IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 7, 2020, accepted October 5, 2020, date of publication October 8, 2020, date of current version October 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029627

A Novel Visual Measurement Method for Three-Dimensional Trajectory of Underwater Moving Objects Based on Deep Learning

TAO LIU[1], NINGNING WANG[1], LEI ZHANG[2], SHANGMAO AI[2], AND HONGWANG DU[1]
[1]College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China
[2]College of Shipbuilding Engineering, Harbin Engineering University, Harbin 150001, China

Corresponding authors: Lei Zhang (cheung103@163.com) and Ningning Wang (1009930623@qq.com)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 3072020CFT0102, in part by the Sub-Project of National Major Scientific and Technological Special Project, Research on Hydrodynamic Analysis and Model Experiment Technology of Tender Assist Drilling (TAD) Multi Floating Body, under Grant 2017ZX05032-005-005, in part by the China National Offshore Oil Corporation (CNOOC) Research Center, in part by the National Natural Science Foundation of China Project under Grant 51409059, and in part by the Heilongjiang Postdoctoral Research Start-Up Foundation Project under Grant LBH-Q16066.

ABSTRACT For deep sea equipment suspension installation used in marine engineering, the multi-camera video motion analysis method is used to calculate the three-dimensional underwater trajectory of the underwater engineering structure. Considering difficulty in underwater modeling caused by the problems of light scattering and refraction under water, the camera imaging model on land is no longer applicable in water, and a new underwater camera imaging model needs to be proposed. This paper introduces an underwater camera imaging model with light refraction, studies the calibration method of the internal and external parameters of the underwater camera, and improves the multiscale rotation dense feature pyramid convolutional neural network to detect the position of the target object in the image. The underwater motion videos of the target produced by three fixed underwater cameras are optimized to fuse and calculate the trajectory of the underwater target. This method is suitable for large-scale motion of underwater objects and can obtain more accurate trajectories. Experimental analysis and data comparison have verified the effectiveness of the method.

INDEX TERMS Deep sea equipment installation, underwater camera calibration, three-dimensional trajectory calculation, BA optimization, multiscale rotation dense feature pyramid networks.

I. INTRODUCTION

Marine engineering equipment is a high-input and high-risk product. With the continuous development of marine engineering equipment, the demand for marine target detection has become a hot research topic. In the past few decades, underwater equipment installation methods have been widely used. For underwater equipment installation on different occasions, people pay more and more attention to its stability and high precision, and this is an urgent problem that needs to be solved for marine engineering equipment. Fortunately, the detection direction of underwater targets has been greatly developed, for example using methods based on acoustic sensors [1]–[4], lidar [5]–[8], geometric scattered waves [9]–[11], sonar pictures [12]–[15], hybrid particle filter tracking [16], wavelet transform [17], [18] and others to detect underwater targets.

These methods can detect underwater target objects to a large extent, but because they rely too much on the information provided by the sensors, the results obtained are unstable, and the performance of different sensors in complex underwater environments will be affected. As a result, their accuracy requirement cannot be fully met. Compared with the environment in the air, the dynamics, unstructural characteristics, backscattering of impurities in the water, and light attenuation in seawater, especially in the deep-sea environment, seriously hinder exploration of the ocean. Therefore, using the image captured by the camera and using computer vision to conduct underwater detection has increasingly become a research hotspot. Underwater vision imaging is an effective method for detecting underwater environments. Great progress has been made in the analysis of target movement

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Zhang.

on land, but the analysis of underwater target movement still faces many difficulties. For example, there are many suspended matters and particles in the underwater environment. The light is scattered under water and the underwater image is degraded. Another major obstacle is the image error caused by the refraction of light [19]. How to accurately detect underwater targets in blurred underwater images is also very challenging. PSYCHARAKIS [20] pre-made a calibration frame, based on several markers and used the DLT [21] equation to obtain its 3D coordinates to perform a three-dimensional swimming analysis, but the DLT algorithm did not consider the refractive coordinate system. KWON [22] proposed considering both refraction and DLT methods, that is, the pixel coordinates used to calculate the refraction point after considering refraction based on Snell's law. ALEXANDRI and DIAMANT [23] proposed an autonomous underwater vehicle (AUV) using an inertial navigation system (INS) and a simultaneous localization and mapping (SLAM) method. The method needs to rely on plenty of external environmental imaging to provide the required ''nodes''. Shen *et al.* [24] used a binocular camera to detect underwater target objects. By improving the Harris operator, the camera can be correctly modeled in high-scattering underwater environments, but this method depends on the clarity of the pictures taken. In addition, the final results need to be optimized for computer vision methods, so these methods have great limitations when used underwater. High-precision target detection in an underwater environment with low illumination and blurred imaging effects is a great challenge. The detection cannot be reliably done relying on traditional feature corner extraction methods. Significant progress has been made in research on the target detection methods based on deep learning in terms of detection accuracy and speed, such as RCNN, SSD, YOLO and others. Some scholars have used deep learning-based target detection and tracking methods to achieve good results for special applications in complex scenarios. Stewart *et al.* [25] made a great contribution to stereo tracking 3D position by studying the connection between kernel-based algorithms and traditional template tracking methods; in order to maintain the speed and accuracy of YOLO, Redmon and Farhadi [26] introduced the YOLOv3 network framework for training and recognition; Kaiming [27] improved the Faster R-CNN framework and introduced the Mask R-CNN framework to effectively detect objects in images; Zhou *et al.* [28] proposed a deep alignment network to effectively solve the misalignment (that is, excessive background and partial loss) and occlusion problems of the detector; Chen *et al.* [29] used a fully convolutional network (FCN)), and proposed a simple and effective visual tracking framework; Zhou *et al.* [30] proposed a novel fine-grained spatial alignment model (FGSAM) to discover fine-grained local information and effectively deal with complex challenging scenes such as pose, inaccurate detection, occlusion and misalignment; Zhong *et al.* [31] proposed a hierarchical tracker to solve the problem of low search efficiency and reduced tracking performance. This paper aims at the detection of special underwater targets, and it is necessary to propose a special non-axis alignment frame for recognition. By building a deep learning detection framework for non-axis alignment frames, we determine the pixel position of the target object in the picture according to the learned model parameters, which effectively prevents the disturbance caused by the irregular self-motion of underwater objects.

In this paper, to simulate the installation process of large-scale offshore oil and gas production equipment based on the deep-sea suspension method, to study the equipment movement characteristics under different layout parameters, the relevant underwater three-dimensional motion trajectory measurement test was carried out in the pool environment, using multiple underwater cameras to produce motion videos to calculate the target's motion trajectory. Previously, relevant research has been done and a paper has been published in the Journal of Graphics [32]. This article will consider issues related to refraction, target recognition, and optimization in camera modeling. The contributions are as follows:

1) We introduce a camera imaging model in the event of refraction. Compared with literature [33], the model is simpler and requires fewer intermediate variables to be calculated, which makes the parameters obtained by the camera calibrated underwater more accurate.
2) With the refraction coordinate system considered in the camera modeling process, the target object needs to be identified. The traditional recognition method relies too much on the clarity of the image, so a convolutional neural network is built to accurately recognize the target object. Considering special target objects, a multiscale rotation dense feature pyramid neural network is built to detect non-axis-alignment target objects. Compared with literature [34], [35], after passing through the RPN layer, non-maximum suppression is performed on the obtained suggestions, filtering out the suggestions that do not meet the requirements, and providing appropriate regional suggestions for the next stage of the network.
3) Introduce multiple cameras to measure large-scale motion of underwater objects, and add BA optimization links to the measurement results to make the measurement results more accurate.

The rest of this article is as follows: section II discusses the method of underwater camera calibration. Section III deals with the details of building a neural network to identify the target object and obtain the center pixel coordinates. Section IV is about the content of 3D measurement and BA optimization. Finally, section V is the conclusions.

II. UNDERWATER CAMERA CALIBRATION

A. UNDERWATER CAMERA IMAGING MODEL

Underwater camera calibration is the key technology of underwater vision measurement. Due to the refraction and scattering of underwater light, the imaging model of the underwater camera is no longer the pinhole imaging model. When the camera captures an underwater object, the

scattering phenomenon will affect its imaging quality. After acquiring the underwater motion datasets of the object, this paper uses image restoration and image enhancement methods to make deconvolution calculation on the degraded image and the point expansion function in the spatial domain to repair the image [36], which facilitate the later processing of the dataset, and the expected effect is achieved. Since scattering does not affect the physical imaging model of the object, but refraction changes the imaging model of the object on land, the method for studying vision on land is not suitable, and a new imaging model needs to be proposed. TREIBITZ [37] analyzed the effects of refraction on camera calibration. When the light at the far end of the interface extends backwards to the camera, it will not intersect at a point, but form a caustic surface (not compatible with single-view camera model), as shown in Figure 1. Therefore, the camera calibration method for the atmospheric environment is no longer applicable, and the camera underwater imaging model needs to be built.

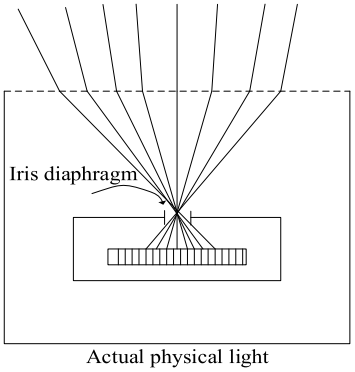


FIGURE 1. Underwater non-single-view camera model.

B. UNDERWATER CAMERA INTERNAL PARAMETER CALIBRATION

The internal parameter calibration technology of the camera for the atmospheric environment is very mature. The existing calibration methods include the TSAI [38] calibration method and the Zhang [39] calibration method. Due to the harsh underwater environment, pictures taken by different types of cameras may have different effects on the final result. If the object trembles during the movement, the image captured by the camera will be blurred and the feature points cannot be identified. The underwater camera used in this article has the characteristics of high-definition signals, minor video delay and Ethernet transmission. The shell material is 316L stainless steel, and the glass sealed camera has strong pressure resistance and corrosion resistance in seawater. This underwater camera can meet engineering needs. The parameters are as follows:

1) RESOLUTION

2592 × 1964, the frame rate is 64 fps.

2) PIXEL SIZE

The horizontal direction is 5μm, and the vertical direction is also 5μm.

3) SENSOR SIZE AND SHUTTER EXPOSURE METHOD

The sensor CMOS size is 1/1.8 inches, and the shutter exposure mode is global.

4) INTERFACE MODE

The camera uses a network interface, and the pictures and videos captured are transmitted to the computer through the local area network.

The internal parameters in the camera are only related to the properties of the camera itself, and will not change in an underwater or atmospheric shooting environment. This paper first uses the Zhang [39] calibration method to calibrate the camera parameters of the three cameras in the air.

Matrix A is the internal parameter matrix of the camera, and the result of camera 1 calibration is

$$\boldsymbol{A} = \begin{bmatrix} 621.5368 & 0 & 320.9654 \\ 0 & 621.7594 & 247.9370 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

Considering the refraction in water, the position of the object point in the underwater image is not the actual position, as shown in Figure 2.

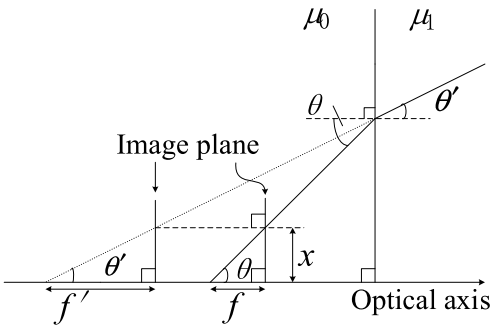


FIGURE 2. Underwater light refraction model.

Due to the refraction of light, the focal length in the imaging model has changed significantly. Figure 3 shows the model of underwater light refraction. The true focal length is f. The light rays refracted after being extended in the opposite direction intersect with the optical axis to obtain a virtual imaging plane and a virtual focal length f'. The angle between the real light in the air and the optical axis is θ, and the angle between the virtual light in the air and the optical axis is θ'. For any point x of the imaging plane:

$$\tan(\theta) = \frac{x}{f} \tag{2}$$

$$\tan(\theta') = \frac{x}{f'} \tag{3}$$

According to the Snell's law and the approximation when θ is small, we can obtain:

$$\frac{f}{f'} \approx \frac{\mu_1}{\mu_0} \tag{4}$$

FIGURE 3. Underwater target.

TABLE 1. Focal length comparison.

Focal length	Air	Underwater	Ratio
f_x	621.54	830.86	1.33
f_y	621.75	830.76	1.33

According to the derivation of LAVEST [40] on the imaging model of the lens in two media with different refractive indexes, the image must be enlarged to a multiple of the refractive index between the two media.

$$n_0(u + du) \tag{5}$$

This article discusses underwater cameras, so the refractive index n_0 of air to water is 1.33. du and du' are pixel distortion values. If the distortion is not considered, it equals the focal length being enlarged by 1.33 times, which is the refractive index of water to air. Then f_x and f_y in the internal parameter matrix calibrated by the Zhang [39] calibration method for the underwater checkerboard image is 1.33 times that in the air.

In this paper, the calibration program of the Zhang [39] calibration method is used to experimentally validate the above-mentioned parameter calibration theory of underwater cameras. 12 underwater checkerboard images of each camera are collected as the input of the calibration program. The length of the checkerboard used for calibration is 25 mm. The image of the underwater target object is shown in Figure 3.

Calibrate the three cameras separately. The calibration result of camera 1 is

$$A = \begin{bmatrix} 830.8608 & 0 & 320.9772 \\ 0 & 830.7591 & 247.9170 \\ 0 & 0 & 1 \end{bmatrix} \tag{6}$$

See Table 1 and Table 2 for the comparison between atmospheric and underwater calibration results of camera 1.

To verify the above-mentioned conclusions, the Zhang [39] calibration method can be used to calibrate the underwater and atmospheric pictures, and the horizontal and vertical focal lengths of the underwater pictures are 1.33 times those of the atmospheric ones. This number happens to be the refractive index of water to air. The principal point coordinates are basically the same in the two cases, which are approximately equal to half of the resolution.

TABLE 2. Comparison of principal coordinates.

Principal coordinates	Air	Underwater
U_0	320.9654	320.9772
V_0	247.9370	247.9170

C. UNDERWATER CAMERA EXTERNAL PARAMETER CALIBRATION

Generally, the optical axis of the camera is perpendicular to the camera lens and passes through the center of the lens of an underwater camera. Matrix A represents the direction vector of the optical axis. The plane where the refraction occurs is defined as α, and the vector perpendicular to α is defined as normal vector n, and the connection between any three-dimensional point P(i) in the air and the optical center of the camera is represented by vector V_0. According to Snell's law, the refracted ray must be coplanar with the incident ray and the normal of the refracting plane. Therefore, the entire optical path should be on α, and the final refracted light should intersect with the axis, as shown in Figure 4.

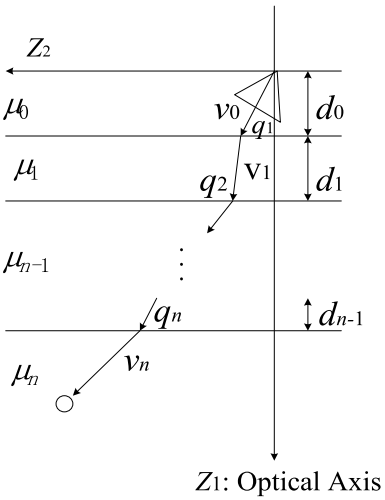


FIGURE 4. Underwater light model.

The point of any three-dimensional world coordinate system converted by the external parameters of the underwater camera is also on the refraction plane, that is:

$$(RP + t)^T (A \times v_0) = 0 \tag{7}$$

where P is any point in the world coordinate system, R and t respectively represent the rotation matrix and translation vector converted from the camera coordinate system to the world coordinate system. It is noteworthy that the coplanar constraints are independent of thickness d of the medium and refractive index μ_i, which depends only on the direction vector of the axis and the camera external parameters (pose).

[A]× is the anti-symmetric matrix of matrix A. The coplanar constraints can be rewritten as:

$$v_0^{\mathrm{T}}(\boldsymbol{A} \times (RP + t)) = \boldsymbol{v}_0^{\mathrm{T}} EP + \boldsymbol{v}_0^{\mathrm{T}} s \tag{8}$$

where E=[A]×R and s=A×t. Matrix E has some similarities with the pose essential matrix E=[t]×R between two frames of the camera. It can use the 5-point method of the essential matrix to decompose and estimate the camera pose R and t to estimate the axis direction vector.

With 8 given points corresponding to the two continuous frames, the above-mentioned equation is expanded to obtain the following linear equation:

$$\underbrace{\begin{bmatrix} P(1)^{\mathrm{T}} \otimes \boldsymbol{v}_0(1)^{\mathrm{T}} & \boldsymbol{v}_0(1)^{\mathrm{T}} \\ \vdots & \vdots \\ P(8)^{\mathrm{T}} \otimes \boldsymbol{v}_0(8)^{\mathrm{T}} & \boldsymbol{v}_0(8)^{\mathrm{T}} \end{bmatrix}}_{\boldsymbol{B}} \begin{bmatrix} \boldsymbol{E} \\ s \end{bmatrix} = 0 \tag{9}$$

where B is an 8 × 12 matrix with a rank of 8; ⊗ is the Kronecker product operation; P(i) is the corresponding three-dimensional coordinate of the *i*th world coordinate system; v_0(i) is the vector connecting the corresponding *i*th point to the optical center of the camera.

The resolution of the image captured by the camera is 2592 × 1964. The experiment uses a plane calibration plate. The z-axis coordinate is 0. Apply them to the above-mentioned coplanar constraint equation to obtain:

$$\begin{bmatrix} P_{(12)}^{\mathrm{T}} & \otimes & v_{(12)}^{\mathrm{T}} & v_{(12)}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} E_{(12)} \\ s_{(12)} \end{bmatrix} = 0 \tag{10}$$

According to the direction of optical axis A of the camera and the Z axis of the world coordinate system, the following formula can be obtained:

$$\boldsymbol{E} = \begin{bmatrix} -r_2 & -r_5 & -r_8 \\ r_1 & r_4 & r_7 \\ 0 & 0 & 0 \end{bmatrix}, \quad s = \begin{bmatrix} -t_2 & t_1 & 0 \end{bmatrix}^{\mathrm{T}} \tag{11}$$

$$\boldsymbol{E}_{(1:2)} = \begin{bmatrix} -r_2 & r_5 \\ r_1 & r_4 \end{bmatrix} \tag{12}$$

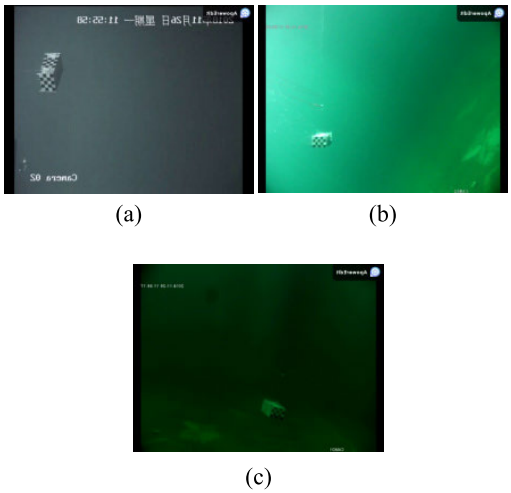
Here r_i and t_i represent the components of R and t respectively. Select the corner world coordinates of 5 or more plane calibration plates, apply them to the above-mentioned formula, get the linear equation, and find the solutions to E and s. The last column of E is solved using the constraint of det(E) = 0 and the unit orthogonality of the rotation matrix.

Since the world coordinate of the plane calibration plate has no Z-axis components, t_3 in the translation vector, that is, the translation component parallel to the Z axis, has not been solved. According to the research results of Agrawal *et al.* [41], the assumption is as follows:

$$t_3 = \alpha \boldsymbol{A} \tag{13}$$

In (13), α is the amplification factor. From the refraction restriction conditions, the following formula can be obtained:

$$\boldsymbol{vp}_1 \times \begin{bmatrix} \boldsymbol{vp}_0 / c_0 & z_p \end{bmatrix} \begin{bmatrix} d_0 \\ \alpha \end{bmatrix} = -\boldsymbol{vp}_1 \times u \tag{14}$$



(a) (b) (c)

FIGURE 5. Video sequence object position. (a) The object captured by the first camera. (b) The object captured by the second camera. (c) The object captured by the third camera.

TABLE 3. Camera external parameters corresponding to different positions.

Position number	Rotation matrix	Translation vector
(a)	$\begin{pmatrix} 0.40065 & 0.15884 & 0.90235 \\ -0.91160 & -0.02966 & 0.40998 \\ 0.09189 & -0.98685 & 0.13292 \end{pmatrix}$	$\begin{pmatrix} -162.41 \\ 111.65 \\ 495.47 \end{pmatrix}$
(b)	$\begin{pmatrix} -0.59700 & 0.37653 & -0.70830 \\ -0.73276 & -0.61538 & 0.29045 \\ -0.32657 & 0.69248 & 0.64330 \end{pmatrix}$	$\begin{pmatrix} -1847.60 \\ 523.03 \\ 4508.60 \end{pmatrix}$
(c)	$\begin{pmatrix} 0.9828 & 0.1757 & -0.0573 \\ -0.1745 & 0.9843 & 0.0253 \\ 0.0608 & -0.0149 & 0.9980 \end{pmatrix}$	$\begin{pmatrix} -1398.50 \\ 113.47 \\ 7978.80 \end{pmatrix}$

vp_i is a two-dimensional vector of v_0; $c_i = vp_i^{\mathrm{T}} z_1$, $z_p = [0; 1]$; u represents the projection vector from the target point in the camera coordinate system to the corresponding refractive layer; d_0 is the vertical distance from the optical center of the camera to the refractive surface. Using two or more matching corner points, get the least square solution of α, and then t_3 can be obtained.

After the underwater cameras shoot the video sequences, the video sequences are intercepted into a picture set with a frame as the unit by the method of image processing, and then 3 positions are selected as shown in Figure 5.

Using the above-mentioned method, the three position cameras in Figure 5 can be obtained, as shown in Table 3.

In order to verify the feasibility of the underwater camera calibration method, this paper uses the same data to calibrate the camera according to the underwater imaging model proposed by Xiaoze *et al.* [33] from China Ocean University, and calculates the reprojection error of the two methods. The reprojection error refers to the difference between the

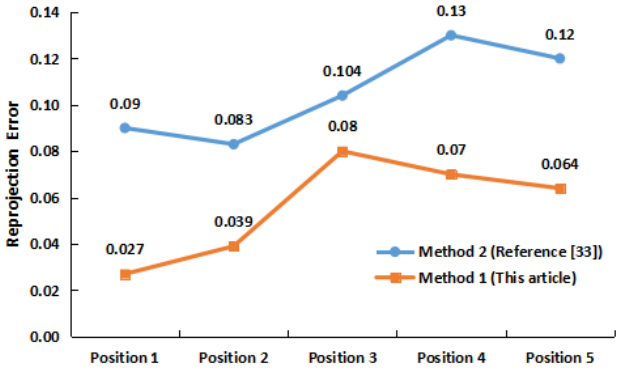


FIGURE 6. Reprojection error comparison.

projection of a real three-dimensional space point on the image plane (i.e. the pixel point on the image) and the reprojection (i.e. the virtual pixel point calculated using the calibration parameter). For a specific position, the differences between the real and virtual coordinates of each corner point on the image plane are accumulated, and finally divided by the total number of accumulated corner points to obtain the reprojection error at this position. In order to explain the problem, five positions of data were selected to calculate the reprojection error values using different methods (the method proposed in this paper is method 1, and the method proposed by Xiaoze *et al.* [33] is method 2), as shown in Figure 6. It can be seen from Figure 6 that the accuracy of the method in this paper has been significantly improved.

III. OBJECT CENTER PIXEL POSITIONING

Through the previous method, we can easily find internal parameter A of the underwater camera and the external parameter corresponding to each picture (rotation matrix R and translation vector T), and then we need to calculate the three-dimensional coordinates of the target object on each picture. The center of the target picture captured by the camera is used as the target point instead of the entire object to describe its trajectory. Although the experiment was carried out in the pool, the poor quality and insufficientillumination caused the target image to be blurred. Accurate detection of the center position of the target picture was the key step to achieving 3D trajectory measurement. It is difficult to use conventional target detection to accurately calculate the center coordinates of the target. With the application of deep learning in target visual detection, significant results have been achieved. In this paper, deep learning is used for target detection. Although the effect of using neural networks to detect target objects is significant, the detection effects of different network models are very different for specific objects. Girshick [42] proposed the Fast R-CNN network framework, which can train and recognize the target object. Although the result is good, it uses the traditional segmentation method for feature map segmentation, resulting in the entire network framework taking too much time compared to other networks. Ren *et al.* [43] introduced the RPN layer on the basis of Fast R-CNN to solve the problem of the traditional feature segmentation taking too much time. However, the recognition frame obtained by the proposed model frame are not all tilting parallel to the edge of the picture. In order to more accurately compute the coordinates of the center point of the checkerboard, it is necessary to perform oblique marking along the movement direction of the target object. Jiang *et al.* [44] proposed the R^2CNN framework to recognize the text in the real-life scene, and the recognition frame was marked according to the tilt direction of the text, in order to meet the requirements of text recognition; Xue *et al.* [34], Yang *et al.* [35] proposed a multi-task rotating region convolution neural network which is used to identify the ships sailing on the sea. When multiple ships appear on the sea port, a multi-task rotating neural network based on dense pyramid is proposed to identify the ships in any direction from the port. Based on this experiment, the deep learning method is used to identify and detect the target surface in each picture, and return the pixel coordinates of the recognition frame obtained through regression. It should be noted that: first, because the world coordinate system is established with the plane where the checkerboard is located as the xoy coordinate system and the direction perpendicular to this plane as the Z axis, the center point obtained by identifying the entire object is not on the checkerboard plane; second, consider directly identifying the checkerboard, but during the actual underwater movement process of the object, the direction of the target surface changes as the object moves. In order to obtain the pixel coordinates of the center of the target surface more accurately, the axis is obtained through classification regression.

The traditional deep learning framework aligned with the recognition frame can no longer meet the needs. In order to solve these problems, this paper proposed an improved CNN algorithm based on the Multiscale Rotation Dense Feature Pyramid Convolutional Neural Network to detect non-axis aligned target surfaces and obtain accurate center pixel coordinates.

A. MODEL ARCHITECTURE

Considering the impact of the rotation of the target surface, a new end-to-end target surface detection deep learning framework consisting of three consecutive parts is constructed: Dense Feature Pyramid Network (DFPN), RPN and Fast-RCNN, the architecture of which is shown in Figure 7.

1) DFPN

For the features learned by the deep learning network, the bottom layer feature semantic information is relatively scarce, but the object position is accurate; by contrast, the learned advanced feature semantic information is rich, but the object position is relatively rough. The bottom position information and high layer semantic information are very important for object detection. The feature pyramid is an effective approach for fusing multi-level information, and it works well for small object detection. The structure based on ResNet is shown in Figure 8, and can be divided into three

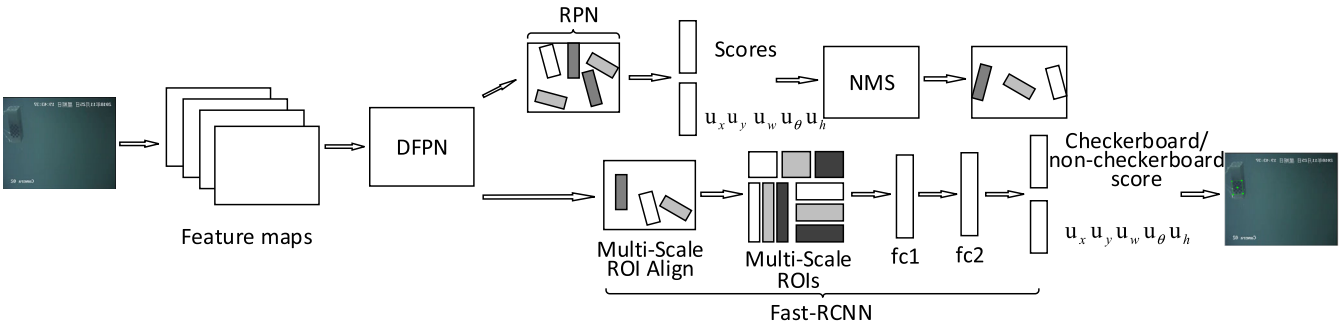


FIGURE 7. Deep learning model architecture based on DFPN.

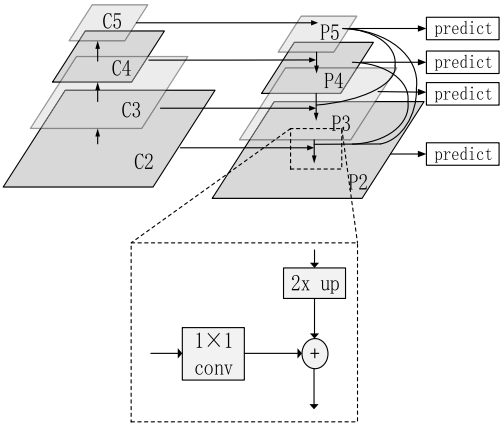


FIGURE 8. DFPN architecture.

parts: bottom-up convolutional neural network (Figure 8 left), top-down process (Figure 8 right) and the side connection between features.

Using ResNet as the backbone of the architecture, the bottom-up part is the forward process of the convolutional neural network. In the forward process, the size of the feature map will change after some layers are passed, but will not change when other layers are passed. The layers that do not change the size of the feature map are classified as a stage, and the features extracted each time are the output of the last layer of each stage {C2, C3, C4, C5}, so that a feature pyramid can be established with a step size of {4, 8, 16, 32} pixels. In order to reduce the number of parameters, the number of channels of all feature maps is set to 256. In the top-down network, horizontal and dense connections {P2, P3, P4, P5} can obtain higher resolution. P5 is 1×1 convolution of C5, Pi ($0<i<5$, i is an integer) uses the nearest neighbor upsampling for all the previous feature maps, and through merging in series, the aliasing effect brought by 3×3 convolutional layer upsampling is eliminated, while reducing the number of channels. DFPN significantly improves the detection performance through feature propagation and feature reuse.

2) RPN

In the RPN stage, in order to realize the detection of rotating objects and improve the recall rate, a method of using rotation anchor points to redefine rotation suggestions is proposed. There are many ways to represent rectangular boxes. This article uses four-point method and five-value method. The four-point method refers to directly using the coordinates of the four vertices of the rectangle, and the five-value method refers to the center point coordinates (x, y), width w, height h and rotation angle θ to express. The schematic diagram of the five-value method (as shown in Figure 9) and the conversion between the five-value method and the four-point method (as shown in Equation 15) are shown below.

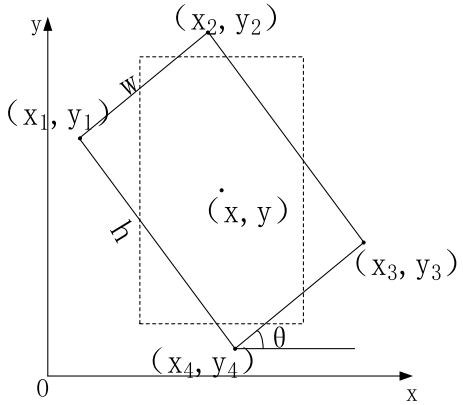


FIGURE 9. Five-valued representation of rotation anchor.

According to the characteristics of the target object checkerboard and the shape in the picture, the checkerboard can be described by using three parameters of scale, ratio and angle. In order to cover the target object more effectively, the anchor ratio is set to {1:1, 2:1, 1:2, 3:1, 1:3}, in the feature

$$
\begin{aligned}
x_1 &= x - \frac{w}{2}\cos\theta - \frac{h}{2}\sin\theta \\
y_1 &= y + \frac{h}{2}\cos\theta - \frac{w}{2}\sin\theta \\
x_2 &= x + \frac{w}{2}\cos\theta - \frac{h}{2}\sin\theta \\
y_2 &= y + \frac{h}{2}\cos\theta + \frac{w}{2}\sin\theta \\
x_3 &= y + \frac{w}{2}\cos\theta + \frac{h}{2}\sin\theta \\
y_3 &= y - \frac{h}{2}\cos\theta + \frac{w}{2}\sin\theta
\end{aligned}
$$

$$x_4 = y - \frac{w}{2}\cos\theta + \frac{h}{2}\sin\theta$$
$$y_4 = y - \frac{h}{2}\cos\theta - \frac{w}{2}\sin\theta \quad (15)$$

map {P2, P3, P4, P5, P6} the scale is set to {50, 150, 250, 350, 500} pixels, and finally the rotation angle of each anchor is set to {15°, 30°, 45°, 60°, 75°, 90°} to include the position of the object at different angles. At that time, each feature point in the feature map will generate 30 (1 × 5 × 6) anchor points, each regression layer will obtain 150 (5 × 30) outputs, and each classification layer will obtain 60 (2 × 30) outputs. In order to provide high-quality regional suggestions for the next stage of the network, the outputs need to be "filtered" to filter out suggestions that do not meet the requirements, so non-maximum suppression is required. In order to further improve the learning of the proposal, a Skew IoU calculation method considering the triangulation is proposed, the anchor with the highest score in the RPN stage is selected to perform NMS non-maximum suppression, and two constraints are set for the NMS: (a) prediction results with IOU>15 shall be discarded; (b) when 0.3≤IOU≤0.5, prediction results with angle difference greater than 15° shall be discarded. In this way, appropriate regional suggestions can be made for the next stage of the network. It should be noted that NMS will also be used to obtain accurate target values after the next stage of prediction.

3) FAST-RCNN

First of all, it is recommended that the regions obtained by the RPN layer be dealt with. In order to expand the feature area and retain the complete feature information, the Multi-Scale ROI Align method is used to obtain a fixed-length feature vector. Use interpolation to obtain the minimum horizontal circumscribed rectangle suggested by the region, and add 3:8 and 8:3 pooling layers to minimize the effects of distortion caused by interpolation. After obtaining the suggestions of the horizontal circumscribed rectangular area, by training the parameters of the fully connected layer model, performing position prediction and classification through regression and classification, the parameter value of the checkerboard represented by the five-value method can be obtained. After the second non-maximum suppression, accurate pixel coordinates of the center point of the target object can be obtained.

B. LOSS FUNCTION

The multi-scale loss objective function is defined as follows:

$$L(P, t, v_i, v_i^*) = \frac{1}{N_{cls}}\sum_i L_{cls}(P_i, t_i) + \lambda\frac{1}{N_{reg}}\sum_i P_i L_{reg}(v_i, v_i^*) \quad (16)$$
$$L_{cls}(P, t) = -\log Pt \quad (17)$$
$$L_{reg}(v, v^*) = smooth_{L_1}(v^* - v) \quad (18)$$
$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & if\ |x| < 1 \\ |x| - 0.5, & otherwise \end{cases} \quad (19)$$

where p represents the probability distribution of the category, t represents the parameterized coordinate vector represented by the predicted five-value method, $L_{cls}(P, t) = -\log Pt$ represents the logarithm of the true value, $v = (v_x, v_v, v_w, v_\theta)$ represents the predicted value, and $v^* = (v_x*, v_v*, v_w*, v_\theta*)$ represents the deviation of the true value from the anchor. λ is the proportionality coefficient between the regression loss function and the classification loss function. Here classification and regression play an equally important role, so λ = 1.

The five-parameter equation that defines a five-value method to represent a rectangular frame is shown in the following formula.

$$t_w = \log(w/w_a), t_h = \log(h/h_a) \quad (20)$$
$$t_x^* = (x^* - x_a)/w_a, t_y^* = (y^* - y_a)/h_a \quad (21)$$
$$t_\theta = \theta - \theta_a \quad (22)$$
$$t_x = (x - x_a)/w_a, t_y = (y - y_a)/h_a \quad (23)$$
$$t_w^* = \log(w^*/w_a), t_h = \log(h^*/h_a) \quad (24)$$
$$t_\theta^* = \theta - \theta_a \quad (25)$$

where (x, y) represents the center coordinates of the recognition frame, (w, h) represents the width and height of the recognition frame, and x (y, w, h), xa (ya, wa, ha) and x * (y *, w *, h *) represent the predicted value, the anchor and the true value respectively. For this particular dataset, w and h will not be reversed, so it is unnecessary to redefine it.

C. TRAINING AND EVALUATION

Take 6000 pictures of the checkerboard's dataset in different depth and set the ratio of the test dataset to the training dataset to 5:3. The training frequency is set to 60k times. In order to improve the learning efficiency, the learning rate needs to be larger when set at the beginning. In order to avoid shocks, the learning rate is then reduced. Therefore, the learning rate of the first 25k times is 0.001, the learning rate of the following 25k times is 0.0001, and the learning rate of the last 10k times is 0.00001.

Different learning frameworks are used to train and identify the checkerboard dataset separately, and the differences between the evaluation parameters of different methods are analyzed, as shown in Table 4. It can be seen from the data in the table that compared with other detection methods, the deep learning framework adopted in this paper has given better performance in terms of detection accuracy. Although the detection time is slightly longer than that of the Faster-RCNN framework, the method is more appropriate for the identification of the checkerboard. By changing the detection confidence score threshold to obtain different recall and precision rates, as shown in Figure 10, the curve in the figure shows that the best performance can be obtained with the set precision.

Through using the trained model framework, it is easy to determine the pixel coordinate value of the center point of the checkerboard in each picture, and use checkerboards of different positions to identify the object, as shown in Figure 11.

TABLE 4. Performance analysis of different methods in CheckerBoard data set.

Approaches	Rotion Anchor	Pooled sizes	Recall	Precision	measure	Time
Faster-RCNN	NO	7×7	65.1%	94.2%	78.3%	**0.33s**
FPN	NO	7×7	79.2%	95.3%	82.9%	0.41s
RRPN	NO	7×7	72.1%	81.1%	72.4%	0.62s
R^2CNN	NO	7×7,16×3,3×16	84.3%	85.6%	83.6%	0.49s
DFPN	YES	7×7,3×8,8×3	**91.1%**	**95.5%**	**88.7%**	0.34s

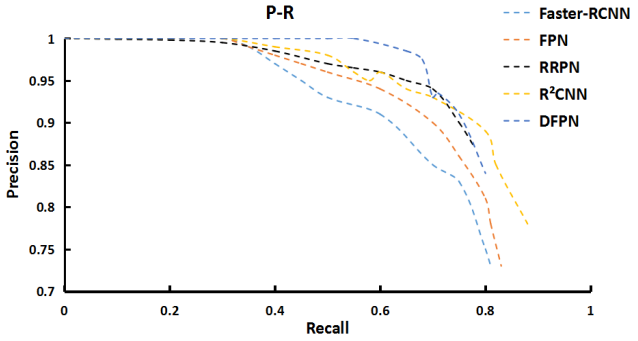


FIGURE 10. P-R curves corresponding to different methods.

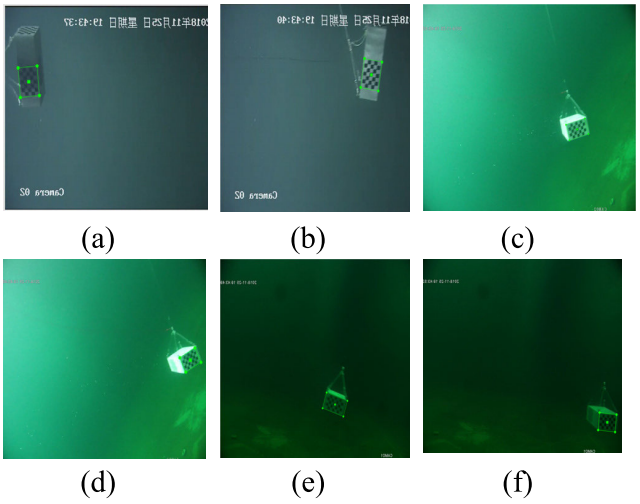


FIGURE 11. CheckerBoard test recognition frame in different positions. (a) Identify the object in the first position captured by the first camera. (b) Identify the object in the second position captured by the first camera. (c) Identify the object in the first position captured by the second camera. (d) Identify the object in the second position captured by the second camera. (e) Identify the object in the first position captured by the third camera. (f) Identify the object in the second position captured by the third camera.

The calculated value and the actual value of the pixel coordinates of the center point are shown in Table 5. The error value can be controlled within 1%, which provides accurate data for the following three-dimensional measurement.

D. DISCUSSION

For this type of network framework, traditional errors appear in false alarms and misjudgments. False alarms refer to the appearance of objects with a similar aspect ratio to the target

TABLE 5. CheckerBoard center point pixel coordinate error analysis.

Serial number	Calculated(x,y)	Actual value(u,v)	Error /%
1	(53,152)	(52.3,153.1)	0.50
2	(75,165)	(75.1,163.8)	0.58
3	(110,221)	(109.1,220.3)	0.43
4	(82,133)	(83.1,131.6)	0.38
5	(92,166)	(91.7,165.5)	0.30
6	(158,196)	(157.6,197.7)	0.42

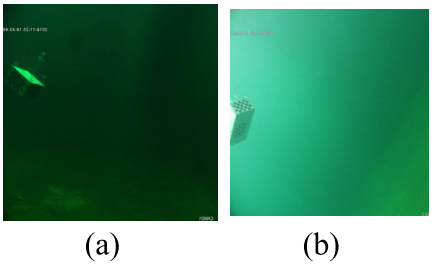


FIGURE 12. Misjudgment. (a) Misjudgment of camera 1. (b) Misjudgment of camera 2.

object in a complex environment, and there are almost no objects similar to the target object in the large-scale test pool established in this paper, so there is basically no error due to false alarms. Misjudgments are caused by objects with IoU overlap and large difference in length and width. The rotation angle of the object's non-axis alignment detection frame has a greater impact on the sensitivity of IoU, and it is easy to make wrong judgments. The target objects in this article are similar in length and width. The chance of this kind of error appearing is small. However, when the object itself moves to some neutral position (a certain edge of the object is facing the camera), a wrong judgment will be made, as shown in Figure 12. However, there are very few pictures with this kind of error, and estimation can be performed based on the data of the previous and following frames. Therefore, such error has little effect on the final result.

IV. THREE-DIMENSIONAL MEASUREMENT

Through the above-mentioned method, we can calculate internal parameter A and the external parameters (rotation matrix R(r_1, r_2, r_3) and translation vector T) corresponding to each image with high precision.

The parameters of underwater camera calibration are optimized before later calculation. After the corresponding external parameters rotation matrix R(r_1, r_2, r_3) and translation vector T of each image are obtained through underwater camera calibration, the external parameters need to be optimized to make the later calculation results more accurate. Good results can be obtained by using the BA optimization method. The so-called BA optimization means that the camera attitude and the position of the feature points are adjusted at the same time, so that the light reflected from each feature point can pass through the camera light center. LOURAKIS, M.I.A [45] proposed SBA and used the principle of least square to estimate the structure and motion of features, and Gao *et al.* [46] aimed to minimize the reprojection error, so as to obtain the camera pose and the coordinates of the feature points.

A. PTIMIZATION FUNCTION

With the reprojection error as the objective function, the BA optimization is constructed as a least square problem, and the camera pose and the coordinates of the feature points are adjusted simultaneously by minimizing the reprojection error. Re-projection is to re-project the points in the world coordinate system to the pixel coordinate system, taking into account the internal and external parameters and the distortion coefficient of the camera, that is:

$$u = h(\xi, p) \tag{26}$$

where h represents the imaging process of the camera, ξ represents the pose of the camera in the world coordinate system (represented by Lie Algebra), p is the coordinate of the feature point in the world coordinate system, u is the theoretical value of the pixel coordinate corresponding to the feature point in the pixel coordinate system, and the objective function (re-projection error function) is defined as:

$$e_r = (z - u)^2 = (z - h(\xi, p))^2 \tag{27}$$

where z is the pixel coordinate of the feature point in the pixel coordinate system. The least square problem is constructed, and the obtained external parameters are considered as the initial values, and the corresponding external parameters are optimized by minimizing the target function. The re-projection error function is minimized by adjusting (ξ, p). The checkerboard has a total of 24 corners, and the target function to be solved can be established as shown in (28), where ξ is the Lie Algebraic representation of the optimized position and attitude, and $p_i = [x_i, y_i, z_i]^T$ is the three-dimensional coordinate of the target point in the world coordinate system.

$$\min \frac{1}{2} \sum_{i=1}^{24} |z - h(\xi, p_i)|^2 \tag{28}$$

The error of a single feature point is:

$$e_i = z - h(\xi, p_i) \tag{29}$$

Then the error of all the feature points of the checkerboard in each image is:

$$f(x) = \begin{bmatrix} e_0 & e_1 & \cdots & e_{24} \end{bmatrix}^T \tag{30}$$

The overall target function can be written as follows:

$$\min \frac{1}{2} \sum_{i=1}^{24} |z - h(\xi, p_i)|^2 = \min \frac{1}{2} |f(x)|^2 \tag{31}$$

Define the state variables to be estimated as:

$$x = \begin{bmatrix} \xi & p_1 & \cdots & p_{24} \end{bmatrix}^T \tag{32}$$

The incremental equation can be obtained from the Gauss-Newton method.

$$H \Delta x = g \tag{33}$$

And:

$$\begin{cases} H = J(x)^T * J(x) \\ g = -J(x)^T * f(x) \end{cases} \tag{34}$$

Corresponding to each feature point in checkerboard, the Jacobian matrix is:

$$J_i(x) = \begin{bmatrix} F_i \; 0 \; \ldots \; E_i \; \cdots \; 0 \; 0 \end{bmatrix} \tag{35}$$

And:

$$\begin{cases} F_i = \dfrac{\partial e_i}{\partial \delta \xi} \\ E_i = \dfrac{\partial e_i}{\partial p_i} \end{cases} \tag{36}$$

Considering all the feature points, the global Jacobian matrix can be written as follows:

$$J(x) = \begin{bmatrix} F_1 & E_1 & 0 & \ldots & \ldots & 0 \\ F_2 & 0 & E_2 & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ F_i & 0 & 0 & E_i & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ F_{24} & 0 & 0 & \ldots & \ldots & E_{24} \end{bmatrix} \tag{37}$$

And, (38) and (39), as shown at the bottom of the next page.

In the formula, f_x and f_y is the original size of the image, (c_x, c_y) is the coordinate of the principal point, and $[X', Y', Z']^T$ is the three-dimensional coordinate of the feature point transformed to the camera coordinate system.

B. ITERRATIVE OPTIMATION

With a given initial value, the external parameters can be optimized through iteration:

a: Set initial parameter values x_0.

b: Assume that the iteration times reach k times and find the current $J(x_k)$ and $f(x_k)$.

c: Solve incremental equation (28) and find Δx_k.

d: If Δx_k is small enough, stop; otherwise, $x_{k+1} = x_k + \Delta x_k$ and return to step b.

The external parameters obtained through underwater camera calibration are set to initial value x_0, and the optimized

FIGURE 13. Comparison of reprojection of feature points before and after optimization. (a) The change process of reprojection in the optimization process of the target object in the first position. (b) The change process of reprojection in the optimization process of the target object in the second position.

external parameters are obtained after 5 iterations. The 24 feature points of different images before and after optimization are re-projected, as shown in Figure 13.

It can be deemed that each image uses the object checkerboard plane facing the camera for calibration, the internal and external parameters are obtained, and the world coordinate system $O_i x_i y_i z_i$ is established. Correspondingly, the checkerboard plane coordinate system is $O_i x_i y_i$ (the default value of z_i is 0). As a result, the corresponding relationship between the coordinates of the target point in the world coordinate system and the image plane coordinate system is as follows:

$$sm = \boldsymbol{H}M \tag{40}$$

where s is a non-zero scale factor, m and M are the image plane coordinate system and the world coordinate system of the target point, respectively. The following can be obtained from the geometric relation and formula (40) of the imaging point:

$$\boldsymbol{H} = \boldsymbol{A}(r_1, r_2, \boldsymbol{T}) = (h_1, h_2, h_3) \tag{41}$$

$$r_1 = s\boldsymbol{A}^{-1}h_1 \tag{42}$$

$$r_2 = s\boldsymbol{A}^{-1}h_2 \tag{43}$$

$$r_3 = r_1 \times r_2 \tag{44}$$

The following conclusions can be drawn by taking modulo on both sides of equations (42) and (43):

$$s = 1/\left\|\boldsymbol{A}^{-1}h_1\right\| = 1/\left\|\boldsymbol{A}^{-1}h_2\right\| \tag{45}$$

If the coordinate of the target point of the checkerboard plane coordinate system is $(x_i, y_i, 1)^T$, and the coordinate of the image plane coordinate system is $(u,v,1)^T$, then according to the Homography matrix [47] (homography matrix H) and

$$F = -\begin{bmatrix} \frac{f_x}{Z'} & 0 & -\frac{f_x X'Y'}{Z'^2} & -\frac{f_x X'Y'}{Z'^2} & f_x + \frac{f_y X'^2}{Z'^2} & -\frac{f_x Y'}{Z'} \\ 0 & \frac{f_x}{Z'} & -\frac{f_y Y'}{Z'^2} & -f_y - \frac{f_y Y'^2}{Z'^2} & \frac{f_y X'Y'}{Z'^2} & \frac{f_y Y'}{Z'} \end{bmatrix} \tag{38}$$

$$E = \begin{bmatrix} \frac{f_x}{Z'} & 0 & -\frac{f_x X'}{Z'^2} \\ 0 & \frac{f_y}{Z'} & -\frac{f_x Y'}{Z'^2} \end{bmatrix} R \tag{39}$$

formula (40), if the image plane coordinate system coordinate of the target point is known, the coordinate in the checkerboard plane coordinate system can be obtained, that is:

$$(x_i, y_i, 1)^T = s\boldsymbol{H}^{-1}(u, v, 1)^T \tag{46}$$

After calculating the coordinate of the checkerboard plane coordinate system of the target point, the three-dimensional homogeneous coordinate is $Q(x_i, y_i, 0, 1)^T$, and the coordinate of the target point in the camera coordinate system can be obtained by rotating the translation matrix, that is, the coordinate of the target point Q_c in the camera coordinate system can be obtained.

$$\boldsymbol{Q}_c = \begin{bmatrix} R & \boldsymbol{T} \\ 0^T & 1 \end{bmatrix} Q \tag{47}$$

Once the pixel coordinates of the ''target point'' are known, the camera coordinate system coordinate can be obtained through the above-mentioned algorithm, and the same ''pixel point'' (target point) can be found in each frame by extracting the corner of the image target. If it is regarded as the above-mentioned ''target point'', the camera coordinate corresponding to the target point in each frame can be obtained, and finally a three-dimensional trajectory can be constructed.

V. UNDERWATER EXPERIMENT AND RESULT ANALYSIS

The motion of the object can be divided into the complete pendulum drop test (initial stage approximate vertical fall, affected by installation cable and water) and the vertical drop stage test (pendulum falling motion). Because the field of view of a single camera is small and the span of the object trajectory is large, it is impossible for a single camera to capture a complete trajectory. In order to solve this problem, this experiment uses three underwater cameras to divide the whole trajectory into three segments, and the calculation of the trajectory can be achieved by reasonably arranging their respective positions. In order to better detect the corner features of the underwater moving object image, the checkerboard is pasted around the object. In the experiment, the distance between the camera and the initial target is about 5 m, and the depth of the test pool water is 10 m.

In the process of the object motion, a specific point can be selected to replace the whole object, so that the trajectory of the object can be calculated conveniently. Using different methods to determine the target point for trajectory calculation will cause different errors. The deep learning method is selected in this paper to determine the pixel coordinates of the target point, which can obtain a high-precision trajectory. For comparison, after obtaining the dataset, the conventional method is used to calculate the pixel coordinates of the target point to calculate the trajectory of the object, that is, a specific corner in the target is selected as the target pixel to calculate it; for the same dataset, the deep learning method is used to determine the coordinates of the target center point to calculate the trajectory. Since the deviation of the viewing angle will lead to different results, after calculating the trajectories using the conventional method and the deep learning method respectively, the two trajectories are placed in the same coordinate system for analysis.

A. TRAJECTORY MEASURED BY A SINGLE CAMERA

The positions captured by camera 1 during part of the time are shown in Figure 14.

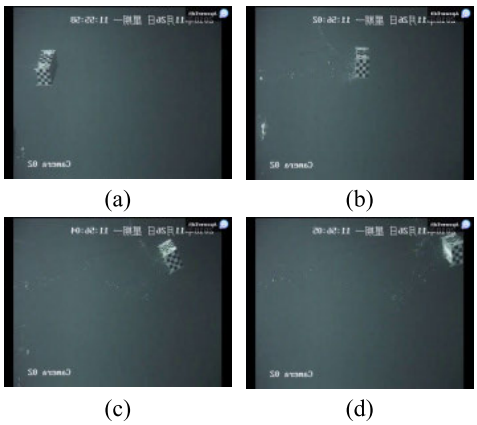


FIGURE 14. Part of the position of the object at the moment measured by camera 1. (a) The first position. (b) The second position. (c) The third position. (d) The fourth position.

The trajectory measured through the conventional method is shown in Figure 15.

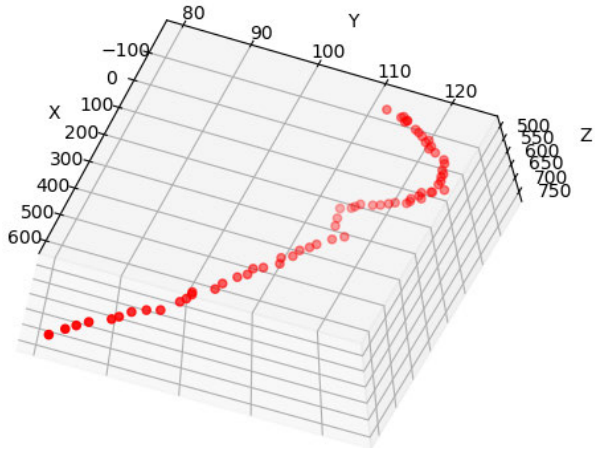


FIGURE 15. The trajectory of the object measured by camera.

The trajectory measured through the deep learning method is shown in Figure 16.

Place the trajectories measured by the two methods in the same coordinate system as shown in Figure 17.

The positions captured by camera 2 during part of the time are shown in Figure 18.

The trajectory measured through the conventional method is shown in Figure 19.

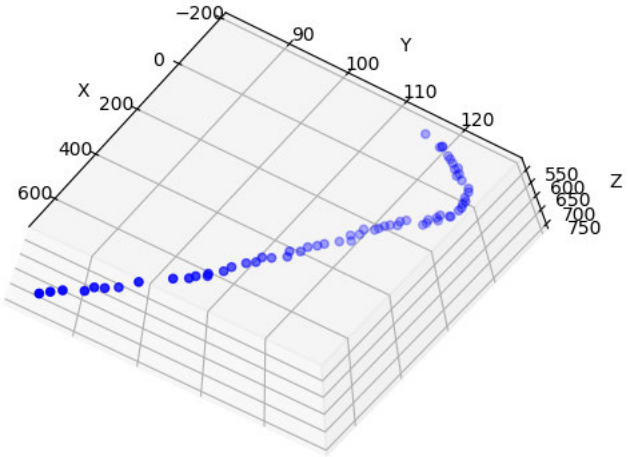


FIGURE 16. The trajectory of the object measured by camera.

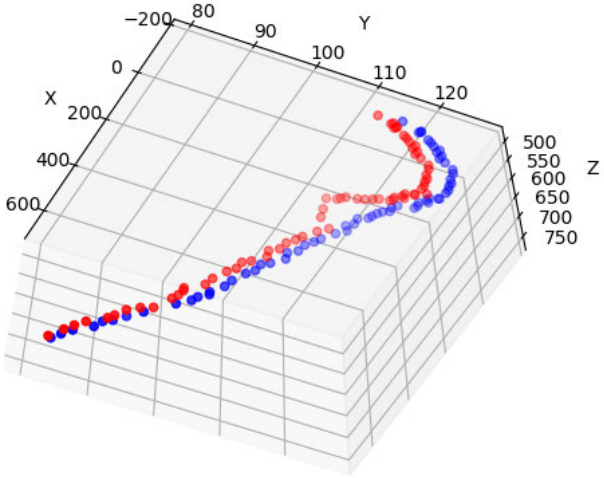


FIGURE 17. The trajectory of camera 1 merged into the same coordinate system (red curve: conventional method; blue curve: deep learning method).

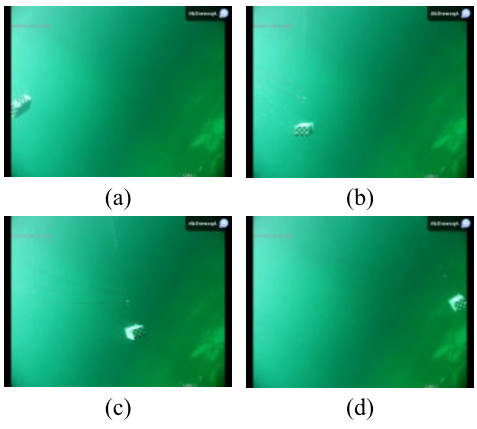


FIGURE 18. Part of the position of the object at the moment measured by camera. (a) The first position. (b) The second position. (c) The third position. (d) The fourth position.

The trajectory measured through the deep learning method is shown in Figure 20.

Place the trajectories measured by the two methods in the same coordinate system as shown in Figure 21.

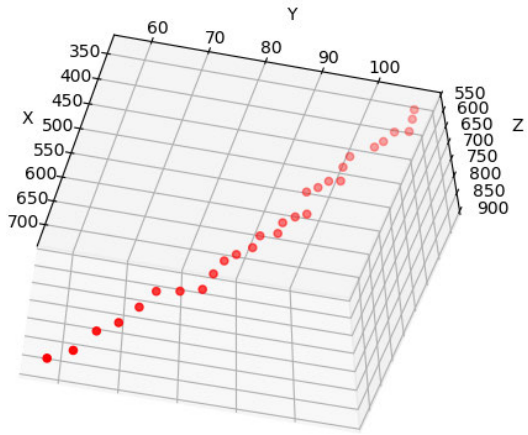


FIGURE 19. The trajectory of the object measured by camera.

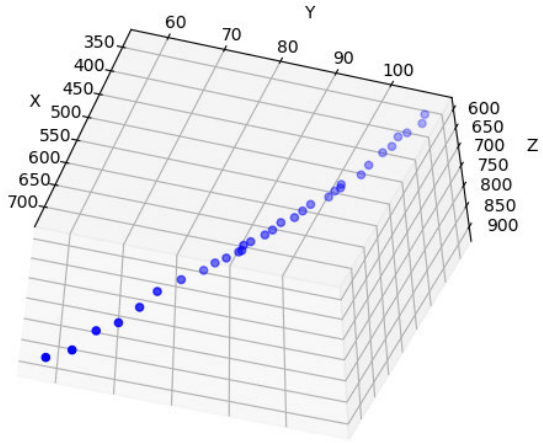


FIGURE 20. The trajectory of the object measured by camera.

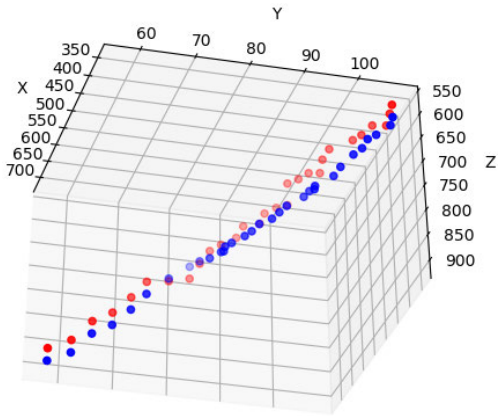


FIGURE 21. The trajectory of camera 2 merged into the same coordinate system (red curve: conventional method; blue curve: deep learning method).

The positions captured by camera 3 during part of the time are shown in Figure 22.

The trajectory measured through the conventional method is shown in Figure 23.

The trajectory measured through the deep learning method is shown in Figure 24.

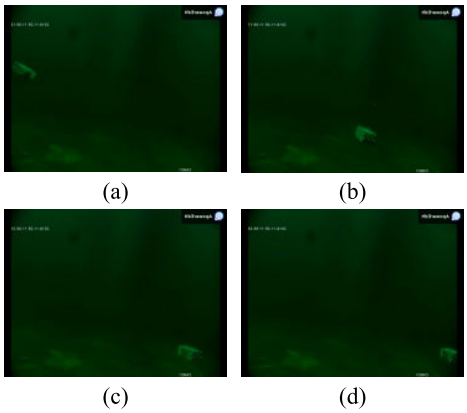


FIGURE 22. Part of the position of the object at the moment measured by camera 3. (a) The first position. (b) The second position. (c) The third position. (d) The fourth position.

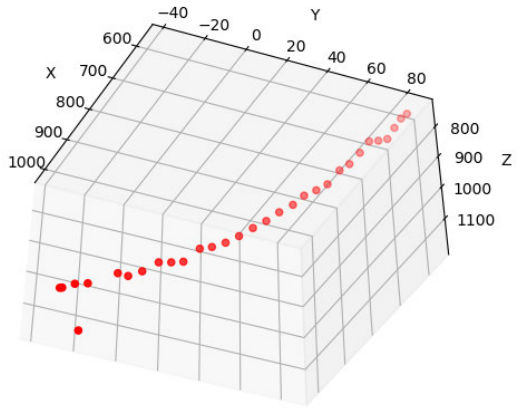


FIGURE 23. The trajectory of the object measured by camera 3.

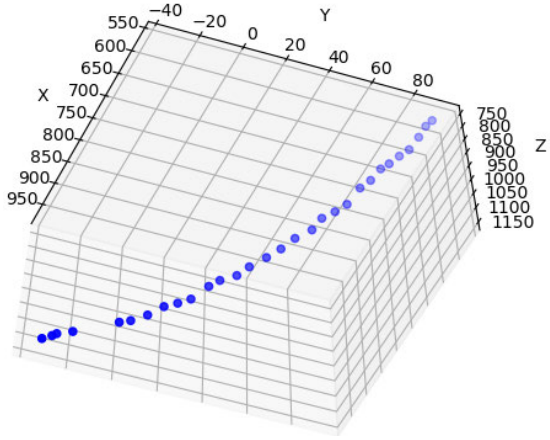


FIGURE 24. The trajectory of the object measured by camera.

Place the trajectories measured by the two methods in the same coordinate system as shown in Figure 25.

B. TRAJECTORY SYNTHESIS

When arranging the camera position during underwater operation, make sure that there are coincident parts in the field of

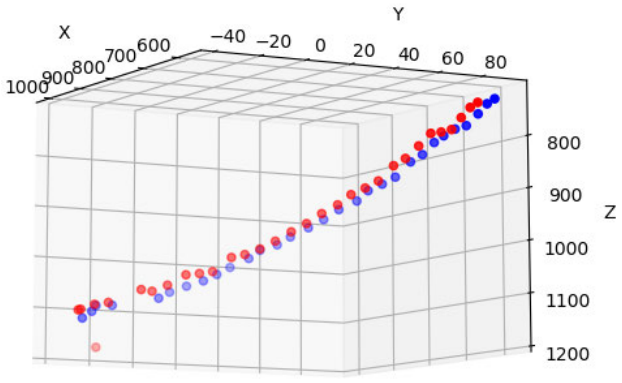


FIGURE 25. The trajectory of camera 2 merged into the same coordinate system (red curve: conventional method; blue curve: deep learning method).

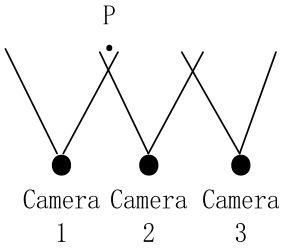


FIGURE 26. Camera field of view coincidence model.

view of the three cameras, and the field of view coincide with the camera model, is shown in Figure 26.

In Figure 26, it is assumed that the coordinates of point P in the camera 1 coordinate system and the camera 2 coordinate system are (x_1,y_1,z_1) and (x_2,y_2,z_2), respectively. Then the following formula is correct:

$$\begin{bmatrix} \boldsymbol{R} & \boldsymbol{T} \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ 1 \end{bmatrix} = \begin{bmatrix} x_2 \\ y_2 \\ z_2 \\ 1 \end{bmatrix} \tag{48}$$

Rotation matrix R and translation vector T can be calculated from the positive definiteness of rotation matrix R and the coordinates of several coincidence points P. Therefore, the trajectories of camera 1 and camera 2 can be merged into the same camera coordinate system through formula (48). Similarly, the trajectories of camera 3 can also be merged into the same camera coordinate system, as shown below.

The merger trajectory measured through the conventional method is shown in Figure 27.

The merger trajectory measured by the deep learning method is shown in Figure 28.

Place the two trajectories in the same coordinate system as shown in Figure 29.

From the above-mentioned comparison, we can see that there are some offset points around the trajectory using the conventional method, because there may be a flip or a great rotation in the process of object motion, resulting

TABLE 6. Coordinate error analysis.

Serial number	True value (mm)	Value calculated by conventional method (mm)	Error (%)	Value calculated by deep learning method (mm)	Error (%)
1	(–162.08, 111.29, 494.36)	(–158.04, 109.31, 489.38)	1.19	(–160.13, 110.92, 493.24)	0.32
2	(–90.16, 120.15, 508.36)	(–86.55, 116.23, 502.33)	1.39	(–89.12, 121.31, 507.46)	0.14
3	(–44.07, 121.14, 515.21)	(–38.21, 116.16, 510.30)	1.20	(–42.92, 125.67, 510.33)	0.71
4	(–2.58, 122.51, 522.98)	(–1.55, 120.54, 518.92)	1.21	(–2.01, 121.26, 521.73)	0.28
5	(31.94, 123.49, 529.19)	(26.85, 115.50, 525.23)	1.09	(33.39, 120.25, 527.86)	0.35
6	(189.39, 111.36, 469.97)	(183.33, 105.32, 467.10)	1.18	(188.17, 109.55, 468.42)	0.43
7	(476.59, 92.85, 656.47)	(468.36, 103.98, 645.19)	1.56	(468.58, 90.77, 659.31)	0.31
8	(882.76, 6.54, 1042.29)	(865.16, 2.77, 1030.10)	1.53	(880.35, 4.49, 1040.76)	0.20

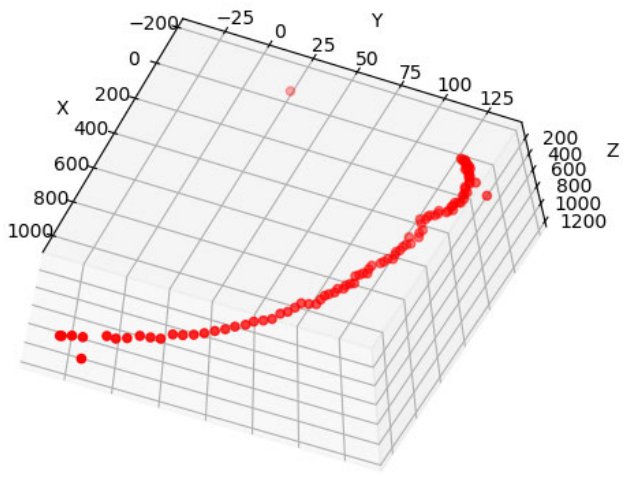


FIGURE 27. Fusing tracjectory.

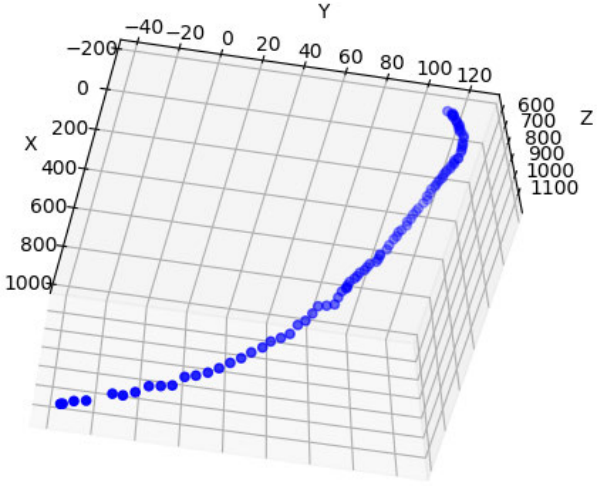


FIGURE 28. Fusing tracjectory.

in another target plane of the target object photographed by the camera. Therefore, it is calculated that the three-dimensional point representing the object deviates to some extent. Such situation can be avoided by using the deep learning method. If two target surfaces appear at the same time, the larger target will be selected for recognition and its central coordinates will be determined, which can reduce a lot of offset points. The measured trajectory is displayed as a smooth curve in the 3D coordinate system, as shown in Figure 30.

C. ERROR ANALYSIS

The commercial video motion analysis software is relatively mature. In order to quantitatively analyze the authenticity of the experimental results, the experimental results are compared with the trajectories calculated by the professional motion video analysis software TEMA in this paper. Error analysis with the coordinates calculated by the TEMA software as the actual value, and the coordinates calculated by the method described in this paper, is shown in Table 6. However, the TEMA software needs to introduce the camera calibration model parameters generated by refraction to analyze the underwater motion video, and the operation is relatively complex, so application in actual underwater engineering is difficult. Through the error analysis of the trajectory calculated through the conventional method and the deep learning method, it can be seen that the method of deep learning and computer vision proposed in this paper

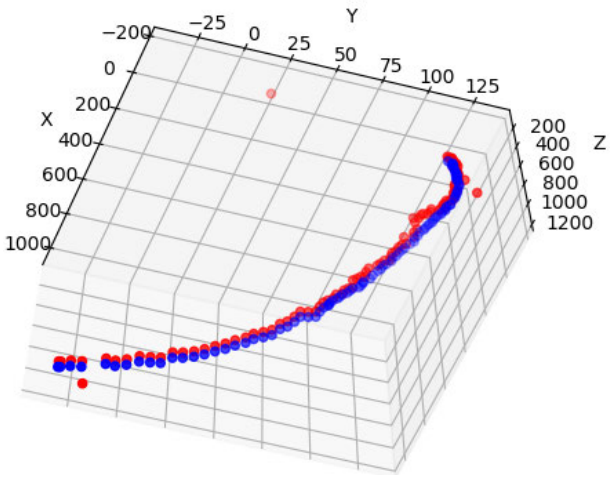


FIGURE 29. The trajectorys merged into the same coordinate system (red curve: conventional method; blue curve: deep learning method).

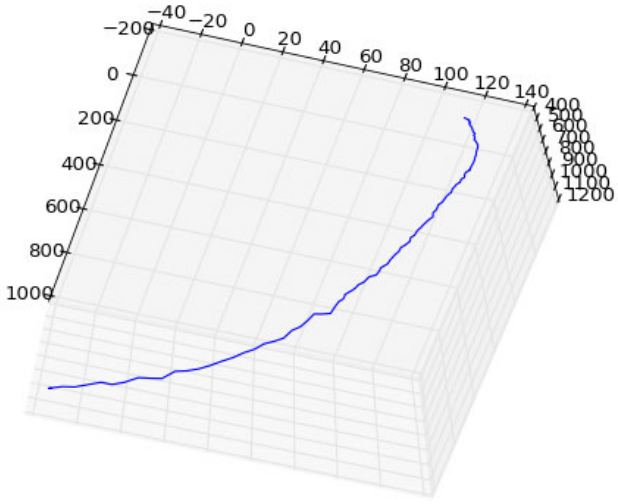


FIGURE 30. Smooth trajectory.

is relatively simple and easy to implement and has high accuracy.

VI. CONCLUSION

In this paper, based on the analysis of the special underwater environment, the internal and external parameters of the underwater camera are calibrated by using the modified underwater camera imaging model, and an algorithm for calculating the three-dimensional trajectory of underwater moving objects based on multiple cameras is proposed and realized. In this paper, the difficulty of being unable to capture the trajectory of an object with large-scale motion is solved experimentally, and an algorithm for registering the trajectory captured by multiple cameras is proposed by using multiple underwater cameras to capture the motion of the object at the same time. Finally, the complete trajectory of underwater large-scale object motion after registration is obtained. The experimental results show that the underwater camera modeling based on refraction can accurately obtain the trajectory data of moving objects in the underwater environment, which provides some reference for the work related to marine engineering equipment.

ACKNOWLEDGMENT

This achievement is inseparable from the joint efforts of all members of the research group. In repeated discussions and experiments, important information about image datasets has been obtained. In addition, the authors sincerely thank the anonymous review experts for their valuable suggestions.

REFERENCES

[1] A. Gunes and M. B. Guldogan, "Joint underwater target detection and tracking with the Bernoulli filter using an acoustic vector sensor," *Digit. Signal Process.*, vol. 48, pp. 246–258, Jan. 2016.
[2] Z. Chen and W. Wang, "Research on underwater target detection using side-scan sonar and multibeam sounding system," *Hydrograph. Surv. Charting*, vol. 33, pp. 51–54, Jul. 2013.
[3] N. H. Klausner and M. R. Azimi-Sadjadi, "Performance prediction and estimation for underwater target detection using multichannel sonar," *IEEE J. Ocean. Eng.*, vol. 45, no. 2, pp. 534–546, Apr. 2020.
[4] A. P. Galusha, G. Galusha, J. M. Keller, and A. Zare, "A fast target detection algorithm for underwater synthetic aperture sonar imagery," *Proc. SPIE*, vol. 10628, Apr. 2018, Art. no. 106280Z.
[5] N. Alem, F. Pellen, G. Le Brun, and B. Le Jeune, "Extra-cavity radiofrequency modulator for a LiDAR radar designed for underwater target detection," *Appl. Opt.*, vol. 56, no. 26, p. 7367, Sep. 2017.
[6] B.-T. Zha, H.-L. Yuan, and Y.-Y. Tan, "Ranging precision for underwater laser proximity pulsed laser target detection," *Opt. Commun.*, vol. 431, pp. 81–87, Jan. 2019.
[7] S. Zhenmin, Z. Tong, W. Yuncai, Z. Yongchao, S. Weidong, W. Bingjie, and L. Jingxia, "Underwater target detection of chaotic pulse laser radar," *Infr. Laser Eng.*, vol. 48, no. 4, 2019, Art. no. 406004.
[8] M. Darwiesh, A. F. El-Sherif, M. F. Hassan, H. S. Ayoub, and Y. H. Elbashar, "Simulation and characterization of underwater target detection using LiDAR system," *J. Opt.*, vol. 49, pp. 416–426, Jul. 2017.
[9] X.-K. Li, X.-X. Meng, and Z. Xia, "Characteristics of the geometrical scattering waves from underwater target in fractional Fourier transform domain," *Acta Phys. Sinica*, vol. 64, no. 6, p. 64302, 2015.
[10] Z. Haisheng, "An optimization algorithm for underwater homing target detection based on beamforming," *Intell. Comput. Appl.*, vol. 8, pp. 137–140, Dec. 2018.
[11] X. Li, M. Liu, and S. Jiang, "Morphological research on geometrical scattering waves of an underwater target," *J. Mar. Sci. Appl.*, vol. 14, no. 2, pp. 208–214, Jun. 2015.
[12] L. I. Huizhou, L. Zhenghong, and M. Dun, "Underwater small target tracking algorithm based on diver detection sonar image sequences," *Ship Electron. Eng.*, vol. 38, pp. 26–34, Feb. 2018.
[13] D. B. Gillis, "An underwater target detection framework for hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1798–1810, 2020.
[14] D. U. Jin-Xiang and X. U. Heng-Bo, "Wideband acoustic imaging of underwater target using spatial time-frequency analysis," *J. Unmanned Undersea Syst.*, vol. 27, pp. 392–397, Aug. 2019.
[15] Y. Zhou, Q. Li, and G. Huo, "Underwater moving target detection based on image enhancement," in *Proc. Int. Symp. Neural Netw.*, 2017, pp. 427–436.
[16] X. U. Feng *et al.*, "Underwater small target tracking based on mixture particle fifilter," *J. Appl. Acoust.*, vol. 34, pp. 297–302, Jul. 2015.
[17] Z. L. Wu, J. Li, and Z. Y. Guan, "Feature extraction of underwater target ultrasonic echo based on wavelet transform," *Appl. Mech. Mater.*, vols. 599–601, pp. 1517–1522, Aug. 2014.
[18] M. Liu, X. Li, and X. Guo, "Approach of elastic scattering extraction of underwater target based on wavelet transform and morphological method," in *Proc. IEEE/OES China Ocean Acoust. (COA)*, Jan. 2016, pp. 1–8.
[19] G. Stockman and L. G. Shapiro, *Computer Vision*, 2nd ed. Beijing, China: Electronic Industry Press, 2017, pp. 2–15.
[20] S. Psycharakis, R. Sanders, and F. Mill, "A calibration frame for 3D swimming analysis," in *Proc. 22nd Int. Symp. Biomech. Sports*, Beijing, China, 2005, pp. 901–904.

[21] L. Puig, Y. Bastanlar, P. Sturm, J. J. Guerrero, and J. Barreto, "Calibration of central catadioptric cameras using a DLT-like approach," *Int. J. Comput. Vis.*, vol. 93, no. 1, pp. 101–114, May 2011.
[22] Y. Kwon, "A camera calibration algorithm for the underwater motion analysis," in *Proc. ISBS-Conf. Proc. Arch.*, 1999, pp. 1–4.
[23] T. Alexandri and R. Diamant, "A reverse bearings only target motion analysis for autonomous underwater vehicle navigation," *IEEE Trans. Mobile Comput.*, vol. 18, no. 3, pp. 494–506, Mar. 2019.
[24] J. Shen, H. Y. Sun, H. B. Wang, Z. Chen, and Y. Wei, "A binocular vision system for underwater target detection," *Appl. Mech. Mater.*, vols. 347–350, pp. 883–890, Aug. 2013.
[25] G. D. Hager, M. Dewan, and C. V. Stewart, "Multiple kernel tracking with SSD," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2004, p. 1.
[26] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767
[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020.
[28] Q. Zhou, B. Zhong, Y. Zhang, J. Li, and Y. Fu, "Deep alignment network based multi-person tracking with occlusion and motion reasoning," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1183–1194, May 2019.
[29] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese box adaptive network for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6668–6677.
[30] Q. Zhou, B. Zhong, X. Lan, G. Sun, Y. Zhang, B. Zhang, and R. Ji, "Fine-grained spatial alignment model for person re-identification with focal triplet loss," *IEEE Trans. Image Process.*, vol. 29, pp. 7578–7589, 2020.
[31] B. Zhong, B. Bai, J. Li, Y. Zhang, and Y. Fu, "Hierarchical tracking by reinforcement learning-based searching and coarse-to-fine verifying," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2331–2341, May 2019.
[32] T. Liu *et al.*, "A method of 3D trajectory measurement of underwater moving objects," *J. Graph.*, vol. 40, pp. 908–917, Oct. 2019.
[33] X. Xiaoze, X. Li, and S. Xin, "Self-scanning three-dimensional measurement technology of underwater structured light," *Chin. Joursal Lasers*, vol. 37, no. 8, p. 2010, 2010.
[34] X. Yang, H. Sun, X. Sun, M. Yan, Z. Guo, and K. Fu, "Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network," *IEEE Access*, vol. 6, pp. 50839–50849, 2018.
[35] X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo, "Automatic ship detection in remote sensing images from Google earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, p. 132, Jan. 2018.
[36] R. Schettini and S. Corchs, "Underwater image processing: State of the art of restoration and image enhancement methods," *EURASIP J. Adv. Signal Process.*, vol. 2010, no. 1, pp. 1–14, Dec. 2010.
[37] T. Treibitz, Y. Schechner, C. Kunz, and H. Singh, "Flat refractive geometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 51–65, Jan. 2012.
[38] R. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE J. Robot. Autom.*, vol. RA-3, no. 4, pp. 323–344, Aug. 1987.
[39] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.
[40] J. M. Lavest, G. Rives, and J. T. Lapresté, "Dry camera calibration for underwater applications," *Mach. Vis. Appl.*, vol. 13, nos. 5–6, pp. 245–253, Mar. 2003.
[41] A. Agrawal, S. Ramalingam, Y. Taguchi, and V. Chari, "A theory of multilayer flat refractive geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3346–3353.
[42] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 1440–1448.
[43] 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.
[44] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*. [Online]. Available: https://arxiv.org/abs/1706.09579
[45] M. I. A. Lourakis and A. A. Argyros, "SBA: A software package for generic sparse bundle adjustment," *ACM Trans. Math. Softw.*, vol. 36, no. 1, pp. 1–30, Mar. 2009.
[46] X. Gao and T. Zhang, *Visual SLAM14 Lecture (From Theory to Practice)*. Beijing, China, 2017, pp. 111–119.
[47] F. Zhou and G. Zhang, "Complete calibration of a structured light stripe vision sensor through planar target of unknown orientations," *Image Vis. Comput.*, vol. 23, no. 1, pp. 59–67, Jan. 2005.

TAO LIU received the Ph.D. degree in control theory and control engineering from the College of Automation, Harbin Engineering University, in 2009, and the Ph.D. degree from Harbin Engineering University, in 2015. From January 2010 to July 2015, he was with the Key Laboratory of Underwater Robots, Harbin Engineering University. He was carried out a Postdoctoral Research Work of underwater robot control and vision detection direction. From November 2016 to November 2017, he was a Visiting Scholar with the Image Processing and Interpretation Research Group, University of Ghent, Belgium. He is currently a Lecturer with the College of Automation, Harbin Engineering University. His main research interests include control theory and its application, visual measurement, visual information processing, and unmanned system control.

NINGNING WANG was born in Pingdingshan, Henan. He is currently pursuing the master's degree in control science and engineering with the College of Automation, Harbin Engineering University. His main research interest includes computer vision.

LEI ZHANG was born in Huaibei, Anhui, China. He is currently an Associate Professor and a Ph.D. Supervisor with Harbin Engineering University. His main research interests include simulation and intelligent control of unmanned marine vehicles.

SHANGMAO AI is currently a Teacher with the College of Shipbuilding Engineering, Harbin Engineering University. His main research interests include laboratory technology of marine riser and pipeline.

HONGWANG DU received the master's and Ph.D. degrees from the School of Mechanical and Electrical Engineering, Harbin Engineering University, in 2005 and 2010, respectively. He is currently a Lecturer with the College of Automation, Harbin Engineering University. His main research interests include control theory and its application, industrial robot control, and perception.