2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

QO-Net: Query Optimization Underwater Object Detection Network

Jiandong Tian, Hongyang Sun, Baojie Fan* and Hongxin Xu

***Abstract*— Underwater object detection has attracted increasing interest for its wide application in various underwater tasks. However, due to underwater image quality degradation and the lack of large-scale underwater object datasets, many underwater detectors suffer from low detection performance. To address the issues, we not only propose a novel underwater transformer detector with multi-scale feature enhancement and query optimization, named QO-Net, but also construct a new underwater object detection dataset, called UODD. Specifically, a Conv-Trans Layer is developed as the unit of QO-Net, which effectively learns multi-scale image feature representation through CNN and simultaneously captures the dependencies among different positions in the sequence data through Transformer, enabling QO-Net to process underwater image sequence information over longer distances. An effective combination can enhance the representation of multi-scale features. Then, QO-Net develops a positional query enhancement strategy to optimize the spatial prior of positional queries, thereby speeding up the convergence of the network training. In addition, UODD also contains more than 20,000 underwater images for training and validation, with a variety of rich underwater categories. Extensive experiments on UODD, Brackish, and TrashCan datasets demonstrate that QO-Net presents favorable detection performance against state-of-the-art methods in terms of robustness and accuracy.**

I. INTRODUCTION

Recent years have witnessed the extensive development of underwater object detection technology, which is widely applied in naval coastal defense, fishery and aquaculture, underwater robot guidance, and so on. However, the complex underwater imaging environment is distance and wavelength dependent and leads to quality degradations such as color and texture distortion, blurred and hazed images, uneven illumination, and low contrast visibility. In addition, there are only a limited number of available underwater object image datasets with incomplete labels. Therefore, how to accurately and robustly locate and classify marine creatures with poor image visibility is a remarkable and challenging task.

In order to resolve these problems, some researchers directly introduce generic object detection frameworks into underwater scenes. Currently, mainstream object detection methods are divided into two categories: two-stage detectors [1], [2] and one-stage detectors [3]–[5]. As we all know, two-stage detectors have higher detection accuracy, while one-stage detectors have faster inference speed. However, directly applying general object detection algorithms to special underwater scenes performs poorly. The primary distinction between underwater object detection and generic object detection is the nature of the various detection contexts. Poor vision, light refraction, absorption, and scattering [6] invariably affect underwater images. This is mainly attributed to the following points: (1) Objects in underwater scenes are of different sizes, including many smaller targets; (2) The effects of low contrast and color distortion of underwater images lead to differences in the color of objects under different lighting; (3) The color and shape of aquatic organisms are similar to the complex underwater environment, making it difficult to distinguish whether they are objects.

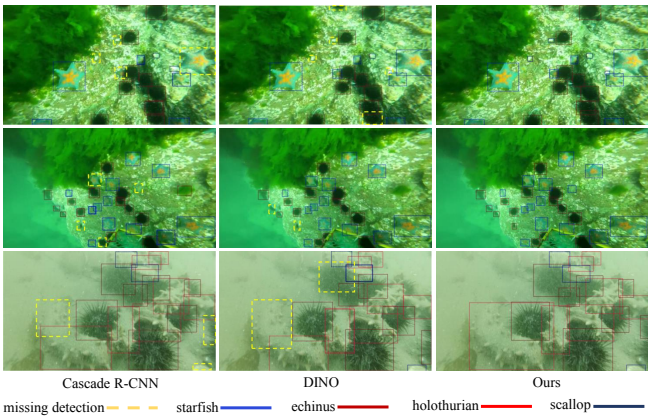


Fig. 1. Comparison of our algorithm with other object detection algorithms for underwater image detection. The left part is the detection result of Cascade R-CNN, the middle part is the baseline detection result, and the right part is the detection result of our algorithm. The yellow dotted line in the figure indicates that the object has not been detected.

With the development of underwater detection algorithms, underwater object detection methods have been significantly improved today. For example, in order to address the issue of sample imbalance, Fan et al. [7] suggested an underwater detection framework with feature enhancement and anchor refinement. Dai et al. [8] proposed a gated cross-domain collaboration network that simultaneously acquires original images and enhanced images in parallel and then interacts with and fuses information from both domains within a single framework. Using hard example mining and uncertainty modeling, Song et al. [9] presented Boostig R-CNN, a two-stage underwater detector, to enhance underwater tar-

*This work is supported by the National Natural Science Foundation of China (No. U2013210, 62103388), and the young and middle-aged leading scholar in Qinglan Project by Jiangsu Province

*corresponding authors. Baojie Fan is with the College of Automation and College of Artificial Intelligence, Nanjing University of Posts and Telecommunications jobfbj@gmail.com

Jiandong Tian is with State Key Laboratory of Robotics, Shenyang Institute of Automation Chinese Academy of Sciences tianjd@sia.cn

Hongyang Sun is with College of Automation and College of ArtificialIntelligence, Nanjing University of Posts and Telecommunications shy886366@163.com

Hongxin Xu is with Delft University of Technology, Landbergstraat 15 2628 CE Delft H.Xu-14@student.tudelft.nl

get identification performance. Enhancing underwater image quality has been the subject of several works [10], [11], all of which show that the images obtained by underwater image improvement are useful for later object detection tasks. However, as several studies [12], [13] have noted, relying solely on improved photos might not improve detection efficacy and might even result in a significant decrease in performance. This is due to the fact that algorithms interpret scenes differently than human eyes, even though these techniques for enhancing underwater images produce greater visual restoration in human perception. The comprehension of an underwater image is not always improved by merely adjusting its color and contrast. Additionally, it is certain that the enhancement process may eliminate or modify important patterns and details, which could cause noise or artifacts to appear in the image [12]. Consequently, the association between target identification and underwater visual enhancement was investigated by [12], [13]. Although existing underwater object detection algorithms have greatly improved their performance, they still perform poorly when faced with small and camouflaged underwater creatures, as well as in low-contrast underwater environments.

Different from existing generic object detection algorithms, to address problems such as blur, low contrast, and texture distortion in underwater images, we propose an underwater object detection framework with feature enhancement and query optimization. It achieves favorable performance, as shown in Figure 1. We will make improvements in the model architecture and spatial location of queries. Regarding the model architecture, we develop the Conv-Trans Layer structure, which effectively combines CNN and Transformer. This fusion enhances the representation ability of the multi-scale features of the network. Additionally, we address the spatial query location aspect by introducing the Positional Queries Enhancement Strategy. This strategy not only expedites the model's convergence but also offers precise spatial priors for positional queries. Finally, we also construct a new underwater detection dataset called UODD.

In summary, our contributions are summarized as follows:

- We propose an underwater object detection framework, QO-Net, and construct a brand-new Underwater Object Detection Dataset, UODD.
- To deal with blurring and texture distortion in underwater images, we propose a Conv-Trans Layer structure to enhance the representation ability of multi-scale features.
- We propose the Positional Queries Enhancement Strategy, which provides precise spatial priors for positional queries, thereby speeding up the convergence of the network.

II. Related Work

A. Object Detection

CNN-based detectors can be divided into two categories: two-stage and one-stage methods. In two-stage detectors such as Faster R-CNN [1] and Cascade R-CNN [2], these networks first generate some candidate regions through Region Proposal Networks (RPN) in the first phase and then refine the generated candidate regions in the second phase. In contrast, one-stage detectors SSD [4], YOLO [14], FCOS [15], and RetinaNet [3] do not generate candidate boxes from RPN but directly predict the offset of the ground-truth box relative to the anchor box.

Transformer-based detectors, such as DETR [16], DAB-DETR [17], and DINO [18], eliminates the need for anchor design and additional post-processing stages and performs end-to-end object detection. Many subsequent papers have considered the problem of decoder cross-attention, making DETR training converge slowly. In order to speed up the convergence speed of DETR, it has been improved in many aspects. For example, Dai et al. [19] proposed a dynamic decoder that focuses on regions of interest by combining multiple attentions between scale-aware feature layers, spatially-aware spatial locations, and within task-aware output channels. Another direction to improve DETR is to understand more deeply the role of queries in the decoder in DETR. For instance, DAB-DETR [17] directly uses dynamically updated anchor boxes to provide both a reference query point (x, y) and a reference anchor size (w, h) to improve the cross-attention computation. Recently, DINO [18] outperformed the DETR-like model in terms of performance and efficiency by using a contrast denoising training method, a hybrid query selection method with anchor initialization, and a look forward twice scheme with box prediction. However, state-of-the-art object detection methods are not suitable for specific underwater object detection because aquatic organisms are usually blurred with the background and often hidden around aquatic plants or rocks. Therefore, solving underwater object detection requires more knowledge of visual perception and attention.

B. Underwater Object Detection

Underwater object detection [12], [20] aims to localize and identify objects in underwater scenes. It has attracted continuous attention due to its wide application in fields such as oceanography and underwater navigation. However, in underwater scenes, the quality of underwater images is greatly degraded by the effects of light, which leads to problems such as low visibility, weak contrast, texture distortion and color variations, and the complexity of the environment in underwater scenes, which makes underwater object detection more difficult. Due to the scarcity of currently available underwater datasets, some works [21]–[23] use data augmentation to try to increase the diversity of the data. For example, Lin et al. [21] proposed an augmentation method called ROIMIX for proposal-level fusion among many images. For instance, Chen et al. [24] proposed SWIPENET, which is made up of hyper-feature maps with high resolution and extensive semantic content. In order to model the object prior probability, Song et al. [9] proposed a two-stage detector called Boosting R-CNN, which can produce high-quality suggestions and take objectness and IoU prediction into account for uncertainty.

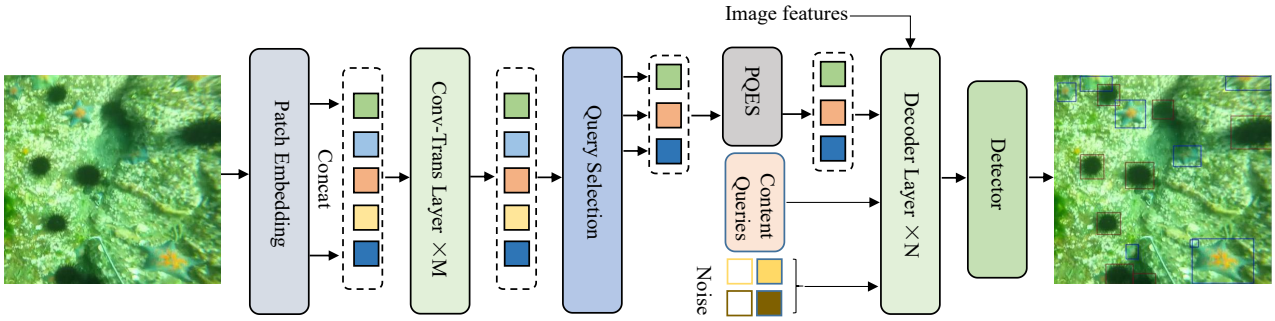


Fig. 2. Overall architecture of our framework. Our improvements are mainly in the Transformer encoder and decoder queries. In the encoder, we developed the Conv-Trans Layer as the basic building block. The positional query enhancement strategy in our decoder optimizes the spatial prior of the positional query.

Due to the challenges inherent in the underwater environment, Such as low contrast and various lighting conditions and changes. In terms of water quality, image enhancement is necessary. Existing methods [12] attempt to use underwater image enhancement as a preprocessing step and perform object detection on the enhanced images. However, the detection performance is unsatisfactory and even leads to a service performance drop. Since the algorithms perceive the scene differently from human eyes, the enhanced image may lose some important details and potential value that are beneficial for object detection. Liu et al. [13] address the problem by considering it a multi-task optimization problem. Others take cross-domain collaboration as their starting point. For example, Dai et al. [8] proposed a gated cross-domain collaboration network that simultaneously acquires original images and enhanced images in parallel and then interacts with and fuses information from both domains within a single framework.

However, the performance of these methods is still limited due to the insufficient extraction of basic features. Different from the above studies, we aim to design a more powerful network for extracting feature representation and optimize spatial priors for decoder position queries.

III. Method

A. Overview

The method proposed is an end-to-end object detector similar to DETR, which is based on DINO [18] improvements. The architecture consists of a CNN backbone, a multi-layer Conv-Trans Layer, a multi-layer Transformer Decoder layer and multiple prediction heads (see Figure 2). Given an underwater image, we use the backbone network to extract the multi-scale features of the image, and then feed the obtained multi-scale features and corresponding positional embeddings into the Conv-Trans Layer. Conv-Trans Layer can effectively enhance the feature representation ability to refine CNN features and deal with underwater image blurring. After feature enhancement, we send it to a mixed query selection strategy to obtain a preliminary representation of the decoder positional queries, and then pass through the Positional Queries Enhancement Strategy to obtain a more accurate spatial prior. For the content queries section, keep it the same as the content queries in DINO. On the decoder side, the positional queries and content queries are input into the decoder, and the object is searched by cross-attention. The final decoder layer output predicts objects with boxes and labels through the prediction head, and then have a contrast denoising branch like DINO to perform denoising training.

B. Conv-Trans Layer

To deal with problems such as blurred underwater images, low contrast, and texture distortion, we propose a powerful Conv-Trans Layer structure, as shown in Figure 3. This structure can extract key feature information and enhance the recognition ability of the classifier. Specifically, our Conv-Trans Layer consists of two key components: Conv Block and Transformer Block. Among them, Conv Block is composed of convolutional neural networks. Because of their inherent advantages, they are naturally suitable for a variety of computer vision tasks. For example, translation equivariance, which introduces inductive bias to CNNs, allowing them to adapt to different sizes of input images. Since convolution operations in CNN can only capture local information, long-distance connections of global images cannot be established. The Conv-Trans Layer also includes Transformer Block, which enables input adaptive, long-range dependencies designed to extract a global understanding of the visual scene, and high-order spatial interactions through multi-scale transformable attention operations, as well as their competing modeling capabilities. The combination of the two can effectively enhance the feature expression ability of the detector. Given an input feature map $Xin \in \mathbb{R}^{C\times H\times W}$, first input into the Conv Block to obtain the local feature representation $Y \in \mathbb{R}^{C\times H\times W}$, and then input the result into the Transformer Block to obtain the global feature representation $Z \in \mathbb{R}^{N\times C}$(N is the result of multi-scale feature H multiplied by W flattening).

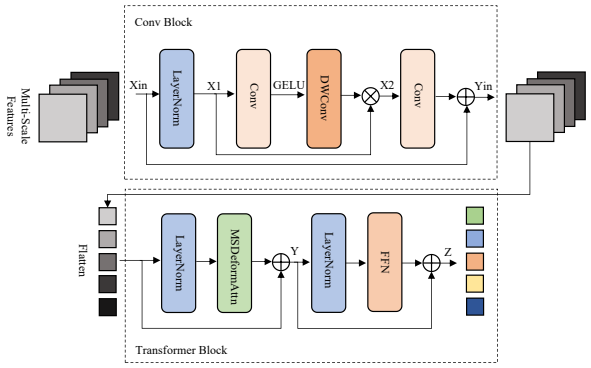


Fig. 3. Conv-Trans Layer(CTL). It consists of two parts: Conv Block and Transformer Block. CTL can enhance the feature representation capabilities of the network.

Conv Block can be expressed as:

$$X1 = Norm\left(Xin\right) \tag{1}$$

$$X2 = DWConv\left(\delta\left(Conv\left(X1\right)\right)\right) \otimes X1 \tag{2}$$

$$Yin = Conv\left(X2\right) \oplus Xin \tag{3}$$

where $\delta\left(\cdot\right)$ refers to the GeLU activation function, Norm(.) denotes layer normalization, DWConv(.) denotes depthwise separable convolution, $\otimes$ denotes element-wise multiplication and $\oplus$ denotes element-wise addition.

Transformer Block can be expressed as:

$$Y = MSDeformAttn\left(Norm\left(Yin\right)\right) \oplus Yin \tag{4}$$

$$Z = FFN\left(Norm\left(Y\right)\right) \oplus Y \tag{5}$$

where MSDeformAttn(.) denotes multi-scale deformable attention and FFN(.) denotes a feed-forward neural network.

C. Positional Queries Enhancement Strategy

In order to make the positional queries in the cross-attention module of the decoder have a better spatial prior, we introduce the Positional Queries Enhancement Strategy after the query selection module, as shown in Figure 4.

Channel transformation. Since each channel of the feature map is considered a feature detector [25], we use the interchannel relationships of features to generate a channel attention map. In order to effectively aggregate spatial information, average-pooling downsampling is commonly used in the field of computer vision. For instance, Hu *et al.* [26] uses the attention module to calculate spatial statistics and Zhou *et al.* [27] uses it to effectively learn the range of target objects. In addition to average-pooling downsampling, we believe that max-pooling downsampling gathers other important information for different object characteristics. Therefore, we use both average-pooling and max-pooling features to improve the representation capability of the network. We first input $U \in \mathbb{R}^{C1\times N}$(C1 is the number of channels) into the depthwise separable convolution to obtain intermediate features. Secondly, the feature is aggregated by average-pooling and max-pooling operations to aggregate the spatial information of the feature map to generate two different

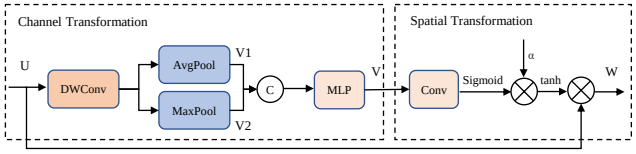


Fig. 4. Positional Queries Enhancement Strategy(PQES). It contains channel transformation and spatial transformation. PQES is used to enhance spatial information for positional queries.

spatial context features $V1 \in \mathbb{R}^{C1\times 1}$ and $V2 \in \mathbb{R}^{C1\times 1}$. Finally, the two features are concatenated and forwarded to a multi-layer perceptron (MLP) to generate our channel map $V \in \mathbb{R}^{C2\times 1}$(C2 is the number of channels). In short, the channel transformation is computed as:

$$V1 = AvgPool\left(DWConv\left(U\right)\right) \tag{6}$$

$$V2 = MaxPool\left(DWConv\left(U\right)\right) \tag{7}$$

$$V = MLP\left(Concat\left(V1; V2\right)\right) \tag{8}$$

where AvgPool(.) denotes average-pooling downsampling, MaxPool(.) denotes max-pooling downsampling, MLP(.) denotes a multilayer perceptron (consisting of two linear layers and a LeakyReLU activation function), Concat(.) denotes tensor concatenation and DWConv(.) denotes depthwise separable convolution.

Spatial transformation. Generating spatial attention maps using inter-spatial relationships of features. Unlike channel transformation, spatial transformation use the spatial relationships of features to generate spatial attention maps. Firstly, on the basis of channel transformation, the spatial transformation is performed.$V \in \mathbb{R}^{C2\times 1}$ is input into convolution, channel information fusion and spatial attention map is generated, and spatial weights are normalized by the Sigmoid activation function. Second, multiply the normalized result by the learnable parameter $\alpha \in \mathbb{R}^{C1\times 1}$ to learn how to control channel activation wisely. Finally, the input feature map and the spatial weights are multiplied to obtain $W \in \mathbb{R}^{C1\times N}$. In short, the spatial transformation is computed as:

$$W = tanh\left(Sigmoid\left(Conv\left(V\right)\right) \otimes \alpha\right) \otimes U \tag{9}$$

where tanh(.) and Sigmoid (.) denotes the activation function, Conv(.) denotes the convolutional layer, α is the learnable parameter, $\otimes$ denotes element-wise multiplication.

IV. Experiments

A. Datasets

We conduct experiments on three challenging underwater object detection datasets to verify the effectiveness of our method.

UODD is an underwater image we collected from the web and then formed a completely new underwater object detection dataset. As shown in Figure 5, our UODD dataset compares quantitatively to other underwater datasets. It can be seen that the UODD dataset has the largest number, reaching 20309 underwater images. It contains 16247 training images, 4062 validation images, and more than 120000 label

TABLE I

COMPARISONS WITH OTHER OBJECT DETECTION MODEL RESULTS ON THE UODD DATASET. OUR METHOD ACHIEVES THE BEST DETECTION PERFORMANCE WHEN USING RESNET50 AS THE BACKBONE NETWORK. THE FPS IS TESTED ON A SINGLE NVIDIA RTX 3090 GPU. '$*$' INDICATES THAT THE MODEL HAS PQES ADDED.

Method	Epochs	AP	AP50	AP75	APS	APM	APL	FPS
Two-Stage Detector:								
Faster R-CNN w/ FPN [1]	12	51.6	87.4	55.0	24.9	49.8	56.2	33.4
Cascade R-CNN [2]	12	53.2	87.3	58.6	24.7	51.2	58.2	28.6
Cascade RPN [28]	12	52.7	85.1	54.8	23.6	49.1	55.7	27.4
Dynamic R-CNN [29]	12	52.6	85.7	57.8	24.3	50.7	57.8	33.0
Sparse R-CNN [30]	12	47.5	80.7	50.4	21.6	46.4	52.6	33.4
Libra R-CNN [31]	12	52.5	86.8	59.2	24.9	51.6	58.4	32.5
RoIAttn [32]	12	52.7	86.7	59.8	25.4	52.3	58.2	28.3
DetectoRS [33]	12	53.8	88.5	57.5	24.9	52.7	57.4	13.6
CenterNet2 [34]	12	53.7	87.9	56.5	23.6	52.3	56.6	**34.7**
One-Stage Detector:								
ATSS [5]	12	53.7	88.2	59.3	24.8	52.4	57.1	31.3
FCOS [15]	12	50.2	87.9	53.6	22.4	48.1	53.8	33.8
RetinaNet [3]	12	50.1	86.3	53.7	21.8	48.0	53.6	32.9
RepPoints [35]	12	50.5	85.9	54.3	23.2	47.4	53.6	32.4
AutoAssign [36]	12	52.5	88.4	56.2	22.7	53.5	56.9	32.2
CenterNet [37]	12	41.7	72.4	42.8	16.4	45.8	46.3	18.9
Transformer:								
Anchor DETR(DC5) [38]	50	52.8	88.9	56.5	25.3	51.7	57.4	-
Deformable DETR(4scale) [39]	50	53.2	89.3	57.1	25.7	52.4	57.9	18.3
Deformable DETR(4scale)* [39]	50	54.0	90.6	58.4	26.6	53.1	58.7	-
DAB-DETR(DC5) [17]	50	51.8	87.7	55.2	24.4	50.2	56.9	13.4
DAB-DETR(DC5)* [17]	50	52.7	88.5	55.9	25.2	51.0	57.8	-
DN-Deformable-DETR(4scale) [40]	50	54.0	89.7	61.5	25.9	53.4	59.8	17.8
DINO(4scale) [18]	36	55.7	89.8	62.1	28.4	53.7	60.5	18.6
Ours:								
QO-Net	12	54.4	89.6	59.6	26.4	52.5	60.1	18.0
QO-Net	36	**57.2**	**90.4**	**63.2**	**30.2**	**55.4**	**62.4**	18.0

TABLE II

RESULTS OF QO-NET AND OTHER DETECTION MODELS TRAINED ON THE BRACKISH DATASET USING RESNET50 AS THE BACKBONE NETWORK.

Method	Epochs	AP	AP50
Faster R-CNN w/ FPN [1]	12	79.3	97.4
Cascade R-CNN [2]	12	80.7	96.9
Boosting R-CNN [9]	12	82.0	97.4
RetinaNet [3]	12	78.0	96.5
CenterNet2 [34]	12	79.3	97.4
DetectoRS [33]	12	81.6	97.0
RoIAttn [32]	12	78.3	91.0
GCC-Net [8]	12	80.5	98.3
Deformable DETR [39]	50	77.5	97.1
DN-DAB-DETR [40]	50	75.8	95.6
DN-Deformable-DETR [40]	50	78.7	97.3
DINO [18]	12	81.3	97.4
QO-Net	12	**82.2**	**98.4**

files. At the same time, the categories are also relatively rich, consisting of four categories: starfish, echinus, holothurian, and scallop. The images contain four resolutions: 704×576, 1920×1080, 3840×2160, and 720×405.

Brackish is the first annotated underwater image dataset captured in temperate brackish waters to be proposed earlier. It includes six categories: big fish, crab, jellyfish, shrimp, small fish, and starfish. The data set contains a total of 12,902 underwater images, in which the training set, validation set, and test set are randomly divided into 9967, 1467, and 1468 images, respectively. The image size is 960×540.

TrashCan is the first underwater trash instance segmentation annotation dataset. The dataset contains 16 categories, including trash, bio, rov, unknown, metal, and plastic. The dataset contains a total of 7212 underwater images, of which the training set and the validation set have 6008 images and 1204 images, respectively. The image size is 480×270.

B. Experimental Details and Evaluation Metrics

Experimental Details. Our method is implemented on MMdetection [41]. In the experiments, we use two training methods. The first method trains for 12 epochs, using AdamW as the optimizer, with a weight decay of 0.0001 and a momentum of 0.9. The initial learning rate is 0.0001, and the learning rate is reduced by 0.1 after the 11th epoch. The second method trains for 36 epochs. The optimizer and its corresponding parameters are the same as the first training method, but the initial learning rate is reduced by 0.1 after the 30th epoch. Here, we use the same data augmentations as DETR and its variants: random cropping and scale augmentation. For the CNN-based detector, we train for 12 epochs. The optimizer is SGD, where the initial learning rate is 0.005, the momentum is 0.9, and the weight decay is 0.0001. The learning rate is reduced by 0.1 at the 8th and 11th epochs. There is no additional data enhancement other than traditional horizontal flipping.

Evaluation Metrics. The main reported results in this paper follow standard COCO-style Average Precision (AP) metrics that include AP, AP50 (IoU = 0.5), AP75 (IoU =

0.75), APS (Small), APM (Medium), and APL (Large). AP is measured by averaging over multiple IoU thresholds, ranging from 0.5 to 0.95 with an interval of 0.05.

C. Comparison with State-of-the-art Models

We compare QO-Net with some state-of-the-art methods on three underwater object detection datasets. The results are shown in Table I, Table II, and Table III.

TABLE III
RESULTS OF QO-NET AND OTHER DETECTION MODELS TRAINED ON THE TRASHCAN DATASET USING RESNET50 AS THE BACKBONE NETWORK.

Method	Epochs	AP	AP50
Faster R-CNN w/ FPN [1]	12	31.2	55.3
Cascade R-CNN [2]	12	33.6	54.3
Boosting R-CNN [9]	12	36.8	57.6
RetinaNet [3]	12	30.4	54.1
DetectoRS [33]	12	36.2	56.1
RoIAttn [32]	12	32.6	57.2
Deformable DETR [39]	50	36.1	56.9
DINO [18]	12	39.8	60.6
QO-Net	12	**41.2**	**61.4**

Results on UODD: As shown in Table I, we evaluate the effectiveness of QO-Net by comparing it with baselines and other state-of-the-art methods. First, in terms of detection accuracy, when our method uses ResNet50 as the backbone network to train for 12 epochs, the accuracy reaches 54.4 AP. It exceeds DetectoRS (53.8AP), ATSS (53.7AP), CenterNet2 (53.7AP), and DN-Deformable-DETR (54.0AP). When trained for 36 epochs, the detection accuracy is 57.2 AP, which is 1.5 AP higher than the baseline method with our proposed method. At the same time, it is better than Anchor DETR (52.8AP), Deformable DETR (53.2AP), and DAB-DETR (51.8AP) trained for 50 epochs. Secondly, in terms of inference speed, QO-Net achieved 18.0 FPS, which is better than two-stage detectors such as DetectoRS (13.6FPS) and Transformer detectors such as DAB-DETR (13.4FPS) and DN-Deformable-DETR (17.8FPS). Finally, the positional query enhancement strategy we proposed is applied to Deformable DETR and DAB-DETR, and the detection accuracy is improved by 0.8 AP and 0.9 AP, respectively. From this, we can conclude the effectiveness of our proposed method.

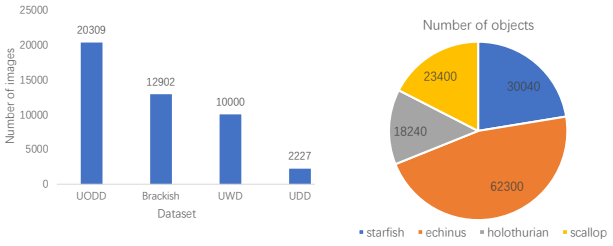


Fig. 5. The figure on the left is a comparison of the number of UODD and other underwater datasets. The figure on the right is the number of each category in the UODD dataset.

Results on Brackish: As shown in Table II, we extensively evaluate QO-Net on the Brackish dataset. Our method achieved 82.2 AP and 98.4 AP50 using ResNet50 as the backbone network. Compared with the baseline method, QO-Net increases 0.9 AP on the Brackish dataset. In terms of detection accuracy, our proposed method exceeds Deformable DETR (77.5AP), DN-DAB-DETR (75.8AP) and DN-Deformable-DETR (78.7AP) trained for 50 epochs using ResNet50, Boosting R-CNN (82.0AP), Cascade R-CNN (80.7AP), DetectoRS (81.6AP), RoIAttn (78.3AP) and GCC-Net (80.5AP) trained for 12 epochs. It is obvious that QO-Net achieves significant performance improvements.

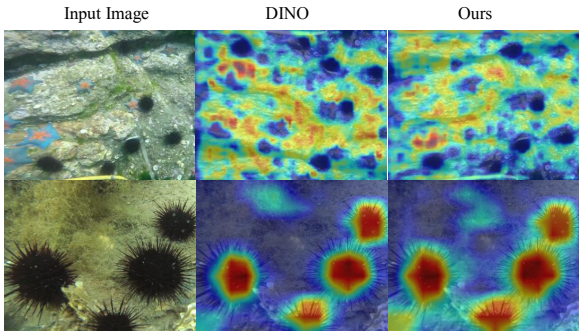


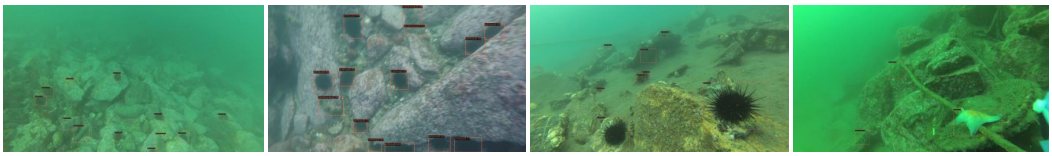
Fig. 6. Visualization comparison of attention maps of input images on baseline and our method.

Results on TrashCan: The TrashCan dataset contains a wide variety of underwater debris. The experimental results of the TrashCan dataset are shown in Table III. QO-Net obtained 41.2 AP, which is 4.4 AP higher than the recently proposed underwater object detection method Boosting R-CNN. At the same time, we also compared with other state-of-the-art object detection methods, the detection accuracy of QO-Net is better than DINO (39.8AP), Deformable DETR (36.1AP), RoIAttn (32.6AP) and Cascade R-CNN (33.6AP). As demonstrated by the results, our method may be tailored to various water bodies and can still operate well even in situations when the water environment undergoes major changes.

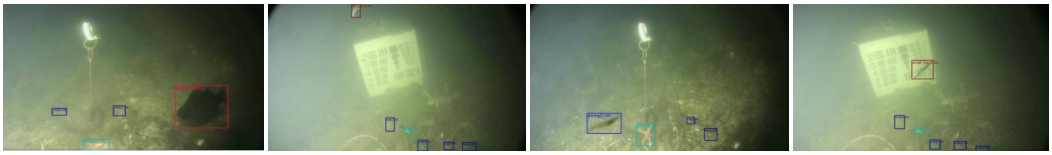
D. Ablation Experiments

To verify the validity of each module proposed in our method, we performed ablation experiments on the UODD dataset. In addition, we also added tricks such as random-rotating and random-erasing to the experiment to participate in the comparison.

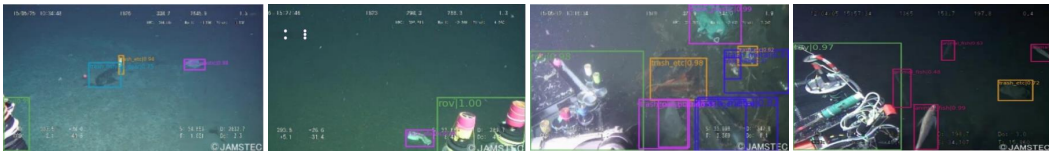
Table IV shows the rationality of our proposed functional modules, which we gradually add to the experiment to observe the changes. On the UODD dataset, first, we add the Conv-Trans Layer (CTL) module to the baseline algorithm, and we can observe that the Conv-Trans Layer function module can improve the accuracy by 0.5 AP. In order to fully utilize the feature representation capabilities of Conv-Trans Layer, we added two tricks, random-erasing and random-rotating, which improved by 0.7 AP compared to the baseline algorithm. Secondly, the Positional Queries Enhancement Strategy (PQES) was added on this basis, and the final accuracy reached 54.4 AP. The effectiveness of our proposed module can be seen from the above ablation experiment.



(a) UODD



(b) Brackish



(c) TrashCan

Fig. 7. Qualitative detection results of QO-Net on the UODD (the first row), Brackish (the second row), and TrashCan (the third row) datasets.

TABLE IV

ABLATION COMPARISON OF THE PROPOSED ALGORITHM MODULES AND ADDED TRICKS ON THE UODD DATASET. WE USE THE TERMS "CTL", "RE", "RR", AND "PQES" TO DENOTE "CONV-TRANS LAYER", "RANDOM-ERASING", "RANDOM-ROTATING", AND "POSITIONAL QUERIES ENHANCEMENT STRATEGY", RESPECTIVELY.

Method	CTL	RE	RR	PQES	AP	AP50	AP75	APS	APM	APL
Baseline(BL)					52.9	88.7	57.2	24.9	50.4	57.9
BL+CTL	√				53.4	89.2	58.2	25.2	51.6	58.8
BL+CTL+RE	√	√			53.5	89.2	58.4	25.3	51.9	59.1
BL+CTL+RR	√		√		53.5	89.1	58.1	25.2	51.7	58.9
BL+CTL+RE+RR	√	√	√		53.6	89.4	58.3	25.4	52.1	59.6
BL+CTL+RE+RR+PQES	√	√	√	√	54.4	89.6	59.6	26.4	52.5	60.1

In summary, the detection accuracy of our proposed functional module is significantly improved on several different datasets. In particular, the position query enhancement strategy not only enhances the spatial prior of the position query in the decoder but also speeds up the convergence speed of the network. The Conv-Trans Layer module produced by the effective combination of CNN and Transformer enables the model to effectively capture different levels of information in the image through the convolution layer and enables the model to process longer-distance sequence information through the Transformer. In this way, the representation of multi-scale features is enhanced.

E. Qualitative Comparisons and Visualization Analysis

Figure 6 shows the visualization of the attention maps of the input image on the baseline and our method. Obviously, our method makes the attention area more complete and has higher attention response to objects than other detectors. Figure 7 shows the qualitative detection results of our proposed method on UODD, Brackish and TrashCan datasets. It is evident that the suggested method is capable of handling a wide range of difficulties in underwater item identification tasks, such as diverse underwater settings, small, low-contrast objects, and objects that are densely organized. As is shown from the image, we applied the detector to underwater blurry, occlusion, and uneven illumination and set the prediction score to 0.05 to maintain high performance. It can be seen from the test results that, our improved method has obvious advantages in dealing with dense small objects and ambiguous objects in underwater environments and is more suitable for specific underwater environments.

V. CONCLUSION

In this paper, we discuss issues affecting object detection performance in underwater environments. Underwater object detection is challenging due to low contrast, texture distortion, and imitation of aquatic life. To address these issues, we propose an end-to-end underwater object detector based on Transformer. Compared to general object detectors, our detector can handle issues such as underwater image blur, low contrast, and color distortion. First, we propose the Conv-Trans Layer structure. It effectively learns multi-scale image feature representation through CNN while capturing the dependencies between different positions in sequence data through Transformer. An effective combination can improve the representation of multi-scale features. Secondly, we introduce a position query enhancement strategy in the decoder cross-attention, which can speed up the convergence to a certain extent and provide accurate spatial priors for

position queries. Experiments on three challenging underwater object detection datasets demonstrate the generalization performance of our method, and the effectiveness of the proposed module is also verified in ablation experiments.

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[2] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "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, 2016, pp. 21–37.

[5] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9759–9768.

[6] M. Jahidul Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *arXiv e-prints*, pp. arXiv–1903, 2019.

[7] B. Fan, W. Chen, Y. Cong, and J. Tian, "Dual refinement underwater object detection network," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 275–291.

[8] L. Dai, H. Liu, P. Song, and M. Liu, "A gated cross-domain collaborative network for underwater object detection," *arXiv preprint arXiv:2306.14141*, 2023.

[9] P. Song, P. Li, L. Dai, T. Wang, and Z. Chen, "Boosting r-cnn: Reweighting r-cnn samples by rpn's error for underwater object detection," *Neurocomputing*, vol. 530, pp. 150–164, 2023.

[10] Z. Fu, W. Wang, Y. Huang, X. Ding, and K.-K. Ma, "Uncertainty inspired underwater image enhancement," in *European Conference on Computer Vision*. Springer, 2022, pp. 465–482.

[11] C. Li, S. Anwar, J. Hou, R. Cong, C. Guo, and W. Ren, "Underwater image enhancement via medium transmission-guided multi-color space embedding," *IEEE Transactions on Image Processing*, vol. 30, pp. 4985–5000, 2021.

[12] X. Chen, Y. Lu, Z. Wu, J. Yu, and L. Wen, "Reveal of domain effect: How visual restoration contributes to object detection in aquatic scenes," *arXiv preprint arXiv:2003.01913*, 2020.

[13] R. Liu, Z. Jiang, S. Yang, and X. Fan, "Twin adversarial contrastive learning for underwater image enhancement and beyond," *IEEE Transactions on Image Processing*, vol. 31, pp. 4922–4936, 2022.

[14] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[15] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 213–229.

[17] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "Dab-detr: Dynamic anchor boxes are better queries for detr," *arXiv preprint arXiv:2201.12329*, 2022.

[18] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022.

[19] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, and L. Zhang, "Dynamic detr: End-to-end object detection with dynamic attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2988–2997.

[20] Y. Cong, B. Fan, D. Hou, H. Fan, K. Liu, and J. Luo, "Novel event analysis for human-machine collaborative underwater exploration," *Pattern Recognition*, vol. 96, p. 106967, 2019.

[21] W.-H. Lin, J.-X. Zhong, S. Liu, T. Li, and G. Li, "Roimix: proposal-fusion among multiple images for underwater object detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2588–2592.

[22] H. Liu, P. Song, and R. Ding, "Towards domain generalization in underwater object detection," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 1971–1975.

[23] C. Liu, Z. Wang, S. Wang, T. Tang, Y. Tao, C. Yang, H. Li, X. Liu, and X. Fan, "A new dataset, poisson gan and aquanet for underwater object grabbing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2831–2844, 2021.

[24] L. Chen, F. Zhou, S. Wang, J. Dong, N. Li, H. Ma, X. Wang, and H. Zhou, "Swipenet: Object detection in noisy underwater scenes," *Pattern Recognition*, vol. 132, p. 108926, 2022.

[25] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 818–833.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[27] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[28] T. Vu, H. Jang, T. X. Pham, and C. Yoo, "Cascade rpn: Delving into high-quality region proposal network with adaptive convolution," *Advances in neural information processing systems*, vol. 32, 2019.

[29] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic r-cnn: Towards high quality object detection via dynamic training," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 260–275.

[30] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang *et al.*, "Sparse r-cnn: End-to-end object detection with learnable proposals," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 454–14 463.

[31] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: Towards balanced learning for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 821–830.

[32] X. Liang and P. Song, "Excavating roi attention for underwater object detection," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2651–2655.

[33] S. Qiao, L.-C. Chen, and A. Yuille, "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 213–10 224.

[34] X. Zhou, V. Koltun, and P. Krähenbühl, "Probabilistic two-stage detection," *arXiv preprint arXiv:2103.07461*, 2021.

[35] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9657–9666.

[36] B. Zhu, J. Wang, Z. Jiang, F. Zong, S. Liu, Z. Li, and J. Sun, "Autoassign: Differentiable label assignment for dense object detection," *arXiv preprint arXiv:2007.03496*, 2020.

[37] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[38] Y. Wang, X. Zhang, T. Yang, and J. Sun, "Anchor detr: Query design for transformer-based detector," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2567–2575.

[39] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[40] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "Dn-detr: Accelerate detr training by introducing query denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 619–13 627.

[41] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.