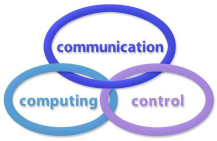
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 21, Issue: 4, Month: August, Year: 2026
Article Number: 6915, https://doi.org/10.15837/ijccc.2026.4.6915



CCC Publications

A Few-shot Learning Algorithm for Underwater Small Object Detection based on the Transformer Architecture

X. Wu, X. Zhang, P. Tan

Xiaobing Wu
College of Intelligent Systems Science and Engineering
Harbin Engineering University, China
No. 145 Nantong Street, Nangang District, Harbin 150009, Heilongjiang
296055178@qq.com
Naval Research Institute, China

Xiaoyu Zhang*
College of Artifitial Intelligence
Nankai University, China
No. 38, Tongyan Road, Haihe Education Park, Jinnan District, Tianjin 300380, Tianjin
*Corresponding author: 819029@nankai.edu.cn

Panlong Tan
Haihe Lab of ITAI
Building 1, Xin'an Entrepreneurship Plaza, Tanggu, Binhai New Area, Tianjin 300459, Tianjin
tanpl@hl-it.cn

Abstract

This paper proposes a few-shot learning approach based on the transformer model to tackle the challenges of underwater small object detection in sonar images using deep learning techniques. We analyze the difficulties encountered in underwater small object detection with deep learning methods when training samples are limited and propose corresponding solutions to address these challenges. To address the challenges of unclear object contours and blurred feature details in sonar images, we employ a transformer-based object detector that leverages the attention mechanism to effectively utilize global image information for precise object localization and identification. To handle the prevalence of small targets, we design a Multi-Branch Feature Extraction Module that aggregates feature maps with different receptive fields, thereby enhancing the effective utilization of small object features. In addition, the core design of the RT-DETR model is incorporated as the baseline architecture, which significantly improves the real-time performance of the detector. To overcome the challenge of limited training data, we apply transfer learning by first training the entire MP-DETR (More Precise DETR) network on a large-scale, general-purpose dataset. Then the IOU-aware query selection module and detection head of the MP-DETR network is fine-tuned using a self-compiled underwater small sample sonar image training dataset. The proposed approach achieves a deep neural network suitable for underwater object detection based on sonar images. Experiments were conducted on a self-constructed underwater small-sample sonar image dataset, and the proposed MP-DETR achieved 98.5% mean average precision (MAP) and 53 frames per second (FPS) real-time performance, which provides higher detection accuracy and real-time performance compared with existing methods.

Keywords: sonar image, underwater small object detection, MP-DETR, transfer learning.

1 Introduction

With the rapid development of the global ocean economy, human development and utilization of the oceans have been increasing, which has led to the frequent occurrence of maritime accidents [16, 20, 23]. In this context, securing key sea routes against threats such as underwater weapons (e.g., mines) has become an important issue of global concern [8, 28]. Underwater object detection technology plays a key role in this process and is not only crucial in monitoring military threats such as mines, but also has indispensable value in maritime emergency response missions such as search and rescue [6, 27, 38].

Sonar, as the primary technology for underwater object detection, utilizes acoustic waves that can travel long distances through water, making it particularly effective for detecting small objects on the seabed [11, 36, 39]. Compared to electromagnetic and light waves, acoustic waves experience less attenuation and scattering in water, allowing for longer detection ranges. However, sonar images are often affected by noise, resulting in blurred or distorted object contours, especially when detecting small targets [29, 41]. Additionally, the limited data transmission rates in underwater environments pose a challenge for remote real-time processing. To address these issues, more small vessels and underwater unmanned submersibles are now processing sonar data locally, reducing transmission load. Therefore, developing a real-time, high-precision object detection system that can be deployed on such platforms is crucial [14, 17, 45]. Traditional machine learning and image processing methods face difficulties in efficiently and accurately detecting distant small objects in sonar images due to challenges such as low image clarity, noise interference, and small object detection [1, 12, 33]. In contrast, deep learning-based methods show great potential in overcoming these limitations, offering higher detection accuracy and improved robustness. Furthermore, deep learning models can be deployed on underwater platforms, such as small vessels and unmanned submersibles, enabling automatic target detection without human intervention [22, 32, 34]. This not only generates smaller data volumes for detection results but also reduces data transmission demands, leading to cost savings and enhanced operational efficiency.

Currently, deep learning-based object detectors fall into two major categories: Convolutional Neural Network (CNN-based) detectors and Transformer-based detectors [3, 30, 40]. CNN-based detectors typically include convolutional layers, pooling layers, and fully connected layers. They have been widely applied in various domains, such as video surveillance, disease detection, threat alerts, and shipwreck searches. However, these detectors often involve complex post-processing steps, such as Non-Maximum Suppression (NMS) and pre-select bounding boxes, which limit their ability to perform true end-to-end detection and hinder real-time performance improvements [7, 10, 25]. On the other hand, Transformer-based object detectors, particularly the DETR series, have gained prominence. DETRs incorporate self-attention processes in their encoders and decoders to capture global context and long-range dependencies, they also eliminate the need for NMS and other manually designed elements, simplifying the target detection process and enabling true end-to-end monitoring [4, 26, 48]. However, DETR faces challenges such as long training times and query optimization difficulties [2]. To address these issues, researchers have proposed optimization algorithms like Deformable-DETR, Conditional-DETR, Anchor DETR, DAB-DETR, DN-DETR, and DINO, which improve training convergence, query optimization, and positional information accuracy [15, 18, 21, 35, 43, 47]. Despite these advancements, the complexity of attention mechanisms in Transformer-based architectures has hindered real-time object detection. With the object of achieving real-time detection, RT-DETR uses intra-scale feature interaction and cross-scale feature fusion of multi-feature maps instead of cross-scale interaction of multi-scale feature maps, which greatly reduces the computational cost associated with the attention mechanism. It outperforms all CNN-based and other Transformer-based detectors in terms of real-time computation and detection accuracy, enabling the practical deployment of Transformer-based object detectors [46]. However, RT-DETR also has its limitations in that its Neck part mainly performs self-attentive operations on the S5 feature layer, which contains richer semantic information. The direct fusion of this layer with lower feature layers (S3, S4) will bring about a huge difference in information levels, which may negatively affect the detection results.These methods have achieved satisfactory performance in visible-light image detection; however, due to the inherent differences between sonar and visible-light imagery, they cannot be directly applied to sonar image object

detection.

Publicly available underwater sonar datasets suitable for training deep neural networks are scarce due to reasons such as expensive equipment and difficult acquisition conditions, which makes training high-performance object detectors a major challenge [19, 37]. In order to overcome this challenge, researchers have proposed a variety of methods, which can be divided into two main categories: optimizing the model structure and utilizing transfer learning techniques. Optimizing the model structure mainly involves improving the architecture of the existing network so that it can learn the image information more efficiently.Fan and colleagues optimized the overall network architecture of the YOLOv4 network and replaced it with a more efficient feature extraction module to improve the feature extraction capability of the detector [5].LE and his team embedded a parameterized Gabor filter module in the deep neural network cascade layer to improve the scale and orientation decomposition of the image and obtain better detection results [13].Phung and colleagues used a generative adversarial network to augment the dataset with additional sonar images and introduced a hierarchical Gaussian process classifier to improve the classification accuracy of the traditional convolutional neural network [24]. Although these methods achieve promising results, most of them are developed on CNN-based architectures, which struggle to effectively aggregate global-scale features. As a result, their performance in small object detection remains limited, and they often fail to account for the blurring or distortion commonly present in sonar images. The main idea of the transfer learning technique is to first pre-train the object detector using large-scale datasets that are publicly available in the related fields, and then use the limited available sonar datasets for fine-tuning, with the training can enable the network to obtain sufficient feature extraction and refinement capability, and apply this capability to the detection of small-sample sonar images. Huo et al. fine-tuned the VGG network using a large amount of real sonar data and semi-synthetic data and applied it to a specific sonar image dataset [9]. Tang froze some of the convolutional layers in the pre-trained YOLOv3 model on the COCO dataset and fine-tuned the remaining convolutional layers on a real sonar dataset [42]. Zhang et al. improved the overall detection framework of the YOLOv5 network and performed the second pre-trained on a real sonar image dataset, which solved the problem of discrepancy between optical image objects and sonar image objects [44]. Transfer learning techniques generally achieve better results for small-sample sonar image learning compared with methods that solely optimize model structures. However, these approaches primarily rely on pre-learned features and do not incorporate mechanisms specifically tailored to the unique challenges of sonar imagery, such as detecting small targets or handling blurred and distorted object contours. Consequently, their performance may still be limited when objects are weakly represented or partially obscured.

In this paper, we aim to develop a high-performance detector capable of accurately identifying small objects with poorly imaged contours and deformation-prone shapes in sonar images. The detector is also designed for deployment on low-computing-capacity platforms, such as small ships and underwater unmanned submersibles. To this end, we propose MP-DETR (More Precise Detection Transformer), a transformer-based framework that leverages the self-attention mechanism to capture global contextual information for precise object localization. To enhance the representation of weak targets, we design a Multi-Branch Feature Extraction Module that aggregates features from multiple receptive fields. For real-time applicability, the core design principles of RT-DETR are integrated into the detector's neck, and its limitations are further addressed through a Reduce Information Refining Differences and Feature-Fusion Module (RIRDFM). Instead of conducting cross-scale interactions across all feature maps, this design applies self-attention only to semantically rich feature maps while employing an optimized path aggregation network for multi-scale feature fusion. This strategy substantially reduces the computational complexity of transformer-based detectors while preserving high detection accuracy. Furthermore, to mitigate the scarcity of annotated sonar data, we adopt a transfer learning paradigm: the network is pre-trained on a large-scale general dataset and subsequently fine-tuned on a small, self-compiled sonar dataset, thereby improving learning efficiency and enhancing detection robustness under data-limited conditions. The main contributions of this work are summarized as follows:

1. We propose MP-DETR, a transformer-based detector trained with a specially designed efficient transfer learning strategy that pretrains on a large-scale general dataset and fine-tunes on a small customized sonar dataset, achieving high detection accuracy while retaining real-time performance.

2. A Reduce Information Refining Differences and Feature-Fusion modules is proposed to address limitations in the design of RT-DETR, which reduces the information refinement differences between feature maps when fusing multi-scale feature maps. This improvement greatly increases the detector's object detection accuracy.

3. A backbone network called Multi-Branch Feature Aggregation Network (MBFANet) is proposed, which utilizes multiple Multi-Branch Feature Extraction Modules (MBFEMs) to extract multi-dimensional features from sonar images, providing higher quality multi-layer feature maps to encoders and decoders, and achieving improved detection of small objects in the image.

The remainder of this paper is organized as follows: Section II provides an overview of existing research in the field of object detection based on underwater sonar images. Section III describes the MP-DETR, the Reduce Information Refining Differences and Feature-Fusion modules (RIRDFM), and the Multi-Branch Feature Aggregation Network (MBFANet) backbone.Section IV conducts extensive experiments and analyzes the results. Section V summarizes the paper.

2 Related Work

2.1 RT-DETR

RT-DETR is the first real-time deployable transformer-based object detector, which is a true end-to-end object detector, surpassing the best CNN-based object detectors and transformer-based object detectors in terms of detection accuracy and real-time performance. RT-DETR has shown through experimental analysis that performing self-attention operations only on the S5 feature layer containing richer semantic information, and then fusing the self-attention processed S5 layer with low-level S3 and S4 features across scales, can achieve better detection performance than directly performing global self-attention operations on the S3, S4, and S5 feature layers. Not only that, this processing method can also greatly reduce the computational consumption of the encoders. This improvement allows RT-DETR to overcome the computational complexity bottleneck that previously limited transformer-based object detectors and become the first true real-time end-to-end object detector. In addition, RT-DETR integrates an IOU-aware query selection module for initializing object queries, which encourages selected tokens to have both high classification scores and high IOU scores by adding the IOU loss to the classification loss, with the top 300 tokens with the highest scores being selected for content query initialization. The position query is then obtained by mapping the content query to the corresponding detection box and encoding it, adding it to the content query, and summed to get the final object query. This object query initialization allows object queries to contain rich information about the objects, which makes it easier to optimize and thus speeds up the convergence of the model.

2.2 Transfer learning

Transfer learning is a machine learning technique that involves using a model developed for a previous task as the starting point and reusing it in the process of developing a model for a subsequent task. In recent years, transfer learning has been frequently applied to few-shot learning tasks to address the challenge of training large-scale models without having enough training samples. The fundamental idea behind transfer learning is that different domains can have inconsistent data distributions, which also leads to inconsistent weight data distributions in the deep neural networks trained on corresponding datasets, but the ability of the backbone to extract image features is generally universal. Based on this concept, the classical implementation process of transfer learning consists of first fully training the network model in a domain rich in training samples, then freezing the weights of the backbone and certain network layers that are not affected by the domain; and finally fine-tuning the weights of the rest of the network modules using a small dataset of the object domain, so as to obtain a deep neural network that can achieve good results in the object domain. In cases where there are significant differences in data distributions between different domains, additional domain adaptation modules will be included to enhance the model's generalization capabilities.

In the case of underwater sonar images, which contain fewer fine-grained details compared to generic visible light images, the feature extraction capabilities of deep neural networks trained on a

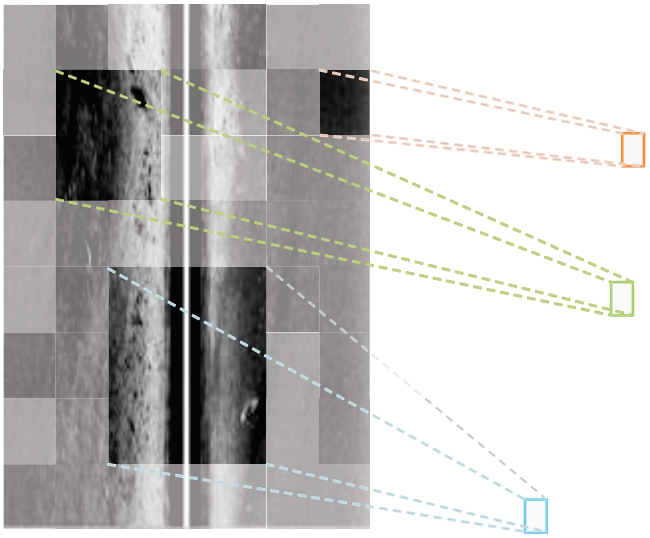


Figure 1: The receptive field of each pixel on the feature map obtained by convolution kernels of different sizes.

large visible light image are sufficient to handle sonar images. Moreover, there isn't a significant data distribution difference between them, so this paper does not include an additional domain adaptation module. Improvements are made based on classic transfer learning methods, which are sufficient to achieve good results.

3 Methodology

Sonar images contain more small objects than optical images, so the design of sonar image detectors must be more oriented toward small object detection. In this paper, we use the architecture of RT-DETR as a baseline. Since the original design of RT-DETR was tailored for object detection in general visible light images, and did not fully consider the problems related to small object detection, we made a series of modularization and training strategy improvements by combining the characteristics of sonar images and the shortcomings of RT-DETR itself. The detailed structure of the proposed MP-DETR and corresponding training scheme, the Reduce Information Refining Differences and Feature-Fusion modules (RIRDFM), and the Multi-Branch Feature Aggregation Network (MBFANet) backbone, are introduced in this section.

3.1 Multi-branch Feature Aggregation Network

Although Transformer can correlate the global information of the feature map and has some advantages in small object detection, the backbone of the general-purpose object detector is not necessarily effective in extracting the small object feature information in the image, and if the backbone cannot extract sufficiently rich small object information in the sonar image and provide it to the encoders and decoders, the detector will still be difficult to achieve satisfactory results in small object detection. Since there is a significant difference between the features of the object itself in the image and those of the surrounding background, which is the basis for the detector to distinguish between them, we hope that the feature maps obtained from the backbone network can fully capture the features of both, while selectively distinguishing between them. As shown in Fig. 1, processing the image with different sizes of convolutional kernels will result in feature maps with different receptive fields, so adjusting the size of the convolutional kernels can make the processed feature maps contain the feature information of different parts of the original image and the objects, and by synthesizing this information, the detector can better distinguish between the foreground information and the objects to be detected in the image.

Based on the above considerations and the unique characteristics of sonar images, we design a new backbone network, MBFANet, to replace the baseline backbone and achieve more efficient

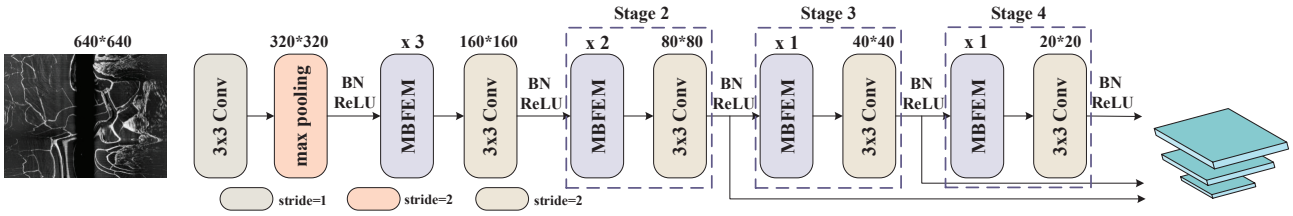


Figure 2: MBFANet Network Architecture.

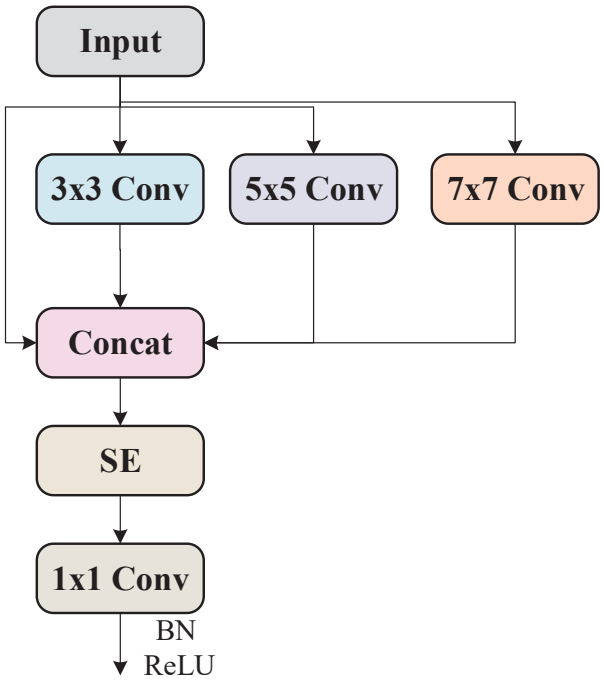


Figure 3: MBFEM Network Architecture.

and discriminative feature extraction. Compared with conventional backbone networks, MBFANet enhances multi-scale feature representation, better captures fine-grained details, and strengthens the detection of small and weak targets that are common in sonar imagery. The core building block of the network is the Multi-Branch Feature Extraction Module (MBFEM), as illustrated in Fig. 3. In this module, the input feature maps are first split into four channels: three channels process the features using convolution kernels of sizes 3×3, 5×5, and 7×7 (all with a stride of 1) to extract multi-scale contextual information, while the fourth channel serves as a shortcut connection. By aggregating the outputs from these four channels, MBFEM effectively captures features across multiple receptive fields, providing richer representations than standard single-branch or traditional multi-convolution modules. A Squeeze-and-Excite (SE) module is then applied to recalibrate channel-wise feature responses, emphasizing informative channels and producing more discriminative and detailed feature maps. To ensure consistency with the input feature dimensions, the concatenated outputs are passed through a 1×1 convolution followed by an activation function to obtain the final output of the module. Furthermore, at the end of each stage, feature maps undergo downsampling for subsequent processing. Considering that conventional max-pooling layers often lead to information loss, we replace them with 3×3 strided convolutions (stride=2), which preserve subtle features critical for detecting small or weakly represented objects. By stacking different numbers of MBFEM modules at different stages, MBFANet generates multi-scale feature maps that are then fed into the subsequent transformer structures, enabling the network to efficiently extract and refine features while maintaining high detection performance on sonar images.

Considering the practical deployment requirements of a sonar image object detector, we would like to design a detector with a smaller number of parameters to achieve faster inference. Inspired by

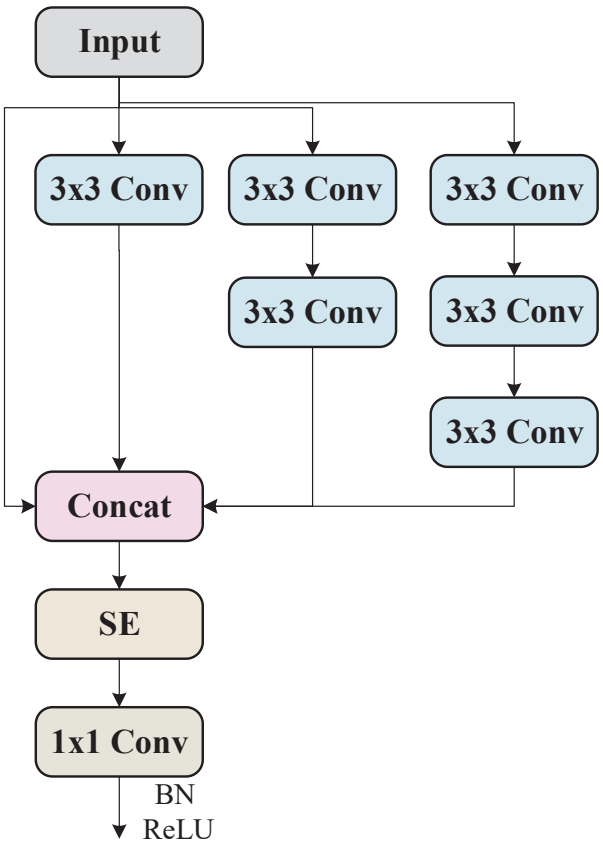


Figure 4: Modified MBFEM Network Architecture.

[31], we replace the originally proposed 5x5 and 7x7 convolutional kernels with two 3x3 convolutional kernels and three 3x3 convolutional kernels, respectively, an operation that effectively reduces the number of trainable parameters while keeping the sensory fields of the feature maps unchanged. The structure of the improved MBFEM module is shown in Fig. 4.

3.2 RIRDFM module

RT-DETR only performs self-attention operation on the S5 feature map in its neck without any processing on S3 and S4 feature maps, and then inputs S3, S4, and the processed S5 feature map (afterward referred to as P5) into the CCFM module for the fusion process, in which to enhance the comprehensive utilization of the feature map information and to improve the detection accuracy of small objects. This design greatly reduces the computational cost of the RT-DETR encoder and improves the overall detection accuracy, but in this paper, we believe that there is still room for further improvement in this design. We believe that there is already some information refinement difference between the S3, S4, and S5 feature maps generated by RT-DETR's backbone network, that is, from S3 to S5 feature maps, the semantic information becomes richer and richer, but the positional information gradually decreases, using the PAN (Path Aggregation Network) structure to fuse the S3, S4, and S5 feature maps can synthesize their positional and semantic information to achieve the purpose of improving the detector and enhancing the detection accuracy. However, RT-DETR has only processed the S5 feature map with self-attention, which makes the difference in the information hierarchy between S5 and S3 and S4 feature maps further widen, and the direct fusion of the refined P5 feature map and the unrefined S3 and S4 feature maps will negatively affect the final detection accuracy.

To address the challenges of feature refinement and fusion in RT-DETR, we design the RIRDFM (Reduced Information Refinement Difference and Fusion Module), illustrated in Fig. 5. Unlike conventional fusion approaches, RIRDFM reduces the refinement differences among multi-scale feature maps before fusion, while preserving critical spatial and semantic details. The module receives up-sampled

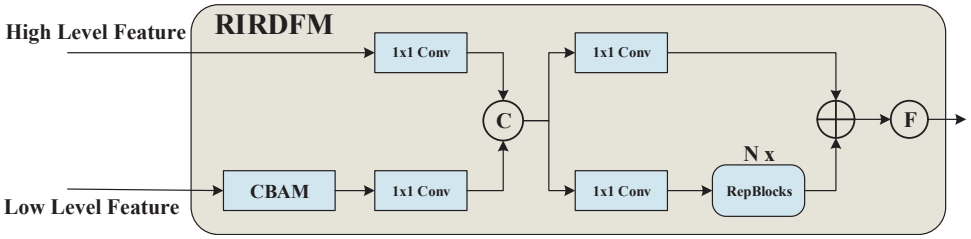


Figure 5: RIRDFM Network Architecture.

high-level feature maps and low-level feature maps for refinement. First, the CBAM module processes the low-level features to selectively emphasize informative channels and spatial regions, which are then connected to the high-level features. The resulting feature maps are processed through two complementary paths: one using a 1×1 convolution layer and the other using a multi-layer stacked RepBlocks module. This dual-path design allows richer and more diverse feature extraction, capturing subtle variations often missed by traditional single-path fusion. The outputs from both paths are then summed, producing refined feature maps. Finally, all RIRDFM outputs are flattened into 2D and concatenated to form the final memory feature, which is fed into the IOU-aware query module for object query initialization. The specific process is shown in Equation 1.

$$\begin{aligned} \mathbf{C} &= Concat\left(\begin{array}{cc} Conv\left(\mathbf{A}\right) & SE\left(Conv\left(\mathbf{B}\right)\right) \end{array} \right) \\ \mathbf{D} &= Flatten\left(Conv\left(\mathbf{C}\right) \oplus RepBlocks(Conv\left(\mathbf{C}\right))^{N}\right) \end{aligned} \tag{1}$$

"RepBlocks" module is consistent with RT-DETR.The purpose of the RIRDFM module is to effectively reduce the differences in information refinement between feature maps of different scales, and then fuse them together to obtain a fused feature map with clearer and richer semantic and location information. This helps to mitigate the negative impact of significant information refinement differences between different feature maps on the final detection accuracy.

3.3 MP-DETR

The overall architecture and implementation of MP-DETR proposed in this paper are shown in Fig. 6, and the data flow and feature map scales through each module are further detailed in Fig. 7. The network consists of a backbone, a neck (an encoder and a RIRDFM module), an IOU-aware query selection module, six decoders, and their corresponding detection heads. The backbone uses an MBFANet, which is responsible for performing multidimensional feature extraction on the raw images fed into the network and sending the S3, S4, and S5 feature maps output from the second, third, and fourth stages to the neck for processing. The neck consists of an encoder and an RFM (Refinement and Fusion feature module) architecture, the latter consists of a PAN structure, multiple upsampling and downsampling processes, and multiple RIRDFM modules. In this paper, only the S5 feature layer is fed into the encoder for self-attention to refine its global features. Considering that the S3 and S4 feature layers are directly output from the backbone, there is a large difference in terms of the degree of information refinement compared to the P5 feature layer, and their direct fusion through the PAN will adversely affect the final results. Therefore, this paper uses the RIRDFM module to process the S3, S4, and P5 feature layers before fusion, and the RIRDFM module can effectively reduce the difference in the degree of information refinement between these feature layers so that the fused feature maps can be utilized more efficiently. The IOU-aware Query Selection Module can map the outputs of the Neck to the predicted results, and this paper selects the highest-scoring tokens according to the category probability scores in the predicted results to initialize the object queries. Specifically, these top 300 tokens with the highest scoring are directly treated as content queries, then their corresponding detection boxes are encoded as location queries in a cosine fashion, and finally, the content queries and their corresponding location queries are summed to obtain the final 300 object queries. This implementation makes the information in the object queries come directly from the feature map, which is closely related to the information of interest in the image, greatly reducing

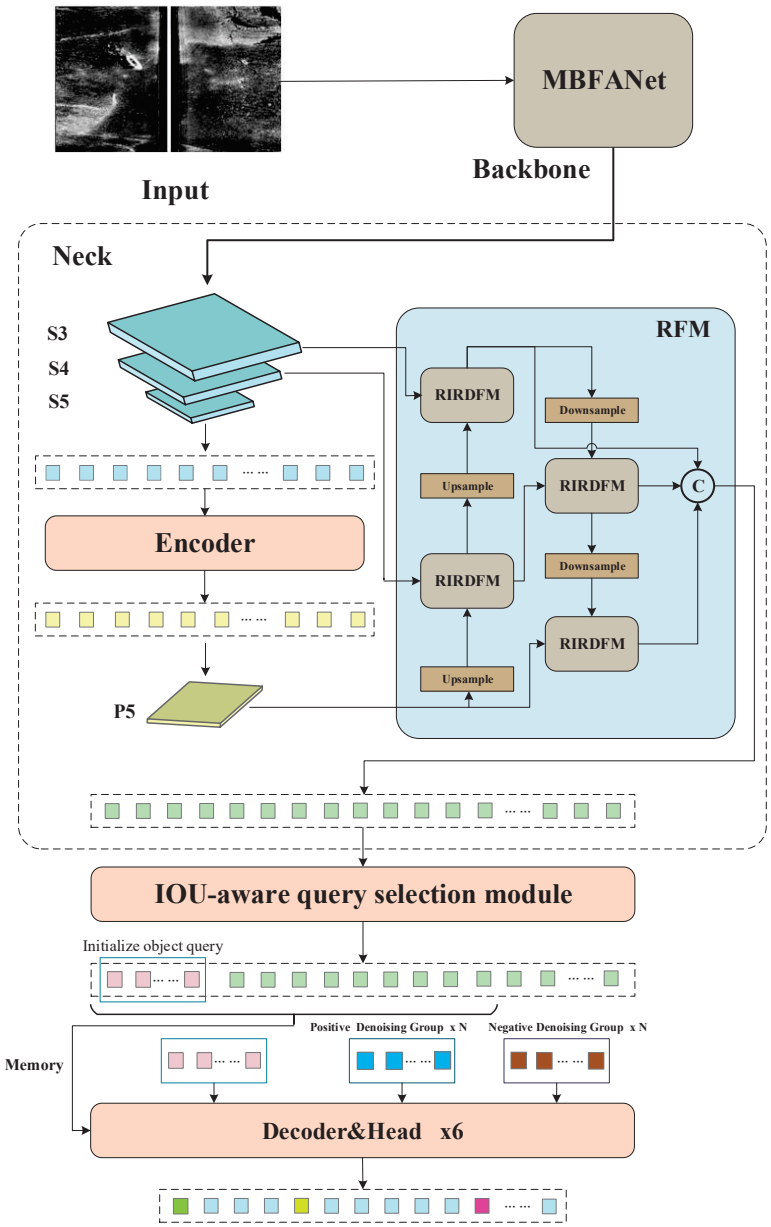


Figure 6: MP-DETR Network Architecture Diagram.

the difficulty of object query optimization and improving the convergence speed of the model. In each decoder, the input object queries are first processed by self-attention and then interact with the output memory of the neck through cross-attention to realize information interaction, from which the key information in the feature map is further extracted. The object queries output from the decoder will be mapped by its corresponding detection head to get the final detection results. In the training mode, the detection result and the ground truth are computed by bipartite graph matching for loss and the model is trained by gradient back propagation. The loss function is calculated as follows

$$\begin{aligned} \mathcal{L}(\hat{y}, y) &= \mathcal{L}_{box}(\hat{b}, b) + \mathcal{L}_{cls}(\hat{c}, \hat{b}, y, b) \\ &= \mathcal{L}_{box}(\hat{b}, b) + \mathcal{L}_{cls}(\hat{c}, c, IoU) \end{aligned} \tag{2}$$

where $\hat{y}$ and y denote detection result and ground truth, $\hat{y} = \left\{\hat{c}, \hat{b}\right\}$ and $y = \{c, b\}$, c and b represent categories and bounding box, respectively.

The training strategy of MP-DETR adopts a transfer learning approach. First, the network is fully pre-trained using the MS COCO 2017 dataset, which consists of more than 118k training images and 5k validation images, and the pre-training allows the network to obtain strong feature extraction and processing capabilities. During the training process, a positive and negative sample denoising

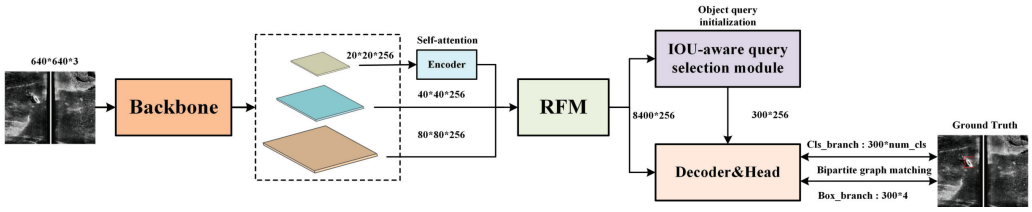


Figure 7: Data flow in MP-DETR Network.

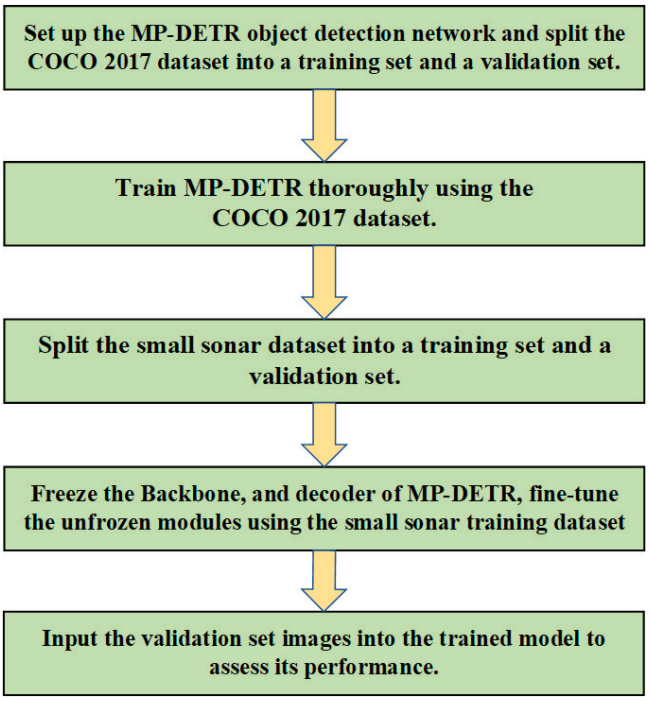


Figure 8: The training process of MP-DETR.

group is introduced into the decoder to optimize the bipartite graph matching and improve the model convergence speed. Then, the network parameters of the backbone, neck, encoder and decoders of the pre-trained model are frozen to retain its powerful feature extraction and refinement capabilities. Finally, a custom underwater sonar image dataset is used to fine-tune the parameters of the IOU -aware query selection module and the detection head for each decoder. The specific implementation process is shown in Fig. 8.

4 Experiment

4.1 Setups

The MP-DETR model is first fully pretrained on the MS COCO 2017 dataset, then the network parameters of the backbone, neck, encoder, and decoders are frozen, and finally, the unfrozen modules are formally trained using a custom underwater sonar image dataset. The input images are resized to 640x640 for both pretraining and formal training.The proposed MP-DETR model is compared with the original RT-DETR [46] and the improved YOLOv5 [44] using AP_{50} in the standard COCO AP evaluation metric and FPS as the detection accuracy metrics and real-time performance metrics, respectively. The RT-DETR uses the HgNetv2 backbone, and it with MP-DETR was pretrained for 72 epochs on the COCO 2017 dataset, and the improved YOLOv5 network was pretrained for 300 epochs using the "L" model. In formal training after partially freezing the network modules, the RT-DETR and MP-DETR were trained for 24 epochs each, and the improved YOLOv5 was trained for 100 epochs. The network models were built based on Baidu's paddle platform, and all the training processes were carried out on two Nvidia GTX 3080 Ti GPUs with a computer operating system of

Ubuntu 20.04.

4.1.1 Evaluation metrics

Due to the significantly higher proportion of small objects in sonar images, the commonly used performance evaluation metric map (IOU=0.5-0.95) for object detection in visible images is no longer applicable. Therefore, in this paper, the map (IOU=0.5) is used as the performance metric for evaluating detector accuracy. The mAP integrally considers the misdetection and omission of the detector, and the calculation of mAP needs to calculate P which reflects misdetection and R reflecting omission, and then calculate the mAP according to P and R, where P and R are obtained by the following way

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

where TP represents the number of correctly predicted positive samples by the detector, FP denotes the number of falsely predicted positive samples, and FN indicates the number of falsely predicted negative samples. Under IoU=0.5, the Precision (P) and Recall (R) for all object categories in the dataset are computed at various confidence thresholds, yielding sets of P-R curves corresponding to different object categories. Projecting these curves onto a two-dimensional Cartesian coordinate system, the area under each class's P-R curve is defined as the Average Precision (AP) value for that class. The formula for calculation is as follows:

$$AP = \frac{1}{m} \sum_{i=1}^{m} (R_i - R_{i-1}) P_i \tag{5}$$

where M is the number of object categories, and map is the mean Average Precision (AP) corresponding to all the categories.In addition, the FPS during the detector inference process is also adopted as a real-time performance metric.

4.1.2 Implementation details

The AdamW optimizer was used for training the object detector with the following hyperparameters: base_learning_rate = 0.0001, weight_decay = 0.0001, global_gradient_clip_norm = 0.0001, and linear_warmup_steps = 2000. Additionally, the Exponential Moving Average (EMA) method was employed with an ema_decay of 0.999. Data augmentation techniques including color distortion, image expansion, cropping, flipping, resizing, etc. are used.

The small sonar image dataset used in this paper is a custom dataset comprising 5000 images. It contains three classes of objects: sea mines, shipwrecks, and crashed airplanes. Among these, there are 3220 images containing sea mines, 2860 images containing crashed airplanes, and 2230 images containing shipwrecks. The majority of the images are of small objects, constituting 92.5% of the entire dataset. Typical examples of these images are illustrated in Figure 9. Additionally, the parameters of the sidescan sonar is shown in Table. 1.

Table 1: Parameters of the sidescan sonar.

Parameters	values
Operating Frequencye	450kHz
Angular Resolution	0.2°
Number of Beams	5
Vertical Beamwidth	45°%
Scan Rate	4
Range Resolution	2.5cm

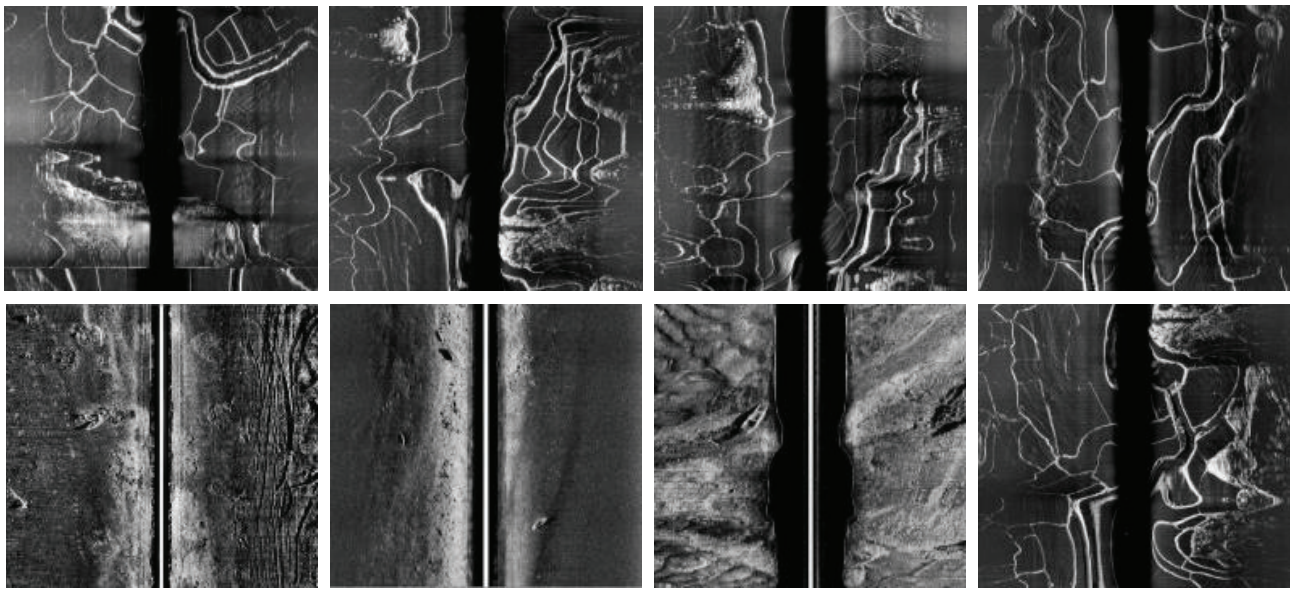


Figure 9: Partial image examples from the custom dataset.

4.2 Comparison with others

Table 2 presents a comparison of the parameters and experimental results of the proposed MP-DETR detector, the baseline RT-DETR detector, and the improved YOLOv5 detector on the custom sonar image dataset. MP-DETR achieves an AP of 98.5 and a frame rate of 102 FPS, outperforming the improved YOLOv5 by 2.5 AP and 20 FPS. Additionally, MP-DETR has a smaller parameter size, making it more suitable for deployment in sonar-based underwater detection tasks, particularly on resource-constrained platforms. While MP-DETR exhibits similar parameter size and inference speed to the baseline RT-DETR, it outperforms RT-DETR by 3.2 AP in detection accuracy. These experimental results highlight the effectiveness of the MP-DETR detector and the proposed training strategy. Further validation of the improved network modules and strategies will be provided in the ablation experiments.

Figure 10 presents a qualitative comparison of different detectors on representative sonar images. It can be observed that the baseline detector, without specific adaptations for sonar image characteristics, performs poorly and often fails to distinguish between objects and their corresponding shadows. The improved YOLOv5 detector achieves better results than the baseline, yet its localization accuracy remains inferior to our proposed method. In contrast, the proposed MP-DETR demonstrates the best visual performance, effectively identifying objecs while providing higher localization precision and clearer contour delineation. These results further validate the effectiveness of the proposed detector and training strategy for sonar image object detection

Table 2: Comparison with other detectors.

Detector	Backbone	Params (M)	$AP_{50}(\%)$	FPS
MP-DETR (proposed)	CSPDarknet-53	31	**98.5**	102
RT-DETR [46]	HgNetv2	32	95.3	103
Improved YOLOv5 [44]	MBFANet	47	96	82

4.3 Ablation Studies

In order to validate the effectiveness of the proposed MBFANet backbone, RIRDFM module, and the designed transfer learning training strategy, we also conducted an ablation study to evaluate the performance impact of the improvements proposed in this paper on MP-DETR, and the experimental results are shown in Table 3. Through the results, it can be found that the MBFANet and RIRDFM modules bring 1.5AP and 1.9AP improvements to the baseline model, respectively, and their combined

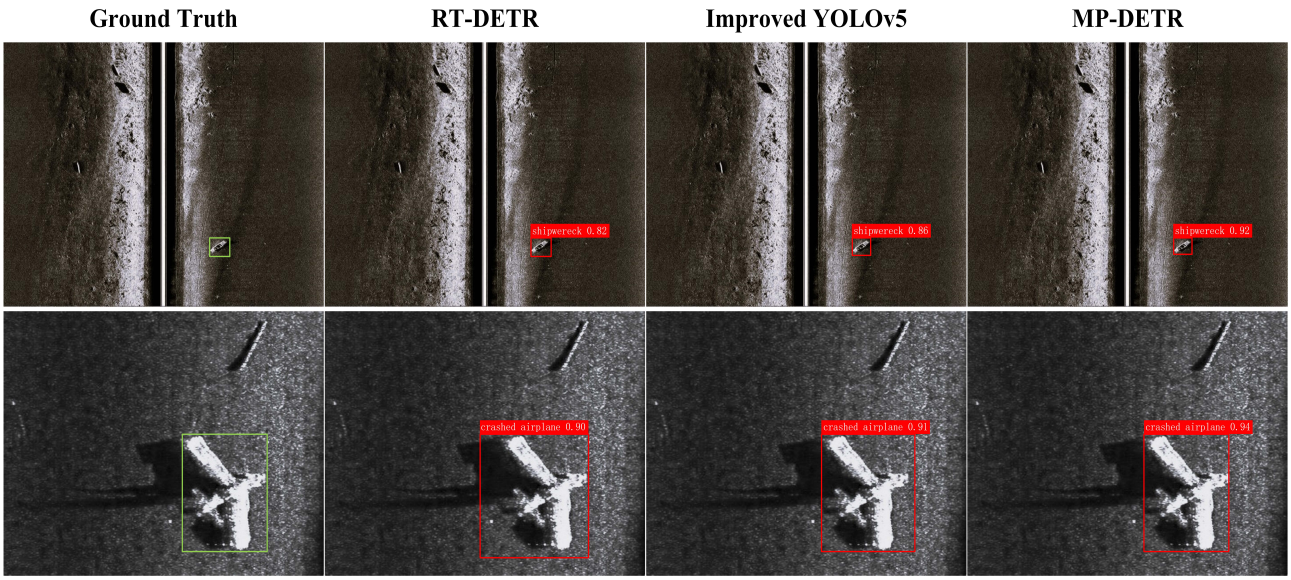


Figure 10: Visual comparison of detection results among detectors

use brings even greater accuracy improvement, which further proves the effectiveness of the network modules proposed in this paper. In addition, Table 3 also gives the results of MP-DETR trained and validated directly on a custom small sonar dataset without pre-training, which only obtains 56.2 AP, illustrating that pre-training on a large dataset enables the network to obtain stronger feature extraction and representation capabilities.

Freezing different modules of the pretrained network also has different effects on the final results of the detector during the transfer learning process. Table 4 shows the results obtained by freezing different network modules after the pretraining is completed and then the second training is performed, when freezing the backbone network, the neck module, the encoder module and the decoder module the network is able to obtain the best detection performance. It can be noticed that when obtaining the best results, the role of the frozen network modules is mainly all about extracting or refining the features, and these modules are insensitive to the object categories, whereas the IOU-aware query selection module is used for result mapping and selecting the high scoring tokens, which is similar to the detection head in terms of its structure and functionality. In addition, since the number of its network parameters is negligible compared to the freezing module, the use of a custom small sonar image dataset in secondary training can train it sufficiently well to adapt it quickly to the object features of sonar images.

Table 3: Comparison of results from ablation experiments.

Detector	$AP_{50}(\%)$
Baseline	95.3
Baseline + MBFANet	96.8
Baseline + RIRDFM module	97.2
MP-DETR (without transfer learning)	56.2

5 Conclusion

In this paper, a transformer-based detector - MP-DETR, which is proposed for detecting small objects in underwater sonar images, which is based on the baseline detector, RT-DETR, to solve the challenge of detecting a large number of deformable and ambiguous small objects in sonar images. The study analyzes the effectiveness and limitations of the Neck of RT-DETR, and while retaining its efficient design, it points out the negative impact of the difference in information refinement between the P5, S3 and S4 feature maps on the detection results. To overcome this problem, we designed the

Table 4: Comparison of results with different frozen modules.

Frozen modules	$AP_{50}(\%)$
Backbone	94.4
Backbone + RFM	94.9
Backbone + RFM + encoder	96.2
Backbone + RFM + decoder	96.5
Backbone + RFM + encoder+decoder	**98.5**
Backbone + RFM + IOU query-aware module	91.6
Backbone + RFM + encoder + decoder + IOU query-aware module	93.3

RIRDFM module, which can reduce the information refinement differences between P5 and S3, S4 feature maps before fusion. We also design an efficient backbone, MBFANet, which can efficiently and multidimensionally extract the information of the original image, provide high-quality feature information for the subsequent transformer structure, and effectively improve the detector's small object detection accuracy. In addition, for the problem of sample scarcity in the underwater sonar image dataset, we also design an efficient transfer learning approach for MP-DETR. First, MP-DETR is fully pretrained on the MS COCO 2017 visible light dataset, then the specific network modules responsible for feature extraction and refinement are frozen, and finally, the unfrozen modules are fully trained using a small sonar image dataset. Extensive experiments demonstrate that the MP-DETR proposed in this paper can achieve higher detection accuracy and real-time performance in the underwater sonar image detection task compared to the existing methods.

However, some limitations remain. The performance of MP-DETR may vary under different underwater conditions, such as changes in water turbidity, sonar sensor characteristics, or object types not present in the training dataset. Future research could focus on developing real-time adaptation mechanisms to handle dynamic underwater environments, integrating advanced sonar image enhancement techniques, or leveraging semi-supervised and self-supervised learning to improve small-object detection with limited labeled data. Addressing these directions would further enhance the robustness and practical applicability of MP-DETR in diverse underwater scenarios.

Funding

This work was supported by the National Science Foundation of China (62103204).

Author contributions

The authors contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Sixian Cai, Guocheng Li, and Yuan Shan. Underwater object detection using collaborative weakly supervision. *Computers and Electrical Engineering*, 102:108159, 2022.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] Linhui Dai, Hong Liu, Hao Tang, Zhiwei Wu, and Pinhao Song. Ao2-detr: Arbitrary-oriented object detection transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(5):2342–2356, 2022.

[4] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2988–2997, 2021.

[5] Xinnan Fan, Liang Lu, Pengfei Shi, and Xuewu Zhang. A novel sonar target detection and classification algorithm. *Multimedia Tools and Applications*, 81(7):10091–10106, 2022.

[6] Sheezan Fayaz, Shabir A Parah, and GJ Qureshi. Underwater object detection: architectures and algorithms–a comprehensive review. *Multimedia Tools and Applications*, 81(15):20871–20916, 2022.

[7] Z Ge. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[8] Stanisław Hożyń. A review of underwater mine detection and classification in sonar imagery. *Electronics*, 10(23):2943, 2021.

[9] Guanying Huo, Ziyin Wu, and Jiabiao Li. Underwater object classification in sidescan sonar images using deep transfer learning and semisynthetic training data. *IEEE access*, 8:47407–47418, 2020.

[10] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, Yonghye Kwon, Kalen Michael, Jiacong Fang, Zeng Yifu, Colin Wong, Diego Montes, et al. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo*, 2022.

[11] Divas Karimanzira, Helge Renkewitz, David Shea, and Jan Albiez. Object detection in sonar images. *Electronics*, 9(7):1180, 2020.

[12] Jaskirat Kaur and Williamjeet Singh. Tools, techniques, datasets and application areas for object detection in an image: a review. *Multimedia Tools and Applications*, 81(27):38297–38351, 2022.

[13] Hoang Thanh Le, Son Lam Phung, Philip B Chapple, Abdesselam Bouzerdoum, Christian H Ritz, et al. Deep gabor neural network for automatic detection of mine-like objects in sonar imagery. *IEEE Access*, 8:94126–94139, 2020.

[14] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022.

[15] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13619–13627, 2022.

[16] Yujie Li, Shinya Takahashi, and Seiichi Serikawa. Cognitive ocean of things: a comprehensive review and future trends. *Wireless Networks*, pages 1–10, 2022.

[17] Kun Liu, Lei Peng, and Shanran Tang. Underwater object detection using tc-yolo with attention mechanisms. *Sensors*, 23(5):2567, 2023.

[18] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.

[19] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021.

[20] Huimin Lu, Dong Wang, Yujie Li, Jianru Li, Xin Li, Hyoungseop Kim, Seiichi Serikawa, and Iztok Humar. Conet: A cognitive ocean network. *IEEE Wireless Communications*, 26(3):90–96, 2019.

[21] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3651–3660, 2021.

[22] Ali Tariq Nagi, Mazhar Javed Awan, Mazin Abed Mohammed, Amena Mahmoud, Arnab Majumdar, and Orawit Thinnukool. Performance analysis for covid-19 diagnosis using custom and state-of-the-art deep learning models. *Applied Sciences*, 12(13):6364, 2022.

[23] Duong Nguyen, Matthieu Simonin, Guillaume Hajduch, Rodolphe Vadaine, Cédric Tedeschi, and Ronan Fablet. Detection of abnormal vessel behaviours from ais data using geotracknet: from the laboratory to the ocean. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pages 264–268. IEEE, 2020.

[24] Son Lam Phung, Thi Nhat Anh Nguyen, Hoang Thanh Le, Philip B Chapple, Christian H Ritz, Abdesselam Bouzerdoum, and Le Chung Tran. Mine-like object sensing in sonar imagery with a compact deep learning architecture for scarce data. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE, 2019.

[25] Joseph Redmon. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[26] Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse detr: Efficient end-to-end object detection with learnable sparsity. *arXiv preprint arXiv:2111.14330*, 2021.

[27] Akshita Saini and Mantosh Biswas. Object detection in underwater image by detecting edges using adaptive thresholding. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 628–632. IEEE, 2019.

[28] Steve Schmit. *Cut and Capture System Technology for Demilitarization of Underwater Munitions*. IEEE, 2022.

[29] Pengfei Shi, Huanru Sun, Xinnan Fan, Qi He, Xuan Zhou, and Liang Lu. An effective automatic object detection algorithm for continuous sonar image sequences. *Multimedia Tools and Applications*, 83(4):10233–10246, 2024.

[30] Pinhao Song, Pengteng Li, Linhui Dai, Tao Wang, and Zhan Chen. Boosting r-cnn: Reweighting r-cnn samples by rpn's error for underwater object detection. *Neurocomputing*, 530:150–164, 2023.

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[32] Hesham Tarek, Hesham Aly, Saleh Eisa, and Mohamed Abul-Soud. Optimized deep learning algorithms for tomato leaf disease detection with hardware deployment. *Electronics*, 11(1):140, 2022.

[33] Yi Wang, Syed Muhammad Arsalan Bashir, Mahrukh Khan, Qudrat Ullah, Rui Wang, Yilin Song, Zhe Guo, and Yilong Niu. Remote sensing image super-resolution and object detection: Benchmark and state of the art. *Expert Systems with Applications*, 197:116793, 2022.

[34] Yingchun Wang, Jingyi Wang, Weizhan Zhang, Yufeng Zhan, Song Guo, Qinghua Zheng, and Xuanyu Wang. A survey on deploying mobile deep learning applications: A systemic and technical perspective. *Digital Communications and Networks*, 8(1):1–17, 2022.

[35] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2567–2575, 2022.

[36] Zhen Wang, Jianxin Guo, Leya Zeng, Chuanlei Zhang, and Buhong Wang. Mlffnet: Multilevel feature fusion network for object detection in sonar images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022.

[37] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3060–3069, 2021.

[38] Shubo Xu, Minghua Zhang, Wei Song, Haibin Mei, Qi He, and Antonio Liotta. A systematic review and analysis of deep learning-based underwater object detection. *Neurocomputing*, 527: 204–232, 2023.

[39] Chao Yang, Yongpeng Li, Longyu Jiang, and Jianxing Huang. Foreground enhancement network for object detection in sonar images. *Machine Vision and Applications*, 34(4):56, 2023.

[40] Honghui Yang, Zili Liu, Xiaopei Wu, Wenxiao Wang, Wei Qian, Xiaofei He, and Deng Cai. Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph. In *European Conference on Computer Vision*, pages 662–679. Springer, 2022.

[41] Li Yuanzi, Ye Xiufen, and Zhang Weizheng. Transyolo: high-performance object detector for forward looking sonar images. *IEEE Signal Processing Letters*, 29:2098–2102, 2022.

[42] Tang Yulin, Shaohua Jin, Gang Bian, and Yonghou Zhang. Shipwreck target recognition in side-scan sonar images by improved yolov3 model based on transfer learning. *IEEE Access*, 8: 173450–173460, 2020.

[43] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[44] Haoting Zhang, Mei Tian, Gaoping Shao, Juan Cheng, and Jingjing Liu. Target detection of forward-looking sonar image based on improved yolov5. *IEEE Access*, 10:18023–18034, 2022.

[45] Lanyong Zhang, Chengyu Li, and Hongfang Sun. Object detection/tracking toward underwater photographs by remotely operated vehicles (rovs). *Future Generation Computer Systems*, 126: 163–168, 2022.

[46] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2024.

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[48] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6748–6758, 2023.

Journal's webpage: http://univagora.ro/jour/index.php/ijccc/

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:

Wu, X.; Zhang, X.; Tan, P. (2026). A Few-shot Learning Algorithm for Underwater Small Object Detection based on the Transformer Architecture, *International Journal of Computers Communications & Control*, 21(4), 6915, 2026.

https://doi.org/10.15837/ijccc.2026.4.6915