Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Underwater Dense Targets Detection and Classification based on YOLOv3

1st Tingchao Shi
School of Marine Engineering
Northwestern Polytechnical University
Xi'an, Shanxi Province, China
shi_tingchao@163.com

2nd Yun Niu*
School of Marine Engineering
Northwestern Polytechnical University
Xi'an, Shanxi Province, China
niuyun010121@nwpu.edu.cn

3rd Mingyong Liu
School of Marine Engineering
Northwestern Polytechnical University
Xi'an, Shanxi Province, China
liumingyong@nwpu.edu.cn

4th Yang Yang
School of Marine Engineering
Northwestern Polytechnical University
Xi'an, Shanxi Province, China
y_yang@126.com

5th Cong Wang
School of Marine Engineering
Northwestern Polytechnical University
Xi'an, Shanxi Province, China
0930wc@mail.nwpu.edu.cn

6th Yuxuan Huang
School of Marine Engineering
Northwestern Polytechnical University
Xi'an, Shanxi Province, China
767805632@qq.com

***Abstract*—In order to meet the requirements of fast detection and classification of underwater targets during intelligent underwater robot operation, an improved YOLOv3 algorithm named YOLOv3-UW algorithm is proposed to improve the detection accuracy and detection speed. Compared with the YOLOv3 algorithm, the YOLOv3-UW algorithm improves the clusters algorithm of data sets, optimizes the network structure, and improves the residual module. The final experimental results show that detection speed and detection accuracy of the YOLOv3-UW algorithm are higher than the YOLOv3 algorithm.**

***Keywords*—YOLOv3, Classification, Detection, Underwater small targets**

I. Introduction

The ocean is the origin of life, an important space for human existence, and contains vast natural resources. Due to the special nature of the marine environment, it is difficult to realize the exploration and development of the ocean by manpower alone[1-2]. Therefore, various underwater robots have emerged. The fast detection and classification of underwater small targets is a key issue in the intelligent operation of underwater robots. With the continuous development of artificial intelligence technology, convolutional neural network has become a hot field of target detection and classification[3-5].

In recent years, along with the rapid development of convolutional neural network technology and computer technology, scholars from different countries have gradually applied convolutional neural network to underwater target detection and classification. Walther et al.[6] developed a system for detecting and tagging suspected underwater targets, which could be used to identify and track objects without backsight. Kamal et al. proposed a deep learning architecture for underwater target detection and classification, using DBN (Deep Belief Networks) architecture for pre-training[7]. Fatan et al. proposed an underwater cable recognition method based on texture features, which extracted image edge information to find the target[8].

In order to balance the detection speed and accuracy, many regression-based detection methods has been proposed. These methods can directly return the coordinate position and confidence score of the detected object, including YOLO[9], SSD[10], YOLOv2[11] and YOLOv3[12]. Among them, YOLO algorithm uses regression method to predict the coordinates of the bounding boxes and achieve the classification of the targets. The detection speed reaches 30 FPS, but YOLO has serious positioning error problems, causing poor detection accuracy. The SSD is based on the VGG network, which combines the feature maps of different convolutional layers to enhance the feature representation ability of the system. However, the SSD method for underwater small target detection has a low recall rate. YOLOv2 optimizes the model structure of YOLO, which significantly improves the detection speed. However, the basic network of YOLOv2 is relatively simple, and it does not improve the detection accuracy. YOLOv3 uses the deep residual network to extract image features and achieve multi-scale prediction. The detection accuracy and speed are relatively high.

Due to the turbid underwater environment and the dark light, the underwater images have problems of blur and color attenuation. In addition, underwater targets are densely distributed and there is a large amount of occlusion. If the YOLOv3 algorithm is used directly, the underwater target detection and

* Corresponding author. This work is supported by National Natural Science Foundation of China (Grant No.51879219; 51679201), Science and Technology on Underwater Information and Control Laboratory (Grant No.6142218051806), Fundamental Research Funds for the Central Universities (Grant No.3102019HHZY030027).

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

classification accuracy is not very high. In this paper, the YOLOv3 algorithm is improved to get better detection and classification results. The main contributions of this paper are listed as follows:

- The parameters of the anchor boxes are determined by the K-means++ algorithm.
- The Darknet-UW network structure is proposed to solve the over-fitting problems caused by too many training parameters in YOLOv3.
- The residual module is improved to increase the expressive ability of network.

II. MODEL DETECTION PROCESS

The YOLOv3 algorithm divides the original input image into SxS grid cells, and predicts N bounding boxes in each grid cell to detect different categories of targets. The network predicts four coordinates for each bounding box, as shown in Fig. 1, including the coordinates of the center point (t_x, t_y) and the size of the bounding box (p_w, p_h) , if the grid cell where the center point is located is offset from the upper left corner of the image by (c_x, c_y) , the coordinates of the predicted bounding box (b_x, b_y, b_w, b_h) are as shown in equation (1):

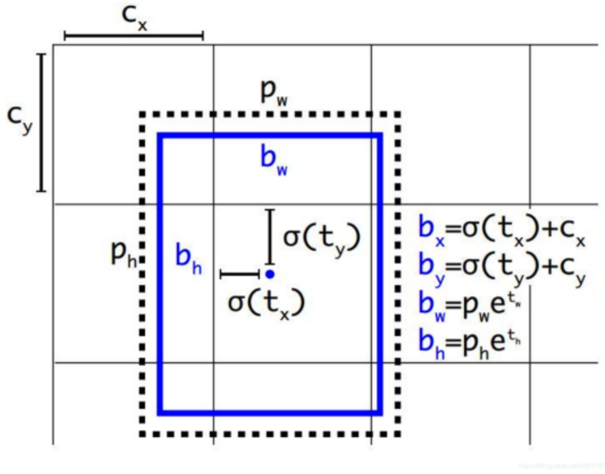


Fig. 1. Positional relationship between anchor box and bounding box.

$$\begin{cases} b_x = \sigma(t_x) + c_x \\ b_y = \sigma(t_y) + c_y \\ b_w = p_w e^{t_w} \\ b_h = p_h e^{t_h} \end{cases} \tag{1}$$

where (b_w, b_h) are the width and height of the predicted bounding box., (b_x, b_y) is the center coordinate of the predicted bounding box.

YOLOv3 uses logistic regression to predict the confidence score for each prediction box. This score is used to determine the confidence of the target in each prediction box, as defined in equation (2):

$$Conf = Pre(object) \times IOU_{pred}^{truth} \tag{2}$$

where $Pre(object)$ is used to judge whether the target object appears in the grid. If it appears, it is set to 1, and if it does not appear, it is set to 0. IOU_{pred}^{truth} is the ratio between the intersection area and union area of the predicted box and the true box, as defined in equation (3):

$$IOU_{pred}^{truth} = \frac{area(box_{pred} \cap box_{truth})}{area(box_{pred} \cup box_{truth})} \tag{3}$$

The target category is needed to be predicted if the target appears. As defined in equation (4):

$$\begin{aligned} C(M) &= Pre(class_M \mid object) \times Pre(object) \times IOU_{pred}^{truth} \\ &= Pre(class_M) \times IOU_{pred}^{truth} \end{aligned} \tag{4}$$

The loss function is used to characterize the degree of inconsistency between the predicted and the true values of the model. It is one of the important parameters to determine the network performance. The loss function of the YOLOv3 algorithm is mainly designed from the prediction error of the bounding box coordinate, the confidence error of the bounding box, and the classification error. The three aspects of prediction error are considered. If the loss value is smaller, the robustness of the model will be better. The loss function of YOLOv3 is defined as equation (5):

$$\begin{aligned} L = {} & \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij}^{obj} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] \\ & + \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij}^{obj} [(\sqrt{\omega_i} - \sqrt{\hat{\omega}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2] \\ & + \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij}^{obj} [(C_i - \hat{C}_i)^2] + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} l_{ij}^{noobj} [(C_i - \hat{C}_i)^2] \\ & + \sum_{i=0}^{S^2} l_i^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2 \end{aligned} \tag{5}$$

III. YOLOv3 ALGORITHM IMPROVEMENTS

A. Clusters Algorithm Improments

The K-Means[13] algorithm is a typical clustering algorithm, and it is widely used in deep learning. The K-Means algorithm selects the centroid position of the K initial points randomly. However, the centroid positions of the K initial points have a great influence on the convergence speed and the final clustering results, if it is selected randomly, the algorithm will converge slowly, and the final clustering results are not good. The K-Means++ algorithm does not select the centroid positions of the K initial points randomly, if we use the K-Means++ algorithm, we will increase the convergence speed and get better final clustering results. In this paper, we chose K-Means++[14] algorithm as the clustering method. In addition, the distance calculation equation of the K-Means++ algorithm is improved. As defined in equation (6):

$$d(a,b) = 1 - IOU(a,b) \tag{6}$$

where box indicates the area of the anchor box, and centroid indicates the center of the cluster. $IOU(a,b)$ is the ratio of the intersection and the union of the anchor box and the ground truth box.

As shown in Fig. 2, we can see that the average IOU value varies with the number of anchor boxes, when the number of anchor boxes increases, the average IOU gradually increases. When the number of anchor boxes is 6, the average IOU is 0.67. The average IOU increases very lowly, when the number of anchor boxes is greater than 6. In order to balance training speed and accuracy, the number of anchor boxes in this paper is six.

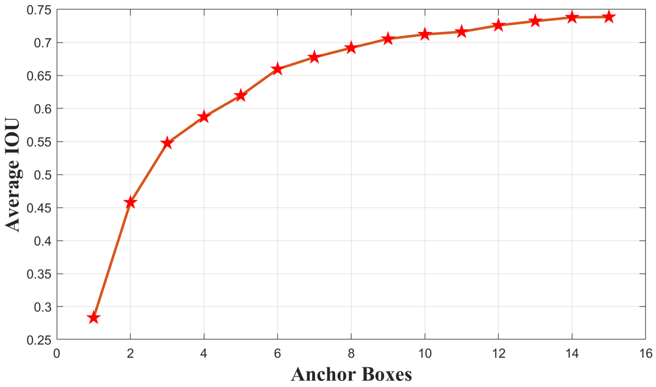


Fig. 2. The average IOU value varies with the number of anchor boxes.

B. Network Structure Improvements

The YOLOv3 algorithm uses the Darknet53[15] network to extract features from the image. The darknet53 network has a total of 53 layers of convolutions, including numerous network parameters. For the dataset of this paper, if we use the darknet53 network directly, it will cause over-fitting problems due to excessive network parameters. In addition, excessive network parameters can slow down the network operation and reduce the speed of detection and classification of underwater targets. Therefore, the network structure of Darknet53 has been improved, and the 1x1 convolution kernels have been added to the transmission module. By improving the network structure, the detection and classification speed of underwater targets is improved. Moreover, the over-fitting problem caused by too many parameters is solved. The improved network structure is shown in Fig. 3.

C. Residual Module Improments

Due to the turbid underwater environment and the dark light, the underwater images have problems of blur and color attenuation. In addition, underwater targets are often dense and there is a large amount of occlusion. If the YOLOv3 algorithm is directly used, the detection and classification accuracy of the underwater target is low and the missed detection rate is high. Therefore, in order to improve the feature extraction ability of the network, the residual module of the YOLOv3 algorithm is improved by adding the Squeeze-and-Excitation(SE) unit[16].

	Type	Filters	Size	Output
	Convolutional	32	3x3	416x416
	Convolutional	64	3x3/2	208x208
1x	Convolutional	32	1x1	
	Convolutional	64	3x3	
	Residual(1)			208x208
	Convolutional	32	1x1	
	Convolutional	64	3x3/2	
	Transition			104x104
2x	Convolutional	64	1x1	
	Convolutional	128	3x3	
	Residual(2)			104x104
	Convolutional	64	1x1	
	Convolutional	128	3x3/2	
	Transition			52x52
8x	Convolutional	128	1x1	
	Convolutional	256	3x3	
	Residual(3)			52x52
	Convolutional	128	1x1	
	Convolutional	256	3x3/2	
	Transition			26x26
8x	Convolutional	256	1x1	
	Convolutional	512	3x3	
	Residual(4)			26x26
	Convolutional	256	1x1	
	Convolutional	512	3x3/2	
	Transition			13x13
4x	Convolutional	256	1x1	
	Convolutional	512	3x3	
	Residual(5)			13x13

Fig. 3. The network structure of Darknet-UW.

By this improvement, the feature extraction ability of the network and the expression ability of the model are improved. The improved YOLOv3 algorithm and residual module are shown in Fig. 4.

IV. Experiments and Analysis

A. Underwater Target Datasets

This paper selects 3000 underwater images from the official dataset of the Chinese Underwater Robot Target Grab Competition in 2018. All images are taken from a real underwater environment, and each image contains different types of targets with uneven target distribution. The entire data set contains four types of targets, such as: holothurian, echinus, scallop, and starfish. Some of the goals in the picture are clear, some are blurred, some are densely distributed, and some are sparsely distributed.

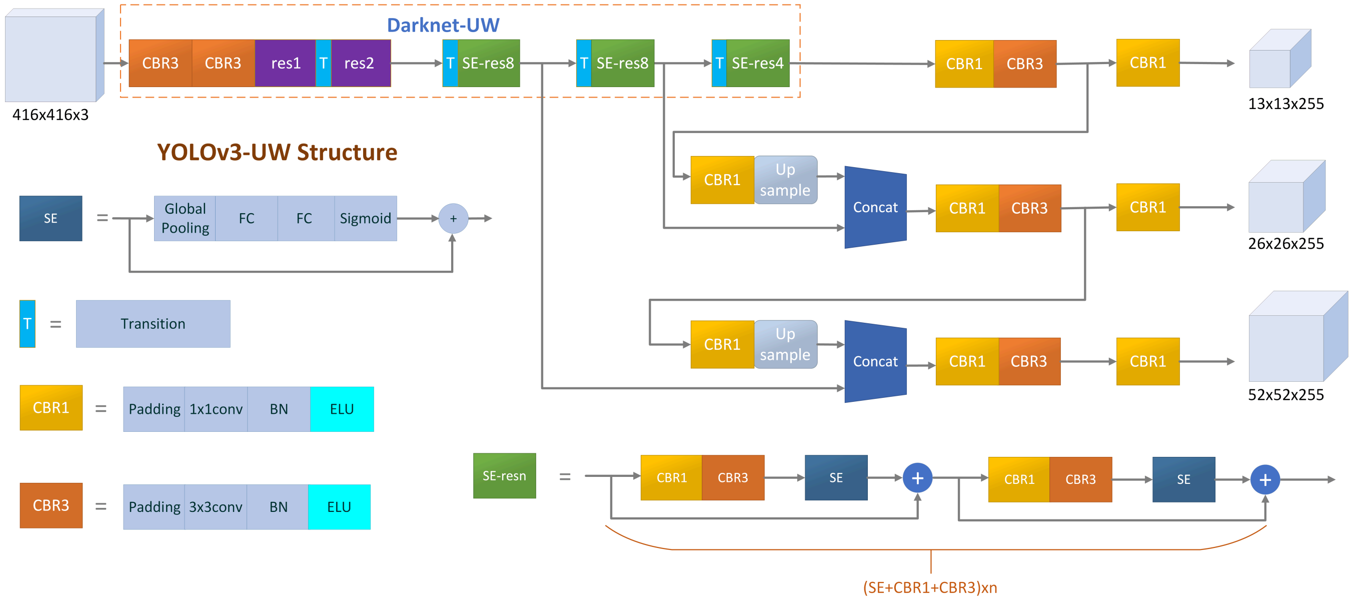


Fig. 4. The description of YOLOv3-UW structure.

B. Experiment Settings

This paper is based on the above data set, and the target category is 4. The initial learning rate during the training process is 0.001, and the learning rate becomes 0.0001 after 10000 steps. After 15000 steps, the learning rate becomes 0.00001. The total number of training steps is 20000. The momentum is 0.9. The experimental configuration is shown in Table 1.

TABLE I
EXPERIMENTAL CONFIGURATION

Parameters	Content
CPU	Intel Core i7 7700K
GPU	Nvidia GeForce GTX 1080Ti
System	Ubuntu 16.04 LTS
Accelerated environment	CUDA 9.0 cuDNN7.0
Training framework	Darknet

C. Training Results and Analysis

Due to the constant changes of parameters such as learning rate and loss function value, the quality of the model is not linear with the number of iterations in the process of training. Evaluating the quality of a model requires not only the accuracy of the detection but also the convergence speed of the loss function. Only when the accuracy of the model is high and the convergence speed is fast, the model can maintain the robustness and stability for the underwater targets detection and classification.

It can be seen from Fig. 5 that the loss function value at the beginning of training is about 6.5. As the number of training iterations increases, the loss value gradually decreases and converges to a lower value. Finally, the loss function converges to 0.4.

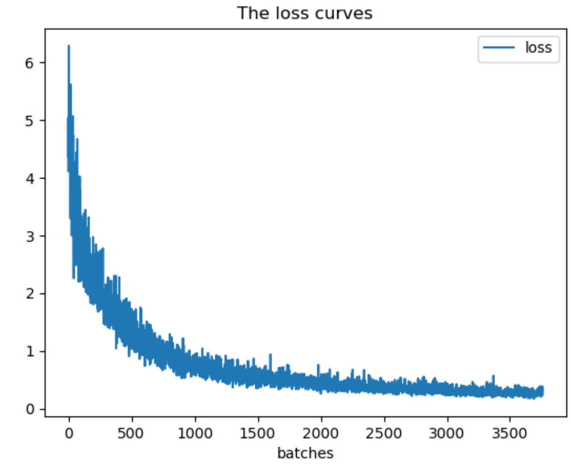


Fig. 5. Loss function curve during the training process.

The performance of different algorithms is usually evaluated by FPS and mAP. FPS represents frame per second and mAP represents mean average precision. The comparison of the detection precision results of the YOLOv3 algorithm and the YOLOv3-UW algorithm for different underwater targets is shown in Fig. 6. Compared with the YOLOV3 algorithm, the mAP of the YOLOV3-UW algorithm increases 6%. The detection speed of the YOLOv3 algorithm and the YOLOv3-UW algorithm is tested. The results show that the detection speed of the YOLOv3 algorithm is 31 FPS, and the detection speed of the YOLOv3-UW algorithm is 46 FPS. The YOLOv3-UW algorithm has improved detection speed by 15 FPS compared with the YOLOv3 algorithm.

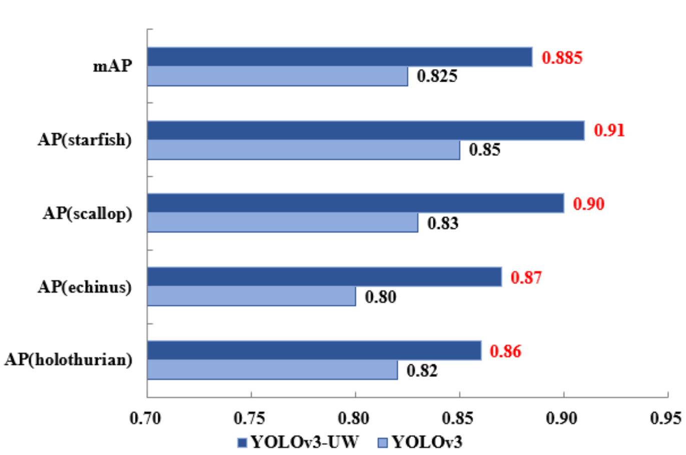
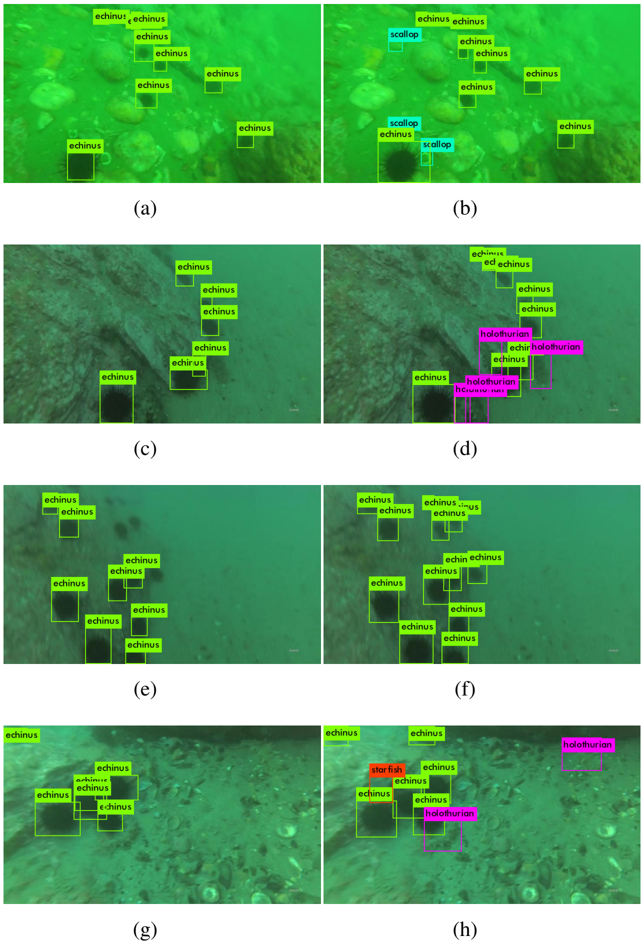


Fig. 6. The comparison of the detection precision results of the YOLOv3 algorithm and the YOLOv3-UW algorithm.

The detection and classification results of different underwater targets by the YOLOv3 algorithm and the YOLOv3-UW algorithm are shown in Fig. 7. Fig. 7(a), Fig. 7(c), Fig. 7(e) and Fig. 7(g) are the detection and classification results of the YOLOv3 algorithm for different underwater targets. Fig. 7(b), Fig. 7(d), Fig. 7(f) and Fig. 7(h) are the detection and classification results of the YOLOv3-UW algorithm for different underwater targets. By comparison, we can see that the YOLOv3 algorithm has a high missed detection rate for the detection of densely distributed targets. In this paper, the residual module is improved by embedding the SE module, which improves the feature extraction ability of the network. Therefore, the detection and classification accuracy of densely distributed targets are improved by the YOLOv3-UW algorithm.

V. CONCLUSIONS

In order to improve the accuracy and speed of detection and classification of underwater densely distributed targets in the case of turbid underwater environment, this paper proposes a underwater target detection and classification algorithm named YOLOv3-UW based on the YOLOv3 algorithm. The YOLOv3-UW algorithm improves the clusters algorithm, optimizes the network structure, and improves the residual module, which improves the detection and classification speed and accuracy of underwater small targets. In the future work, the Generative Adversarial Networks (GAN) will be used for underwater image denoising to further improve the detection accuracy.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Fig. 7. The comparison of the detection and clissification results of the YOLOv3 algorithm and the YOLOv3-UW algorithm..

REFERENCES

[1] ZHANG, Weikang, et al. ”Development of control system in abdominal operating ROV.” CHINESE JOURNAL OF SHIP RESEARCH 12.2 (2017): 124-132.

[2] Apgar, Joshua F., et al. ”ORCA-IX: An Autonomous Underwater Vehicle.” Massachusetts Institute of Technology, Massachusetts, US, Tech. Rep. (2006).

[3] Zhang, Lei, Da Peng Jiang, and Jin Xin Zhao. ”The basic control system of an ocean exploration AUV.” Applied Mechanics and Materials. Vol. 411. Trans Tech Publications, 2013.

[4] Ferri, Gabriele, Andrea Munafo, and Kevin D. LePage. ”An autonomous underwater vehicle data-driven control strategy for target tracking.” IEEE Journal of Oceanic Engineering 43.2 (2018): 323-343.

[5] Jaffe, Jules S., et al. ”Underwater optical imaging: status and prospects.” Oceanography 14.3 (2001): 66-76.

[6] Walther, Dirk, Duane R. Edgington, and Christof Koch. ”Detection and tracking of objects in underwater video.” Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.. Vol. 1. IEEE, 2004.

[7] Kamal, Suraj, et al. ”Deep learning architectures for underwater target recognition.” 2013 Ocean Electronics (SYMPOL). IEEE, 2013.

[8] Fatan, Mehdi, Mohammad Reza Daliri, and Alireza Mohammad Shahri. ”Underwater cable detection in the images using edge classification based on texture information.” Measurement 91 (2016): 309-317.

[9] Redmon, Joseph, et al. ”You only look once: Unified, real-time object detection.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[10] Liu, Wei, et al. ”Ssd: Single shot multibox detector.” European conference on computer vision. Springer, Cham, 2016.

[11] Redmon, Joseph, and Ali Farhadi. ”YOLO9000: better, faster, stronger.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[12] Redmon, Joseph, and Ali Farhadi. ”Yolov3: An incremental improvement.” arXiv preprint arXiv:1804.02767 (2018).

[13] Clausi, David A. ”K-means Iterative Fisher (KIF) unsupervised clustering algorithm applied to image texture segmentation.” Pattern Recognition 35.9 (2002): 1959-1972.

[14] Arthur, David, and Sergei Vassilvitskii. ”k-means++: The advantages of careful seeding.” Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007.

[15] Redmon, Joseph. ”Darknet: Open source neural networks in c.” (2013).

[16] Hu, Jie, Li Shen, and Gang Sun. ”Squeeze-and-excitation networks.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.