Benthic Organism Detection, Quantification and Seamount Biology Detection Based on Deep Learning

Yuhai Liu, Yu Xu, Haining Wang, and Xiaofeng Li

1 Overview

1.1 Backgrounds

Deep-sea organisms are those living below the ocean belt, and they can be divided into three categories according to their living styles, including plankton, swimming organisms and benthos. Deep-sea biological resources are an essential part of the marine ecosystem and play a vital role in the formation, maintenance, and development of marine ecosystem. The deep-sea biological resources are the foundation of marine ranch construction and aquatic development [34]. The problems such as the

Y. Liu
Dawning International Information Industry Co., Ltd., Qingdao 266101, China

Sugon Nanjing Institute, Co., Ltd., Nanjing 211100, China

Y. Xu
Laboratory of Marine Organism Taxonomy and Phylogeny, Shandong Province Key Laboratory of Experimental Marine Biology, Center for Ocean Mega-Science, Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, China

Y. Xu · H. Wang
University of Chinese Academy of Sciences, Beijing 100049, China

H. Wang
Key Laboratory of Marine Ecology and Environmental Sciences, Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, China

Deep Sea Research Center, Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, China

X. Li (✉)
CAS Key Laboratory of Ocean Circulation and Waves, Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, China
e-mail: lixf@qdio.ac.cn

X. Li and F. Wang (eds.), *Artificial Intelligence Oceanography*,
https://doi.org/10.1007/978-981-19-6375-9_16

establishment of deep-sea protection areas, the sustainable utilization of resources, and the maintenance of vulnerable marine ecosystems based on species diversity have become the hot spots in global deep-sea research. The research on the distribution and diversity of deep-sea organisms is helpful to promote human cognition of the ecosystem and plays a vital role in the maintenance of the marine ecosystem. Due to the year-round darkness of the deep-sea area where the sunlight is difficult to penetrate, the high salinity, the considerable pressure, the low water temperature, the number of biological species is relatively small. In contrast, the biological quantity is numerous in some intensive biological areas. Therefore, it is crucial to solving the practical problems by using modern technology.

Species discovery and identification are crucial ways to explore deep-sea biodiversity. To better protect marine ecology, we can monitor the health status and biodiversity of the benthos ecosystem by analyzing the species, quantity, and growth. Traditional methods of marine biological identification are based on morphology and molecular genetics and sometimes even need to use the advanced DNA sequencing technology supported by electron microscope. Although this method is accurate, there are still two main problems for marine species classification. On the one hand, it costs a large amount of human and financial resources to cultivate professional taxonomy experts for marine species, and artificial identification has low efficiency. On the other hand, the special ocean environment is unsuitable for in-situ detection during scientific research using molecular and electron microscopy methods, and heterotopic detection can lead to biological inactivation and species death. To solve the problems above, the application of the target detection technology based on the deep neural network in marine species identification and quantitative analysis emerged.

Considering the problems, including the difficulties in underwater target recognition caused by complex marine imaging environment, brutal penetration of sunlight, high salinity, the high similarity of some detected targets, and uneven distribution of biological density [26], the static counting of dense marine biological communities, and automatic real-time dynamic detection and counting algorithm of marine benthos were explored and studied in this paper. It is significant in helping marine biologists identify marine species, evaluate the population density, improve the operational performance of underwater autonomous robots and promote the underwater operation and the development and ecological protection of marine resources.

Seamounts are relatively isolated conical peaks or groups of peaks in the various oceans and are also an essential part of the marine environmental system. Seamounts rise from the seafloor but do not protrude from sea level. There are an estimated 30,000 seamounts worldwide, but only a few have been studied. However, seamounts have become one of the most popular systems in deep-sea research in recent years because of their unique topographic and hydrological features and their unique ecosystems, rich biodiversity, and excellent resource value.

Recently, our country has successfully carried out a series of seamount explorations represented by ‘Jiaolong’ manned HOV and ‘FaXian’ ROV, and obtained many first-hand submarine image data and samples in the South China Sea, Western and Central Pacific. It not only significantly improves the level of deep-sea detection

of our country but also provides data support for automatic detection of benthos in seamounts. Section 4 will use deep learning technology to identify and detect the giant benthos in seamounts.

1.2 Related Works

Due to the influence of the medium, the propagation distance of light and radio waves is very limited in seawater. In contrast, the propagation performance of sound waves in water is much better, covering a wider sea area. However, the acoustic signal will propagate along different paths because of the reflection, refraction, and other phenomena, so the underwater target recognition based on sonar echo technology has many interferences and low accuracy. The target recognition method of sonar image has the characteristics of high resolution and real-time performance. Therefore, in underwater target recognition and detection, the current research mainly focuses on the underwater target detection of sonar images for a long time.

Traditional sonar image detection algorithm mainly extracts features from sonar images and then classifies and locates the target. Extracting enough information features is the key to detect the underwater target. In this way, researchers proposed a series of hand-designed feature extraction methods, such as Scale Invariant Feature Transform (SIFT) [24, 25], Histogram of Oriented Gradient (HOG) [5]. The features are extracted effectively, and then recognized by algorithms like Morphology, Fuzzy Clustering and Markov Random Field [29]. The manual feature extraction, classification, and detection methods have a good recognition effect in specific application scenarios. Still, these algorithms have poor scalability and low generalization ability, which different features need to be designed for different problems. Therefore, the application value of the algorithm is limited.

With the development of underwater high-definition imaging technology such as ROVs (Remote Operated Vehicles) and AUVs (Autonomous Underwater Vehicles), the data of close-range targets collected by optical imaging equipment can be analyzed by computer vision algorithm without sonar images. It makes the feature information of the target more fully retained and used, and the accuracy and efficiency of target detection are greatly improved. In this trend, Fish4Knowledge [7] project has collected 115 TB of underwater high-definition image/video data and proposed many methods to detect fish in the underwater video for assessing fish biodiversity [39]. In fish detection, SIFT [1, 28, 37] or SC (Shape Context) [35] algorithms have been widely used to calculate marker features. But in the reference [47], the author concludes that HOG algorithm is better than SIFT and SC algorithm. Marcos et al. [27] used Normalized Chromaticity Coordinates (NCC) histogram to extract color features, and Local Binary Pattern (LBP) feature descriptor to extract texture features of the coral image. Stokes and Deane [40] proposed coral classification Discrete Cosine Transform and K-Nearest Neighbor classifier algorithm. Although the upgrades of underwater acquisition devices improve the quality of data, the analysis algorithm continues to use SIFT and HOG to extract features and then use Support

Vector Machine (SVM) [4] and Adaptive Boosting (AdaBoost) [42] to classify. The problem of poor robustness of manual feature extraction has not been effectively solved.

Due to the continuous development of deep learning in recent years, all aspects of computer vision are have been promoted. Especially, Convolutional Neural Network (CNN) algorithm, Fast R-CNN algorithm, and Feature Pyramid Network (FPN) algorithm, which are widely used in image classification, image annotation, and multi-target detection, enable people to obtain rich deep semantic information of images and improve the accuracy of image classification and recognition significantly. The target detection and recognition method based on deep learning benefits from CNN's strong feature autonomous learning ability on large-scale data sets and can effectively solve feature extraction in the above method. Therefore, it has been successfully applied to many underwater target detection and recognition scenes [18]. Kratzert and Mader [16] used the marine fish channel monitoring platform based on CNN algorithm to detect targets without using any artificial features, and the final fish classification accuracy reached 93%. Huang et al. [15] applied Faster R-CNN to detect and identify marine organisms, expanded a small number of samples through three data enhancement methods and verified the effectiveness of Faster R-CNN in biological detection in different marine turbulent environments. Xia et al. [43] proposed a sea cucumber detection scheme based on YOLOv2 model, which has a good detection effect on sea cucumbers with a regular shape or simple natural scene coverage. Although these methods have achieved some success, they are applied to specific target scenarios and do not include the study of target quantity statistics.

1.3 Research Content and Innovation

With the change of economic and marine environment, the value of marine biological resources is enormous. To further improve the ability of marine resources utilization and marine ecological protection, advanced information technology, and data analysis ability are needed to provide accurate data support and decision support for relevant personnel. After in-depth research on big data technology, artificial intelligence, and other technologies, combined with the existing business needs, the deep-sea biological identification quantitative model was designed and realized in this chapter.

The key to the success of the deep-sea biometric quantitative system is the extraction and application of data. The fast extraction of data and big data requires a stable and reliable algorithm basis. Data acquisition, extraction, conversion, cleaning, and data loading are used to enter the data storage layer. A deep-sea biometric quantitative database is formed by deep learning technologies such as data analysis. The computing layer can provide robust image classification and recognition, realize deepseated analysis of deep-sea biological data, fully excavate the hidden value of data, and provide support for quantitative recognition of deep-sea organisms. In this

chapter, Faster R-CNN and SSD are adopted respectively to achieve marine biometric identification and quantification for different scenarios.

Considering the current situation of deep-sea biological resources, and in order to realize the requirements of automatic identification and quantitative analysis of deep-sea organisms and detection of giant benthic organisms in seamounts, the following functions are studied and realized mainly in this chapter:

1. The deep-sea biological recognition and quantitative analysis system is constructed to process a large number of deep-sea biological image data, analyze the deep-sea biological data, extract biological features, classify and quantitatively analyze the deep-sea biological recognition by using deep learning and other artificial intelligence technology.
2. According to the high-definition seamount image data taken by the research ship during the investigation of a seamount in the Western Pacific Ocean, the seamount biological training library is constructed. On this basis, the SSD target detection model is trained. The feasibility of automatic real-time seamount species detection and counting was studied under the condition of the trade-off between speed and accuracy. 63 high-quality images of seamount macrobenthos in the Western Pacific are constructed and manually labeled. They can be used to train various deep learning models, which alleviates the lack of training data for marine species to a certain extent, and is helpful for other people in the same field.

2 The Target Detection Techniques

2.1 Introduction on Target Detection

In computer vision and image processing, Target Detection is an image segmentation technology that scans and searches for specific semantic targets (such as people, buildings or cars) in digital images or videos and marks them. Generally speaking, it is not only to identify which category the target belongs to, but also to get its specific position in the picture. Target Detection is widely used in computer vision tasks, such as automatic image annotation, behavior recognition, face recognition and video target segmentation. It can also be used for target trackings, such as the ball in a football match or the players on the court.

Traditional target detection is usually based on the traditional machine learning method, which is generally divided into two stages: firstly, SIFT, HOG, and other methods are used to extract features, and then, SVM, AdaBoost, and other algorithms are used for classification. However, there are two main problems in traditional target detection methods: (a) feature extraction is not targeted, and time complexity is high; (b) the features designed manually are not robust to the change of diversity. Therefore, when the detection task changes, the features need to be redesigned.

In recent years, with the help of Deep Neural Networks (DNN), the target detection algorithm based on DNN has gradually replaced the traditional target detection

algorithm. In computer vision tasks, DNN based target detection and recognition algorithms are mainly divided into two categories: one is a region proposal-based target detection algorithm, that is, a two-stage detection algorithm. In the first step, a series of sparse candidate regions are generated by a certain method, and in the second step, the candidate regions are further classified and regressed. Typical representatives of such algorithms are R-CNN [9], Fast R-CNN [8], Faster R-CNN [33], Mask R-CNN [14], etc. Due to the low recognition error rate and missing recognition rate, the two-stage target detection algorithm has achieved excellent performance on several challenging benchmarks including Pascal VOC [6] and MS COCO [21]; The other is the single-stage target detection algorithm, which skips the stage of generating candidate regions and directly generates the class probability and position coordinate value of targets. The final detection result can be obtained through a single detection. Therefore, compared with the two-stage algorithm, it has a faster detection speed. There are many typical algorithms, such as YOLOv1 [32], YOLOv2 [30], YOLOv3 [31], YOLOv4 [2], SSD(Single Shot Multibook Detector) [23], RetinaNet [22], RefineDet [46], CornerNet [17], etc. The advantage of a single-stage target detection algorithm is high detection efficiency, but its detection proficiency often lags behind a two-stage algorithm.

2.2 The Single-Stage Target Detection

The core idea of the single-stage target detection algorithm is to take the whole image as the input of the network, and apply regression on the position and category of Bbox in the output layer directly. The primary representative is SSD [23] and YOLO (You Only Look Once). In this paper, we use the SSD to complete the detection of seamount macrobenthos. As a result, the model has a simple structure and fast speed. The following focuses on the SSD algorithm and illustrates the principle of single-stage target detection.

SSD (single shot multibox detector) [23] is the first single-stage detector of a single shot. It abandons the practice of Faster R-CNN using RPN to generate boundary boxes and classify them and puts forward the ideas of multi-scale features and default boxes. Similar to other single-stage detectors, its speed is better than two-stage detectors. SSD algorithm is an algorithm with high speed, high accuracy, and high robustness to scale change. Its main feature is to use multi-layer convolution features with different scales and receptive fields for target detection and recognition.

SSD algorithm is based on a feedforward convolutional neural network. The algorithm first generates a series of fixed number of default boxes. It then uses the corresponding feature graphs of different levels to predict the location and category based on these default boxes. For all the predicted bounding boxes of each category, the redundant and low probability bounding boxes are removed by the non-maximum suppression algorithm. Finally, the detection results are generated. This method is a target detection algorithm based on regression, which imultaneously predicts the location and category within a network framework. Compared with R-CNN series

algorithms, SSD is a single-stage, end-to-end target detection algorithm, and the detection speed is greatly improved. Moreover, multi-layer convolution layers with different scales are used for target detection and recognition due to their unique design. As a result, the detection performance has been improved to a certain extent.

SSD network framework is divided into the base net and extra feature layers. The basic network is a truncated VGG network. The additional layer is the CNN layer with a gradually decreasing scale, and the detection of targets is carried out simultaneously on these characteristic maps with different scales. Feature maps of different scales are used to predict targets of different scales.

The input of the SSD is a 3 channel RGB image. Firstly, the algorithm will map a series of default bounding boxes (default boxes), according to the size of the feature map, and then convolute through a series of convolution cores. Each layer will produce a fixed number of predictions, including 4 position predictions and several category predictions. The default box mechanism is similar to the anchor boxes mechanism in the Region Proposal Network (RPN) in Faster R-CNN. For a p-channel feature map with $m \times n$ size, the convolution kernel with scale $3 \times 3 \times p$ is used to predict the category and location information at each location $m \times n$. Category prediction will predict a score value for each category, representing the category target's possibility in the corresponding box. The position prediction will predict the scale scaling and displacement change based on the corresponding default box, which is the position adjustment based on the default box according to the characteristics of CNN. The default box is a series of rectangular default boxes corresponding to each position $m \times n$ on the original map according to the scale of different levels of the feature map. These default boxes have different sizes and aspect ratios to adapt to the scale transformation of the target to be detected.

For the K default boxes of each position, the SSD algorithm uses convolution operation to predict $c + 1$ category scores (including C target category and a background category) and 4 coordinate positions. That is $(c + 1 + 4) \times K$, each position needs a convolution kernel, so for a scale of $m \times n$. The characteristic graph of n needs a convolution kernel corresponding to $(c + 1 + 4) \times Kmn$ prediction output. Each location corresponds to a fixed number of default boxes, which have different sizes and aspect ratios according to the location and scale of the layer.

During training, you need to match the truth value with the default box to produce positive and negative samples. SSD matches the positive and negative samples by calculating the Jaccard overlap of the default box and the truth box. The threshold is 0.5. If the Jaccard overlap in the truth box is greater than 0.5, it is set as a positive sample, otherwise it is a negative sample. A truth box can match multiple default boxes.

SSD has the following main features:

1. Inherit the idea of transforming detection into expression from Yolo to complete target positioning and classification at one time.

2. Based on anchor in Fast RCNN, a similar prior box is proposed.
3. Add the detection method based on the feature pyramid, that is, predict the target on the feature map of different receptive fields.

2.3 The Two-Stage Target Detection

Among the two-stage target detection algorithms, the R-CNN series is the most famous. This chapter mainly focuses on Faster R-CNN, and its predecessor is Fast R-CNN and R-CNN. We first briefly introduce R-CNN and Fast R-CNN target detection principles, and then focuses on the Faster R-CNN target detection algorithm.

Given the two problems existing in traditional target detection algorithms (see Sect. 2.2), Girshick proposed the R-CNN algorithm in 2014 [9]. Its principle is elementary, mainly by extracting multiple candidate regions to determine the target's position. The R-CNN target detection process is shown in Fig. 1.

Because the traditional algorithm for detecting each sliding window is a way of wasting resources, the R-CNN model uses SS (selective search) image segmentation algorithm [41] to extract 1k-2k candidate regions from the bottom to up. These regions are converted into fixed-size images and sent to CNN respectively to extract the features of each candidate area. Then, the SVM classifier is used to classify the feature vectors extracted by CNN. Then the regression of the coordinates of the upper left and right lower corner of the candidate region is made to modify the location of the candidate region to achieve the target classification and get the boundary. R-CNN uses SS algorithm to generate higher quality ROI and CNN instead of the sliding window used in traditional target detection as ROI and manual feature design. It makes the target detection field achieve a significant breakthrough and open the upsurge of deep learning target detection.

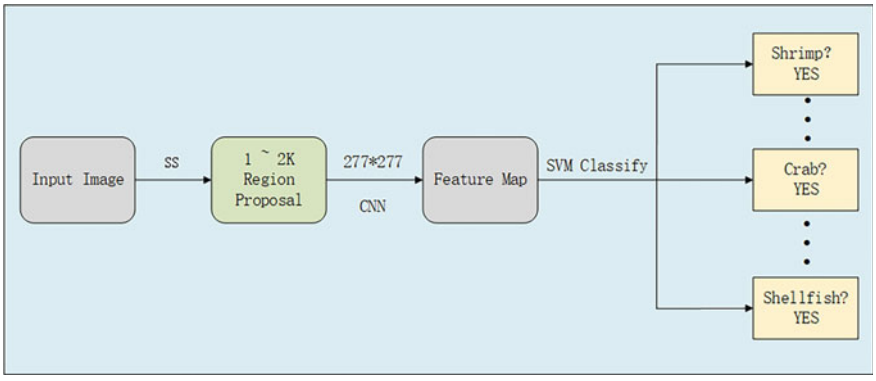

Fig. 1 The target detection process of R-CNN

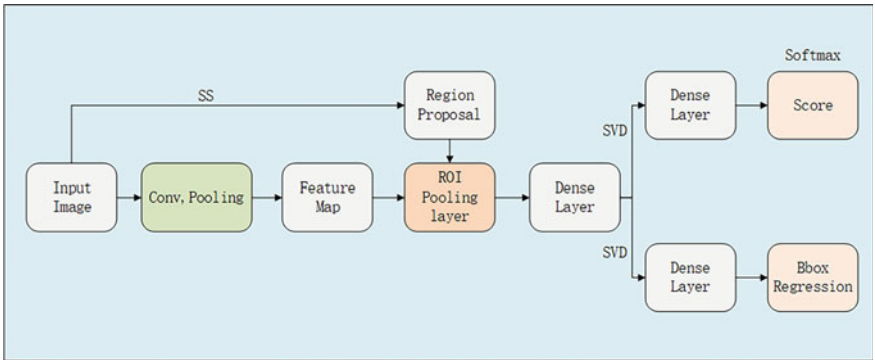

Fig. 2 Fast R-CNN model structure

But the classical R-CNN has the following problems:

1. Due to the need of calculating features for each candidate region, the amount of calculation is very tremendous.
2. The candidate regions are highly overlapped and there are too many repeated calculations.
3. Not end-to-end.
4. Strict size requirement for the input image.

In this case, SPPNet proposed by He et al. [12] successfully solves the problem of repeated convolution in R-CNN. However, the problems of multi-step training and large memory consumption still exist. Therefore, Girshick proposed the Fast R-CNN target detection algorithm in 2015 [8], and the target detection process is shown in Fig. 2.

Fast R-CNN can input any size of pictures into CNN and get the feature map by convolution and pooling operation, which avoid the time-consuming operation of generating candidate regions before convolution in R-CNN. Like R-CNN, Fast R-CNN also uses an SS algorithm to obtain about 2K candidate regions, and then find the corresponding feature boxes of each candidate region in the feature map. However, different from that, Fast R-CNN introduces ROI (Region of Interest) pooling operation. Its input is the feature map and the frame of candidate regions with different sizes obtained by CNN. The size of the output is fixed. The role of the ROI pooling layer is to pool the corresponding region into a fixed-size feature vector in the feature map according to the position coordinates of candidate regions, to carry out the following softmax classification and Bbox (Bounding box) regression.

Fast R-CNN abandons multiple SVM classifiers and Bbox regressors in RCNN and combines classification and regression in one network using a multi-task loss function. It also trains the whole network end-to-end and outputs the target's Bbox value and category label, which improves the model's accuracy. In addition, Fast R-CNN solves the problem of repeatedly extracting features by R-CNN, so the training speed has been significantly improved.

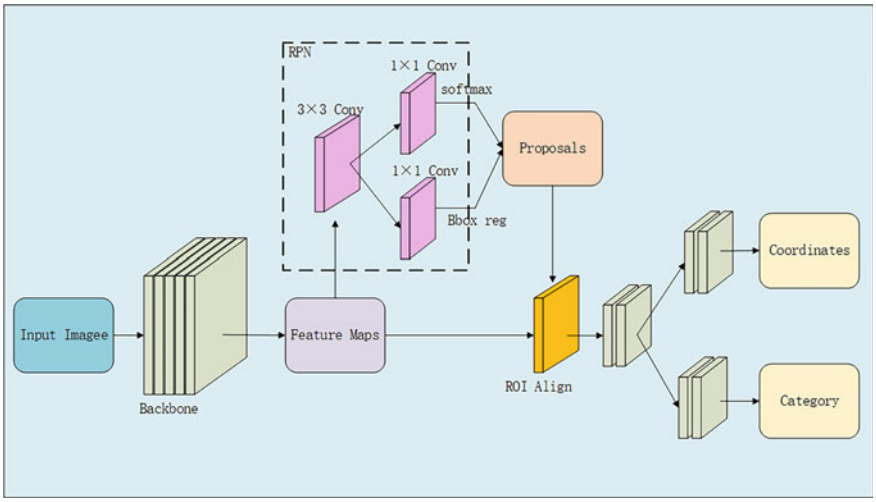

Fig. 3 Faster R-CNN model structure

In 2016, Ren et al. [33] proposed the Faster R-CNN target detection algorithm based on Fast R-CNN. Compared with Fast R-CNN, the most critical point of Faster R-CNN is using RPN (Region Proposal Network) instead of the SS segmentation algorithm to generate candidate frames, which significantly improves the speed of detection frame generation. In addition, Faster R-CNN integrates feature extraction, candidate region extraction, Bbox regression, and softmax classification into one network, which significantly improves speed and accuracy. Generally speaking, the improvement of Faster R-CNN to Fast R-CNN is that the speed of obtaining candidate regions is much faster. The Faster R-CNN target detection process is shown in Fig. 3.

The network structure of Faster R-CNN is similar to that of Fast R-CNN. Firstly, the backbone network is used to extract the features of the input image. The backbone network can use ResNet [13], VGG16, etc. Then, the RPN network is used to obtain the offset of the candidate box relative to the anchor box and the probability of containing targets. The specific operation is: the RPN takes the output feature map of the backbone network as the input and convolutes it using the kernel of 3×3, and then performs 2 times of 1×1 convolution. The number of output channel is $2 \times k$ and $4 \times k$ respectively. Among them, k represents the number of prior frames anchor on each grid point, and RPN uses this k anchor to make k predictions; The output $2 \times k$ is the target score, which represents whether the predicted candidate box on each grid point contains the target and the probability of containing the target; The output $4 \times k$ is coordinate information, which represents the offset of the predicted candidate frame on each grid point relative to the anchor frame; In Faster R-CNN,k is usually taken as 9. Finally, neural network and maximum pooling are used to calculate the pooled ROI feature map, and the result is reshaped into a vector $1 \times n$. Two fully connected layers are used for classification and regression to obtain the target location and classification information.

2.4 Summary

This section introduces the theory of target detection firstly, then focuses on two-stage R-CNN series target detection algorithm and single-stage YOLO series target detection algorithm, especially Faster R-CNN algorithm and SSD algorithm, and introduces the advantages and disadvantages of single-stage algorithm and two-stage algorithm. Combining the different characteristics of the two algorithms, this chapter provides the basis for the subsequent discussion on target detection counting models. Specifically, in Sect. 3, the Detection and Quantification of Benthic Organisms (DQBO) is introduced. Section 4 presents Detection of Macrobenthos in Seamounts (DMS).

3 DQBO Based on Faster R-CNN with FPN

3.1 Introduction on DQBO

Benthic density has always been an indispensable part of benthic target detection. By analyzing the images of marine benthic density, we can understand the social habits of organisms, help estimate the number of organisms and carry out a series of applications such as aquaculture and biotope protection. With the development of artificial intelligence technology and the depth of computer vision theory, intelligent image processing has become a critical research area. Although CNN-based target detection algorithms are widely used in many scenarios, the detection results do not meet all requirements and usually require more in-depth exploration. As shown in Fig. 4, the number of organisms is dense and numerous. Counting the number is cumbersome and has a high labor cost, so it is of great practical significance to automatically count the image targets.

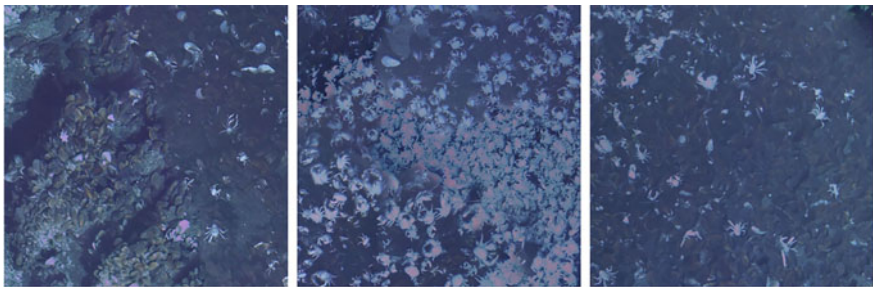

Fig. 4 Benthic organisms density images

3.2 The Faster R-CNN with FPN Framework for DQBO

How to deal with the large-scale change of objects is a fundamental problem in applying target detection. Whether the RPN in Faster R-CNN or Fast R-CNN, it is both based on a single high-dimensional feature, which generally has a poor effect on small object detection. FPN mainly solves the problem of detecting small and medium-sized objects in object detection scenes. Connecting high-dimensional features with low resolution and high semantic information and low dimensional features with high resolution dramatically improves the performance of small object detection.

This section embeds the FPN structure into the Faster R-CNN, combining it with the high-dimensional and low-dimensional feature extraction. Without increasing the amount of calculation of the original model, we successfully solve large-scale change and small object missing detection problems. The FPN network structure is shown in Fig. 5.

Figure 5 ① shows the forward propagation process of the neural network from bottom to top. After convolution operation, the size of the feature map becomes smaller and smaller, and more and more abstract. A pyramid level is defined for each stage of FPN. The output of the last layer of each stage is selected as the reference set of the feature graph because the deepest layer of each stage has more robust semantic information. ② is a top-down process, making the higher-level feature graph more abstract and more semantic to enhance the higher-level feature. Because the feature maps used in each layer are fused with features of different resolutions and semantic intensities, it can detect objects with corresponding resolutions and

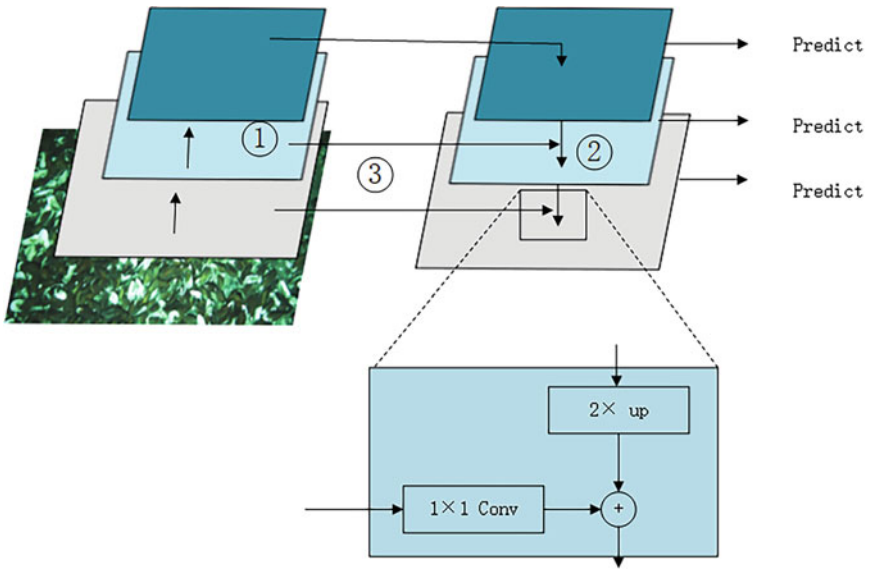

Fig. 5 The structure of FPN

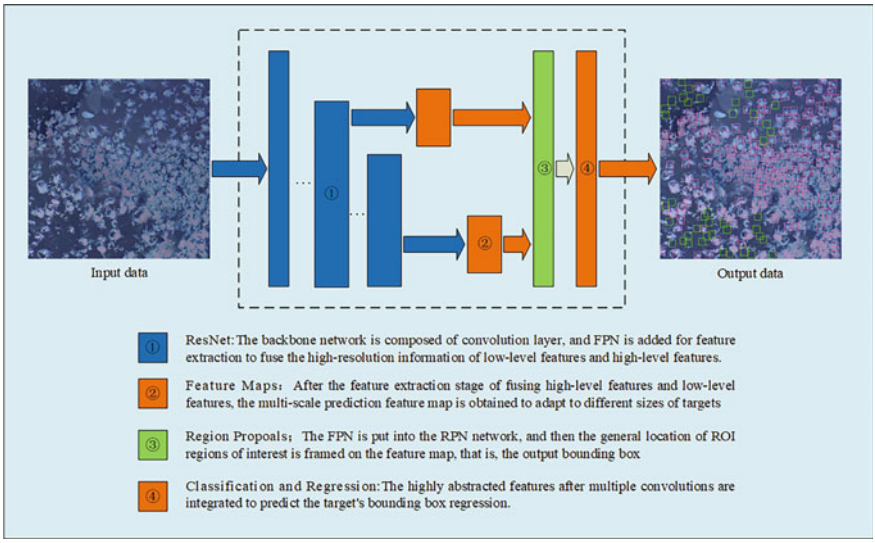

Fig. 6 The structure of Faster R-CNN with FPN

ensure that each layer has appropriate resolution and solid semantic features. ③ is a horizontal connection process, which uses a convolution kernel 1×1 to fuse the result of ② with the output feature graph of ① without changing the size of the feature graph.

In the detection process, the FPN structure is embedded in the Faster R-CNN feature extraction part. The framework of the target detection and counting model based on Faster R-CNN consists of the following three parts: feature extraction, candidate region generation and classification, and Bbox regression. The network structure of Faster R-CNN with FPN is shown in Fig. 6.

1. FPN: Feature extraction
   To improve the recognition accuracy of different sizes of organisms in the image, the backbone in Faster R-CNN is replaced by ResNet50 which combines FPN instead of VGG16. Feature maps of different scales are obtained by the FPN and then sent to the RPN to generate candidate regions. The fusion of deep and shallow features makes the FPN structure effectively improve the detection rate of small targets. The multi-resolution feature map detection design makes Faster R-CNN have a better detection effect for different scale targets.
2. RPN: Get candidate region
   RPN is a complete convolution network, which can be trained end-to-end, to generate the suggestion bounding box, which can predict not only the boundary of the object but also the probability score of the object. The network structure contains two types of outputs: Softmax classifier and Bounding box, a multi-task model. The core of RPN is the anchor. RPN is mainly used to generate candidate regions. However, the different sizes and aspect ratios of targets make it necessary

to set multiple-scale windows. RPN generates a large number of anchor boxes firstly. After clipping and filtering, Softmax further determines whether these anchors belong to the foreground or background, whether there are objects in the box. The foreground represents containing objects and the background represents not containing objects. At the same time, the other branch begins to modify the anchor frame to form more accurate candidate regions.

The implementation process of RPN is as follows: firstly, a small network is used to perform sliding scanning operation on the feature image obtained by convolution, and it is connected fully with the window on the feature image, then it is mapped into a low dimensional vector, and finally, the vector is fed into the Bbox regression layer (reg) and Bbox classification layer (cls). The reg layer is mainly used to estimate the candidate output (x, y, w, h) corresponding to the candidate anchor. The cls layer is used to judge whether the candidate region is foreground or background.

3. Target classification and Bbox regression

Before the target classification and bounging-box regression, we need to carry out the pooling operation. This layer uses the candidate regions generated by RPN and the feature maps of different scales generated by the backbone network to get the fixed-sized candidate feature maps and inputs them into the subsequent network. We can use the full connection operation to identify and locate the target. In the classification process, Softmax is used as the classification function to classify the fixed-size feature image formed by the ROI pooling layer according to the specific category. At the same time, the L1 loss is used to complete the candidate regression operation on the bounding box for position verification to obtain the accurate position of the object. The loss function equations of the whole network is shown in Eq. 1.

$$L\left(\{p_i\},\{t_i\}\right) = \frac{1}{N_{cls}}\sum_i L_{cls}\left(p_i, {p_i}^*\right) + \lambda\frac{1}{N_{reg}}\sum_i {p_i}^* L_{reg}\left(t_i, {t_i}^*\right) \tag{1}$$

P_i is the probability of the category of anchor calculated by the Softmax. When the IOU between the anchor and the target window is greater than 0.7, the value of ${p_i}^*$ is 1, and when the IOU is less than 0.3, the value is 0. ${t_i}^*$ is a scaling parameter, which is the real scaling value for regression, including coordinate scaling and size scaling. t_i is used to represent the scaling value predicted by the network in the training process. Faster R-CNN completes the regression task by learning the scaling value. The loss function consists of two parts: classification loss and regression loss. See Eq. 2 for the calculation of classified loss:

$$L_{cls}\left(t_i, {t_i}^*\right) = -\log\left({p_i}^* p_i + \left(1 - {p_i}^*\right)\left(i - p_i\right)\right) \tag{2}$$

See Eqs. 3 for the calculation of regression loss:

$$L_{reg}\left(t_i, {t_i}^*\right) = R\left(t_i - {t_i}^*\right) \tag{3}$$

Where is the loss value is calculated by smooth L1 function, see Eqs. 4:

$$Smooth_{L1} = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 others \end{cases} \tag{4}$$

The general implementation process of the whole network is as follows: First, the input image is represented as $Height \times Width \times Depth$. The tensor is processed by the backbone network with an FPN structure to obtain feature maps of different scales. Then, the RPN is used to extract candidate regions. After obtaining the possible related objects and their corresponding positions in the original image, the features extracted from the backbone network and the bounding box containing the related objects are pooled by ROI, and the features of the related objects are extracted to obtain new vectors. Then it is sent to the subsequent classification and regression network to complete the target recognition and positioning.

3.3 Experimental Results and Discussions

In this experiment, the iterations are 120 and the batch size is 1. We set the learning rate and weight decay to 0.0001. Set the size and scale of anchor to (8, 16, 32) and (0.5, 1, 2). The impulse gradient descent method is used to reduce overfitting, and the impulse is set to 0.9.

For the static image data, a total of 630 samples were obtained. Labeling image annotation tool is used to calibrate these samples manually, and then we divide the training, validation and test sets according to the ratio of 7:2:1. We used the test set to evaluate our model.

In this paper, Recall, Precision and AP were used to evaluate the results of this experiment. The results are shown in Table 1.

The definition of recall and precision is as follows: Eq. 5:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{5}$$

Among them, the definition of TP, TN, FP, FN is shown in Table 2, which respectively represents true positive, true negative, false positive, and false negative recognition, $Precision$ represents the correct proportion of all predicted targets, $Recall$

Table 1 The experimental result

Class	Recall	Precision	AP
Mussel	0.745	0.719	0.720
Shinkaia	0.876	0.755	0.756

Table 2 Obfuscation matrix for specified categories

Ground truth	Predictive	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

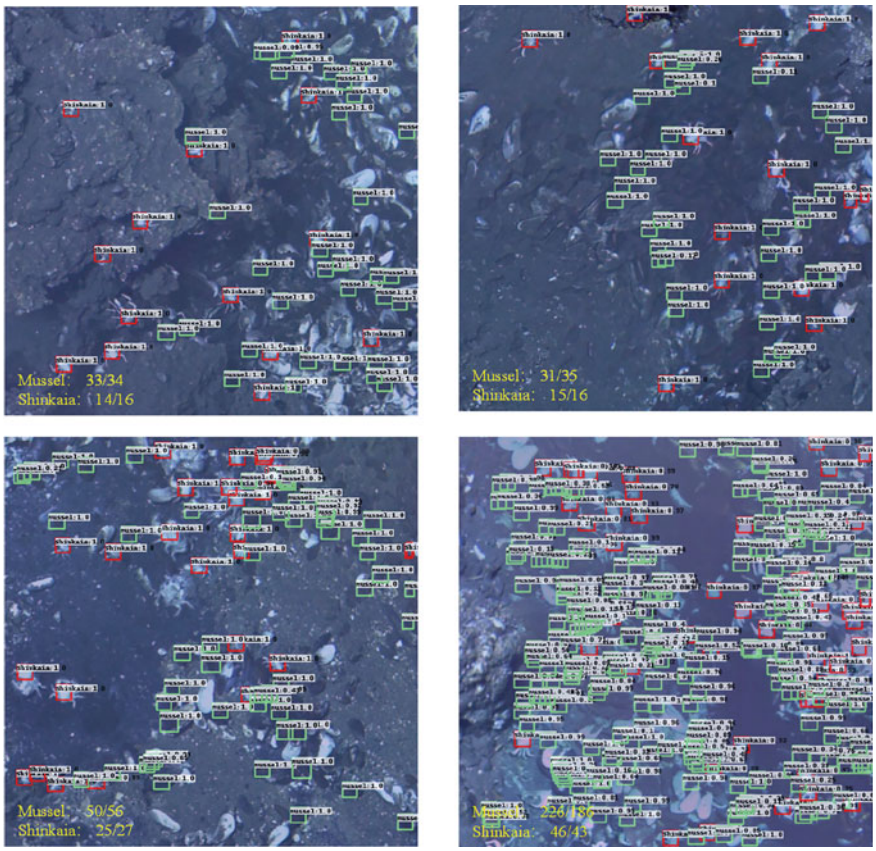

Fig. 7 Marine benthos detection and quantification results. The tag 33/34 indicates that the real quantity of mussel is 34 and the detection quantity is 33

represents the proportion of correctly located and recognized targets in the total number of targets.

Finally, the mean average precision (mAP) of 73.8% is obtained on the marine biological data set. Among them, the accuracy of FPN method for mussels recognition is 72.0%, and the accuracy of shinkaia is 75.6%. The visualization of experimental results is shown in Fig. 7. It can be seen that FPN is an excellent static image counting model for marine organisms.

3.4 Summary

This section first introduces the overall structure of the Faster R-CNN model and describes the network structure in detail. Next, the structure of the convolution neural network used for feature extraction is introduced, and 441 images are trained by this network. The experimental results show that the Faster R-CNN recognition has good effect, and can be applied to the quantitative analysis of marine biological recognition.

4 DMS Based on SSD

4.1 Introduction on DMS

In a narrow sense, seamount refers to the submarine uplift with a height of more than 1000 m below sea level. In a broad sense, the sea-knolls with a height of 500–1000 meters and hills below 500 m are called seamounts. Seamounts are the significant ecological landscapes in the deep ocean. It is estimated that the global seamounts account for 21% of the global seabed area [3, 45]. With the unique topography and hydrological characteristics, as well as the unique ecosystem, abundant biodiversity, and colossal resource value, seamounts have become one of the most concerning areas in deep-sea research. Compared with the surrounding deep-sea area, seamounts have high productivity, high biomass, and high biodiversity.

With the change of water depth and sediment types, the biological communities of seamounts show obvious biota replacement, and different sediment types often distribute different biota. For example, in soft bottom sediments, sea gills, starfish, sea urchins, and sea cucumbers are more common, while in hard rock bottom sediments, sponges, black corals, gorgonians, and sea anemones are dominant. The research on seamounts primarily focuses on the macrobenthos, whose individual is more than 2 cm and can be identified through the seabed image.

With their unique biological communities, rich biodiversity, and huge resource value, seamounts have become the focus of deep-sea biodiversity protection. At present, the protection of marine Biodiversity Beyond National Jurisdiction (BBNJ) has become an issue of global concern. Scientific understanding of the biological composition and distribution of seamounts is the key to the development, utilization, and protection of this fragile deep-sea ecosystem. It is the most concentrated among the seamounts globally and has the most significant number in the Western Pacific Ocean. The Western Pacific is the area with the most densely distributed seamounts and the most developed trench-arc-basin system globally. The cross-linking area of the Yapu Trench, Mariana Trench, and Caroline Ridge is the most representative. It is also one of the areas with minor research on seamounts in the world.

The most considerable difficulty in studying deep-sea biodiversity lies in the acquisition of deep-sea specimens and data. Due to the complex topography of seamounts, biological sampling is more complicated than the general deep-sea. Among the more

than 30000 seamounts globally, only 1% of them have been carried out in biological sampling, and only about 50 seamounts have been sampled comprehensively. As a result, seamounts are still one of the "least known biological habitats" for humans. The research on the biodiversity of seamounts is limited to the macrobenthos, and most of the research only focuses on the species composition, while a few focus on community structure and distribution. Due to the limitation of sample acquisition and insufficient sampling, a considerable part of the classification and identification of seamount organisms is based on the analysis of video and image of benthic organisms. Many novel organisms cannot be identified due to the lack of samples.

In recent years, our country has successfully carried out some seamount explorations represented by 'Jiaolong' HOV and 'FaXian' ROV and obtained many first-hand submarine image data in the South China Sea, Western and Central Pacific. It significantly improves the deep-sea detection level of our country and provides data support for automatic detection of benthos in seamounts.

4.2 Seamount Macrobenthos Dataset

Supported by the strategic leading science and technology project of the Chinese Academy of Sciences(A) "material and energy exchange and its impact on the tropical western Pacific Ocean system", the Institute of Oceanology, Chinese Academy of Sciences has established a research system for the detection of marine biodiversity in seamounts through the construction of technical platform and team. A comprehensive survey of the deep-sea environment, biodiversity, and ecosystem structure of three seamounts in the cross-linking area of Yapu and Mariana Trench and Caroline Ridge in the Western Pacific Ocean was carried out (as shown in Fig. 8). More than 1000 giant and large biological samples were collected through the sampling of seamount detection by using "FaXian" ROV, and more than 880 GB in situ imaging data of seabed organisms were obtained. In Fig. 8, the peak of Yapu seamount (Y3) is located in $8°51'N$, $137°47'E$; the peak of the Mariana seamount (M2) is located in $11°19'N$, $139°20'E$; the peak of the Caroline seamount (M4) is located in $10°29'N$, $140°8'E$.

Based on the in-situ image data of macrobenthos obtained from the above three seamounts' surveys, the 63 in-situ image data were labeled as Paskal VOC format data by LabelImg, an image annotation tool. This images data include *Pheronemoides fungosus* Gong & Li, 2017 [10], *Paragorgia rubra* Li, Zhan & Xu, 2017 [20], *Chrysogorgia ramificans* Xu et al., 2019 [44], *Paraphelliactis tangi* Li & Xu, 2016 [19], *Poloipogon distortus* Gong & Li, 2018 [11] and *Chrysogorgia binata* Xu et al., 2019 [44]. These six species are newly discovered in recent years. Then we check all data manually to ensure that all image resolutions are 1920×1080.

In computer vision, the typical annotation method is annotating the object on the image with a rectangular bounding box. In seamount data, the bounding box is labeled with $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$, and $(x_{\min}, y_{\min})$ and $(x_{\max}, y_{\max})$ are the two vertices on the diagonal of the rectangular label box. The whole macrobenthos data set of seamounts are collected, and each image corresponds to an XML annotation

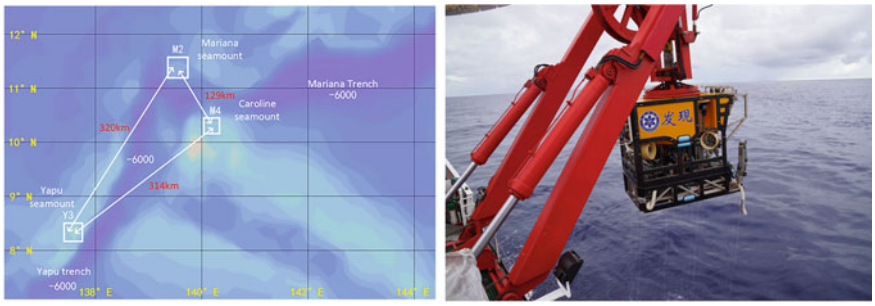

Fig. 8 Seamount data acquisition location and discovery ROV

file. Finally, stratified random sampling is used to divide the labeled data into the training set and test set according to the ratio of 8:2.

4.3 The SSD Framework for DMS

The experimental framework of macrobenthos detection in seamounts with the SSD model is shown in Fig. 9.

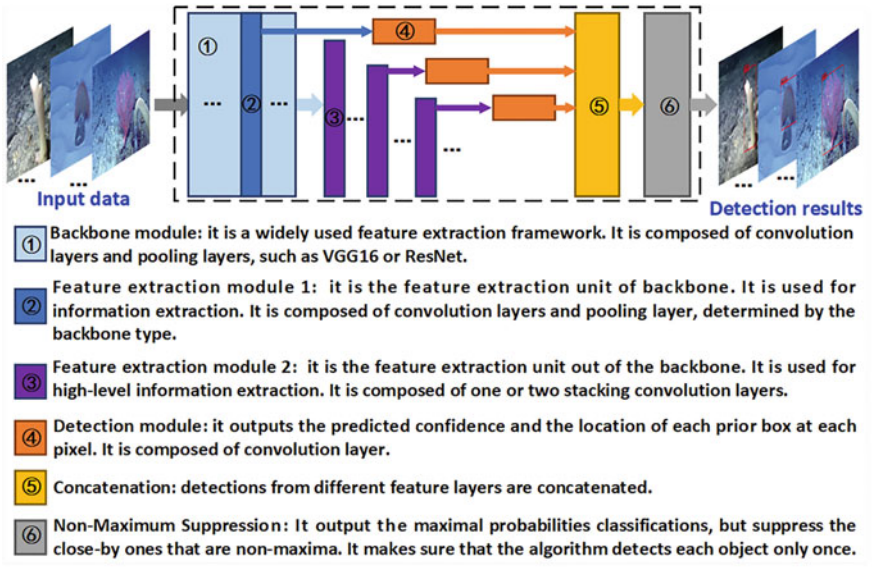

Fig. 9 The entire process of underwater species detection by SSD

SSD is based on VGG16 [38], which is pre-trained on the ILSVRC CLS-LOC dataset [36]. We convert FC6 (sixth fully connected layer) and FC7 to convolutional layers, subsample parameters from FC6 and FC7, after that remove all the dropout layers and the FC8 layer using SSD Weighted Loss Function [23]. We adjust the outcome model using SGD with initial learning rate 10^{-4}, 0.9 momentum, 0.0005 weight decay, and batch size 32. The entire process can be seen in Fig. 9.

4.4 Experimental Results and Discussions

Part of the output of our SSD model is shown in Fig. 10. Among all the six different marine species, our SSD model achieved 98.04% mAP (mean Average Precision) and the average value of IOU (Intersection-over-Union) over 0.8 on the test dataset with 63 images.

In our experiment, although we have verified that the implementation of SSD on our marine species data is feasible, SSD still often fails to detect small objects. Besides, our sample size and marine species categories are not enough. In the future, we will improve the SSD, and improve the ability to detect small objects in the camera further. Our ultimate vision is to build an AI system that can identify hundreds of thousands of marine species in real-time.

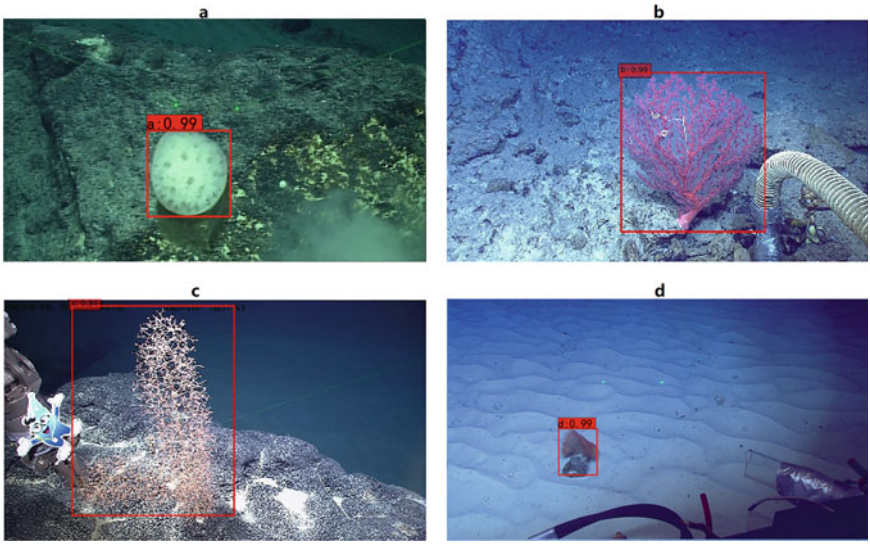

Fig. 10 **a** *Pheronemoides fungosus* Gong & Li, 2017, **b** *Paragorgia rubra* Li, Zhan & Xu, 2017, **c** *Chrysogorgia ramificans* Xu et al., 2019, **d** *Paraphelliactis tangi* Li & Xu, 2016

4.5 Summary

This section first introduces the overall structure of the SSD model applied and describes the network structure in detail. Next, the structure of the convolution neural network used for feature extraction is introduced, and 63 images are trained by this network. The experimental results show that the SSD recognition effect is good and it can be applied to the detection macrobenthos in seamounts.

5 Conclusions and Future Works

Target detection is one of the three major tasks in the field of computer vision. Computer vision algorithms are gradually applied to underwater scenes with the continuous development of deep learning technology and its wide application on land. In this chapter, the application of deep learning algorithm in target detection is extended to detecting marine organism. The detection and counting of marine organisms are studied in detail. First of all, in marine static image biological detection and counting, we explore the network architecture based on Faster R-CNN with FPN. Then, we verify the feasibility of SSD in the detection of giant benthic organisms in seamounts.

This chapter has completed the critical technologies of quantitative analysis system of artificial intelligence for marine benthos, realized the integrated development of marine big data, marine artificial intelligence and marine Internet of things, promoted the comprehensive development of marine artificial intelligence application, and filled the lack of artificial intelligence application in the deep-sea field partly.

Based on the above research, our subsequent work includes the following aspects:

1. Further expand the species richness. In addition, manual tagging is time-consuming and labor-consuming. The active learning method will be used for tagging in the subsequent expansion of the deep-sea biology training database.
2. The following research will focus on the dynamic video object detection and counting algorithm based on the static image.
3. Promote the AI algorithm model's landing and start developing the deep-sea macro-organism recognition and quantitative analysis system.

Acknowledgements The work was supported by the Key Program of National Natural Science Foundation of China (No. 41930533), the Senior User Project of RV KEXUE, man-aged by the Center for Ocean Mega-Science, Chinese Academy of Sciences (KEXUE2019GZ04), and the Key R & D Project of Shan-dong Province (2019JZZY010102).

References

1. Blanc K, Lingrand D, Precioso F (2014) Fish species recognition from video using SVM classifier. In: Proceedings of the 3rd ACM international workshop on multimedia analysis for ecological data, pp 1–6
2. Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: optimal speed and accuracy of object detection. arXiv:2004.10934
3. Clark MR, Rowden AA, Schlacher T, Williams A, Consalvey M, Stocks KI, Rogers AD, O'Hara TD, White M, Shank TM et al (2010) The ecology of seamounts: structure, function, and human impacts. Ann Rev Marine Sci 2:253–278
4. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
5. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol 1. IEEE, pp 886–893
6. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal Visual Object Classes (VOC) challenge. Int. J. Comput. Vis. 88(2):303–338
7. Fisher RB, Chen-Burger YH, Giordano D, Hardman L, Lin FP et al (2016) Fish4Knowledge: collecting and analyzing massive coral reef fish video data, vol 104. Springer
8. Girshick R (2015) Fast R-CNN. Int J Comput Vis 1440–1448
9. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
10. Gong L, Li X (2017) A new genus and species of Pheronematidae (Porifera: Hexactinellida: Amphidiscosida) from the western pacific ocean. Zootaxa 4337(1):132–140
11. Gong L, Li X (2018) A new species of Pheronematidae (Porifera: Hexactinellida: Amphidiscosida) from the Northwest Pacific Ocean. Acta Oceanologica Sinica 37(10):175–179
12. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916
13. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
14. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. Int J Comput Vis 2961–2969
15. Huang H, Zhou H, Yang X, Zhang L, Qi L, Zang AY (2019) Faster R-CNN for marine organisms detection and recognition using data augmentation. Neurocomputing 337:372–384
16. Kratzert F, Mader H (2017) Advances of FishNet towards a fully automatic monitoring system for fish migration. In: EGU general assembly conference abstracts, p 7932
17. Law H, Deng J (2018) Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European conference on computer vision (ECCV), pp 734–750
18. Li X, Liu B, Zheng G, Ren Y, Zhang S, Liu Y, Gao L, Liu Y, Zhang B, Wang F (2020) Deep-learning-based information mining from ocean remote-sensing imagery. Nat Sci Rev 7(10):1584–1605
19. Li Y, Xu K (2016) *Paraphelliactis tangi* n. sp. and *Phelliactis yapensis* n. sp. two new deep-sea species of Hormathiidae (Cnidaria: Anthozoa: Actiniaria) from a seamount in the tropical Western Pacific. Zootaxa 4072(3):358–372
20. Li Y, Zhan Z, Xu K (2017) Morphology and molecular phylogeny of *Paragorgia rubra* sp. nov. (Cnidaria: Octocorallia), a new bubblegum coral species from a seamount in the tropical Western Pacific. Chinese J Oceanol Limnol 35(4):803–814
21. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: European conference on computer vision. Springer, pp 740–755
22. Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
23. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) SSD: single shot multibox detector. In: European conference on computer vision. Springer, pp 21–37

24. Lowe DG (1999) Object recognition from local scale-invariant features. In: Proceedings of the seventh IEEE international conference on computer vision, vol 2. IEEE, pp 1150–1157
25. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
26. Lu H, Li Y, Uemura T, Kim H, Serikawa S (2018) Low illumination underwater light field images reconstruction using deep convolutional neural networks. Futur Gener Comput Syst 82:142–148
27. Marcos MSA, David L, Peñaflor E, Ticzon V, Soriano M (2008) Automated benthic counting of living and non-living components in Ngedarrak Reef, Palau via subsurface underwater video. Environ Monit Assess 145(1):177–184
28. Matai J, Kastner R, Cutter Jr G, Demer D (2010) Automated techniques for detection and recognition of fishes using computer vision algorithms. In: Williams, K., Rooper, C., Harms, J., (eds.), NOAA technical memorandum NMFS-F/SPO-121, report of the national marine fisheries service automated image processing workshop, Seattle, Washington, 4–7 Sept 2010
29. Mignotte M, Collet C, Pérez P, Bouthemy P (2000) Markov random field and fuzzy logic modeling in sonar imagery: application to the classification of underwater floor. Comput Vis Image Underst 79(1):4–24
30. Redmon J, Farhadi A (2017) YOLO9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271
31. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv:1804.02767
32. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
33. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. arXiv:1506.01497
34. Rohwer F, Youle M, Vosten D (2010) Coral reefs in the microbial seas, vol 1. Plaid Press Granada Hills
35. Rova A, Mori G, Dill LM (2007) One fish, two fish, butterfish, trumpeter: recognizing fish in underwater video. In: MVA
36. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
37. Shiau YH, Lin SI, Chen YH, Lo SW, Chen CC (2012) Fish observation, detection, recognition and verification in the real world. In: Proceedings of the international conference on image processing, computer vision, and pattern recognition (IPCV). The Steering Committee of The World Congress in Computer Science, Computer, p 1
38. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
39. Spampinato C, Giordano D, Di Salvo R, Chen-Burger YHJ, Fisher RB, Nadarajan G (2010) Automatic fish classification for underwater species behavior understanding. In: Proceedings of the first ACM international workshop on analysis and retrieval of tracked events and motion in imagery streams, pp 45–50
40. Stokes MD, Deane GB (2009) Automated processing of coral reef benthic images. Limnol Oceanogr: Methods 7(2):157–168
41. Uijlings JR, Van De Sande KE, Gevers T, Smeulders AW (2013) Selective search for object recognition. Int J Comput Vis 104(2):154–171
42. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition, vol 1. IEEE, pp I–I
43. Xia C, Fu L, Liu H, Chen L (2018) In situ sea cucumber detection based on deep learning approach. In: 2018 OCEANS-MTS/IEEE Kobe Techno-Oceans (OTO). IEEE, pp 1–4
44. Xu Y, Li Y, Zhan Z, Xu K (2019) Morphology and phylogenetic analysis of two new deep-sea species of *Chrysogorgia* (Cnidaria, Octocorallia, Chrysogorgiidae) from Kocebu Guyot (Magellan seamounts) in the Pacific Ocean. Zookeys 881:91

45. Yesson C, Clark MR, Taylor ML, Rogers AD (2011) The global distribution of seamounts based on 30 arc seconds bathymetry data. Deep Sea Res Part I: Oceanogr Res Pap 58(4):442–453
46. Zhang S, Wen L, Bian X, Lei Z, Li SZ (2018) Single-shot refinement neural network for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4203–4212
47. Zhu Q, Yeh MC, Cheng KT, Avidan S (2006) Fast human detection using a cascade of histograms of oriented gradients. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), vol 2. IEEE, pp 1491–1498