Vision-Based Target Detection and Positioning Approach for Underwater Robots

Yanli Li, Weidong Liu, Le Li, *Member, IEEE*, Wenbo Zhang, Jingming Xu, and Huifeng Jiao

***Abstract*—The accurate target detection under different environmental conditions and the real-time target positioning are vital for the successful accomplishment of underwater missions of Remotely operated vehicles (ROVs). In this paper, we propose a vision-based underwater target detection and positioning approach to detect and estimate the position and attitude of artificial underwater targets. The proposed approach is composed of an underwater target detection algorithm YOLO-T and a target positioning algorithm. Firstly, we modify the structure of YOLOv5 algorithm using Ghost module and SE attention module to improve the calculation time of target detection. Secondly, a series of image processing operations are performed on the improved YOLOv5 detection results to increase the detection accuracy. Thirdly, a cooperative marker is designed as the artificial underwater target, and the corresponding positioning algorithm is presented to calculated the position and attitude of the target according to the geometric information of the designed marker. We validate our approach through experimental tests respectively in a water tank, an anechoic tank, and the sea trial in Huanghai Sea in China. The results demonstrate the accurate performance of the proposed detection and positioning method.**

***Index Terms*—YOLO-T, Underwater target detection, Target positioning.**

I. Introduction

With the rapid development of manufacture, the requirement for identifying and locating the target accurately becomes more and more increasing. In recent years, accurate detection and positioning of target has been realized in the air [1], [2], [3]. Given the dynamic and uncertain ocean environments, the accurate target detection and the real-time target positioning under different environmental conditions are vital but challenging for the successful accomplishment of underwater missions such as sampling, archaeological surveys and underwater facility inspections for Remotely operated vehicles (ROVs). Under this circumstance, lots of efforts have been made to realize underwater target measurements like target detection and positioning.

Manuscript received 10 October 2022; revised 5 December 2022; accepted 6 December 2022. Date of publication 9 December 2022; date of current version 30 December 2022. This work work was supported in part by the National Science Foundation of China under Grant 61903304, in part by the Fundamental Research Funds for the Central Universities under Grant 3102020HHZY030010, and in part by the 111 Project under Grant B18041. *(Corresponding author: Le Li.)*

Yanli Li, Weidong Liu, Le Li, Wenbo Zhang, and Jingming Xu are with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: liyl@mail.nwpu.edu.cn; liuwd@nwpu.edu.cn; leli@nwpu.edu.cn; 15619339046@163.com; 2018260584@mail.nwpu.edu.cn).

Huifeng Jiao is with the Taihu Lake Laboratory of Deep Sea Technology and Science, Wuxi 214082, China (e-mail: jiaohuifeng@163.com).

Digital Object Identifier 10.1109/JPHOT.2022.3228013

Generally, vision-based underwater target detection methods can be divided into two categories: traditional methods and deep learning based methods. Traditional underwater target detection methods include image feature matching recognition [4], [5], general image segmentation [6], [7], and detection and recognition based on color and shape [8], [9], [10]. Sun et al. [8] designed an autonomous recognition system based on the color and shape of the target for the real-time detection and tracking of robotic fish. Coincidentally, Yahya MH et al. [9] proposed a color based tracking method for underwater tracking. They designed a marker using LEDs as the artificial docking target. The target was identified through color threshold and morphological operations. Meanwhile, the traditional target detection methods in [11], [12], [13] also used artificial targets for real-time underwater tests. While these traditional target detection methods have fast processing times for simple underwater environment, they are still not suitable for dynamic environments.

Compared with traditional target detection methods, the target detection methods based on deep learning are faster and more robust in complex circumstances, such as the partial occlusion of targets. Therefore, these deep learning based methods have gradually become a mainstream method for target detection. At present, the algorithms based on deep learning can be divided into end-to-end algorithms and region proposal algorithms. YOLO [14] and SSD [15] are the typical end-to-end algorithms, which have fast processing speed and can reach 45 FPS. The region proposal algorithms combine region suggestion with convolutional neural network to perform target detection, such as Faster R-CNN [16] and R-FCN [17]. Li et al. [18] designed a vision based remote control vehicle for autonomous capture and absorption of marine organisms, and improved the R-FCN algorithm from two aspects, e.g., small object recognition and dynamic biometrics. Chen et al. [19] combined optical transmission information, image features and illumination with ROI (Region of Interest) in the deep learning process, and realized the target segmentation on the basis of the target detection. These deep learning based target detection methods can detect target in dynamic environments. However, these methods obtain only the rectangular bounding boxes of the targets but not the precise bounding information of the targets, which cannot be used for the accurate position estimation and attitude calculation.

Target positioning is also a crucial research topic for ROVs, especially in the navigation operations and docking missions. In most of these operations, cooperative targets or artificial objects are used to improve the positioning efficiency. The frequently

used underwater artificial objects are usually with regular shape and specific bright color, such as special underwater patterns like trapezoid [20], active laser modules [21] and 3D Markers [22], [26]. The common positioning methods include geometric-based methods [22], [23], curvature-based methods [24] and PnP-based methods [25]. Maki et al. [22] proposed a docking method for hovering type AUVs based on both the acoustic and visual positioning. The short-ranged visual positioning was calculated through a series of geometric relationship calculation of the 3D Markers. Meanwhile, Ghosh et al. [24] proposed a scene invariant approach to estimate the pose of a circular station during underwater docking using single camera. The station was arranged with LED lights on periphery and was detected through the binary captured images and then the pose estimation was realized by curvature-based methods. However, the method was computationally complicated and had not been applied to the docking mission. Similarly, Lwin et al. [26] proposed a real-time position and pose estimation system for the ROV docking and charging using artificial 3D Markers. The whole system applied a multi-step genetic algorithm to suppress the bubble noise in the visual serving control process and located the 3D Markers using the binocular vision.

The focus of this paper is to accurately detect the underwater target and to calculate the position and attitude of the target. A vision-based underwater artificial target detection and positioning approach is proposed in this paper. The proposed approach consists of a YOLO-T underwater target detection algorithm and a target positioning algorithm. Firstly, the structure of YOLOv5 is modified to decrease the detection time cost by means of the replacement of network backbone with GhostBottleneck and the adding of SE attention module. Then, a serial of traditional image processing methods based on shape and color are adopted based on the deep learning framework to carry out the accurate target detection. Finally, a cooperative marker is designed as the artificial underwater target, and the position and attitude relative to the camera are calculated according to the detection results and the geometric information of the target.

The main contributions in this paper are summarized as follows:

1) An underwater target detection algorithm YOLO-T is proposed in this paper, which not only solves the problem that traditional detection methods are prone to errors in complex environments, but also can detect more accurate target contours compared with the deep learning-based detection algorithms. In the target detection tests compared with the series of YOLO algorithms, the detection error is 10%-45.8% lower than that of the YOLO series, which can prove the accuracy of YOLO-T.

2) This paper presents an artificial target positioning algorithm based on feature points sorting. Compared with the above positioning practices, on the one hand, the algorithm can achieve accurate target positioning by monocular camera through feature points sorting and PNP-based coordinate transformation model. On the other hand, this algorithm has low requirements for artificial targets and does not need 3D markers such as the above LEDs. Moreover, the size of targets and the distance between feature points are flexible, and stable positioning can be achieved even when feature points are missing.

The rest of this paper is organized as follows. Section II proposes the YOLO-T underwater target detection algorithm. Section III presents the characteristics of the designed artificial target and the proposed target positioning algorithm. Section IV presents the experimental comparisons and validation of the target detection and positioning algorithm. Section V concludes this paper and discusses directions for future work.

II. YOLO-T Target Detection Algorithm

YOLOv5 algorithm is one of the most popular algorithms in YOLO series, which adds lots of tricks to YOLOv4 [27] to improve the performance. To further improve the target detection efficiency, YOLO-T algorithm mainly modifies YOLOv5 from the aspects of detection time cost and detection accuracy. The network structure of YOLO-T is shown as Fig. 1.

A. The Decreasement in Detection Computational Time

Reducing model parameters is the best way to reduce time consumption in target detection. Therefore, we introduce Depth-wise(DW) convolution module and Ghostbottleneck into the network.

1) DW Convolution Module: In the conventional convolution module, each convolution kernel operates all channels of the input image simultaneously. In contrast, in DW convolution module, each of the kernel is responsible for a single channel, and one channel is convolved by only one convolution kernel.

Assume that the input layer is a three-channel target image with a size of 64×64 pixels. In conventional convolution, the image passes through a convolution layer containing 4 filters, and finally outputs 4 feature maps with the same size as the input layer. There are 4 filters in the convolution layer, each filter contains 3 kernels, and the size of each kernel is 3×3, so the number of parameters in the convolution layer is 4*3*3*3. While in DW convolution, the image also goes through the first convolution operation. The difference is that the convolution is completely carried out in a two-dimensional plane, and the number of filters is the same depth as the previous layer. Therefore, a three-channel image is generated into three feature maps after operation. One filter contains only a kernel with a size of 3×3, and the number of parameters in the convolution part is 3*3*3. The number of model parameters has been greatly reduced.

2) Ghostbottleneck: Ghostbottleneck is a modular structure that is stacked by two GhostNet modules [28]. The core idea of GhostNet is to design a phased convolution calculation module. Based on a small number of feature maps obtained by nonlinear convolution, linear convolution is carried out again to obtain more feature maps, so as to eliminate redundant features and obtain a lighter model.

There are two versions of Ghostbottleneck with the size of stride is $1(s = 1)$ and the size of stride is $2(s = 2)$ as shown in Fig. 2. When $s = 1$, it can be used to replace the BottleneckCSP module in YOLOv5 with the same input and output dimensions. When $s = 2$, a DW convolution of with stride size 2 is added between the two GhostNet modules to reduce the size of the feature graph to half of the input, which can replace the lower

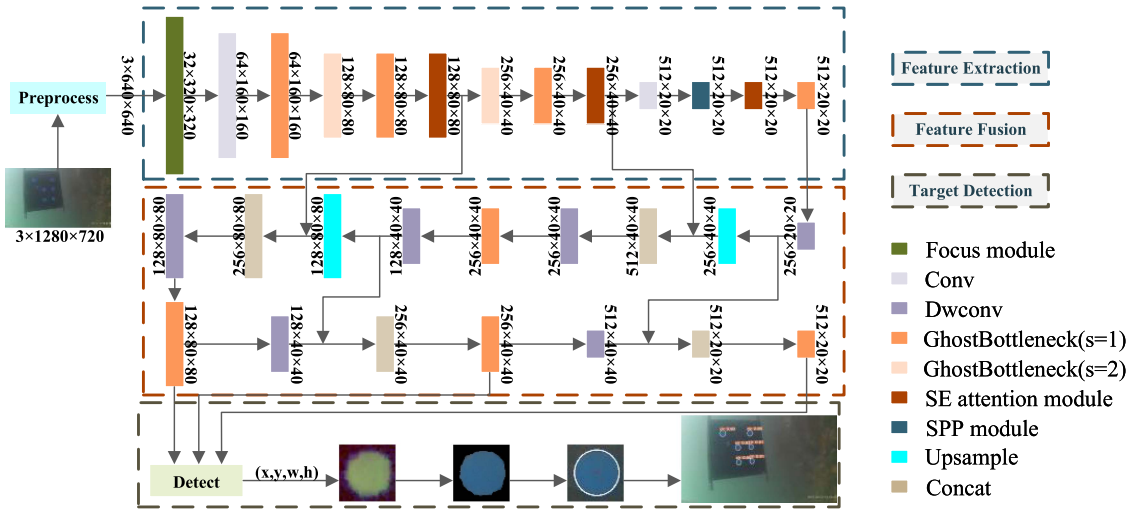
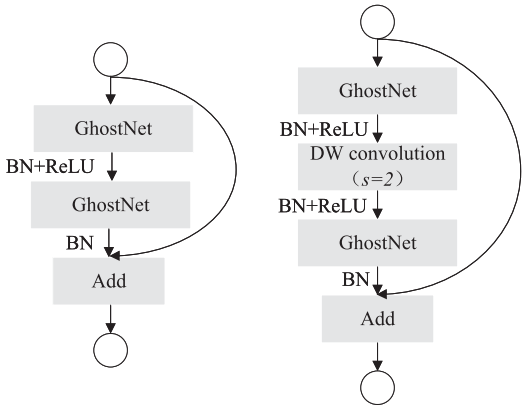


Fig. 1. The network structure of YOLO-T. In the Feature Extraction stage, the target image size is first processed into 3*640*640. Then the features of the target are extracted through several down-sample modules. In the Feature Fusion stage, the network integrates the extracted features and finally obtains target feature maps of three different scales. In the Target Detection stage, the detection module is used to output the bounding box (x, y, w, h) of the target, and then the accurate contour and center coordinate (x_c, y_c) of the target are obtained through image processing operation.



(a) Ghostbottleneck (*s=1*) (b) Ghostbottleneck (*s=2*)

Fig. 2. The structure of Ghostbottleneck.

sampling layer or conventional convolution operation in the network.

In order to prevent the reduction of model parameters from affecting the extraction of target features, we use SE attention module [29] to solve the problem of loss caused by different importance of feature map channels in the process of convolutional pooling, so as to avoid affecting the accuracy of the improved network.

B. The Increasement in Detection Accuracy

To improve the detection accuracy, a series of traditional processing methods are applied on the bounding box results obtained from the Detect module.

The rectangular area of target in the image is extracted for the further processes. In order to avoid the situation that the bounding box detected does not fully contain target, the target bounding box should be extended with a certain number of pixels, which is determined by the detection error of the YOLOv5 target detection algorithm. According to the color features of the designed target, the extracted target image is converted from RGB color space to HSV color space, and the image regions conforming to the target color are screened firstly. Then, the gray scale processing and adaptive binarization are performed on the filtered image to extract the edge contour information of the target image. Finally, the ellipse fitting method is used to screen the extracted contour, and the edge contour of the target and the central point of the target is detected.

To summarize, the YOLO-T algorithm performs conventional target detection based on the improved network structure. On the one hand, it avoids the problem that traditional target detection methods are not accurate in detecting small targets in complex scenes; on the other hand, the YOLO-T algorithm also reduces the accuracy requirements for annotations of the data sets, which takes up the most time in data preparation.

III. Target Positioning Algorithm

This section presents the proposed target positioning algorithm in detail.While using a monocular camera, this algorithm not only calculates the position (x, y, z) of the target, but also estimates the attitude angles (ψ, θ, ϕ) of the target relative to the camera. In other words, this algorithm performs a 6DoF pose estimation of the target. The proposed target positioning algorithm contains two parts. First of all, an underwater artificial target is designed and the feature points of the target are detected using YOLO-T target detection. Then the sorting algorithm is used to sort the coordinates of the detected feature points so that they can match the feature points. Secondly, the relationship among the image coordinate system, the camera coordinate system and the target coordinate system as well as the visual

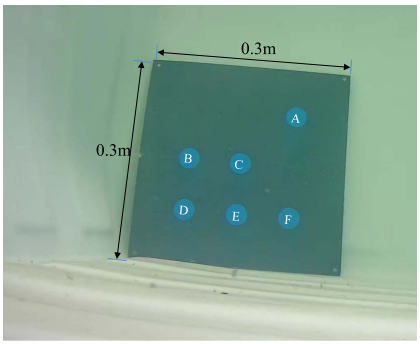


Fig. 3. The arrangement of 6 artificial markers.

Algorithm 1: The Sorting Algorithm of the 6 Feature Points.

Input: Coordinates of the detected feature points $\boldsymbol{p}\{p_i(x_i, y_i), (i = 1, 2, \ldots, 6)\}$.

Output: Coordinates of the matched feature points in order $\boldsymbol{P}\{P_j(X_j, Y_j), (j = A, B, \ldots, F)\}$.

1: Choose 3 points from 6 points respectively to combine. The two sets having collinearity of 3 points are denoted as $\boldsymbol{U_{l1}}$ and $\boldsymbol{U_{l2}}$, then $\boldsymbol{U_{l1}}$ and $\boldsymbol{U_{l2}}$ are the sets of $\{P_A, P_C, P_D\}$ and $\{P_D, P_E, P_F\}$;

2: $P_D = \boldsymbol{U_{l1}} \cap \boldsymbol{U_{l2}}, P_B = \boldsymbol{p} \setminus (\boldsymbol{U_{l1}} \cup \boldsymbol{U_{l1}})$;

3: Choose 2 points from 5 points respectively to combine, and calculate the distance of these 2 points in each set. The set with the largest length is denoted as $\boldsymbol{U_d}, \boldsymbol{U_d}$ is the set of $\{P_A, P_D\}$. Then, $P_A = \boldsymbol{U_d} - P_D$;

4: Find the intersection of $\boldsymbol{U_{l1}}$ and $\boldsymbol{U_d}$ and the intersection of $\boldsymbol{U_{l2}}$ and $\boldsymbol{U_d}$. The set which have one point in the intersection is the set of $\{P_D, P_E, P_F\}$ and is denoted as $\boldsymbol{U_{DEF}}$. Then, the other set is composed of $\{P_A, P_C, P_D\}$ and is denoted as $\boldsymbol{U_{ACD}}$, and $P_C = \boldsymbol{U_{ACD}} - P_A - P_D$.

5: The set composed of P_E and P_F is $\boldsymbol{U_{EF}}$, and $\boldsymbol{U_{EF}} = \boldsymbol{U_{DEF}} - P_D$. Calculate the distance between 2 points in $\boldsymbol{U_{EF}}$ and P_D, with the closer point being P_E and the farther point being P_F.

pinhole imaging model are used to calculate the position and attitude angles of the underwater target.

A. The Artificial Target

The artificial target in this paper is a board of $0.3\,m * 0.3\,m$ with six artificial markers forming a target pattern as shown in Fig. 3. Each of the marker is a blue circle, and the size of the makers and the distance between the individual marker are flexible.

To calculate the position and attitude of the target, at least 4 points on the target are needed by the positioning algorithm. Since the circle is easier to detect and process, the circle center of the markers are chosen as the key points for positioning.

In the artificial pattern in Fig. 3, the first four center points of the markers (A, B, C, D) are used in the positioning calculation, and the center point (E) and (F) are auxiliary points which can

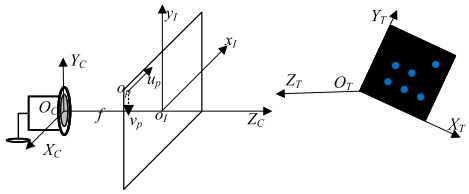


Fig. 4. Coordinate system of the positioning.

order the points to ensure that the coordinates of the detected key points match the feature points on the target.

B. The 6 DoF Pose Estimator

The purpose of the 6 DoF pose estimator is to calculate and estimate the three-dimensional position and the attitude angles of the artificial target relative to the camera. Firstly, the feature points of the target have been detected using YOLO-T algorithm and sorted by the sorting algorithm. Then the pose estimator is applied to realize the target positioning. For the 6D pose estimation, the coordinate system is as Fig. 4.

The 4 coordinate systems in Fig. 4 are the camera coordinate system $O_C - X_C Y_C Z_C$, the pixel coordinate system $o_p - u_p v_p$, the image coordinate system $o_I - x_I y_I$, and the target coordinate system $O_T - X_T Y_T Z_T$ from left to right.

The coordinate vector of an arbitrary point in the target coordinate is $[X_T, Y_T, Z_T]$, and the corresponding camera coordinate vector is $[X_C, Y_C, Z_C]$, the relationship is as follows:

$$\begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} = R \begin{bmatrix} X_T \\ Y_T \\ Z_T \end{bmatrix} + T \tag{1}$$

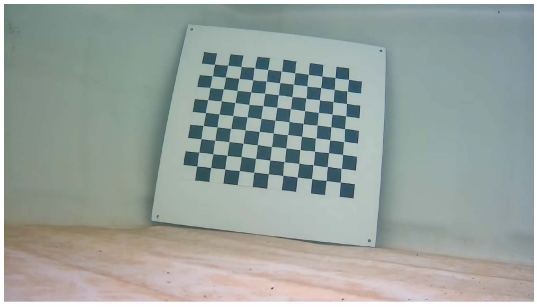
Where, T is the translation matrix of the target and R is the rotation matrix which respectively represent the position and attitude of the target relative to the camera. Thus, for the accurate pose estimation, the camera coordinate vectors of the markers will be required first. The relationship between the pixel coordinate and the camera coordinate is in (2):

$$Z_C \begin{bmatrix} u_p \\ v_p \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} = R_C \begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} \tag{2}$$

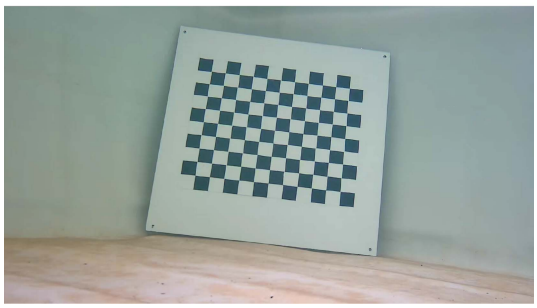
Where, R_C is the internal matrix of underwater camera, f_x and f_y are focal lengths of the camera in the x and y directions respectively, and c_x and c_y are the offsets of the origin of the image relative to the optical imaging point.

The coordinate Z_C is unknown in the monocular camera scene. For the camera coordinate vectors of the markers, the pinhole camera model in Fig. 6 is adopted for the 6D pose estimator.

To calculate the camera coordinates of target feature points, we take O as the original point of the camera coordinate system, A, B, C, D as the center points of the markers, and a, b, c, d as the corresponding image points. OA, OB, OC are used to calculate the coordinates of A, B, C. Taking O, A, a as examples for the illustration, the notations are shown in Table I.



(a) Calibration board before distortion correction



(b) Calibration board after distortion correction

Fig. 5. The underwater calibration board.

TABLE I
NOTATIONS IN POSITIONING ALGORITHM

Notation	Description
$A(Ax, Ay, Az)$	coordinate of A on the camera coordinate system
$a(ax, ay)$	coordinate of A on the image coordinate system
$\vec{OA}$	vector that goes from O to A
$\|OA\|$	length of vector $\vec{OA}$
$\vec{a}$	direction of $\vec{Oa}$ in the camera coordinate system
$\cos < \vec{Oa}, \vec{Ob} >$	the cosine of the angle between $\vec{Oa}$ and $\vec{Ob}$

The cosine equations of the geometric relationships are (3), shown at the bottom of this page:

When, $x = \frac{\|OA\|}{\|OC\|}$, $y = \frac{\|OB\|}{\|OC\|}$, then:

$$\begin{cases} x^2 + y^2 - 2x \cdot y \cdot \cos < \vec{Oa}, \vec{Ob} > = \frac{\|AB\|^2}{\|OC\|^2} \\ x^2 + 1 - 2x \cdot \cos < \vec{Oa}, \vec{Oc} > = \frac{\|AC\|^2}{\|OC\|^2} \\ y^2 + 1 - 2y \cdot \cos < \vec{Ob}, \vec{Oc} > = \frac{\|BC\|^2}{\|OC\|^2} \end{cases} \tag{4}$$

To make $l = \frac{\|AB\|^2}{\|OC\|^2}, ml = \frac{\|AC\|^2}{\|OC\|^2}, nl = \frac{\|BC\|^2}{\|OC\|^2}$, then, $m = \frac{\|AC\|^2}{\|AB\|^2}, n = \frac{\|BC\|^2}{\|AB\|^2}$. $\|AB\|, \|AC\|, \|BC\|$ represent respectively the length between A, B and C, which can be obtained from the geometric information of the arrangement of feature points on the target. Therefore, m and n are known values. Then a binary quadratic equation about x and y can be devised as follows shown at the bottom of this page:

Since a, b, c are the image coordinates detected from the YOLO-T detection algorithm, $\cos < \vec{Oa}, \vec{Ob} >$, $\cos < \vec{Oa}, \vec{Oc} >$ and $\cos < \vec{Ob}, \vec{Oc} >$ can be calculated using The Law of Cosines. Therefore, OA, OB and OC can be obtained by solving (5), shown at the bottom of this page, and then the camera coordinates A, B and C can be calculated. Since there are four groups of solutions to this equation, it is necessary to use the point D to calculate the reprojection error respectively, and the group with the least error can be used to calculate the real pose.

IV. EXPERIMENTAL VALIDATION

To validate the performance of the target detection and positioning approach proposed in this paper, underwater experiments are carried out in a water tank to verify the detection results and positioning results, respectively. First of all, the camera is calibrated underwater to get its internal matrix R_C. Next, the YOLO-T algorithm and s series of YOLO algorithms are used to detect the same group of underwater target images under the same conditions to analyze the results of YOLO-T target detection algorithm. Then, underwater tests are conducted to verify the accuracy of the positioning algorithm in the aspect of position and attitude angle. Finally, several sets of tests are carried out in an anechoic tank, including scenes with different distances and scenes where feature points are blocked. The stability of the proposed method is verified by analyzing the positioning results of fixed targets within a period of time.

A. The Camera Calibration

Although the camera has been calibrated in advance in the air, in order to ensure the accuracy of the parameters, it must be calibrated again before the underwater experiments due to the difference between the underwater environment and the air. One 11x9 chessboard printed on A4 size cardboard is chosen as the calibration board, and the size of each grid of the chessboard is 20 mm. The calibration board is fixed on the wall of the

$$\begin{cases} \|OA\|^2 + \|OB\|^2 - 2\|OA\| \cdot \|OB\| \cdot \cos < \vec{Oa}, \vec{Ob} > = \|AB\|^2 \\ \|OA\|^2 + \|OC\|^2 - 2\|OA\| \cdot \|OC\| \cdot \cos < \vec{Oa}, \vec{Oc} > = \|AC\|^2 \\ \|OB\|^2 + \|OC\|^2 - 2\|OB\| \cdot \|OC\| \cdot \cos < \vec{Ob}, \vec{Oc} > = \|BC\|^2 \end{cases} \tag{3}$$

$$\begin{cases} (1-m) \cdot x^2 - m \cdot y^2 - 2x \cdot \cos < \vec{Oa}, \vec{Oc} > + 2\,m \cdot x \cdot y \cdot \cos < \vec{Oa}, \vec{Ob} > + 1 = 0 \\ (1-n) \cdot y^2 - n \cdot x^2 - 2y \cdot \cos < \vec{Ob}, \vec{Oc} > + 2n \cdot x \cdot y \cdot \cos < \vec{Oa}, \vec{Ob} > + 1 = 0 \end{cases} \tag{5}$$

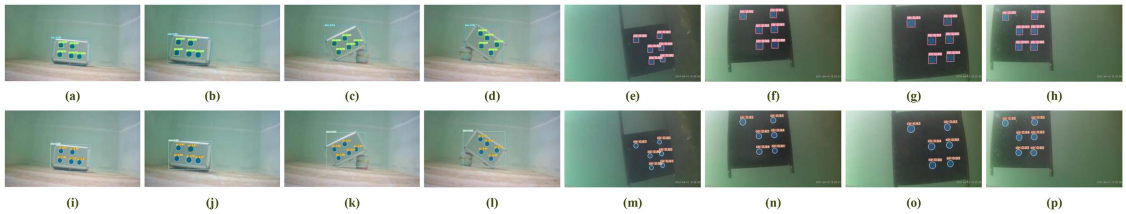


Fig. 6. Detection results of YOLO-T and YOLOv5. The first eight images, i.e. (a)-(h) are the YOLOv5 target detection results, and the latter eight images, i.e., (i)-(p) are the corresponding YOLO-T target detection results.

water tank, and the distance and angle between the camera and the calibration board are changed to obtain different calibration pictures.

A number of pictures in different distances and angles are taken in the water tank. The internal matrix of the underwater camera is calculated using Zhang's calibration algorithm [30]. Due to the radial distortion of the underwater camera, the distortion correction is needed to update the internal matrix for the precise detention and positioning. The underwater calibration board is shown in Fig. 5.

B. Target Detection Accuracy Analysis

This section qualitatively assesses the performance of the YOLO-T target detection algorithm by comparing it with other versions of the YOLO algorithm, including YOLOv3, YOLOv4, YOLOv5, YOLOv6 and YOLOv7. The effectiveness of the proposed YOLO-T algorithm in this paper is verified by analyzing the target detection results under the same scenes both in a water tank in the laboratory and the Huanghai sea in China.

To detect the target using the YOLO-T algorithm and the other versions of YOLO algorithm, the first step is to collect datasets and to conduct network training. In this paper, a blue circle is selected as the target. Considering the simple classification of detected target, 80 target images collected in the laboratory and 400 target images collected in the Huanghai sea are selected, and the number of images are then extended to 880 through data augmentation operation as the data set. Where, the training set, the verification set and the test set contain 634, 158 and 88 images respectively, and the corresponding proportions of the whole data set are 0.72, 0.18 and 0.1. After 1000 epochs of training, the weight with the best mAP is selected as the target detection weight of the YOLO-T algorithm and the series of YOLO algorithms. Since the mean square error of YOLOv5 detection results is 1.6069 pixels, the number of extension pixels used in YOLO-T to select the rectangular area target is 5, which can ensure sufficient detection error range without increasing the amount of calculation.

In order to evaluate the rapidity of the YOLO-T algorithm, the model parameters are compared first, as shown in Table II. The number of model parameters and the weight of the model obtained by training of YOLO-T algorithm are about one-third of that of YOLOv5. Meanwhile, the average detection time of YOLO-T is also less than that of YOLOv5 under the same CPU processor, which is conducive to the rapid detection of underwater targets.

TABLE II
COMPARISONS OF TRAINED MODELS

	YOLO-T	YOLOv5
Amount of network parameters	$2.54 * 10^6$	$7.25 * 10^6$
Size of model weight	$5.12M$	$14.0M$
Mean time to detect	$0.110s$	$0.136s$

TABLE III
COMPARISONS OF THE TARGET DETECTION RESULTS

	MSE of water tank tests / $pixel$	MSE of huanghai sea tests / $pixel$
YOLOv3	1.8250	1.8362
YOLOv4	1.3220	2.2401
YOLOv5	1.4431	1.6069
YOLOv6	1.5318	1.8548
YOLOv7	1.2068	1.3479
YOLO-T	**0.9283**	**1.2132**

Underwater target detection results are shown in Fig. 6. Since the detection results of several versions of YOLO algorithms have little difference in vision, only the target detection results of YOLOv5 are shown in the figure for comparison with those of YOLO-T. As shown in Fig. 6, the first four columns are the results of the water tank target detection tests and the last four columns are the results of Huanghai sea target detection tests. The detection results of the target images include the surrounding contours of the targets and the confidence of the detection result. It can be seen that the target contours detected by YOLO-T algorithm are more consistent with the real edge of the targets than that detected by YOLOv5 algorithm. To further validate the efficiency of YOLO-T, the quantitative comparisons of the target detection results are shown in Table III.

In Table III, MSE is the mean square error of coordinates between the detected center of the target and the true center of the target.

In the water tank detection tests, YOLO-T algorithm has the smallest MSE of detection, which is only 0.9283 pixels. Among the detection results of several versions of YOLO algorithms, the MSE of YOLOv3 algorithm is the largest while that of YOLOv7 is the smallest. Compared with them, the MSE of YOLO-T algorithm is reduced by $23.1\% - 49.1\%$.

The results of the sea trial in the Huanghai Sea are not quite the same as those of the tank tests. First of all, the detection accuracy of these algorithms in the sea tests decreases to some extent compared with those in the water tank tests, which can be attributed to the poor image quality because of the large

TABLE IV
CONDITIONS OF ACCURACY VALIDATION TESTS

Condition	Specification
water tank	1m * 1.5m * 1m
camera	SENU SW01
image size	1920 pixel *1080 pixel
computer	Legion R7000P2020H/AMD/32GB RAM

number of suspended particles and the dim light in marine environment. Secondly, YOLOv4 algorithm has the largest detection error than that of the other versions of YOLO algorithms. Nevertheless, the detection error of the YOLO-T algorithm is still smaller than that of the YOLO series, reducing by about $10\% - 45.8\%$.

The above results show that the YOLO-T algorithm proposed in this paper not only greatly reduces the detection time compared with YOLOv5, but also has higher accuracy than the YOLO algorithm of these versions.

C. Target Positioning Accuracy Analysis

To validate the accuracy of the target positioning algorithm in this paper, positioning tests are conducted in a water tank. In the positioning tests, the camera is fixed to the end of the tank and the underwater target is moved in the direction of x, y, and z to validate the accuracy in the position aspect. Then the target is rotated around the X, Y, and Z axes of the camera coordinate system respectively to validate the accuracy in the attitude aspect. The conditions of the tests are shown in Table IV.

Due to the influence of camera definition, range of camera field of view and underwater environment, the area where underwater target can be located is limited. Therefore, it is necessary to determine the effective range of target positioning through effective field of view estimation. In this paper, the positioning range of x and y is $(-1.25\,\mathrm{m}, 1.25\,\mathrm{m})$, and the positioning range of z is $(0.3\,\mathrm{m}, 2\,\mathrm{m})$. In the aspect of attitude of the target, the positioning range of ψ is $(0, 360\,^\circ)$, the positioning range of θ and the positioning range of ϕ are both $(-60\,^\circ, 60\,^\circ)$.

1) Position Accuracy Analysis: This section tests the position accuracy in the direction of x, y, and z, respectively. The positioning results are shown in Table V.

The positioning error of x is shown in Fig. 7(a) in red line, and the error of y and z is shown in Fig. 7(b) and (c) in green line and blue line respectively. The minimum positioning error of x is 0.01 m, the maximum error is 0.05 m, and the average error is 0.025 m. The minimum positioning error of y is 0, the maximum error is 0.02 m, and the average error is 0.009 m. The minimum error of z is 0, the maximum error is 0.06 m, and the average error is 0.023 m. During the tests in water tank, due to the limitation of water tank materials, the water depth in the water tank is limited to a certain extent, which makes the positioning test range of y is smaller than that of x.

2) Attitude Accuracy Analysis: This section changes the relative angles between the underwater target and the Z, X and Y axes of the camera, i.e., ψ, θ and ϕ, to conduct the attitude accuracy verification. The positioning results are shown in Table VI.

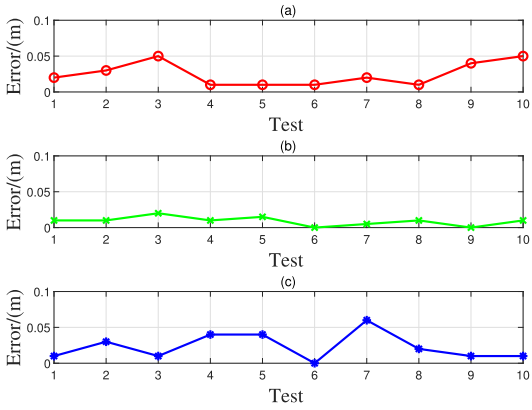


Fig. 7. The position accuracy tests.

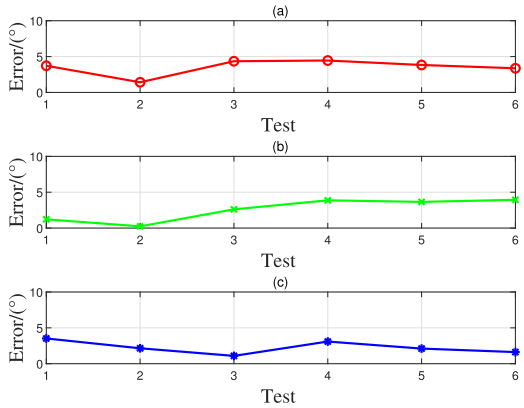


Fig. 8. The attitude accuracy tests.

The positioning error of ψ is shown in Fig. 7(a) in red line, and the error of θ and ϕ is shown in Fig. 7(b) and (c) in green line and blue line respectively. The attitude angle errors in the positioning tests results are analyzed respectively as shown in Fig. 8. In the aspect of ψ, the minimum error is $1.42\,^\circ$, the maximum is $4.45\,^\circ$, and the average error is $3.52\,^\circ$. The minimum positioning error of θ is $0.24\,^\circ$, the maximum error is $3.94\,^\circ$, and the average error is $2.59\,^\circ$. The minimum error of ϕ is $1.09\,^\circ$, the maximum is $3.52\,^\circ$, and the average is $2.26\,^\circ$.

3) Sea Trial: The proposed vision-based underwater target detection and positioning approach has also been validated in the Huanghai sea in China. The experiments are carried out in the ocean environment 2 meters from the shore and 5 meters from the sea surface. The underwater target in the sea trail is a black board, which is held by two vertical rods and moves in accordance with the experimenters' movements or oscillates in small amplitude with the waves in the ocean. The underwater camera captures images at a frame rate of 12 fps and the detected results during the movement is shown as Fig. 9.

The duration of the test is about 91 s. In the first half of the test, the test personnel mainly swing the target board from side to side, which indicates the fluctuations mainly occurred in the x

TABLE V
POSITIONING TESTS IN THE DIRECTION OF x, y, z

Position	x									
True value(m)	-0.85	-0.75	-0.65	-0.55	-0.45	0	0.45	0.55	0.65	0.75
Calculated value(m)	-0.83	-0.72	-0.7	-0.56	-0.44	0.01	0.47	0.56	0.69	0.8
Position	y									
True value(m)	-0.55	-0.35	-0.25	-0.15	-0.075	0	0.075	0.15	0.25	0.35
Calculated value(m)	-0.54	-0.34	-0.23	-0.14	-0.06	0	0.08	0.16	0.25	0.36
Position	z									
True value(m)	0.3	0.5	0.7	0.9	1.1	1.3	1.5	1.7	1.9	2.0
Calculated value(m)	0.31	0.53	0.71	0.94	1.14	1.3	1.56	1.72	1.89	2.01

TABLE VI
POSITIONING TESTS IN THE DIRECTION OF ψ, θ, ϕ

Position	ψ						θ		
True value(°)	315	330	360(0)	30	45	60	-45	-30	0
Calculated value(°)	318.72	331.42	364.35(4.35)	34.45	48.83	63.37	-46.22	-30.24	2.61
Position	θ			ϕ					
True value(°)	30	45	60	-60	-45	0	30	45	60
Calculated value(°)	33.87	48.65	63.94	-63.52	-42.85	-1.09	33.09	47.11	61.62

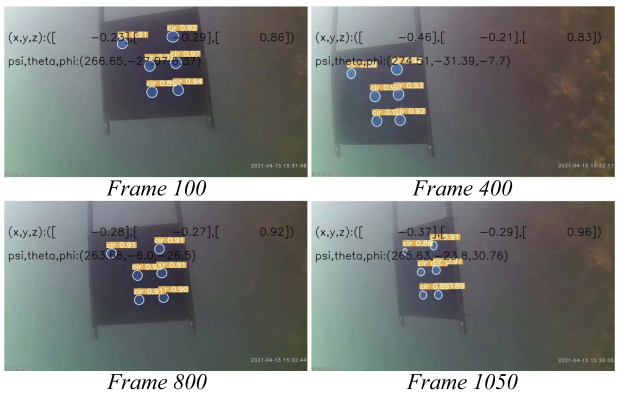


Fig. 9. Ocean experimental detection.

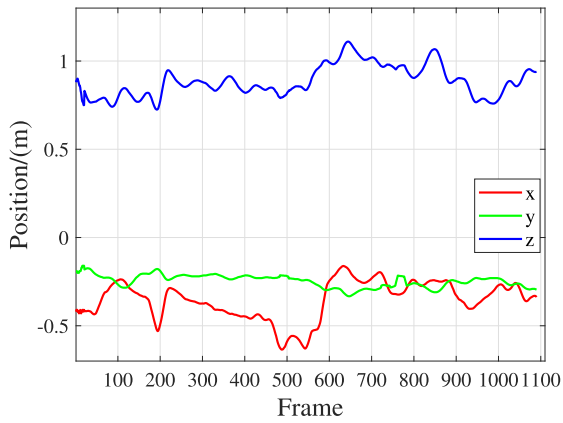


Fig. 10. Time evolution of position of the underwater target.

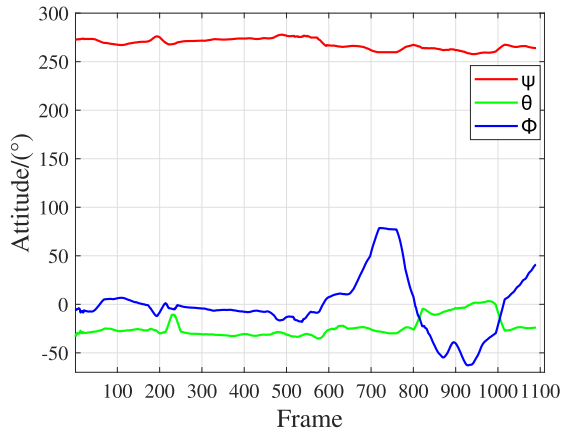


Fig. 11. Time evolution of attitude of the underwater target.

direction in the target position. However, since the two vertical rods are held by a human, it is inevitable that the target position in the y and z directions will fluctuate during the swing process. In the second half of the test, the test personnel mainly rotates the target around the Z axis of the camera. Accordingly, the roll angle ϕ of the target should change significantly, while the yaw angle ψ and pitch angle θ should only fluctuate within a small range.

The curves in Figs. 10 and 11 represent the evolution process of the position and attitude of the underwater target contain 1090 frames of data. The data of target from the positioning algorithm are filtered by KF algorithm to provide stable continuous positioning information. The filtered data will be further applied to underwater missions to grasp underwater target.

As shown in Fig. 10, the vertical distance z of the target from the camera fluctuates within the range of $(0.7\,m, 1.1\,m)$, while the longitudinal deviation y only changes slowly within the range of $(-0.3\,m, -0.2\,m)$. In terms of the attitude of the target, since the two vertices of the target board are fixed by the vertical rods, the yaw angle ψ of the target should be stabilized near a fixed value in each experiment. Compared with the setting of the target image shown in Fig. 9 and the target coordinate system, the yaw angle value in this experiment should be 270°, which is consistent with the red curve in Fig. 11. The pitch angle stabilized at $-30°$ in the early stage of the experiment, changes to near 0 between Frame 800 and Frame 1000, and then returns to $-30°$ finally. By comparing the video data collected, the target

(a) Metal shelf used to fix camera and target

(b) Target lifting scene

Fig. 12. The experimental scene in the tests.

board swings forward in the rotation process between Frame 800 and Frame 1000, and the distance z between the target and the camera also decreases correspondingly at the same time.

It can be seen from the change curves of the position and attitude of the target, the position data x of the target changes greatly before Frame 600, while the attitude data during this period of time fluctuates a little. After Frame 600, the target's roll angle ϕ changes $(-60\,^{\circ}, 60\,^{\circ})$, while the target's position data alters only in a small range. These phenomena are consistent in the experimental process, which proves the effectiveness of the target positioning algorithm.

D. Target Detection and Positioning Stability Analysis

In the actual underwater applications, the target detection and positioning should not only be accurate, but also have continuous stability to ensure the safe navigation of the ROV and the completion of underwater missions. Therefore, several tests are conducted in an anechoic tank to demonstrate the stability of the proposed vision-based underwater target detection and positioning approach. The performance of the approach is verified by analyzing the position and attitude change curve of the fixed target in a period of time.

The experimental scene is shown in Fig. 12. A metal shelf used to hold underwater targets in the tests is shown in Fig. 12(a). The camera is fixed at the top of the shelf to ensure the field of view. The overall size of the shelf is $2\,\mathrm{m} * 1\,\mathrm{m} * 2\,\mathrm{m}$, and extra metal rods can be flexibly installed in the middle of the shelf to adjust the position of the target. As shown in Fig. 12(b), during the tests, the shelf was dropped into the anechoic tank by a crane.

Before each test, the true value of the positioning result can be obtained by measuring the mounting position of the target. However, there may be a certain tilt in the camera installation process, it is impossible to get the accurate pose truth value of the target relative to the camera through simple measurement. In this case, the rigid-body coordinate transformation relation is employed to calculate the position and attitude of the target in the medal shelf coordinate system whose true value can be easily measured.

$$\begin{aligned} P_C &= R_{CS} \cdot P_S + T_{CS} \\ P_C &= R_{CT} \cdot P_T + T_{CT} \end{aligned} \tag{6}$$

Where, S, C, T is the shelf coordinate, the camera coordinate and the target coordinate, respectively. For an arbitrary point on target, P_S, P_C, P_T are the corresponding coordinates in the three coordinate systems. T_{CS} and R_{CS} represent the the position and attitude of the camera relative to the metal shelf, which are fixed values and are easily to be calculated through calibration in advance; T_{CT} and R_{CT} are the pose position and attitude of the camera relative to target, which can be extracted as illustrated in Section III. From this, the position and attitude of target relative to the shelf are shown as follows:

$$\begin{aligned} T_{CT} &= R_{CS}^{-1} \cdot (T_{CT} - T_{CS}) \\ R_{ST} &= R_{CS}^{-1} \cdot R_{CT} \end{aligned} \tag{7}$$

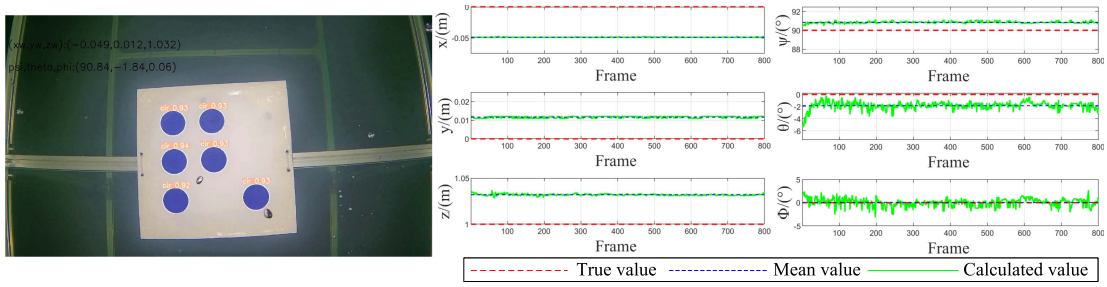
In these experiments, the feature points were patterned on a $0.3m * 0.3\,m$ white board. The specific test process is as follows: (1) The board is fixed in a certain place in the metal shelf, and its position information is measured; (2) Lift the metal shelf into the water and keep it for a period of time to record the change curve of the target position and attitude; (3) Change the board position and repeat the first two steps.

In order to prove the effectiveness of the target positioning, experiments were carried out under different distances, different positions and the scene with missing feature points. Some of the results are shown in Figs. 13–17.

Figs. 13, 14 and 15 are the test results when the vertical distance between the target board and the top of the metal shelf is 1 m, 1.6 m and 2 m respectively. In the figure, the green curve is the calculated value of the target position (attitude), the blue curve is the average value of the calculated value, and the red curve represents the true value of the position (attitude). Fig. 16 shows the detection and positioning results of the target when a feature point is missing. In this test, the distance between the target and the top of the shelf is 2 m. The approach proposed in this paper can identify the target and sort the 5 feature points to calculate the pose of the target. Fig. 17 shows the test results when two target feature points are missing, from which it can be seen that there is just little difference between the calculated results and the real values.

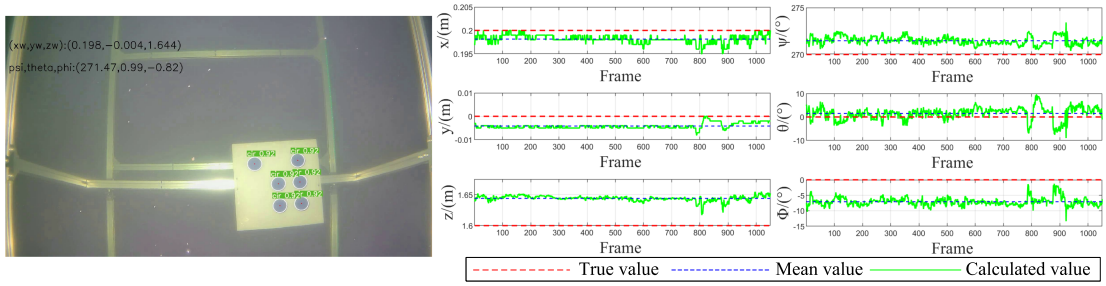
In Fig. 13, the calculated position curved of the target board fluctuate slightly, but the pitch angle θ tends to rise in the early stage. By comparing the video records obtained, it can be found that the target board is not stable in the first few seconds after entering the water, and oscillates with the water pressure in the aspect of θ. Once the target plate is in equilibrium with the water pressure, the calculated attitude data gradually remains stable.

In Fig. 14, the position curves and attitude curves of the target board change around Frame 750, which is caused by the crane lifting at this moment. Since the connection between the metal rod and the metal frame is not fixed at $z = 1.6\,\mathrm{m}$, the metal rod moves in a small range during the lifting process, which leads to the change of the target pose.



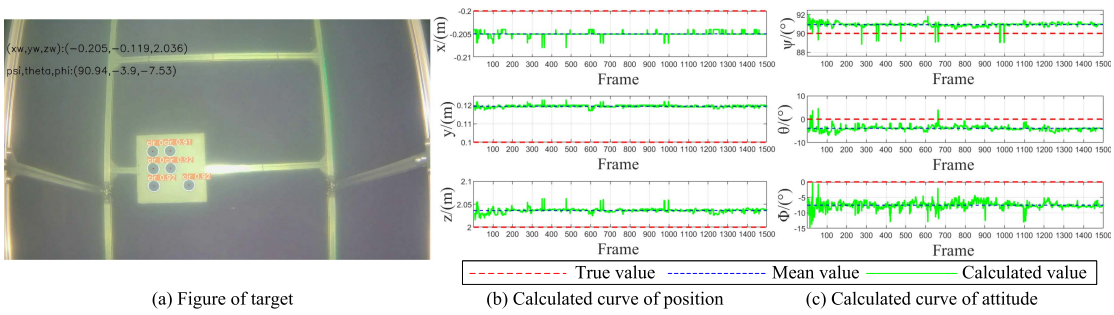
(a) Figure of target (b) Calculated curve of position (c) Calculated curve of attitude

Fig. 13. Target positioning stability test: $z = 1$ m.



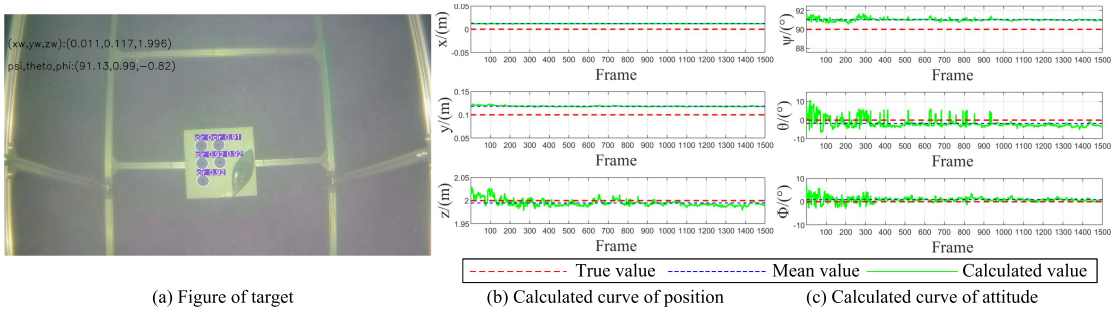
(a) Figure of target (b) Calculated curve of position (c) Calculated curve of attitude

Fig. 14. Target positioning stability test: $z = 1.6$ m.



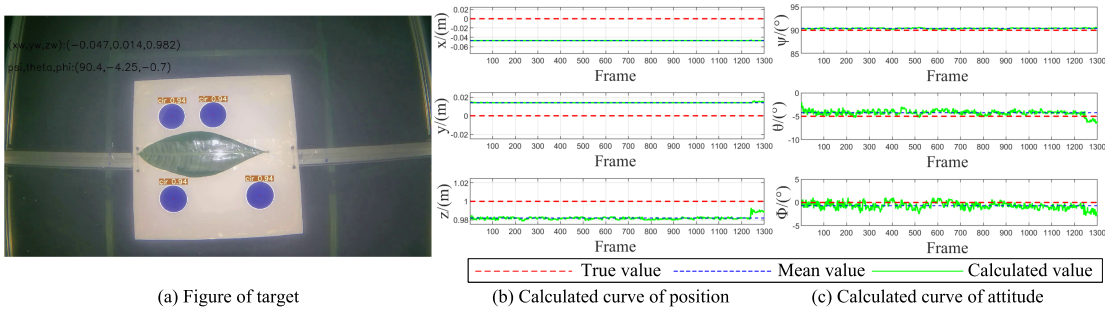
(a) Figure of target (b) Calculated curve of position (c) Calculated curve of attitude

Fig. 15. Target positioning stability test: $z = 2$ m.



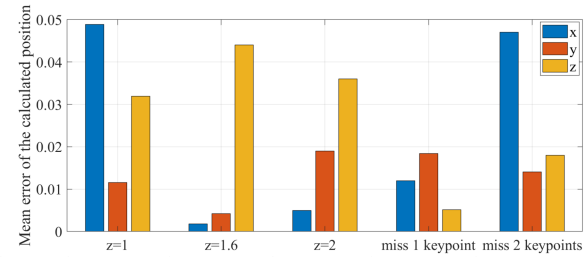
(a) Figure of target (b) Calculated curve of position (c) Calculated curve of attitude

Fig. 16. Target positioning stability test: miss 1 feature point.



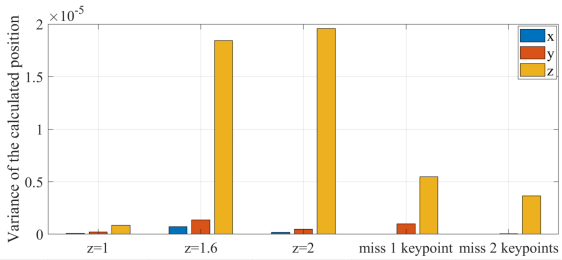
(a) Figure of target (b) Calculated curve of position (c) Calculated curve of attitude

Fig. 17. Target positioning stability test: miss 2 feature points.



True value	(0, 0, 1)	(0.2, 0, 1.6)	(-0.2, 0.1, 2)	(0, 0.1, 2)	(0, 0, 1)
x	0.0489	0.0018	0.0050	0.0120	0.0470
y	0.0116	0.0042	0.0190	0.0184	0.0141
z	0.0319	0.0440	0.0360	0.0051	0.0180

Fig. 18. Mean error of the calculated target position.

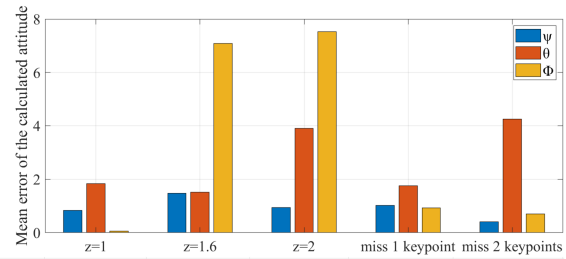


True value	(0, 0, 1)	(0.2, 0, 1.6)	(-0.2, 0.1, 2)	(0, 0.1, 2)	(0, 0, 1)
x	8.6514*10^(-8)	7.3804*10^(-7)	1.8289*10^(-7)	2.3918*10^(-8)	5.3393*10^(-9)
y	2.3520*10^(-7)	1.3489*10^(-6)	4.7536*10^(-7)	1.0025*10^(-6)	5.7785*10^(-8)
z	8.4272*10^(-7)	1.8425*10^(-5)	1.9593*10^(-5)	5.4760*10^(-5)	3.6433*10^(-6)

Fig. 19. Variance of the calculated target position.

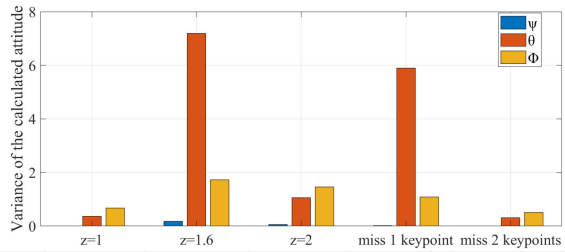
The first 100 frames of data shown in Fig. 15 also fluctuate greatly compared with the following data which appears because in the early stage of the test, the LED has not be turned on and the target images collected are green, which results in the failure to extract the feature circle during target detection and further affect the positioning results.

The reason for the frequent change in the early data in Fig. 16 is similar to the reason of test at $z = 2$. As shown in Fig. 17, in the test with two missing feature points, the target board fluctuates at the end, leading to changes in the calculated results. It can be seen from Fig. 18 that the mean error between the calculated target position and the true value of the target position are no more than 0.05 m in all tests and the variance of the calculated values tend to be 10^{-5} orders of magnitude as shown in Fig. 19. As illustrated in Fig. 20 , the average error bars of target attitude calculation show that the test results at $z = 1.6$ m and $z = 2$ m are worse than those in the other three groups. The main reason for this is the fluctuation of data caused by environmental influences during the tests. As for the variance of attitude data, the variance of pitch angle tested at 1.6 m and scene with a feature point missing is more than twice of other variances, which can also be seen in the data curves in Figs. 14 and 16. In these two sets of data, the range of angle change is about 20°, and the frequency of fluctuation data appear more than other data. Except for these two sets of data, the variance of the calculated values of other attitude is all within 2, which can prove the stability of target detection and positioning method.



True value	(90, 0, 0)	(270, 0, 0)	(90, 0, 0)	(90, 0, 0)	(90, -5, 0)
ψ	0.8412	1.4700	0.9400	1.0208	0.4034
θ	1.8391	1.5137	3.9000	1.7575	4.2517
Φ	0.0636	7.0824	7.5300	0.9226	0.7025

Fig. 20. Mean error of the calculated target attitude.



True value	(90, 0, 0)	(270, 0, 0)	(90, 0, 0)	(90, 0, 0)	(90, -5, 0)
ψ	0.0094	0.1864	0.0614	0.0172	0.0062
θ	0.3655	7.1883	1.0580	5.8974	0.3165
Φ	0.6701	1.7230	1.4666	1.0883	0.5201

Fig. 21. Variance of the calculated target attitude.

V. Conclusion and Future Work

In this paper, an underwater target detection and positioning approach with a monocular camera is proposed. The proposed approach is composed of an underwater target detection algorithm YOLO-T and a target positioning algorithm. Firstly, we modify the structure of YOLOv5 algorithm using Ghost module and SE attention module to improve the calculation time of target detection. Secondly, a series of image processing operations are performed on the improved YOLOv5 detection results to increase the detection accuracy. Thirdly, a cooperative marker is designed as the artificial underwater target, and the corresponding positioning algorithm is presented to calculate the position and attitude of the target according to the geometric information of the designed marker. We conduct water tank tests and Huanghai sea tests to verify the accuracy of target detection and target positioning separately. Finally, the stability performance of the proposed detection and positioning method is demonstrated through the pool tests.

In the future research, the network structure of the YOLO-T target detection algorithm will be adjusted according to the features of the underwater artificial target to improve the detection results. Meanwhile, research efforts will also extend the target detection and positioning approach for the real-time tracking of underwater cooperative targets in different underwater missions. At the same time, the lack of feature points does not lead to poor target positioning result in terms of position.

References

[1] Y. Shen et al., "Rapid detection of camouflaged artificial target based on polarization imaging and deep learning," *IEEE Photon. J.*, vol. 13, no. 4, pp. 1–9, Aug. 2021.

[2] J. Yang et al., "A deep learning-based surface defect inspection system using multiscale and channel-compressed features," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 8032–8042, Oct. 2020.

[3] C. Xia and H. Zhang, "Unsupervised salient object detection by aggregating multi-level cues," *IEEE Photon. J.*, vol. 10, no. 6, pp. 1–11, Dec. 2018.

[4] S. Cui, Y. Wang, S. Wang, R. Wang, W. Wang, and M. Tan, "Real-time perception and positioning for creature picking of an underwater vehicle," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 3783–3792, Apr. 2020.

[5] A. Elibol, J. Kim, N. Gracias, and R. Garcia, "Efficient image mosaicing for multi-robot visual underwater mapping," *Pattern Recognit. Lett.*, vol. 46, pp. 20–26, 2014.

[6] J. Shen, Z. Xu, Z. Chen, H. Wang, and X. Shi, "Optical prior-based underwater object detection with active imaging," *Complexity*, vol. 2021, 1–12, 2021.

[7] K. Srividhya, "Intelligent object recognition in underwater images using evolutionary-based Gaussian mixture model and shape matching," *Signal Image Video Process.*, vol. 14, no. 5, pp. 877–885, 2020.

[8] F. Sun, J. Yu, S. Chen, and D. Xu, "Active visual tracking of free-swimming robotic fish based on automatic recognition," in *Proc. IEEE 11th World Congr. Intell. Control Automat.*, 2014, pp. 2879–2884.

[9] M. F. Yahya and M. R. Arshad, "Tracking of multiple markers based on color for visual servo control in underwater docking," in *Proc. IEEE 5th Int. Conf. Control Syst., Comput. Eng.*, 2015, pp. 482–487.

[10] D. Ji, H. Li, C. W. Chen, W. Song, and S. Zhu, "Visual detection and feature recognition of underwater target using a novel model-based method," *Int. J. Adv. Robot. Syst.*, vol. 15, no. 62018, Art. no. 1729881418808991.

[11] A. Nikolovska, "AUV based flushed and buried object detection," in *Proc. IEEE OCEANS*, 2015, pp. 1–5.

[12] D. Lee, G. Kim, D. Kim, H. Myung, and H. T. Choi, "Vision-based object detection and tracking for autonomous navigation of underwater robots," *Ocean Eng.*, vol. 48, pp. 59–68, 2012.

[13] G. J. Hou, X. Luan, D. L. Song, and X. Y. Ma, "Underwater man-made object recognition on the basis of color and shape features," *J. Coastal Res.*, vol. 32, no. 5, pp. 1135–1141, 2016.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 779–788.

[15] W. Liu et al., "SSD: Single shot multiBox detector," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[17] J. Dai and K. R. F. Li, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, vol. 29, pp. 379–387.

[18] L. Ji -yong, Z. Hao, H. Hai, Y. Xu, W. Zhaoliang, and W. Lei, "Design and vision based autonomous capture of sea organism with absorptive type remotely operated vehicle," *IEEE Access*, 2018, vol. 6, pp. 73871–73884, 2018.

[19] Z. Chen, Z. Zhang, F. Dai, Y. Bu, and H. Wang, "Monocular vision-based underwater object detection," *Sensors*, vol. 17, no. 8, 2017, Art. no. 1784.

[20] P. Trslic et al., "Vision based autonomous docking for work class ROVs," *Ocean Eng.*, vol. 196, 2020, Art. no. 106840.

[21] K. Holak, P. Cieslak, P. Kohut, and M. Giergiel, "A vision system for pose estimation of an underwater robot," *J. Mar. Eng. Technol.*, vol. 21, no. 4, pp. 234–248, 2020.

[22] T. Maki, R. Shiroku, Y. Sato, T. Matsuda, T. Sakamaki, and T. Ura, "Docking method for hovering type AUVs by acoustic and visual positioning," in *Proc. IEEE Int. Underwater Technol. Symp.*, 2013, pp. 1–6.

[23] Y. Li, Y. Jiang, J. Cao, B. Wang, and Y. Li, "AUV docking experiments based on vision positioning using two cameras," *Ocean Eng.*, vol. 110, pp. 163–173, 2015.

[24] S. Ghosh, R. Ray, S. R. K. Vadali, S. N. Shome, and S. Nandy, "Reliable pose estimation of underwater dock using single camera: A scene invariant approach," *Mach. Vis. Appl.*, vol. 27, no. 2, pp. 221–236, 2016.

[25] Y. Deng and H. Wang, "Underwater circular object positioning system based on monocular vision," in *Proc. IEEE 19th Int. Symp. Signal Process. Inf. Technol.*, 2019, pp. 1–5.

[26] K. N. Lwin et al., "Visual docking against bubble noise with 3-D perception using dual-eye cameras," *IEEE J. Ocean. Eng.*, vol. 45, no. 1, pp. 247–270, Jan. 2020.

[27] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[28] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1580–1589.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[30] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2010.