb{R}^{K\times N}$ respectively. For fully supervised object localization, the pixel-level localization mask $Y \in \mathbb{R}^{K\times N}$ is adopted as supervision for $\hat{Y}$ to learn $f(\cdot)$ and $e(\cdot)$. It is noted that the element $Y_{k,i}$ identifies whether or not pixel i belongs to the object of the class k.

Inspired by the one-stage YOLO networks, the detection problem is formulated as a regression problem that predicts the offsets and confidence of each anchor box and suppresses overlapping predictions with non-maximum suppression. Here, the input image is divided into $s \times s$ grid, each grid predicts B anchor box and the corresponding confidence score, which is defined as $confidence = Pr\left(Object\right) \times IoU\left(GT, pred\right)$, where$Pr\left(Object\right) \in [0, 1]$. Thus, the pixel-level features are transformed into grid-level features and the pixel-level localization mask $Y \in \mathbb{R}^{K \times N}$ is the anchor box, the score estimator $e\left(\cdot\right)$ is the K class probabilities. The feature vector of $f\left(\cdot\right)$ can be described as follows:

$$\hat{f} = B \times \{confidence, x, y, h, w, \phi\} \tag{1}$$

where B is the number of prior anchor boxes, $\left(x, y\right), h, w$ represents the center coordinates, height, and width of prior anchor boxes predicted, respectively, and ϕ_k is the probability that the prior anchor box belongs to the category k. The ϕ should satisfy the constraint $\sum_{k=1}^{K} \phi_k = 1$. The object function of the detection pipeline can be formulated as follows:

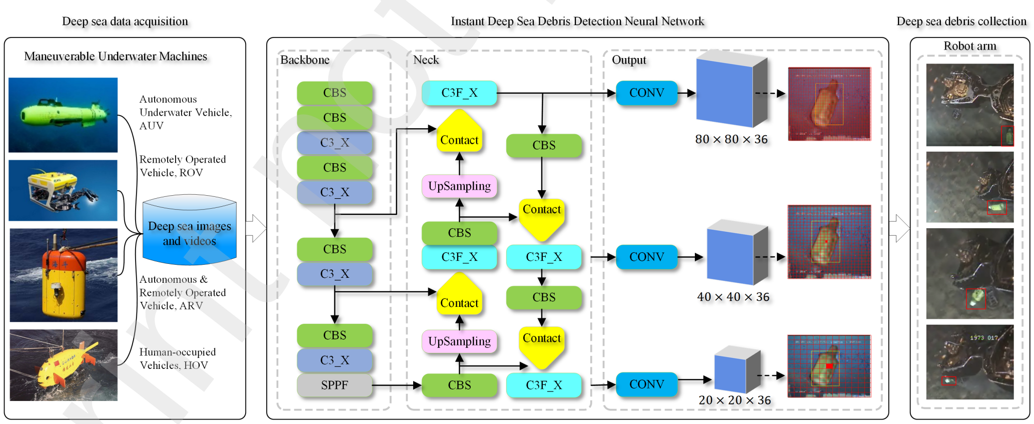

Figure 3: The workflow of the proposed methodology to detect and clean deep sea debris with DSDebrisNet for maneuverable underwater machines.

$$\mathcal{L} = \mathcal{L}_{GIoU} + \mathcal{L}_{confidence} + \mathcal{L}_{classification} \tag{2}$$

As the loss function of bounding box regression, the GIoU method can overcome the shortcomings of IoU Wang et al. [2022a] and makes full use of the advantages of IoU. Supposing $B = (x, y, h, w)$ is the prediction box, $B^{gt} = (x^{gt}, y^{gt}, h^{gt}, w^{gt})$ is the ground truth box, and C represents the smallest convex closed box containing B and B^{gt}, the calculation of GIoU loss can be formulated as follows:

$$\begin{gathered} IoU = \frac{\left|B \cap B^{gt}\right|}{\left|B \cup B^{gt}\right|} \\ \mathcal{L}_{GIoU} = 1 - IoU + \frac{\left|C - (B \cup B^{gt})\right|}{|C|} \end{gathered} \tag{3}$$

The basic components of $\mathcal{L}_{confidence}$ and $\mathcal{L}_{classification}$ are binary cross-entropy (BCE) loss. $\mathcal{L}_{confidence}$ reflects the confidence error between the ground truth box and the predicted box, as shown as follows:

$$\begin{aligned} &\mathcal{L}_{confidence} \\ &= \sum_{i=0}^{s\times s}\sum_{j=0}^{B} \ell_{ij}^{obj}\left[{C_i}^{gt} log\left(C_i\right) + \left(1 - {C_i}^{gt}\right) log\left(1 - C_i\right)\right] \\ &- \sum_{i=0}^{s\times s}\sum_{j=0}^{B} \ell_{ij}^{noobj}\left[{C_i}^{gt} log\left(C_i\right) + \left(1 - {C_i}^{gt}\right) log\left(1 - C_i\right)\right] \end{aligned} \tag{4}$$

where ${C_i}^{gt}$, C_i are the confidence of the ground truth box and predicted box, respectively. and the value of ℓ_{ij}^{obj} is 1 if the jth prior anchor box in the ith grid cell contains the object to be detected, and 0 otherwise. While ℓ_{ij}^{nonobj} is the opposite of ℓ_{ij}^{obj}.

Classification loss can evaluate the classification ability of the model through binary cross-entropy, the calculation of $\mathcal{L}_{classification}$ can be formulated as follows:

$$\mathcal{L}_{classification} = \sum_{i=0}^{s\times s}\sum_{j=0}^{B}\ell_{ij}^{obj}\sum_{k\in classes} \left[{p_i}^{gt}(k)\,log\,(p_i(k)) + \left(1-{p_i}^{gt}(k)\right)log\,(1-p_i(k))\right] \tag{5}$$

where $p_i(k)$ is the predicted probability that the predicted box can be categorized as the class k, while ${p_i}^{gt}(k)$ is the label of the ground truth box, ${p_i}^{gt}(k) \in \{0, 1\}$.

3. Methodology

As the state-of-the-art object detection system, the superior flexibility of YOLO Wang et al. [2022b] facilitates rapid deployment in the mobile hardware platforms of maneuverable underwater machines. According to different depths and widths, the YOLOv5 network is divided into four structures Wang and Liu [2022]. Considering the lightweight requirements, the focus of this study is to improve the design of YOLOv5s architecture. Specifically, the adopted YOLOv5s is the smallest model with 14.10M memory size of the YOLO series. However, the identification accuracy and response time cannot meet the requirements of maneuverable underwater machines, especially when moving at a higher speed.

Following the architecture of YOLOv5s, the basic framework DSDebrisNet also includes input, backbone, neck, and output. The input videos or images are degraded by light scattering and absorption in underwater situations, so the input is first improved with the slide stretching approach, the augmentation process has low computational cost and requirements for hardware devices. The backbone part is composed of CBS (Convolution

+ BatchNorm + SiLU), C3_X, and Spatial Pyramid pooling - fast (SPPF) modules. Feature pyramid network (FPN) Lin et al. [2017] and path aggregation network (PAN) Ni et al. [2020] modules are used to realize multi-scale information fusion Yu et al. [2022]. Finally, the network predicts both the category and position of target boxes. The architecture of the DSDebrisNet is illustrated in Fig.3. The detailed structure of different modules is explained in the following subsections.

3.1. Deep sea debris image enhancement

Mosaic, mixup, and slide stretching approaches are employed to achieve data enhancement. Mosaic splices four different images by randomly zooming, clipping, and arranging them, enriching the detection dataset and increasing the number of small targets to improve the robustness of the network. Mixup randomly reduces the transparency of two different images and superimposes them, complicating the target and also improving the robustness. The two different data enhancement methods can process multiple images at the same time, reducing the number of graphics processing units (GPUs) and improving the training speed. Meanwhile, adaptive image scaling is used instead of traditional unified scaling to improve the detection reasoning speed.

3.2. Feature extractor backbone layer

The backbone layer consists of CBS, C3_X, and SPPF modules, the detailed structures are shown in Fig.4. The basic CBS module is composed of convolution, batch normalization, and SiLU activation function

221 $x \times sigmod(x)$, and the SiLU function Elfwing et al. [2018] can be expressed
222 in Eq.6.

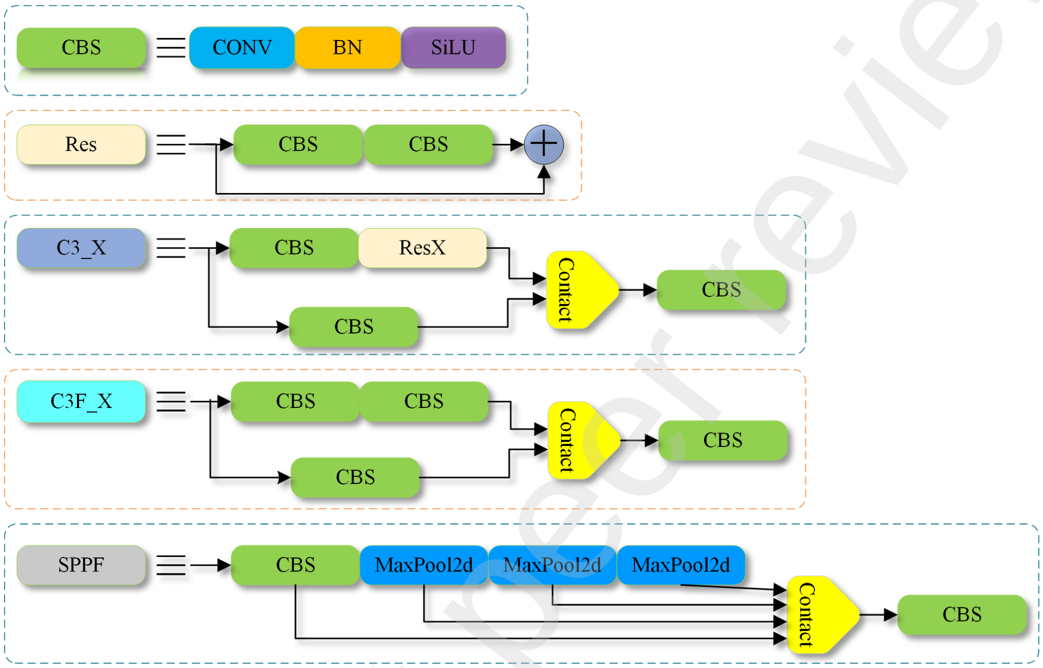

Figure 4: The detailed structure of CBS, C3_X, C3F_X, and SPPF modules in the Backbone layer.

$$SiLU = \frac{x}{1 + e^{-x}} \tag{6}$$

223 The traditional focus structure is replaced with the first CBS structure
224 with a 6×6 convolution layer. Although both can theoretically achieve the
225 same effect while the CBS structure is more efficient. Increasing or decreasing
226 the number of 1×1 convolution kernels in the C3_X module can control the
227 number of channels. The downsampling process uses a CBS structure with a
228 step size of 2 to prevent information loss, replacing the traditional maxpool

or avgpool. All of the five CBS structures in the backbone adopt convolution with a step size of 2 to make the original image subsampling 32 times, and the size of the feature matrix decreases from 640×640 to 20×20. The first CBS can increase the number of channels from 3 to 32, and the last four CBS can double the number of channels. While the C3_X module keeps the size of the feature map and the number of channels. Finally, the SPPF module is adopted to improve efficiency instead of the previous SPP module. SPPF uses the convolution kernel size of 5×5, 9×9, and 13×13 to make MaxPool and Concat with the original feature graph. The structure is parallel while more parameters are generated and the speed is slower due to the larger convolution kernel size. SPPF uses two 5×5 to replace 9×9 and three 5×5 to replace 13×13, although the structure is serial, the parameters are fewer. The feature map size of the backbone is $20 \times 20 \times 512$.

3.3. Multi-scale fusion neck layer

FPN module and PAN module are combined in the Neck structure. FPN module can realize the fusion of multi-scale information, that is, low-level detail information and high-level semantic information are fused to increase the receptive field of low-level, thus enabling low-level to obtain more context information when realizing small target detection. The bottom-up PAN module is added after the top-down path of FPN to ensure the integrity and diversity of the feature, moreover, the detection efficiency can be enhanced by preserving spatial information. The structure of the FPN and PAN module in the neck layer is illustrated in Fig.5 in detail. There are four C3F_X modules in the neck layer for extracting detailed features. C3F_X module is similar to C3_F module in the Backbone layer, the difference is that there

254 is no shortcut connection in ResX. The step size of the last two CBS structures in Neck is 2 instead of maxpool to realize downsampling. After passing through the Neck layer, three feature maps of different scales are obtained for detection, with sizes of $80 \times 80 \times 128$, $40 \times 40 \times 256$, and $20 \times 20 \times 512$, respectively.

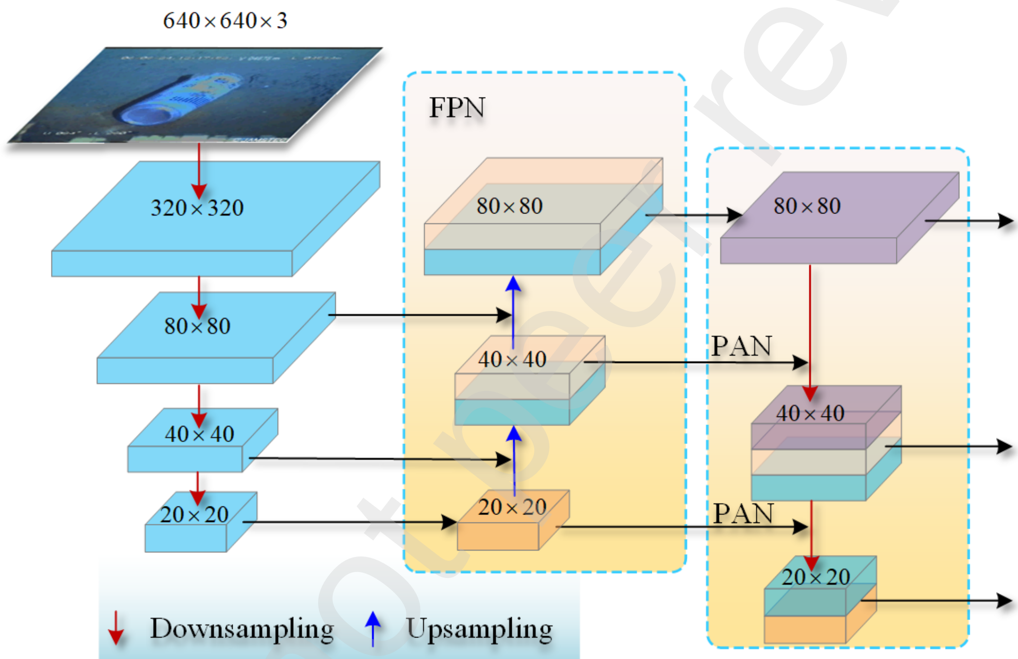

Figure 5: The detailed structure of the FPN and PAN modules in the Neck layer.

3.4. Detection output layer

Three feature maps of different scales, respectively $80 \times 80 \times 36$, $40 \times 40 \times 36$, and $20 \times 20 \times 36$, are obtained through convolution operation in the output layer, as shown in Fig.6. Different feature maps can detect objects of different sizes, which facilitates compound-scaled deep sea debris detection. Specifically, larger feature maps detect small objects because of their smaller receptive fields, and smaller feature maps detect larger objects. The value of the third dimension 36×36 represents the prediction of 3 anchors, each anchor will generate the category probabilities, 4 position coordinates, and a confidence score. Finally, the non-maximum suppression operation is implemented to screen the target boxes.

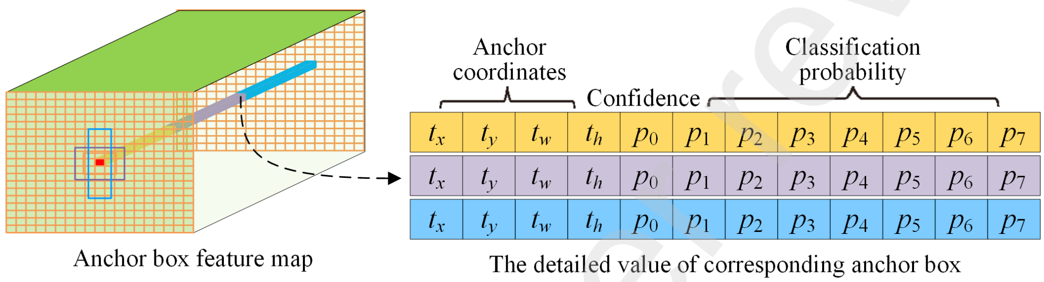

Figure 6: The anchor box feature map of the output layer. The red grid is responsible for predicting debris, and there are three anchor boxes for each grid.

4. Experimental results and analysis

In this study, the loss function adopted in the training of DSDebrisNet includes classification loss $\mathcal{L}_{classification}$, confidence loss $\mathcal{L}_{confidence}$ and boundary box position loss $\mathcal{L}_{GIoU}$. Based on $GIoU$ loss **?**, DSDebrisNet can optimize the distance losses for regressing the parameters of a bounding box to maximize the metric value.

4.1. Evaluation metrics

Deep sea debris detection determines whether the bounding box contains debris, thereby adjusting the location and size of the bounding box to accommodate the debris. The debris is to be detected if it has an intersection over union (IoU) value greater than 0.5 for any bounding boxes generated by the detector. The performance of the detector is mainly evaluated by mean average precision (mAP). Specifically, precision can reflect the closeness between the detection results and the ground truth. Recall is the evaluation and prediction of whether all debris has been identified, which reflects the proportion of correctly determined positive cases in the total positive samples. The precision-recall (P-R) curve Cui et al. [2019] can measure the quality of the object detection method, the specific formulas are as follows:

$$Precision = \frac{area\ of\ Intersection}{area\ of\ Detected\ box} \tag{7}$$

$$Recall = \frac{area\ of\ Intersection}{area\ of\ Object} \tag{8}$$

With the P-R curve, AP can be obtained by calculating the average value of the precision value corresponding to each recall value. Before AP calculation, to smooth the P-R curve and reduce the influence of curve fluctuation, interpolation is first carried out on the P-R curve. Given a recall value r_i, the $P_{interp}(r_{i+1})$ used for interpolation is the maximum precision value between the next recall value r_{i+1} and the current r_i value. Usually, AP is calculated by averaging the precision over a set of evenly uniformly recall levels $\{0,\ 0.1,\ 0.2, \cdots,\ 1.0\}$. Here $\mathrm{mAP}_{0.5}$, and $\mathrm{mAP}_{0.5:0.95}$ are chosen as the specific metrics, $\mathrm{mAP}_{0.5:0.95}$ is the mean of ten AP over the recall levels $\{0.5,\ 0.55,\ \cdots,\ 0.95\}$. The AP and $\mathrm{mAP}_{0.5}$ can be described in the following equations:

$$AP = \sum_{i=1}^{n-1} (r_{i+1} - r_i) P_{interp} (r_{i+1}) \quad (9)$$

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K} \quad (10)$$

4.2. Implementation details

The proposed DSDebrisNet is implemented by the PyTorch framework with Inter(R) Core (TM) i7- 11700K @3.60GHz, 64GB memory, NVIDIA GeForce RTX 3090. 80% of the DSDebris dataset is used for training and the remaining 20% is employed as the test dataset. To increase the diversity of the dataset, hue, saturation, and exposure are set to 0.015, 0.7, and 0.4, respectively. At the same time, mosaic data augmentation combines four images into one image by random scaling, random cropping, and random arrangement is adopted to identify the smaller-scale objects Bedford and Hanson [2022]. The optimizer is stochastic gradient descent (SGD) and the momentum factor is set to 0.937. The learning rate adopted the one_cycle method to slowly increase and then decrease. In the last part of the training, the learning rate decreased lower than the previous minimum value. Using this strategy not only speeds up training but also helps to prevent the model from falling into steep regions of the loss plane, making the model more inclined to look for minima in flatter parts, thus alleviating overfitting. Considering the similar distribution between the DSDebris dataset and the COCO dataset Yang et al. [2022], we employed the pre-trained model of YOLOV5 for fine-tuning instead of K-means clustering for anchors recalculation to speed up the convergence of the DSDebrisNet. Meanwhile, all the images are transformed into $640 time 640$ by adaptive image scaling and then fed into the DSDebrisNet. The batch size and CPU thread are set to 64 and 4, respectively. During the test procedure, non-maximum suppression (NMS) was used to remove some repeated prediction boxes, the IoU of NMS was set to 0.45, and 300 epochs are sufficient for the proposed network to converge.

The trends of loss value in the training and testing process of DSDebrisNet, including bounding box loss, confidence loss, and classification loss, represent the differences between the predicted value and the true value, as shown in Fig.7. The fast loss decrease of DSDebrisNet in the training process indicates that the proposed DSDebrisNet can locate the target position with high convergence speed. At the beginning of training, the loss function value of 30 epochs drops sharply. When the epoch reaches 100, the loss value tends to be stable. DSDebrisNet updates the network parameters with the loss function. The convergence positions of each loss function in the training and testing process are less than 0.03, which demonstrates that the model has good robustness, thus realizing the effective prediction of the model.

The detailed values of $\mathrm{mAP}_{0.5}$, and $\mathrm{mAP}_{0.5:0.95}$ in the training process are shown in Fig.8(a). It can be seen that the curve rises steadily without significant oscillations. Meanwhile, the confidence-precision and confidence-recall curves, as shown in Fig.8(b) and Fig.8(c) respectively, visualize how the DSDebrisNet predicts the positive class. It can be seen that the precision and recall indexes of rubber debris are high because of the distinguishing features to identify. While plastic debris is the worst, the reason for this phenomenon is that the size of plastic debris is very different, resulting in

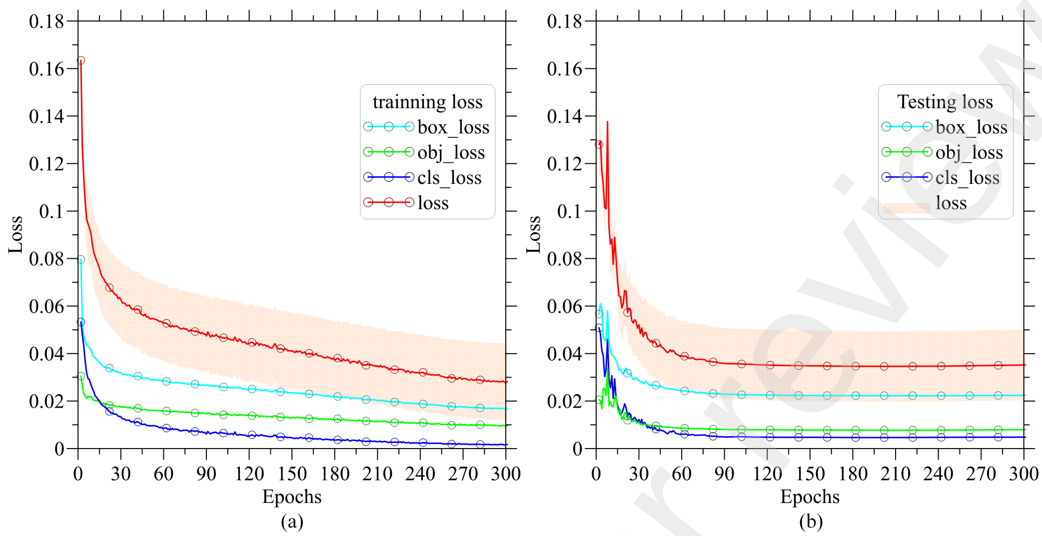

Figure 7: Loss curve during the (a) training and (b) testing process of DSDebrisNet.

low precision and recall.

4.3. Comparisons with other methods

The Faster R-CNN, which is the originator of the two-stage detection network, and SSD, which is known as a one-stage monitoring network with the pyramid-shaped feature hierarchy, and YOLOV3, which is the most important in the YOLO series of one-stage monitoring networks are selected as comparative detection methods. Previous studies have demonstrated that two-stage detection networks should be combined with an efficient classification backbone to improve the detection performance and ResNet 50 can do. Consequently, three competing networks are Resnet50-Faster R-CNN, Resnet50-SSD, and YoloV3.

The $mAP_{0.5}$ and $mAP_{0.5:0.95}$ of different models are shown in Table. 2 in detail. It can be seen that Resnet50-Faster R-CNN, as a two-stage detection

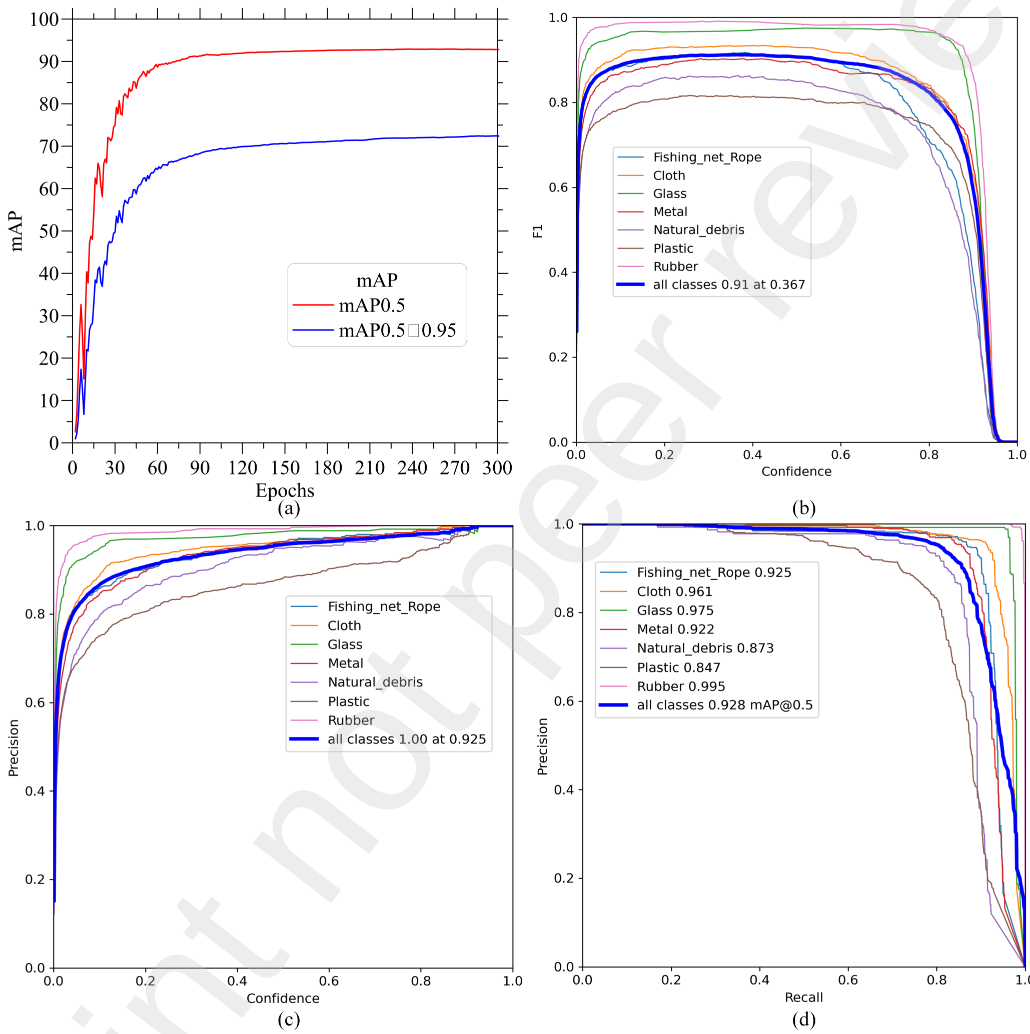

Figure 8: Evaluation indexes of the detection results. (a) mAP curve, (b) F1 score curve, (c) Precision curve, and (d) PR curve.

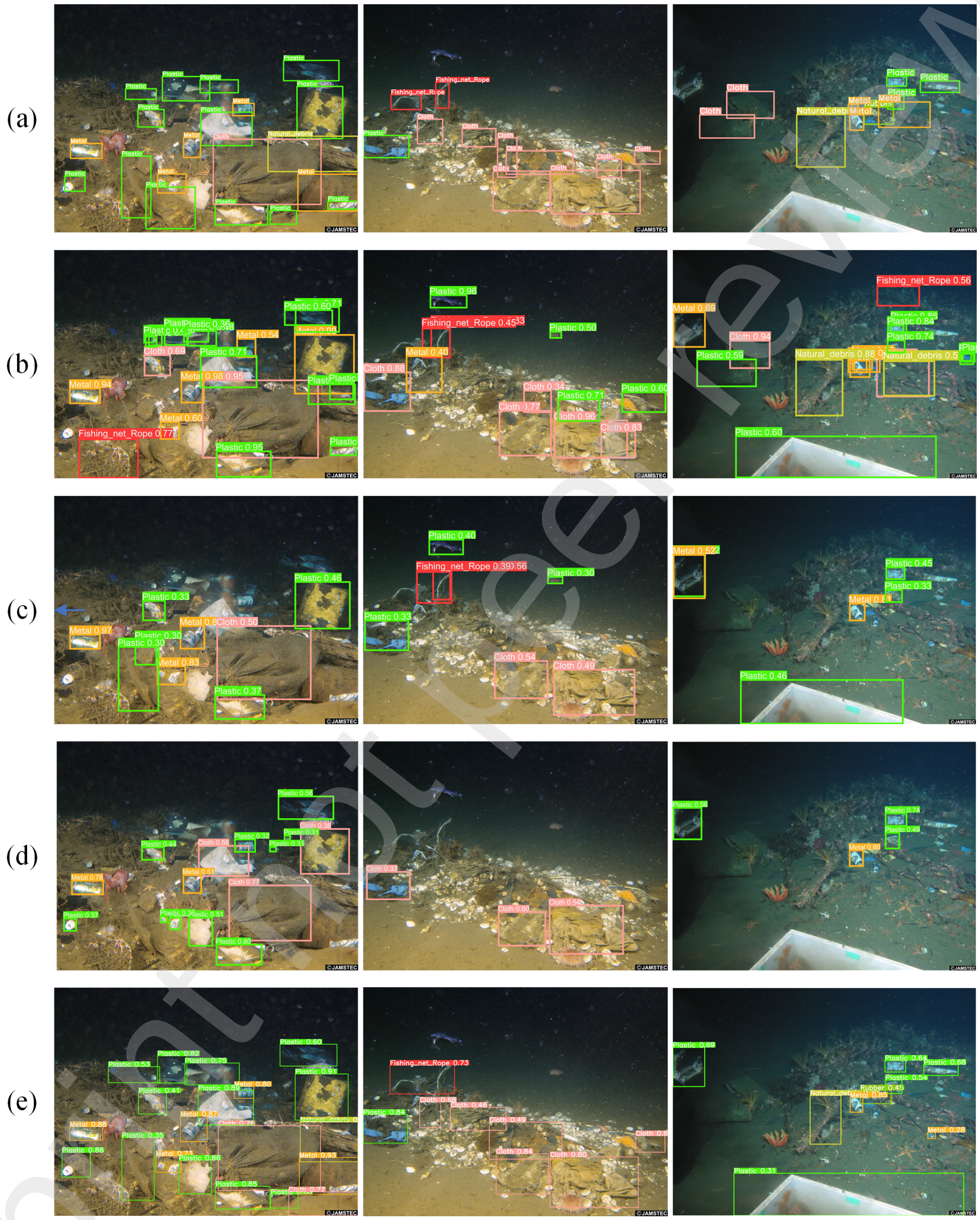

Figure 9: Visualized detection results of different methods. (a) Ground truth, (b) ResNet50-Faster R-CNN, (c) ResNet50-SSD, (d) Yolov3, and (e) DSDebrisNet.

network, cannot compare with the one-stage detection network DSDebrisNet in both accuracy and speed. Specifically, the $mAP_{0.5}$ is 92.8, $mAP_{0.5:0.95}$ is 92.8 72.4, and the number of frames per second (FPS) is up to 60. Consequently, the performance of the proposed one-stage network DSDebrisNet can reach the expectation of video detection.

Table 2: Comparison of evaluation indicators and speed for different detection methods.

Methods \ Indicators	$mAP_{0.5}$	$mAP_{0.5:0.95}$	FPS
ResNet50-Faster R-CNN	71.9	42.3	24
ResNet50-SSD	78.7	47.7	34
YOLOV3	83.4	48.4	52
DSDebrisNet	92.8	72.4	60

For intuitive comparison, the detection results obtained by the DSDebris along with comparative methods are illustrated in Fig.9. The first column is the ground truth, and the second, third, fourth, and last columns are the detection results of ResNet50-Faster R-CNN, ResNet50-SSD, YOLOV3, and DSDebrisNet, respectively. From the comparison, it can be seen that the proposed method can achieve promising detection results with the providential bounding box. Other methods generate more bounding boxes and have a higher recall, while the detection accuracy is lower.

Furthermore, the quantitative evaluations of detection results are summarized in Table.3 for more efficient comparisons, tabulating the subclass $AP_{0.5}$ in detail. It can be inferred that the proposed DSDebrisNet can achieve the highest value. In contrast, the $AP_{0.5}$ value of Rubber is 99.5, indicating that Rubber objects are the easiest to detect because of their single shape. Reciprocally, plastic debris had the lowest $AP_{0.5}$ values for all methods, indicating that despite the large size of the data, detection is more difficult due to deep sea pressures that squeeze plastic into different patterns.

Table 3: Quantitative performance comparison among the different methods in terms of $AP_{0.5}$ for each class.

Methods / Category	ResNet50-Faster R-CNN	ResNet50-SSD	YOLOV3	DSDebrisNet
Cloth	45.5	52.3	61.7	96.1
Fishing net Rope	76.3	76.9	86.0	92.5
Glass	87.3	92.8	91.6	97.5
Metal	68.9	83.4	85.2	92.2
Natural	70.7	74.5	82.5	87.3
Plastic	62.0	74.4	79.4	84.7
Rubber	92.5	96.6	97.6	99.5

As aforementioned, three different scaled feature maps of 17 layer, 20 layer and 23 layer are feed to the output layer. Thus facilitates compound-scaled deep sea debris detection. Here, the detection heatmap of the first and second examples in Fig.9 are drawn in Fig.10 to illustrate the detailed information. It can be concluded that larger feature maps detect smaller objects because they have smaller receptive fields, and smaller feature maps detect larger objects.

4.4. Instant detection of video

As previously identified, the performance of the proposed DSDebrisNet can meet the requirement of real-time video detection. As shown in Fig.11, two different deep sea debris videos with single objects and two videos with multiple objects are selected from JAMSTEC dataset, which can display the

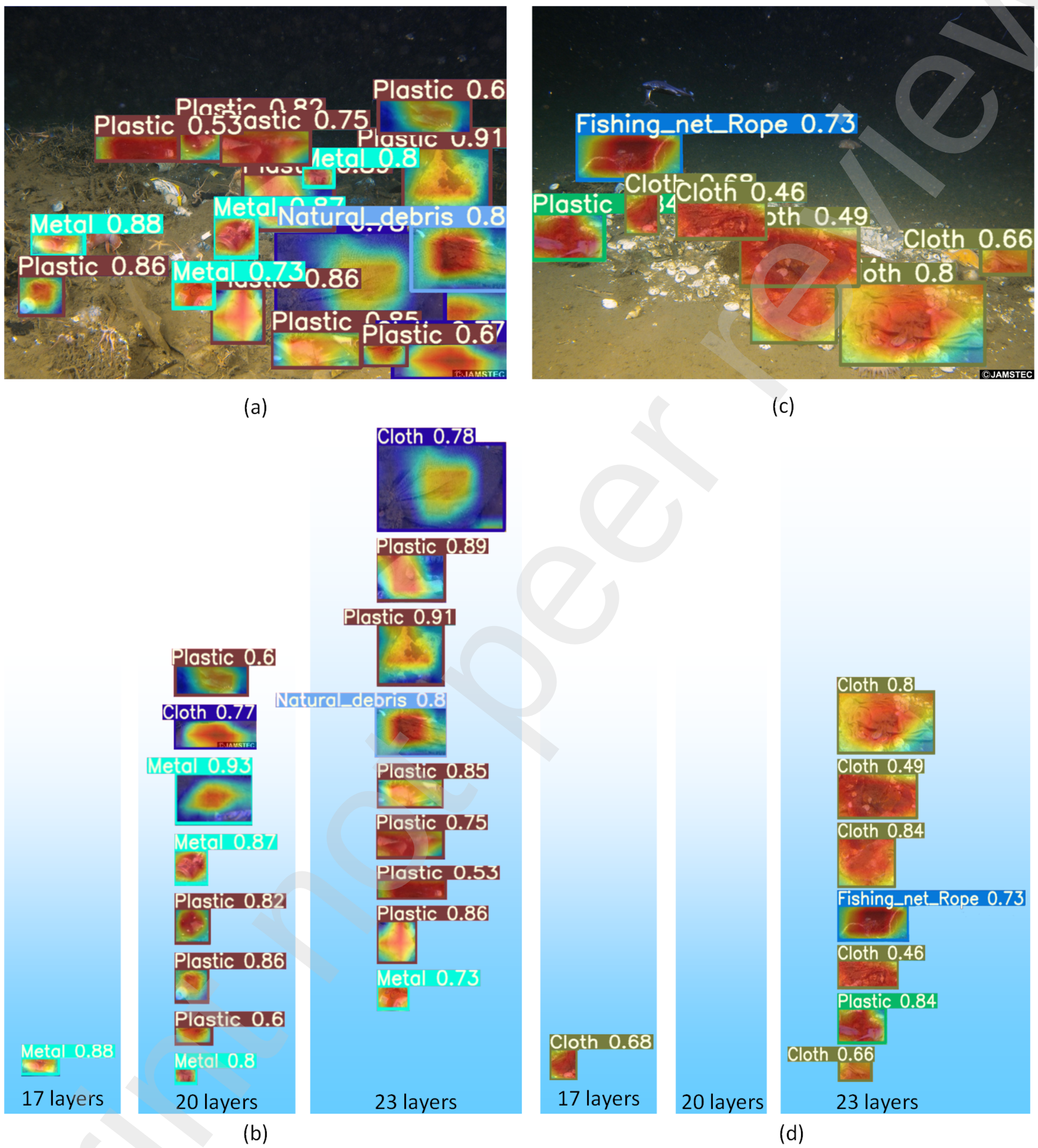

Figure 10: Detection heatmap of the first and second examples in Fig.9 with DSDebrisNet. (a) and (c) are the heatmap of the first and second examples, (b) and (d) are the statistical information of different scales.

real scene of deep sea debris with maneuverable underwater machines. By detecting the debris in the videos, the stability and generalization performance of the proposed DSDebrisNet method are verified. Although the temporal channel is not utilized to extract features for the sake of being lightweight, the proposed DSDebrisNet can accurately and efficiently detect debris without latencies in the video. Consequently, the DSDebrisNet is competent to detect deep sea debris from videos of maneuverable underwater machines, building sustainable oceans Lincoln et al. [2022] feasibly and practically.

5. Conclusion and future work

This paper proposes a deep sea debris detection methodology that directly detects debris via a lightweight DSDebrisNet. Firstly, the DSDebris dataset was built for the training and testing of the proposed DSDebrisNet based on the JAMSTEC dataset by extracting video frames and combining them with the original images. In this process, the imbalance of the dataset was analyzed and overcome, and the bounding boxes are annotated. Then, the DSDebrisNet was constructed following the encoder-decoder architecture. In addition, a hybrid loss function considering the illumination and detection problem was also introduced to improve performance. The benefit of the DSDebrisNet is that it requires relatively few epochs to achieve satisfactory detection results with high speed. Its superior performance has been verified in experimental results. Furthermore, the detection experiments on the videos were also conducted to prove that the real-time capability of DSDebrisNet can support the building sustainable ocean with the maneuverable underwater machine.

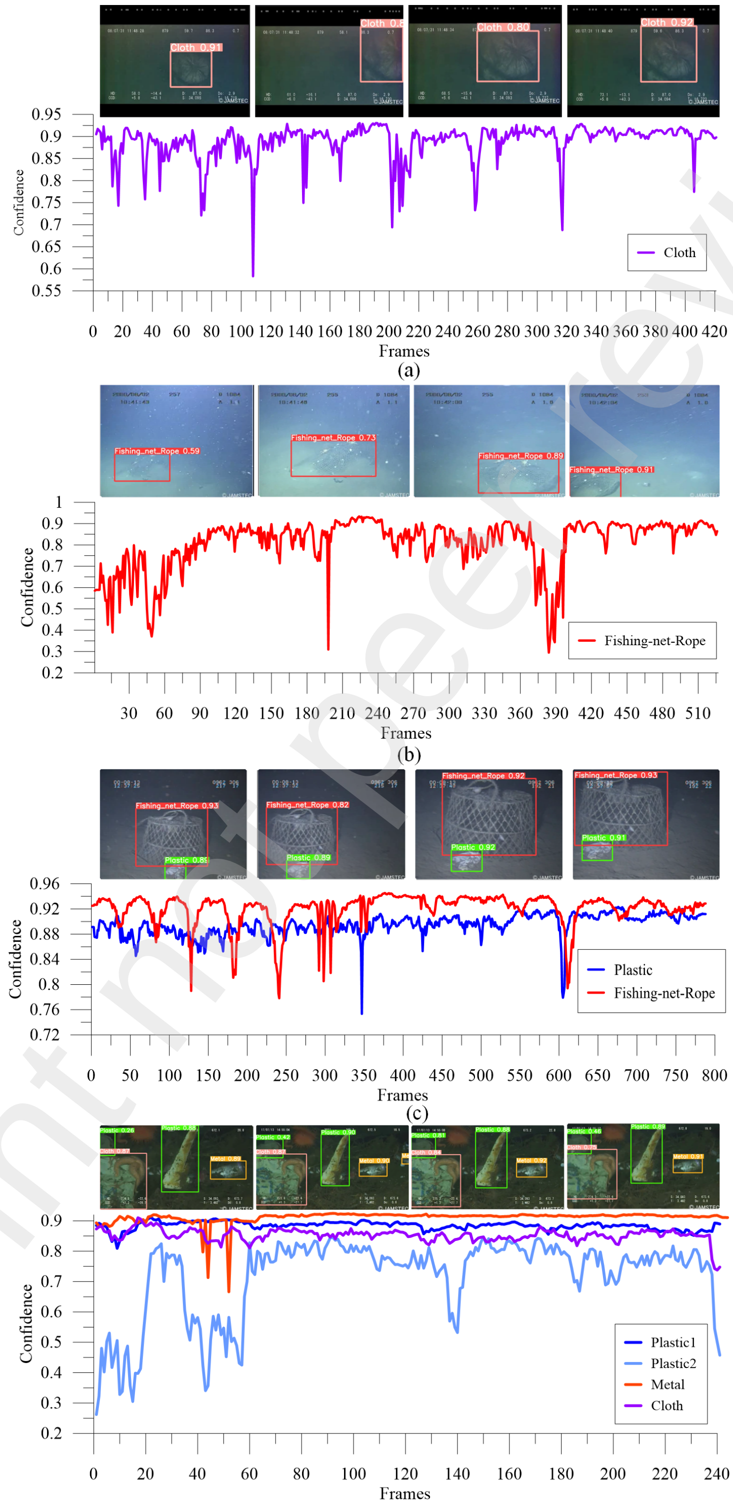

Figure 11: Instant deep sea debris detection of video through trained DSDebrisNet.

Although the effectiveness of DSDebrisNet method has been demonstrated, it does not always guarantee satisfactory detection results. The proposed method can detect debris in the case of incorrect classification because of the indistinguishable feature. However, it is still worthy of consideration since it achieves better results than other competitive methods. In future work, the authors will focus on the utilization of multiple frames on the temporal channel to improve detection performance. The authors believe that there is significant potential to exploit the proposed method to realize the detection task of maneuverable underwater machines' video.

6. Acknowledgments

This research was financially supported by the Marine S&T Fund of Shandong Province for Pilot National Laboratory for Marine Science and Technology (Qingdao) (No.2022QNLM050301), the National Natural Science Foundation of China (No.42276203), and the Natural Science Foundation of Shandong Province (No.ZR2021MD001). In the spirit of reproducibility, the dataset and Python code of the proposed DSDebrisNet methodology can be accessed with the linkage DSDebrisNet.

References

References

Agostini, L., Moreira, J.C.F., Bendia, A.G., Kmit, M.C.P., Waters, L.G., Santana, M.F.M., Sumida, P.Y.G., Turra, A., Pellizari, V.H., 2021. Deep-sea plastisphere: Long-term colonization by plastic-associated bacterial and archaeal communities in the southwest atlantic ocean. Science of

The Total Environment 793, 148335. doi:https://doi.org/10.1016/j.scitotenv.2021.148335.

Amon, D.J., Kennedy, B.R.C., Cantwel, K., Suhre, K., Glickson, D., Shank, T.M., Rotjan, R.D., 2020. Deep-sea debris in the central and western pacific ocean. Frontiers in Marine Science 7. doi:10.3389/fmars.2020.00369.

Bedford, J.L., Hanson, I.M., 2022. A recurrent neural network for rapid detection of delivery errors during real-time portal dosimetry. Physics and Imaging in Radiation Oncology 22, 36–43. doi:https://doi.org/10.1016/j.phro.2022.03.004.

Botero, C.M., Zielinski, S., Pereira, I, C., Leon, J.A., Duenas, L.F., Puentes, V., 2020. The first report of deep-sea litter in the south-western caribbean sea. Marine Pollution Bulletin 157. doi:10.1016/j.marpolbul.2020.111327.

Cau, A., Bellodi, A., Moccia, D., Mulas, A., Porcu, C., Pusceddu, A., Follesa, M.C., 2019. Shelf-life and labels: A cheap dating tool for seafloor macro litter? insights from medits surveys in sardinian sea. Marine Pollution Bulletin 141, 430–433. doi:https://doi.org/10.1016/j.marpolbul.2019.03.004.

Chen, G., Huang, B., Chen, X., Ge, L., Radenkovic, M., Ma, Y., 2022. Deep blue ai: A new bridge from data to knowledge for the perpetual ocean. Deep Sea Research Part I: Oceanographic Research Papers , 103886doi:https://doi.org/10.1016/j.dsr.2022.103886.

Cui, J., Yu, Z., Mi, M., He, L., Sha, Z., Yao, P., Fang, J., Sun, W., 2020. Occurrence of halogenated organic pollutants in hadal trenches of the western pacific ocean. Environmental Science & Technology 54, 15821–15828. doi:10.1021/acs.est.0c04995.

Cui, Z., Li, Q., Cao, Z., Liu, N., 2019. Dense attention pyramid networks for multi-scale ship detection in sar images. IEEE Transactions on Geoscience and Remote Sensing 57, 8983–8997. doi:10.1109/TGRS.2019.2923988.

Elfwing, S., Uchibe, E., Doya, K., 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Networks 107, 3–11. doi:https://doi.org/10.1016/j.neunet.2017.12.012. special issue on deep reinforcement learning.

Fulton, M., Hong, J., Islam, M.J., Sattar, J., 2019. Robotic detection of marine litter using deep visual detection models, in: 2019 INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION (ICRA), pp. 5752–5758. IEEE International Conference on Robotics and Automation (ICRA), Montreal, CANADA, MAY 20-24, 2019.

Galgani, F., Souplet, A., Cadiou, Y., 1996. Accumulation of debris on the deep sea floor off the french mediterranean coast. Marine Ecology Progress Series 142, 225–234. doi:10.3354/meps142225.

Geyer, R., 2020. Chapter 2 - production, use, and fate of synthetic polymers, in: Letcher, T.M. (Ed.), Plastic Waste and Recycling. Academic Press, pp. 13–32. doi:https://doi.org/10.1016/B978-0-12-817880-5.00002-5.

Han, G., Cen, H., Jiang, J., Gao, X., Jiang, X., Zhou, S., Xie, W., Ji, J., Bethel, B.J., Dong, C., 2022. Applying machine learning in devising a parsimonious ocean mixing parameterization scheme. Deep-Sea Research Part II-Topical Studies in Oceanography 203. doi:10.1016/j.dsr2.2022.105163.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. doi:10.1109/CVPR.2016.90.

Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv: Computer Vision and Pattern Recognition .

Huang, B., Ge, L., Chen, G., Radenkovic, M., Wang, X., Duan, J., Pan, Z., 2021. Nonlocal graph theory based transductive learning for hyperspectral image classification. Pattern Recognition 116, 107967. doi:10.1016/j.patcog.2021.107967.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269. doi:10.1109/CVPR.2017.243.

Jones, E.S., Ross, S.W., Robertson, C.M., Young, C.M., 2022. Distributions of microplastics and larger anthropogenic debris in norfolk canyon, baltimore canyon, and the adjacent continental slope (western north atlantic margin, usa). Marine Pollution Bulletin 174. doi:10.1016/j.marpolbul.2021.113047.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning 521, 436–444. doi:10.1038/nature14539.

Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9.

Lincoln, S., Andrews, B., Birchenough, S.N., Chowdhury, P., Engelhard, G.H., Harrod, O., Pinnegar, J.K., Townhill, B.L., 2022. Marine litter and climate change: Inextricably connected threats to the world's oceans. Science of The Total Environment 837, 155709. doi:https://doi.org/10.1016/j.scitotenv.2022.155709.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. Ssd: Single shot multibox detector, in: European Conference on Computer Vision (ECCV), Springer International Publishing. pp. 21–37.

for Marine Earth Science, J.A., Technology, 2018. Deep-sea debris database URL: https://www.godac.jamstec.go.jp/dsdebris/e/index.html.

Ni, Z.L., Bian, G.B., Wang, G.A., Zhou, X.H., Hou, Z.G., Chen, H.B., Xie, X.L., 2020. Pyramid attention aggregation network for semantic segmentation of surgical instruments. Proceedings of the AAAI Conference on Artificial Intelligence 34, 11782–11790. doi:10.1609/aaai.v34i07.6850.

Nurlatifah, Yamauchi, T., Nakajima, R., Tsuchiya, M., Yabuki, A., Kitahashi, T., Nagano, Y., Isobe, N., Nakata, H., 2021. Plastic additives in deep-sea debris collected from the western north pacific and estimation for their environmental loads. Science of The Total Environment 768. doi:10.1016/j.scitotenv.2020.144537.

Peng, X., Chen, M., Chen, S., Dasgupta, S., Xu, H., Ta, K., Du, M., Li, J., Guo, Z., Bai, S., 2018. Microplastics contaminate the deepest part of the world's ocean. Geochemical Perspectives Letters 9, 1–5. doi:10.7185/geochemlet.1829.

Politikos, D.V., Fakiris, E., Davvetas, A., Klampanos, I.A., Papatheodorou, G., 2021. Automatic detection of seafloor marine litter using towed camera images and deep learning. Marine Pollution Bulletin 164, 111974. doi:https://doi.org/10.1016/j.marpolbul.2021.111974.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788. doi:10.1109/CVPR.2016.91.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 1137–1149. doi:10.1109/TPAMI.2016.2577031.

Ren, Y., Li, X., Zhang, W., 2022. A data-driven deep learning model for weekly sea ice concentration prediction of the pan-arctic during the melting season. IEEE Transactions on Geoscience and Remote Sensing 60, 1–19. doi:10.1109/TGRS.2022.3177600.

Schlining, K., von Thun, S., Kuhnz, L., Schlining, B., Lundsten, L., Stout, N.J., Chaney, L., Connor, J., 2013. Debris in the deep: Using a 22-year video annotation database to survey marine litter in monterey canyon, central california, usa. Deep-Sea Research Part I-Oceanographic Research Papers 79, 96–105. doi:10.1016/j.dsr.2013.05.006.

Schmaltz, E., Melvin, E.C., Diana, Z., Gunady, E.F., Rittschof, D., Somarelli, J.A., Virdin, J., Dunphy-Daly, M.M., 2020. Plastic pollution solutions: emerging technologies to prevent and collect marine plastic pollution. Environment International 144, 106067. doi:https://doi.org/10.1016/j.envint.2020.106067.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations (ICLR), pp. 1–14.

Soliño, L., Vidal-Liñán, L., Pérez, P., García-Barcelona, S., Baldó, F., Gago, J., 2022. Microplastic occurrence in deep-sea fish species alepocephalus bairdii and coryphaenoides rupestris from the porcupine bank (north atlantic). Science of The Total Environment 834, 155150. doi:https://doi.org/10.1016/j.scitotenv.2022.155150.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning, in:

Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI Press. p. 4278–4284.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826. doi:10.1109/CVPR.2016.308.

Topouzelis, K., Papageorgiou, D., Suaria, G., Aliani, S., 2021. Floating marine litter detection algorithms and techniques using optical remote sensing data: A review. Marine Pollution Bulletin 170, 112675. doi:https://doi.org/10.1016/j.marpolbul.2021.112675.

UNEP, 2021. From pollution to solution: A global assessment of marine litter and plastic pollution. Nairobi: UNEP .

Wang, G., Tao, B., Kong, X., Peng, Z., 2022a. Infrared small target detection using nonoverlapping patch spatial–temporal tensor factorization with capped nuclear norm regularization. IEEE Transactions on Geoscience and Remote Sensing 60, 1–17. doi:10.1109/TGRS.2021.3126608.

Wang, K., Liu, M., 2022. Yolo-anti: Yolo-based counterattack model for unseen congested object detection. Pattern Recognition 131, 108814. doi:https://doi.org/10.1016/j.patcog.2022.108814.

Wang, Q., Cheng, M., Huang, S., Cai, Z., Zhang, J., Yuan, H., 2022b. A deep learning approach incorporating yolo v5 and attention mechanisms for field real-time detection of the invasive weed solanum rostratum dunal seedlings.

Computers and Electronics in Agriculture 199, 107194. doi:https://doi.org/10.1016/j.compag.2022.107194.

Woodall, L.C., Sanchez-Vidal, A., Canals, M., Paterson, G.L.J., Coppock, R., Sleight, V., Calafat, A., Rogers, A.D., Narayanaswamy, B.E., Thompson, R.C., 2014. The deep sea is a major sink for microplastic debris. Society Open Science Royal 1. doi:10.1098/rsos.140317.

Xue, B., Huang, B., Chen, G., Li, H., Wei, W., 2021a. Deep-sea debris identification using deep convolutional neural networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14, 8909–8921. doi:10.1109/JSTARS.2021.3107853.

Xue, B., Huang, B., Wei, W., Chen, G., Li, H., Zhao, N., Zhang, H., 2021b. An efficient deep-sea debris detection method using deep neural networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14, 12348–12360. doi:10.1109/JSTARS.2021.3130238.

Yang, G., Wang, B., Qiao, S., Qu, L., Han, N., Yuan, G., Li, H., Wu, T., Peng, Y., 2022. Distilled and filtered deep neural networks for real-time object detection in edge computing. Neurocomputing 505, 225–237. doi:https://doi.org/10.1016/j.neucom.2022.07.008.

Yu, J., Ye, X., Li, H., 2022. A high precision intrusion detection system for network security communication based on multi-scale convolutional neural network. Future Generation Computer Systems 129, 399–406. doi:https://doi.org/10.1016/j.future.2021.10.018.

Zhang, C., Liu, Q., Li, X., Wang, M., Liu, X., Yang, J., Xu, J., Jiang, Y., 2021. Spatial patterns and co-occurrence networks of microbial communities related to environmental heterogeneity in deep-sea surface sediments around yap trench, western pacific ocean. Science of The Total Environment 759, 143799. doi:https://doi.org/10.1016/j.scitotenv.2020.143799.

Zhang, X., Peng, X., 2022. How long for plastics to decompose in the deep sea? Geochemical Perspective Letters 22, 20–25. doi:10.7185/geochemlet.2222.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: