Underwater Image Object Detection based on Multi-scale Feature Fusion

Chao Yang
Southeast University
Ce Zhang
Southeast University
Longyu Jiang
JLY@seu.edu.cn

Southeast University
Xinwen Zhang
Southeast University

Research Article

Keywords: underwater image object detection, deep learning, feature fusion, data augmentation

Posted Date: March 20th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4114831/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Machine Vision and Applications on September 2nd, 2024. See the published version at https://doi.org/10.1007/s00138-024-01606-3.

Underwater Image Object Detection based on Multi-scale Feature Fusion

Chao Yang[1], Ce Zhang[1†], Longyu Jiang[1*], Xinwen Zhang[1]

[1*]School of Computer Science and Engineering, Southeast University, No.2 Sipailou, NanJing, 210096, Jiangsu, China.

*Corresponding author(s). E-mail(s): JLY@seu.edu.cn;
Contributing authors: chuckyang@seu.edu.cn; 827135598@qq.com; 220212110@seu.edu.cn;
†These authors contributed equally to this work.

Abstract

Underwater object detection and classification technology is one of the most important ways for humans to explore the oceans. However, existing methods are still insufficient in terms of accuracy and speed, and have poor detection performance for small objects such as fish. In this paper, we propose a multi-scale aggregation enhanced (MAE-FPN) object detection method based on the feature pyramid network, including the multi-scale convolutional calibration module (MCCM) and the feature calibration distribution module (FCDM). First, we design the MCCM module, which can adaptively extract feature information from objects at different scales. Then, we built the FCDM structure to make the multi-scale information fusion more appropriate and to alleviate the problem of missing features from small objects. Finally, we construct the Fish Segmentation and Detection (FSD) dataset by fusing multiple data augmentation methods, which enriches the data resources for underwater object detection and solves the problem of limited training resources for deep learning. We conduct experiments on FSD and public datasets, and the results show that the proposed MAE-FPN network significantly improves the detection performance of underwater objects, especially small objects.

Keywords: underwater image object detection, deep learning, feature fusion, data augmentation

1 Introduction

The classification and detection of underwater objects is a fundamental part of the exploration and exploitation of marine resources. Through underwater object detection technology, locating underwater objects in the ocean and obtaining object information can provide basic information for human exploration of the ocean. However, in practical applications, due to the diversity of underwater objects (e.g. the high variability and different sizes of individual fish), underwater object detection has higher detection complexity and difficulty compared to natural images.

Currently, underwater object detection methods are mainly traditional machine learning-based detection methods: such as SVM[2], decision tree[14], principal component analysis[24] and random forest[3]. Most of the machine learning-based underwater object detection methods require manual intervention for object feature extraction, followed by object classification and localisation. However, these detection methods usually have lower detection accuracy and slower speed, and often have problems such as missing and incorrect classification in the actual detection process.

With the rise of deep learning technology, more and more scholars try to solve the problem of underwater object detection with deep learning technology, typical CNN-based underwater object detection algorithms are [36],[25],[41],[39],[38]. These methods are mainly used to train convolutional neural networks to complete the classification by underwater image features, which is mainly divided into region-based object detection methods and regression-based detection methods. The region-based object detection method is mainly divided into two stages: first, extracting the candidate region, and then detecting according to the candidate region. The regression-based object detection method no longer generates the proposed region branch for underwater target feature extraction separately, but uses the anchor frame mechanism to directly return the candidate frame position to classify targets at multiple locations in the underwater image, thus greatly speeding up the target detection speed.

Fig.1 shows the scenarios that can occur in underwater object detection. Although the above deep learning based object detection algorithm can perform the underwater object detection task more accurately, there will be some miss detection and false detection in the underwater object detection task in practical applications due to the lower quality of underwater images compared to optical images and the greater variety of underwater objects. In addition, the lack of underwater image data also hinders the training of deep learning algorithms.

Therefore, we propose the feature fusion network of MAE-FPN and create the underwater fish object segmentation and detection dataset FSD, which effectively solves the problem of lack of underwater image data while achieving the high efficiency and high accuracy of underwater image object detection. The main contributions of this thesis are as follows:

First, we design the Multi-scale Convolutional Calibration module. This module adaptively selects different sensory field features based on the channel attention mechanism and assigns weights to the features at different scales separately, so that the feature information of multiple objects at different scales can be fully extracted.

Then, we innovatively propose the Feature Calibration Distribution Module. By fusing the high-level feature information more effectively and then distributing the output, we make the multi-scale information fusion more adequate, in order to reduce the information loss in the process of network feature fusion and alleviate the problem of small target feature loss.

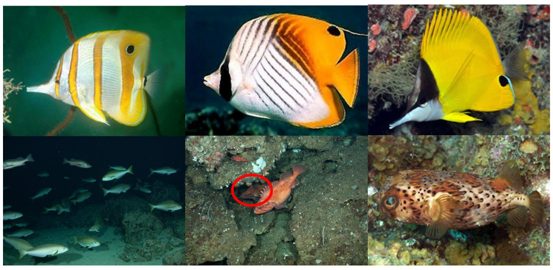


Fig. 1 The samples of underwater objects. The first row shows different fish species that have extremely high similarity, making classification difficult. The second row shows several difficult detection scenarios for underwater objects, such as dense alignment, overlapping, and blending with the background.

Next, we construct Fish Target Segmentation and Detection (FSD) dataset through data crawler, data expansion and enhancement, and manual screening, etc., which can alleviate the problem of lack of target detection dataset in underwater images to some extent. The dataset is well labeled, rich and challenging, and suitable for underwater target detection algorithm research.

Finally, we conduct exhaustive experiments and achieve significant performance improvement in both the FSD dataset and the open dataset, which verifies the effectiveness of the proposed method for underwater object detection.

2 Related Works

Deep Learning-based Object Detection are mainly divided into two categories: two-stage detectors and one-stage detectors. Two-stage detectors, such as the Faster RCNN[29] family, consist of two parts: one part is the proposal of candidate object bounding boxes, and the other part is the task of classifying and regressing the bounding boxes using features extracted from each candidate object. One-stage detectors, such as YOLO series[28], [37], SSD[20], and RetinaNet[18], propose the prediction boxes directly from the input image without the region proposal step. In general, two-stage detectors have higher localization accuracy and object recognition accuracy, while single-stage detectors have higher inference speed.

The above algorithms rely on a set of predefined anchor frames and perform object detection based on these previous anchors. We also refer to these algorithms as anchor-based object detection methods. Recently, major breakthroughs have been made in anchor-free object detection algorithms to overcome the computational challenges posed by the introduction of anchor points. Major anchor-free object detection algorithms include FCOS[34], CornerNet[15], and CenterNet[7]. In addition, some researchers have introduced the transformer mechanism from natural language processing to object detection, and proposed new object detection methods such as DETR[4] and Swin Transformer[22].

Overall, deep learning-based object detection algorithms have developed rapidly and achieved good results in both scientific research and industrial applications[42],[19],[38],[40],[31]. However, these common object detection algorithms are mainly designed for optical natural images rather than underwater images, and their direct application to underwater images often results in poor performance, so special designs for underwater object detection are still needed.

Object Detection In Underwater Images.

Traditional underwater object detection methods mainly rely on manual feature extraction. For example, Kim[13] proposed the underwater object detection algorithm with multi-channel Haar class features, Villion[35] proposed the fish behaviour feature extraction and analysis based on machine learning. These methods extract clearer features and are more widely used in aquaculture. However, such methods are unable to maintain high robustness and generalisation ability in different environmental situations due to the use of artificially preset features, and the higher human involvement consumes excessive human and financial costs.

With the maturity of deep learning technology, object detection algorithms for underwater images have shifted from traditional manual feature extraction methods to deep learning-based target detection methods. Ahsan et al[12] developed a hybrid solution for estimating species richness and population changes in underwater fish habitats by combining a YOLO network with optical flow and a Gaussian mixture model to detect dynamic fish in the background. Sung et al[33] extracted fish images from underwater videos and used the YOLO algorithm model to detect them and achieved a performance balance between detection accuracy and speed. Qi et al[27] proposed a two-stage network underwater small object detection method using a deformable convolutional pyramid structure to solve the object deformation problem. structure to solve the object deformation problem.

The above methods improve the performance of underwater object detection to a certain extent, but due to the diversity of underwater object types, the limited nature of the dataset, and the complexity of the underwater environment and occlusions, the underwater target detection task is subject to a number of omissions and misjudgments in practice. In addition, for some small targets, such as fish, the existing network has an overall poor detection performance.

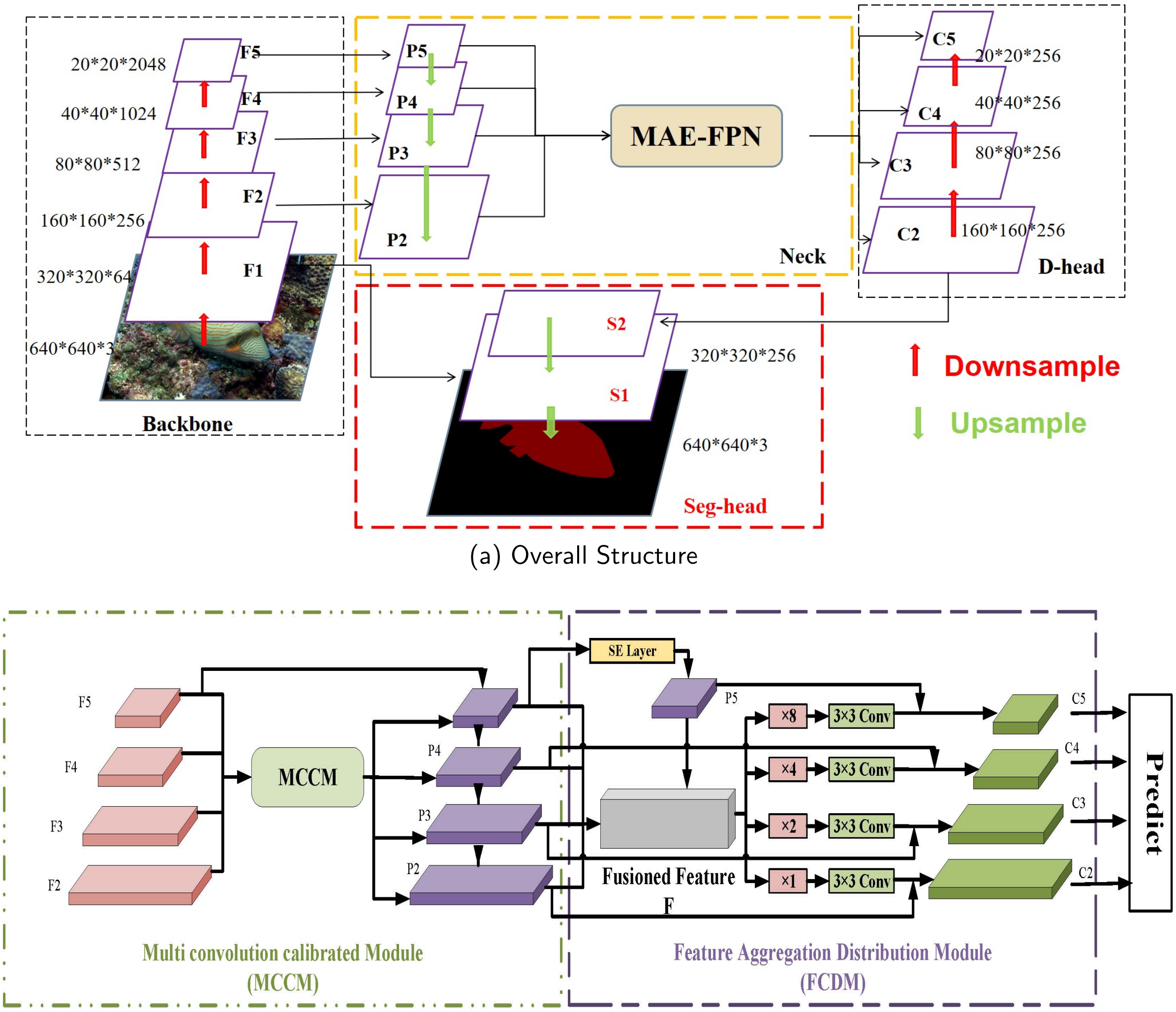
3 Methods

In this section, we provide a detailed description of the approach proposed in this paper. Combining the consideration of detection accuracy and speed, we adopt YOLO_v5 as the base detector. First, we take the YOLO_v5 object detector as an example to itemize the overall network structure. Then, we will elaborate the overall structure and individual modules of the proposed Multi-scale Aggregation Enhanced Feature Pyramid (MAE-FPN). Finally, we introduce the newly proposed Fish Segmentation and Detection (FSD) dataset, which provides a detailed description of the collection and production process.

3.1 Overall network structure

Based on YOLO_v5, we propose a multi-scale aggregated object detection network as shown in Fig.2a. It consists of four main parts: the input (Input), the backbone network (Backbone), the neck network (Neck) and the predictive head output (Head).

Assuming that an RGB image with a size of 640×640×3 is input, firstly, multiple feature maps with different scales are obtained by upsampling in the Backbone network noted as F1, F2, F3, F4, F5. Then MAE-FPN constructs lateral connections based on the Multi-scale Convolutional Calibration Module (MCCM) to obtain the



(a) Overall Structure

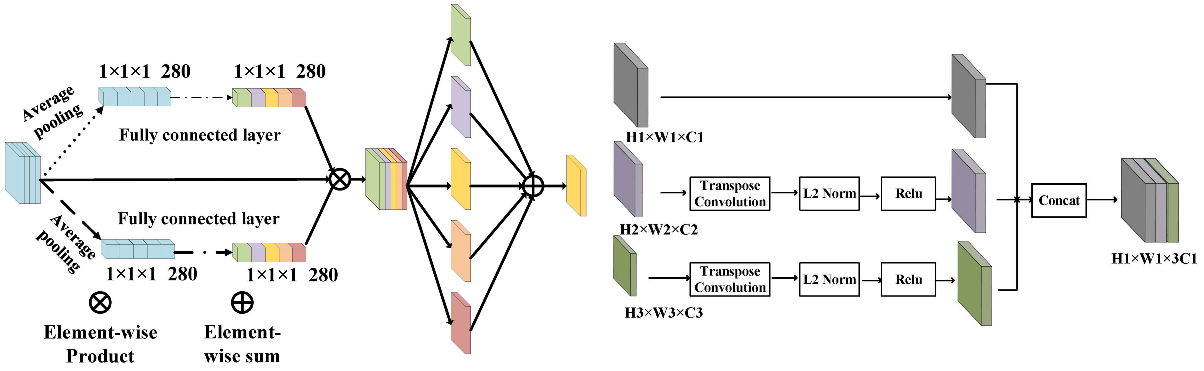
(b) MAE-FPN structure

Fig. 2 The architecture of our proposed network. Figure (a) shows the overall network structure with yolo_v5 as the detector. Figure (b) illustrates the module composition in MAE-FPN.

feature layers P2,P3,P4,P5. Next, the adaptive feature convergence distribution module (FCDM) is used to further fuse the highest layer feature P5 with the lower layer features P2,P3,P4 to obtain the fused feature F. Finally, the fused feature F enhances P4,P3,P2 layer-by-layer, and obtains the final feature level C2,C3,C4,C5 for object detection. The C2 layer feature map is directly output into the Seg-head, and the input feature map is recovered to 640×640×3 size after two upsampling.

3.2 MAE-FPN

In order to improve the detection level of small underwater objects, existing object detection networks usually adopt the feature pyramid (FPN)[17] structure and its improvements like CB-FPN[21], Nas-FPN[9], Aug-FPN[10], CE-FPN[23]. However, the existing FPN structure has the following problems: (1) the lateral connection of FPN has only limited non-dynamic receptive fields, which makes the extracted target feature information insufficient; (2) the feature information within the highest layer of the FPN network is not effectively fused, and the top-down fusion is



(a) MCCM

(b) FCDM

Fig. 3 The detail of our proposed method. In Fig (a) the process of convolutional calibration at different scales in the MCCM network is shown in detail. Fig (b) demonstrates the process of aggregating features from different layers in the FCDM structure.

relatively single, which makes the semantic information of the higher layers easily lost. The above problems will lead to insufficient feature extraction of underwater targets by the feature pyramid structure, and even cause some target features to be lost, resulting in poor detection of small targets such as fish.

Therefore, we deeply analyze the existing FPN structure and propose the MAE-FPN structure, as is shown in Fig.2b. It consists of Multi-scale Convolutional Calibration Module (MCCM) and Feature Calibration and Distribution Module (FCDM) from left to right.

3.2.1 Multi-scale Convolutional Calibration Module

The structure of the MCCM network is shown in Fig.3a. In the MCCM structure, the larger receptive field convolution in the original FPN structure is first decomposed and equivalently replaced by N multiple 3×3 convolutions to reduce the computational overhead of the network. These N 3×3 convolutions can represent a total of N different receptive fields ranging from the 3×3 scale to the $(2N+1)\times(2N+1)$ scale. By decomposing and sharing convolution kernels in this way, the MCCM network is able to cover multiple receptive fields consecutively without introducing the large kernel convolutions and with less computational overhead. The multi-scale convolution module is expressed as follows:

$$f(2N+1) = \begin{cases} F_{\mathbf{1},\mathbf{1}}(x), \;\; N=0 \\ F_{\mathbf{1},\mathbf{3}}(x), \qquad N=1 \\ F_{N,\mathbf{3}}(f(2N-1)), \;\; N>=2 \end{cases} \tag{1}$$

Where, $f(2N+1)$denotes the feature map with the receptive field of $(2N+1)\times(2N+1)$, $F_{i,j}(x)$ denotes the ith receptive field of $j\times j$ convolution, f represents the input feature map, N represents the number of 3 × 3 convolution kernels in the convolution group, and the larger value of N represents the larger maximum receptive field of the convolution group.

After extracting multiple receptive field feature maps at different scales, the scale calibration module in the MCCM network selects different receptive field features based on the channel attention mechanism[11] to adaptively select different receptive field features and weight the channels of the feature maps of interest among them, assigning higher or lower weights to the features at different scales. This is mainly achieved by the following four steps.

1) Feature splicing: After extracting the multi-scale sensory field feature maps within each feature map, the MCCM module splices the feature maps into a feature map F with global multi-scale feature information.

2) Feature encoding: after compressing the feature maps of each channel space based on the channel attention mechanism and encoding them

in the channel dimension, they are concatenated to generate a valid channel identifier for description. In order to be able to accurately label each channel, the MCCM module employs Global Maximum Pooling Operation (GMP) and Global Average Pooling (GAP)[43] to compress the input feature maps in the spatial dimensions and to generate the channel identifiers mc and ac, respectively. Assuming that the number of channels, height and width of the input features are [C, H, W], the spatial feature encoding process for each channel is shown in Eq.2:

$$\begin{gathered} m_c = \operatorname{Max}\left(f_c(i,j)\right), 0 < i < H \& 0 < j < W, \\ a_c = \frac{\mathbf{1}}{H \times W} \sum_{i=\mathbf{1}}^{H} \sum_{j=\mathbf{1}}^{W} f_c(i,j), \end{gathered} \tag{2}$$

where m_c is the global maximum pooling, a_c is the global average pooling, and $f_c(i,j)$ is the feature map with the number of channels, height and width of [C, H, W], respectively.

3) Channel calibration: the module interacts with the identifiers m_c and a_c through two fully connected layers and calculates the weights of each dimension of the feature, corresponding to the weights of each channel of the original feature. Finally, the weights of each channel are multiplied with the original feature by channel. Finally, the weights of each channel are multiplied with the original features by channel, and the calculation process is expressed as Eq.3.

$$f_c = f_c \times \sigma\left(W_{\mathbf{1}}\left(\delta\left(W_{\mathbf{2}}\left(m_c\right)\right)\right) + W_{\mathbf{1}}\left(\delta\left(W_{\mathbf{2}}\left(a_c\right)\right)\right)\right), \tag{3}$$

4) Feature separation and fusion: after the above weighting operation and calibration, segmentation and reduction by channel dimensions are performed into a multi-scale feature map before stitching. After completing feature separation, additive fusion is performed to realize multi-scale feature fusion.

3.2.2 Feature Calibration and Distribution Module

The highest feature level F5, which contains rich semantic information, is not fused sufficiently and effectively, and the top-down fusion method is relatively single, so its improvement is considered. In the proposed FCDM module, a new fusion path is added by introducing SELayer, which enhances the highest layer P5 layer features so that its output features have stronger semantic information. Then the lower layers P2, P3, and P4 are fused through the Feature Fusion layer.The structure of the Feature Fusion layer is shown in the following Fig.3b.

In the Feature Fusion structure, Transpose Convolution is implemented in the context features before fusing the features to give them the same spatial size with the target features. Since different feature values from different layers have different scales, batch normalization and ReLU activation functions are performed after each layer to normalize their scales.

Specifically, the multiscale feature aggregation module first sends the F (Feature Fusion) layer to the average pooling layer according to a pyramidal downsampling rate to transform the aggregated features into different scale spaces, which are 1, 2, 4, 8 for the nth layer (n ∈ 2, 3, 4, 5), respectively. The dimensions were then passed through a 3 × 3 convolutional layer, batch normalization, and ReLU activation function after each downsampling, respectively, in order to regenerate the dimensions that match the original channels.

The MAE-FPN network after the above steps outputs features as in Eq.

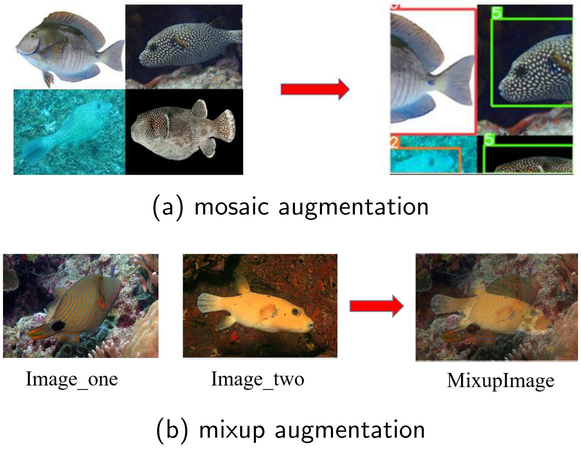
$$C_i = SE_i \times P_i + F \times D(i), \tag{4}$$

where C_i is the output feature layer after fusion redistribution, SE_i is the attention mechanism, F is the Feature Fusion layer multi-scale feature aggregation processing, $D(i)$ is the downsampling rate on the fusion path.

Due to our proposed MAE-FPN, the feature mapping of each fusion path contains both semantic information and feature details, and more complementarities can be retained on the fusion paths, so the fusion effect can be enhanced to obtain more superior feature extraction performance.

3.3 Fish Segmentation and Detection Dataset

One of the main problems of underwater image object detection is the scarcity of data resources. Specifically, compared with natural images, the amount of existing publicly available and well-labeled underwater object detection data is small, and most of them are obtained by capturing in



(a) mosaic augmentation

(b) mixup augmentation

Fig. 4 Data augmentation process. We use the above two data augmentation means to make our constructed dataset, with a wide variety of categories and diverse scenarios, closer to the actual scenarios, which is more suitable for the research of object detection algorithms.

a certain environment, and the detection categories and scenes are relatively single. Therefore, underwater target detection algorithms often do not have domain adaptation. Based on the above problems, we constructed an underwater fish object segmentation detection dataset (FSD) based on web crawling as well as fusion of different data enhancement means.

In order to make our proposed dataset more representative and closer to the actual underwater object detection environment, we use a fusion of data enhancement means, focusing on the enhancement of small objects.

First, Mosaic data enhancement was used to improve the ratio of underwater small object data. The process is shown in Fig. As can be seen from the figure, Mosaic data enhancement takes the four fish images on the left and randomly scales and splices them together to obtain the new image on the right, which increases the number of small targets in the image, which is equivalent to learning many small objects at the same time. Many small objects are added to the image, which is equivalent to learning the features of the four fish images at the same time. After enhancement, the distribution of small targets in the data set is more uniform, which makes the network model more sufficiently trained on the small targets in the images and enhances the robustness of the network.

Many small targets are added to the images by Mosaic data enhancement method in a randomized scaling and splicing manner, which enhances the robustness of the network. However, during the training process of fish data samples, it leads to overfitting problem from time to time due to limited training samples. Therefore, we use Mixup data enhancement approach to solve the above problem.Mixup is a data linear enhancement method, the core idea of which is to randomly select two images in the training sample and mix them proportionally to generate a new image. The computational process is as follows:

$$\begin{cases} \tilde{x} & = \lambda x_i + (1-\lambda)x_j \\ \tilde{y} & = \lambda y_i + (1-\lambda)y_j \end{cases}, \lambda \in [0.1] \tag{5}$$

where (x_i ,x_j) is the original input vector of the two samples i,j; (y_i,y_j) is the one hot labels of the two samples. From the above equation, Mixup data augmentation is used to generate a set of new samples by fusing positive and negative samples to the input vectors in the original samples through fusion coefficients λ. The new samples ($\tilde{x}$, $\tilde{y}$) retain the feature information of the original samples but are not identical to them, doubling the overall sample information and thus expanding the sample capacity of the dataset.

4 Experiments

4.1 Experimental data and environment

Datasets

The experiments are based on our proposed FSD dataset and the URPC2021 dataset, which is an underwater optical image data for underwater object detection, and the data are collected from the real marine environment. The dataset contains a total of 7600 images of underwater objects classified into four categories: scallop, spiny fish, holly and starfish.

The FSD dataset is obtained by the method described in Section.5a. After data filtering, data enhancement, data augmentation and other operations, 38 categories with a total of 6242 fish images were finally obtained. The sample images and corresponding numbers of some of the major fish species are shown in Fig.5a.

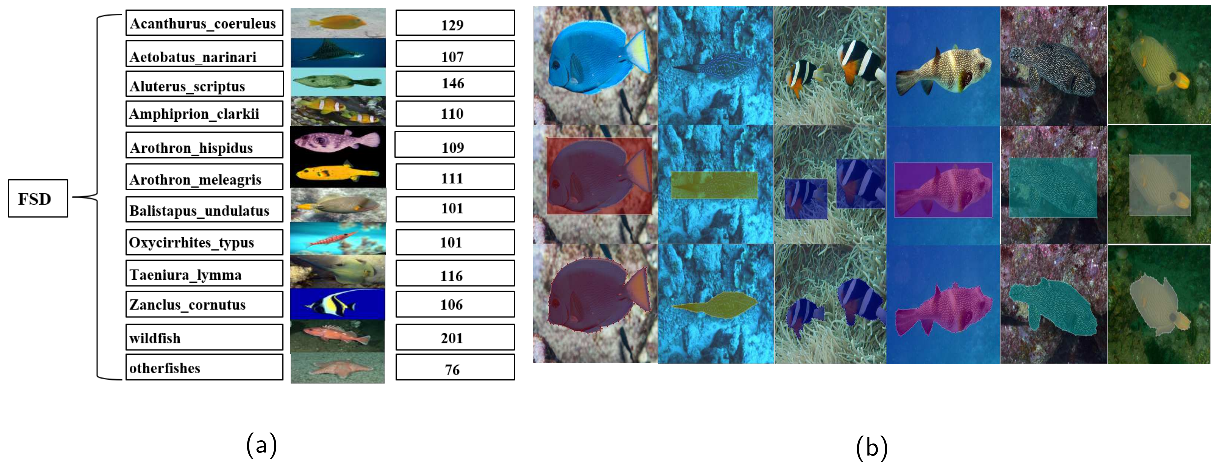


Fig. 5 The samples of our proposed FSD dataset. Figure (a) shows a portion of the fish names and the corresponding numbers. Figure (b) presents the image data and the corresponding object detection, instance segmentation labelling results.

Table 1 The Dataset division strategy

Dataset	Train	Valiation	Test
FSD	4370	624	1248
URPC2021	5320	938	1875
Total	9690	1562	3123

We conduct experiments based on two datasets respectively to validate the performance of our proposed method. The dataset divide strategy is shown in the Table.1.

Evaluation Metrics

The evaluation metrics are the number of parameters of the model (Params, millions); the computational volume of the model (FLOPs, billions); the Average Precision (AP); and the average precision of the targets of different scales, such as AP_s (small targets, with size less than 32×32), AP_M (medium targets, with size between 32*32 and 96×96), and AP_L (large targets, with size more than 96×96).

4.2 Implement Details

To validate the effectiveness of the proposed method, we use Pytorch as a deep learning framework for network model training.The GPU is NVIDIA RTX 3090 with 24GB memory. All object detection networks are running based on MMDetection

Training parameter settings: For the experimental parameter settings, the optimizer we used is stochastic gradient descent optimizer (SGD), and the batch-size is set to 2. The initial learning rate during the training process is set to 0.001, and the epoch is 120, and the learning rate decreases to 0.0004 and 0.0002 at the 40th epoch and the 80th epoch. input resolution of the image is 640 × 640, Momentum is set to 0.9, and the N value of the MCCM module is set to 4.

4.3 Comparison Study

To demonstrate the effectiveness of the MAE-FPN network optimisation, we conducted experiments based on basic networks such as ResNet, YOLOv5, Faster RCNN and compared them with state-of-the-art detectors such as Faster R-CNN[29], YOLOX[8], Libra R-CNN[26], Spares R-CNN[32], FCOS[34], DiffusionDet[6], VF[1], as well as the newest underwater detectors RoiAtt[16], Boosting R-CNN[30] and SWIPENet[5], specifically designed for underwater images.

The results of the comparison with these state-of-the-art methods are shown in Table.2. From this, we can conclude that our method has a great improvement in detection accuracy compared to the existing methods, and the network structure of Faster RCNN+MAE-FPN achieves the best performance. In addition, our MAE-FPN structure combined with different detectors, such as YOLO_v5 and Faster RCNN structure,

Table 2 Comparison of mAP metrics for underwater image data object detection. The best results are shown in bold. The experimental results show that the MAE-FPN network designed in this paper combined with different detectors can achieve improved detection performance and inference speed with significant improvement compared to existing underwater object detectors and pre-processing object detection methods.

method	backbone	mAP	Parameters(M)	FPS
SOTA Detectors:				
YOLO_V5	ResNet-50	0.406/0.421	**54.2**	58.4
YOLOX	CSPDarkNet	0.381/0.432	60.3	54.8
Faster R-CNN	ResNet-50	0.401/0.428	122.14	26.4
Faster R-CNN	ResNet-101	0.414/0.441	131.6	24.5
Libra R-CNN	ResNet-50	0.420/0.452	146.60	22.4
Libra R-CNN	ResNet-101	0.435/0.457	164.9	19.4
Sparse R-CNN	ResNet-50	0.425/0.443	212.5	16.6
DiffusionDet	ResNet-50	0.428/0.453	314.4	12.2
FCOS	ResNet-50	0.394/0.438	117.4	27.8
VFNet	ResNet-101	0.416/0.443	141.4	23.1
Underwater Object Detector:				
Boosting R-CNN	ResNet-50	0.433/0.442	143.55	23.0
Boosting R-CNN	ResNet-101	0.443/0.445	160.55	20.3
RoiAtt	ResNet-50	0.462/0.447	183.45	18.1
RoiAtt	ResNet-101	0.473/0.451	192.45	17.8
SWIPENet	ResNet-50	0.451/0.449	166.7	19.3
SWIPENet	ResNet-101	0.466/0.452	179.4	18.9
Ours:				
MAE-YOLO_V5	ResNet-50	0.451/0.463	84.4	**46.8**
MAE-Faster-RCNN	ResNet-50	**0.466/0.478**	144.7	22.4
MAE-Faster-RCNN	ResNet-101	**0.479/0.484**	164.6	20.1

Table 3 Comparison results of the number of parameters and accuracy of each feature pyramid network.

method	backbone	Parameters(M)	AP(%)	AP$_S$(%)	AP$_M$(%)	AP$_L$ (%)
FPN	YOLO_v5-ResNet-50	54.2	0.421	0.378	0.426	0.429
Nas-FPN		92.6	0.456	0.434	0.461	**0.473**
Aug-FPN		91.8	0.457	0.433	0.466	0.472
CE-FPN		89.6	0.449	0.423	0.465	0.459
MAE-FPN		84.4	**0.463**	**0.446**	**0.472**	0.471
FPN	YOLO_v5-ResNet-101	73.9	0.436	0.391	0.451	0.466
Nas-FPN		113.6	0.466	0.440	0.473	**0.485**
Aug-FPN		114.5	0.465	0.451	0.476	0.468
CE-FPN		123.6	0.470	0.451	0.481	0.478
MAE-FPN		114.2	**0.480**	**0.469**	**0.489**	0.482

can achieve performance improvement, which also shows the generality of our method.

In addition, we compare with improved FPN-based structures such as Nas-FPN[9], Aug-FPN[10], CE-FPN[23], etc., and conduct experiments on the FSD dataset to verify in detail the performance of our proposed MAE-FPN for object detection at different scales.

The comparison results are shown in the Table.3. It can be concluded that in the experiments based on YOLOv5 and ResNet-50, the average detection accuracy of MAE-FPN is 46.3%, which is 4.1% better than the original model using

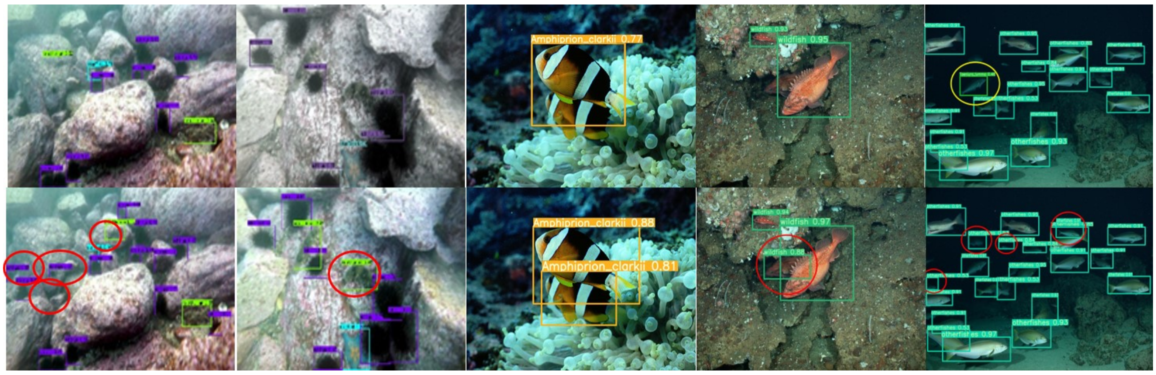


Fig. 6 The visual detection results of our proposed MAE-FPN. The first line shows the yolo_v5 detection results for the original setup and the second line shows the detection results for our MAE-FPN structure. In comparison, our method has stronger performance for some dense small targets, occluded targets, etc. (as shown in the circled note on the figure).

FPN. In the detection of small, medium and large targets, the average detection accuracy of MAE-FPN is improved over the original FPN network. In particular, MAE-FPN has an average detection accuracy of 44.6% for small underwater objects, which is better than the other four detection models, and the overall performance is also better. When ResNet-101 is used as the backbone network, the APS value of small object detection accuracy increases from 39.1% to 46.9%, which is an improvement of 7.8%. Comparing the above experimental results, it can be concluded that the overall detection performance of MAE-FPN is better than the mainstream feature pyramid network model, and can be effectively used for the detection of the same features. The tower network model can effectively improve the detection performance of small fish objects.

4.4 Ablation Study

To verify the effectiveness of the different modules, we also set up ablation experiments to test the multi-scale convolutional calibration module as well as the feature clustering distribution module. The benchmark model chosen for the experiment is YOLOv5, the backbone network is ResNet-50, and the dataset used is the validation set of the FSD dataset. The dataset used is the validation set of the FSD dataset.

Our ablation experiments are set up as follows: the first row uses only the original FPN model; the second row uses only the MCCM module to investigate the effect of the multi-scale convolutional calibration module on the detection performance. The third row uses only the FCDM module to investigate the effect of augmenting the P5 level features and the distribution of feature aggregation. The fourth row uses both modules. The results of the ablation experiments are shown in Table.4.

From the Table.4, we can see that after using the MCCM module, the detection accuracy of the detection model as a whole, as well as the detection accuracy of small, medium and large targets are improved, and the proportion of improvement in the detection accuracy of small targets is the highest. After using the FCDM module, the detection performance of small objects is mainly improved, and the value of APS is increased from 31.2% to 33.1%, which is 1.9%.

In addition, in order to select the best N value in MCCM module to get better small object detection effect, this paper conducts a comparison experiment under different N values, and the results are shown in Table 5.

From Table.5, it can be seen that as the value of N increases, the overall average detection accuracy of MAE-FPN shows a tendency to first increase and then decrease, and the average accuracy of MAE-FPN reaches the highest when the value of N is 4 or 5. Considering the number of model parameters, the computational volume and the target detection accuracy at each scale, in this chapter, the N value of the MCCM module is set to 4, which means that the detection field areas are 1×1, 3×3, 5×5, 7×7 and 9×9.

4.5 Discussions

For our proposed MAE-FPN network architecture, we conduct experiments on the publicly

Table 4 Detection average precision of ablation study. MCCM represents Multi-scale Convolutional Calibration Module and FCDM represents Feature Calibration and Distribution Module.

method	backbone	AP(%)	AP$_S$(%)	AP$_M$(%)	AP$_L$(%)
FPN	ResNet-50	0.421	0.378	0.426	0.429
FPN+MCCM	ResNet-50	0.442	0.426	0.458	0.442
FPN+FCDM	ResNet-50	0.435	0.419	0.455	0.431
FPN+MCCM+FCDM	ResNet-50	**0.463**	**0.446**	**0.472**	**0.471**

Table 5 N comparison results. According to the results, the optimum performance is achieved when the value of N is taken as 4.

N	Parameters	FLOPS	AP(%)	AP_s(%)	AP_m(%)	AP_L(%)
1	41.3	226.7	0.445	0.431	0.468	0.436
2	43.8	279.6	0.449	0.435	0.472	0.440
3	46.4	327.5	0.456	0.442	0.472	0.454
4	49.0	378.2	**0.463**	**0.446**	0.472	**0.471**
5	52.1	428.6	0.460	0.439	**0.474**‘	0.467
6	54.9	472.5	0.459	0.441	0.470	0.466

available URPC dataset and our own proposed FSD dataset, respectively, to fully validate the model performance.

Through comparative experiments, our model outperforms existing target detection networks on the URPC and FSD datasets. Meanwhile, by comparing with the existing improved FPN-based structure, our model has improved in terms of algorithmic accuracy as well as model computational complexity.

We also conduct ablation experiments to gradually validate the modular performance of the proposed MCCM and FCDM. Through the experimental results, it is found that compared with MCCM, FCDM has more obvious improvement for the model, which is due to the fact that the FCDM module enhances the features of the highest P5 layer and then enhances the features of each fusion path, which makes the semantic information of the feature maps of each layer complementary, and is more conducive to the detection of small targets. The fusion of MCCM and FCDM results in a more obvious model improvement, especially for the detection of small underwater targets, which demonstrates the effectiveness of our scheme.

Finally, we also compare the N-value of the convolution group in the MCCM module, and when model complexity and performance are considered together, the performance of MCCM is optimal when the N-value is 4.

5 Conclusion

In this paper, we propose a new underwater image target detection network, MAE-FPN, which is a model that solves the problem of poor performance of existing methods for recognizing small underwater targets by fusing the high-level feature information more efficiently and then distributing the outputs in order to fuse the multi-scale information more adequately. Then, in order to solve the problem of scarce data resources for underwater image target detection, we constructed a richly varied and well-labeled fish segmentation detection dataset FSD, and at the same time, in the process of construction, we fused a variety of data enhancement means to expand the proportion of small samples and difficult-to-detect samples, and further enriched the sample information on the basis of the original sample features, which made the dataset diversified. Our method is validated on the public dataset as well as the FSD dataset, and the experimental results prove that our method achieves good performance in terms of speed as well as accuracy, and outperforms the existing FPN and its improved structure.

Declarations

- Funding. The project supported by the National Natural Science Foundation of China(No: 61871124 and 61876037), The national defense Pre-Research foundation of

China, by the fund of Science and Technology on Sonar Laboratory (No: 6142109KF201806), by the Stable Supporting Fund of Acoustic Science and Technology Laboratory (No:JCKYS2019604SSJSSO12).

- Conflict of interest. The authors have no conflicts of interest to declare relevant to this article's content
- Ethics approval. This article does not contain any studies with animals performed by any of the authors.
- Availability of data and materials. The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

[1] Ahmed A, Tangri P, Panda A, et al (2019) Vfnet: A convolutional architecture for accent classification. In: 2019 IEEE 16th India Council International Conference (INDICON), IEEE, pp 1–4

[2] Bai X, Wang W (2016) Principal pixel analysis and svm for automatic image segmentation. Neural Computing and Applications 27:45–58

[3] Biau G, Scornet E (2016) A random forest guided tour. Test 25:197–227

[4] Carion N, Massa F, Synnaeve G, et al (2020) End-to-end object detection with transformers. In: European conference on computer vision, Springer, pp 213–229

[5] Chen L, Liu Z, Tong L, et al (2020) Underwater object detection using invert multi-class adaboost with deep learning. In: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, pp 1–8

[6] Chen S, Sun P, Song Y, et al (2022) Diffusiondet: Diffusion model for object detection. arXiv preprint arXiv:221109788

[7] Duan K, Bai S, Xie L, et al (2019) Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6569–6578

[8] Ge Z, Liu S, Wang F, et al (2021) Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:210708430

[9] Ghiasi G, Lin TY, Le QV (2019) Nas-fpn: Learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7036–7045

[10] Guo C, Fan B, Zhang Q, et al (2020) Augfpn: Improving multi-scale feature learning for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12595–12604

[11] Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141

[12] Jalal A, Salman A, Mian A, et al (2020) Fish detection and species classification in underwater environments using deep learning with temporal information. Ecological Informatics 57:101088

[13] Kim B, Yu SC (2017) Imaging sonar based real-time underwater object detection utilizing adaboost method. In: 2017 IEEE Underwater Technology (UT), IEEE, pp 1–5

[14] Kotsiantis SB (2013) Decision trees: a recent overview. Artificial Intelligence Review 39:261–283

[15] Law H, Deng J (2018) Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European conference on computer vision (ECCV), pp 734–750

[16] Liang X, Song P (2022) Excavating roi attention for underwater object detection. In: 2022 IEEE International Conference on Image Processing (ICIP), IEEE, pp 2651–2655

[17] Lin TY, Dollár P, Girshick R, et al (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125

[18] Lin TY, Goyal P, Girshick R, et al (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988

[19] Liu M, Lin K, Huo W, et al (2023) Feature enhancement modules applied to a feature pyramid network for object detection. Pattern Analysis and Applications 26(2):617–629

[20] Liu W, Anguelov D, Erhan D, et al (2016) Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, pp 21–37

[21] Liu Z, Cheng J (2023) Cb-fpn: object detection feature pyramid network based on context information and bidirectional efficient fusion. Pattern Analysis and Applications 26(3):1441–1452

[22] Liu Z, Lin Y, Cao Y, et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022

[23] Luo Y, Cao X, Zhang J, et al (2022) Ce-fpn: Enhancing channel information for object detection. Multimedia Tools and Applications 81(21):30685–30704

[24] Maćkiewicz A, Ratajczak W (1993) Principal components analysis (pca). Computers & Geosciences 19(3):303–342

[25] Nakashima Y, Babaguchi N, Fan J (2012) Intended human object detection for automatically protecting privacy in mobile video surveillance. Multimedia Systems 18:157–173

[26] Pang J, Chen K, Shi J, et al (2019) Libra r-cnn: Towards balanced learning for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 821–830

[27] Qi S, Du J, Wu M, et al (2022) Underwater small target detection based on deformable convolutional pyramid. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 2784–2788

[28] Redmon J, Divvala S, Girshick R, et al (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

[29] Ren S, He K, Girshick R, et al (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28

[30] Song P, Li P, Dai L, et al (2023) Boosting r-cnn: Reweighting r-cnn samples by rpn's error for underwater object detection. Neurocomputing 530:150–164

[31] Song W, Fu C, Zheng Y, et al (2022) Protection of image roi using chaos-based encryption and dcnn-based object detection. Neural Computing and Applications pp 1–14

[32] Sun P, Zhang R, Jiang Y, et al (2021) Sparse r-cnn: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14454–14463

[33] Sung M, Yu SC, Girdhar Y (2017) Vision based real-time fish detection using convolutional neural network. In: OCEANS 2017-Aberdeen, IEEE, pp 1–6

[34] Tian Z, Shen C, Chen H, et al (2019) Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9627–9636

[35] Villon S, Chaumont M, Subsol G, et al (2016) Coral reef fish detection and recognition in underwater videos by supervised machine learning: Comparison between deep learning and hog+ svm methods. In: International Conference on Advanced Concepts for Intelligent Vision Systems, Springer, pp 160–171

[36] Wang H, Song Y, Huo L, et al (2023) Multiscale object detection based on channel and data enhancement at construction sites. Multimedia systems 29(1):49–58

[37] Xianbao C, Guihua Q, Yu J, et al (2021) An improved small object detection method based on yolo v3. Pattern Analysis and Applications 24:1347–1355

[38] Xu F, Wang H, Peng J, et al (2021) Scale-aware feature pyramid architecture for marine object detection. Neural Computing and Applications 33:3637–3653

[39] Xu F, Wang H, Sun X, et al (2022) Refined marine object detector with attention-based spatial pyramid pooling networks and bidirectional feature fusion strategy. Neural Computing and Applications 34(17):14881–14894

[40] Yang C, Li Y, Jiang L, et al (2023) Foreground enhancement network for object detection in sonar images. Machine Vision and Applications 34(4):1–14

[41] Ying L, Zhang T, Xu C (2015) Multi-object tracking via mht with multiple information fusion in surveillance video. Multimedia Systems 21:313–326

[42] Zhang SX, Zhu X, Hou JB, et al (2023) Graph fusion network for multi-oriented object detection. Applied Intelligence 53(2):2280–2294

[43] Zhou B, Khosla A, Lapedriza A, et al (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929