UNDERWATER SMALL TARGET DETECTION BASED ON DEFORMABLE CONVOLUTIONAL PYRAMID

Shuhan Qi[1,3], *Jianjun Du*[1], *Mingyan Wu*[1], *Hong Yi*[2], *Linlin Tang*[1], *Tao Qian*[*1] , *Xuan Wang*[*1,3]

[1]School of Computer Science and Technology, Harbin Institute of Technology Shenzhen, China
[2]Ricoh Software Research Center, Beijing, China
[3]Peng Cheng Laboratory, Shenzhen, China

ABSTRACT

Due to the problem of severe deformation, occlusion, diversified scenarios, general object detection methods cannot achieve satisfactory results in underwater object detection tasks. In this paper, we propose a two-stage Underwater Small Target Detection (USTD) network. In the proposed USTD, the Deformable Convolutional Pyramid(DCP) is proposed to deal with the problems of deformation, occlusion, and various object sizes effectively. Besides, we also propose a strategy of domain generalization based on curriculum learning to improve generalization in multi-domain environments, which is named as Phased Learning. Afterward, we construct an underwater target detection set (UTDS) to evaluate the accuracy of our method in underwater target detection tasks. Our method shows superior detection performance in experiments and reaches state-of-the-art for underwater target detection. Finally, in the 2020 China Underwater Robot Professional Contest (URPC), our method reached third place in terms of accuracy.

Index Terms— Underwater small target detection, deformable convolutional pyramid, curriculum learning

1. INTRODUCTION

Underwater object detection is a technology that adopts hardware equipment and algorithms to automatically recognize marine organisms and locate their position in the underwater environment[4][5]. Traditional underwater object detection technology is usually realized by underwater equipment such as sonar[11][8], laser[18], and camera[3]. With the development of computer vision, applying object detection algorithms[14][17][15] to the underwater scene is prevalent.

As shown in Figure 1, several challenges limit the application of general object detection algorithms in underwater scenarios[20]. Firstly, there are occlusion and deformation of targets in underwater images caused by the gathering habits of marine organisms. Secondly, marine organisms are too small to be detected, and there is a lack of real underwater object detection data sets for evaluation. Thirdly, the diversity of underwater scenes and the inconvenience of collecting underwater images in various sea areas lead to the problem of multiple domain environments with unbalance sample distribution.

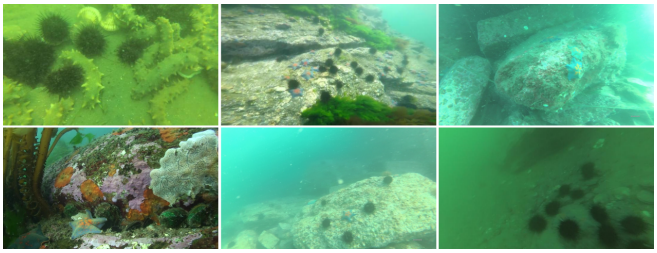

Fig. 1. Illustrative the challenges of underwater object detection. Images with the marine life gathering scene (left column); small marine organisms (middle column); blue and green shift (right column).

We propose a two-stage Underwater Small Target Detection(USTD) network, which aims to solve the above problems. In the USTD framework, we propose a novel deformable convolutional pyramid model to deal with occlusion, deformation, and various target sizes. Besides, we propose a strategy of domain generalization via phased learning to improve the generalization of the model in multi-domain environments. As a result, the USTD network has achieved SOTA in the evaluation of the Underwater Target Detection Set and achieved third place in the 2020 China Underwater Robot Professional Contest (URPC). The main contributions of this work can be summarized as:

a. A two-stage Underwater Small Target Detection (USTD) network is proposed, which is oriented to the complicated underwater environment and detects marine targets with various sizes.

b. Deformable convolutional pyramid (DCP) is proposed to deal with occlusion, appearance deformation, and various object sizes. Meanwhile, phased learning strategy is proposed to improve the generalization of model in multi-domain environments.

c. We annotated an Underwater Target Detection Set

* is Corresponding author.

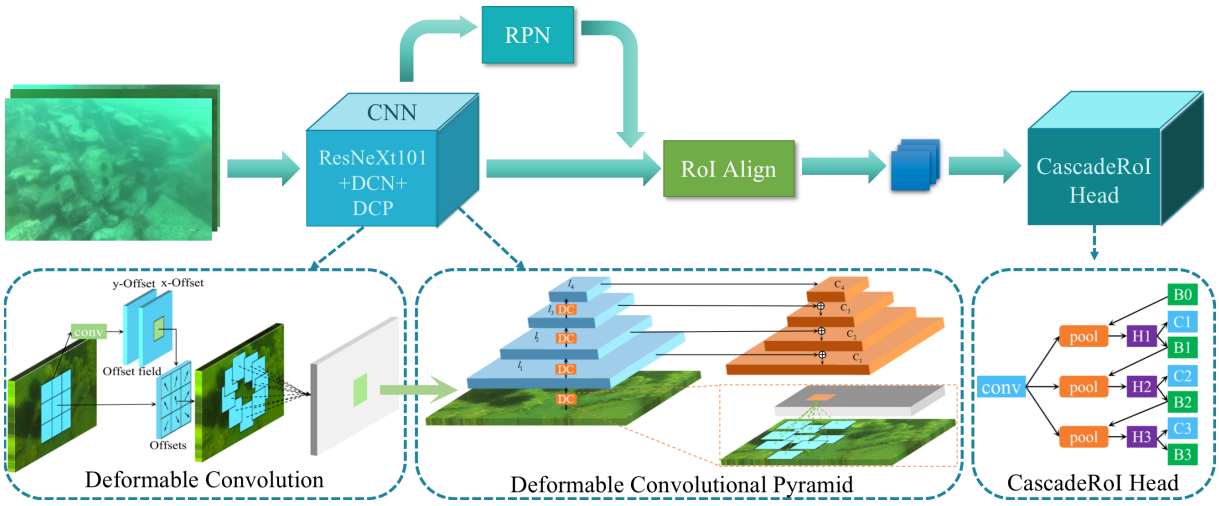

Fig. 2. The framework of the proposed USTD network. In CascadeRoI Head, "H" is network head, "B" is bounding box, "C" is classification and "B0" is proposals.

(UTDS) to evaluate our method, which consists of natural ocean underwater images from the network. Our method achieves the SOTA in the UTDS set.

2. METHOD

2.1. Overview of USTD Network

Figure 2 shows an overview of the USTD network. Deformable convolutional layers are employed to boost the ability to detect objects with deformed appearances. Then multiple deformable convolutional layers of different scales are fused to form the deformable convolutional pyramid (DCP), which significantly improves the ability to detect underwater objects with various sizes. Finally, the feature flow is sent to the Region Proposal Network (RPN) to obtain proposals. CascadeRoI Head will further adjust the position of the proposals and label proposals with category labels to give the final result of prediction.

2.2. Deformable Convolution

The USTD employs deformable convolution[6] to enhance the ability of the network to represent deformed objects.

As shown in the dashed box of Deformable Convolution in Figure 2, the deformable convolution learns by adding an extra pixel-level offset to the feature map about to perform the convolution operation. Through the learned offset, the original image pixels calibrate adaptively to adjust the position of the sampling point for expanding the receptive field.

The grid G defines the size of the receptive field, for example, $G = (-1,-1),(-1,0),...,(0,1),(1,1)$ defines a 3×3 sampling grid. As shown in Equation (1), in deformable convolution, the position of sampling point is adjusted by adding an offset $\{\Delta p_n | n = 1,...,N\}$, where $N = |G|$, to the regular sampling grid G.

$$y(p_0) = \sum_{p_n \in G} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{1}$$

After adding an offset to the regular grid G, the sampling grid is no longer a regular rectangle but an irregular, adaptively adjustable sampling grid. The adjusted sampling point coordinates $p_0 + p_n + \Delta p_n$ are usually non-integer, so the bilinear interpolation algorithm is applied to obtain the pixel value of the non-integer coordinate point.

2.3. Deformable Convolutional Pyramid

The USTD network adopts the hierarchical deformable features fusion to detect small marine effectively and further enhance the ability of the network to represent deformed objects.

As shown in the dashed box of Deformable Convolutional Pyramid in Figure 2, the multi-level features are extracted by different deformable convolutional layers are fused to form a deformation convolutional pyramid (DCP) module. The pyramid is a two-way structure composed of a bottom-up feature extraction branch, a top-down feature up-sampling branch, and a horizontal connection in the middle.

The bottom-up branch is shown via the blue branch in Deformable Convolutional Pyramid in Figure 2. Bottom-up hierarchical features have more accurate spatial location information but less semantic information. Features of each level come from the feature extraction backbone. Specifically, for ResNeXt[23] used in the USTD, it mainly includes four major convolution stages, con2v, con3v, conv4, and conv5. Each stage performs a down-sampling of feature maps with a step length of 2. Take the features extracted from the last residual block of each stage as the hierarchical feature output of this stage.

The top-down branch is shown via the orange branch in Deformable Convolutional Pyramid in Figure 2. Top-down

hierarchical features have richer semantic features, but carry less spatial location information. Therefore, it is necessary to up-sample top-down hierarchical features, and merge them with bottom-up corresponding hierarchical features through horizontal connections.

In the process of bottom-up sampling, DCP uses deformable convolution. As described in 2.2, deformable convolution can adaptively calibrate the original image pixels through the learned offset. DCP achieves the purpose of enhancing the receptive field by adjusting the position of the sampling point.

The multi-layer feature maps obtained through DCP incorporate both semantic information from high-level features and position information from low-level features. Hierarchical features fusion effectively overcomes the shortcomings of small object feature loss, which enhances the model's ability to detect small underwater targets.

2.4. Domain Generalization via Phased Learning

As shown in Figure 1, compared with general object detection tasks, object detection in underwater faces a multiple domain challenge. In addition, the difficulty of collecting images in various water areas and depths is different, resulting in a severe imbalance in the number of samples collected in various scenes. Such multiple domain environments with unbalance sample distribution will seriously harm the generalization ability of the model.

We propose a domain generalization strategy based on curriculum learning[12] to achieve domain generalization in various underwater scenes, which is named as Phased Learning. Assume the training sample is composed of the samples from 3 different underwater scenarios. Then, the training data can be divided into three subsets Sub_1, Sub_2, and Sub_3. Three training subsets correspond to three training courses. In our case, the training sets are divided according to resolution because images with the same resolution usually come from the same waters, and underwater scenes have similar hydrologic conditions.

As shown in Equation (2), the difficulty of the course is determined by calculating the average loss on the subset, and the loss is composed of classification loss and bounding box regression loss. The smaller the loss, the easier the course is to learn. As shown in Figure 3, we first train a general model on the entire training set, and then implement course learning. The courses are learned from easy to hard, and the models learned in multiple courses are used for ensemble detection by the NMS[19] algorithm.

$$P_{Sub_i} = \frac{1}{|Sub_i|} \sum_{j=1}^{|Sub_i|} L_{cls_j} + \lambda L_{loc_j} \tag{2}$$

2.5. Loss Function based on CascadeRoIHead

As shown in Figure 2, the USTD network adopts a multi-threshold detection head as the detector named CascadeRoIHead to achieve high-precision positioning. The loss function of USTD network consists of the classification loss L_{cls} and the bounding box regression loss L_{loc}. In stage t, CascadeRoIHead uses classifier h_t and regressor f_t for IoU threshold u^t ($u^t > u^{t-1}$). In our work, IoU thresholds of CascadeRoIHead are respectively in [0.5, 0.6, 0.7]. USTD network converges to the optimum by minimizing Equation (3). In Equation (3), in the t stage, $h_t(x^t)$ is the category label given to regional feature x^t by classifier h_t, and $f_t(x^t, b^t)$ is result of regression between regional feature x^t and position coordinate b^t by regressor f_t.

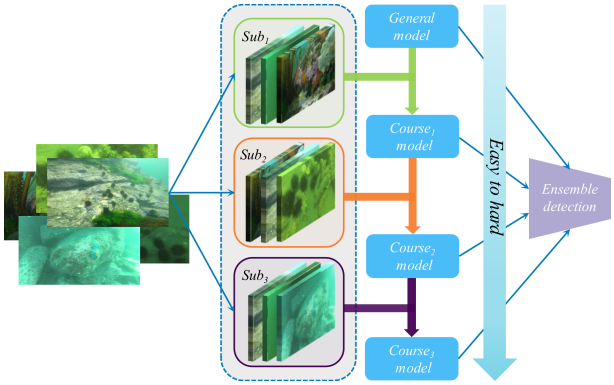

Fig. 3. Illustration of domain generalization via phased learning.

$$L(x^t, g) = L_{cls}(h_t(x^t), y^t) + \lambda[y^t \geq 1] L_{loc}(f_t(x^t, b^t), r) \tag{3}$$

In the USTD network, L_{cls} uses cross-entropy loss. x^t and g are respectively the candidate regional feature of stage t and its corresponding ground truth. In Equation (4), y^t is the category label of x^t determined by IoU threshold u^t, and g_y is the category label of region g.

$$y^t = \begin{cases} g_y & , IoU(x^t, g) \geq u^t \\ 0 & , otherwise \end{cases} \tag{4}$$

At stage t, when the IoU of candidate region is not less than u^t, it is considered as a foreground. As shown in Equation (5), b^t is the bounding box optimized by regressor f_{t-1} of stage $t-1$.

$$b^t = f_{t-1}(x^{t-1}, b^{t-1}) \tag{5}$$

In Equation (3), r is the position of ground-truth object corresponding to x^t. The regression loss L_{loc} of bounding box uses the $smooth_{L1}$ loss[9]. Combining the classification loss L_{cls} and the regression loss L_{loc}, the total loss function of USTD network is shown in Equation (6).

$$L = \sum_{t}^{T} L_{cls}(h_t(x^t), y^t) + \lambda[y^t \geq 1] L_{loc}(f_t(x^t, b^t), r) \tag{6}$$

3. EXPERIMENT

3.1. Dataset

We annotated an Underwater Target Detection Set (UTDS) to evaluate our method, which consists of natural ocean underwater images from the network and are labeled with the labelImg tool[1]. UTDS data set includes about 9K underwater images. About 8K images are used for training and verification, and 1200 images are used for testing. The dataset is annotated with marine life in four categories: holothurian, echinus, scallop, and starfish.

3.2. Detection Performance

Table 1 is the experimental[1] results of USTD network and each comparative method. It can be seen from the table that our method reaches 51.57mAP on the UTDS data set. USTD network outperforms the current conventional object detection algorithms in the field of underwater target detection.

Table 1. mAP performance on UTDS data set. USTD-PL (Phased Learning) indicates the use of phased learning strategy.

Method	Backbone	Image Size	mAP
RetinaNet[16]	ResNeXt101	[1000,1400]	44.20
YOLOv5x[7]	CSP-Net	[1000,1400]	45.08
FoveaBox[13]	ResNeXt101	[1000,1400]	45.41
GA-Faster R-CNN[22]	ResNeXt101	[1000,1400]	45.60
Faster R-CNN[21]	ResNeXt101	[1000,1400]	46.10
Cascade R-CNN[2]	ResNeXt101	[1000,1400]	48.40
USTD(ours)	ResNeXt101	[1000,1400]	**49.50**
USTD-PL(ours)	ResNeXt101	[1000,1400]	**51.57**

Table 2. Accuracy ranking of 2020 China Underwater Robot Professional Contest.

Teams	1	2	**3(ours)**	4	5
mAP	53.24	52.62	**51.67(USTD)**	51.53	51.47

USTD network can achieve SOTA in underwater target detection tasks. The main reasons include the following aspects. First, the USTD network introduces a deformable convolutional module, which can effectively improve the problem of occlusion and deformation of marine organisms. Secondly, the DCP in the USTD network enhances the fine-grained detection capabilities of the model, which significantly enhances the detection performance of small target marine organisms. Meanwhile, based on the USTD network, we also propose a strategy of domain generalization via phased learning, which solves the problem of multiple domain environments with unbalance sample distribution. The phased learning strategy further improves the USTD network's generalization ability in various underwater scenes.

[1]https://github.com/tzutalin/labelImg

We participated in the 2020 China Underwater Robot Professional Contest (URPC)[2]. The ranking results of the competition are shown in Table 2.

3.3. Ablation

Furthermore, to verify the USTD network's effectiveness, we conducted ablation experiments on the USTD network These comparison modules include Backbone, DCP, Deformable Convolution Network (DCN), Phased Learning, and Random Rotate 90°. Table 3 shows the results of USTD network ablation experiment. Based on the baseline model, we replaced the backbone with ResNeXt101 with deeper layers, increasing 0.7 mAP. After adding the DCN module, there is a 0.6 mAP improvement to the accuracy. The phased learning strategy enhances the generalization of the model in multi-domain environments. The final evaluation result of the USTD network is increased by 2.07 mAP by the phased learning strategy

Table 3. Ablation study of the USTD model.

Backbone	DCP	DCN	Phased Learning	Rotate	mAP
ResNet50[10]	✓	-	-	-	47.7
ResNeXt101	✓	-	-	-	48.4
ResNeXt101	✓	✓	-	-	49.0
ResNeXt101	✓	✓	-	✓	49.5
ResNeXt101	✓	✓	✓	✓	**51.57**

4. ACKNOWLEDGEMENT

This research is funded by National Natural Science Foundation of China(No.61902093), Key-Research and Development Program of Guangdong (2020B0101380001), Guangdong Major Project of Basic and Applied Basic Research (NO.2019B030302002), Natural Science Foundation of Guangdong (No. 2020A15150652), Ricoh-HITsz Joint Research Center(HX20190061), PINGAN-HITsz Intelligence Finance Research Center (YH-OAS-TICS-2020- 5900101), CCF-Tencent Open FundCCF-Tencent RAGR20210105

5. CONCLUSION

In this work, we propose a two-stage underwater small target detection(USTD) network. Deformable convolutional pyramid (DCP) is proposed to enhance the network's ability to detect small underwater targets. In addition, to solve the problem of multiple domain environments with unbalance sample distribution, we also propose a strategy of domain generalization via phased learning. We annotated an underwater target detection set (UTDS) to evaluate our method, and our method achieves the SOTA in the UTDS set.

[2]http://2020.cnurpc.org/index.html

6. REFERENCES

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[3] Long Chen, Zhihua Liu, Lei Tong, Zheheng Jiang, Shengke Wang, Junyu Dong, and Huiyu Zhou. Underwater object detection using invert multi-class adaboost with deep learning. In *2020 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2020.

[4] Zhe Chen, Zhen Zhang, Fengzhao Dai, Yang Bu, and Huibin Wang. Monocular vision-based underwater object detection. *Sensors*, 17(8):1784, 2017.

[5] Yang Cong, Baojie Fan, Dongdong Hou, Huijie Fan, Kaizhou Liu, and Jiebo Luo. Novel event analysis for human-machine collaborative underwater exploration. *Pattern Recognition*, 96:106967, 2019.

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[7] Glenn Jocher et al. Yolov5 `https://github.com/ultralytics/yolov5`, 2021.

[8] Enric Galceran, Vladimir Djapic, Marc Carreras, and David P Williams. A real-time underwater object detection algorithm for multi-beam forward looking sonar. *IFAC Proceedings Volumes*, 45(5):306–311, 2012.

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Lars Henriksen. Real-time underwater object detection based on an electrically scanned high-resolution sonar. In *Proceedings of IEEE Symposium on Autonomous Underwater Vehicle Technology (AUV'94)*, pages 99–104. IEEE, 1994.

[12] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020.

[13] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, and Jianbo Shi. Foveabox: Beyond anchor-based object detector. *arXiv preprint arXiv:1904.03797*, 2019.

[14] Chongyi Li, Saeed Anwar, and Fatih Porikli. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognition*, 98:107038, 2020.

[15] Sen Lin and Kaichen Chi. Underwater image enhancement based on structure-texture reconstruction. *arXiv preprint arXiv:2004.05430*, 2020.

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

[17] Hong Liu, Pinhao Song, and Runwei Ding. Wqt and dg-yolo: towards domain generalization in underwater object detection. *arXiv preprint arXiv:2004.06333*, 2020.

[18] Linda J Mullen, V Michael Contarino, Alan Laux, Brian M Concannon, Jon P Davis, Michael P Strand, and Bryan W Coles. Modulated laser line scanner for enhanced underwater imaging. In *Airborne and In-Water Underwater Imaging*, volume 3761, pages 2–9. International Society for Optics and Photonics, 1999.

[19] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 850–855. IEEE, 2006.

[20] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[22] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2965–2974, 2019.

[23] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.