E3S Web of Conferences **603**, 04014 (2025) https://doi.org/10.1051/e3sconf/202560304014
ISGST 2024

Efficient-Deformable-DETR: Enhancing underwater small target detection in complex environments

Haozheng Xie[1], *Guanghong* Xin[1,2*], *Shuo* Kang[1], and *Qiongdan* Xie[1,2]

[1]University of Sanya, Sanya, Hainan, 572000, China
[2]University of Sanya Academician Zhai Mingguo Workstation, Sanya, Hainan, 572000, China

Abstract. The marine environment is facing multiple challenges, and the development of underwater small target detection technology is crucial for protecting marine ecology. This paper focuses on the problem of small target detection in complex underwater environments and proposes an improved model based on the Deformable-DETR model. By introducing global response normalization and a fully convolutional mask autoencoder framework, the model's feature aggregation ability and detection accuracy are enhanced. At the same time, a gradient harmonization mechanism is used to solve the problem of positive and negative sample imbalance, which enhances the model's learning performance. Experimental results show that the improved model achieves a mAP value of 84.5% on the URPC2020 dataset, an increase of 1.7% over the original model. This technical optimization not only improves the accuracy of underwater small target detection but also contributes to marine ecological monitoring, rational utilization of resources, and environmental protection, promoting the realization of green environment and sustainable development goals.

1 Introduction

The development of computer vision and underwater robotics has facilitated the widespread application of underwater target detection in various fields, including underwater exploration and biological monitoring [1-2]. However, conventional detection models often struggle to cope with the complexity of the underwater environment and the diverse scales of targets [3]. To tackle this challenge, deep learning technology has been increasingly employed, showcasing excellent performance in target detection [4]. Transformer-based approaches, especially the DEtection TRansformer (DETR) [5] and its variants, have demonstrated promising results. Although DETR-based methods achieve true end-to-end detection, they suffer from long training times and slow convergence rates. To overcome these limitations, the Deformable Transformers for End-to-End Object Detection (Deformable-DETR) was proposed [6], leading to the emergence of newer and relatively mature categories of detection methods based on DETR. However, a fundamental question remains: can the Deformable-DETR model accurately capture the intricate features and dynamic changes of small targets

*Corresponding author: guanghongxin@sanyau.edu.cn

in complex underwater environments? Building upon this progress, this paper introduces an enhanced model called Efficient-Deformable-DETR for underwater target detection. By incorporating global response normalization, a fully convolutional mask autoencoder framework, and a gradient harmonization mechanism to address the imbalance between positive and negative samples, our model enhances detection performance. It contributes to the advancement of underwater target detection technology, promoting environmental sustainability and the accurate identification of underwater microorganisms, microplastics, and other targets. This enables better monitoring of marine health, addressing pollution and ecological concerns.

2 Our approach

This paper proposes an enhanced model, Efficient-Deformable-DETR, for underwater small target detection, based on the Deformable-DETR architecture. As depicted in Fig. 1, the model architecture consists of four main modules: Backbone, Encoder, Decoder, and Head. The process begins with the Backbone module, where a 224x224 image is inputted, and feature extraction is performed using the ConvNeXt V2-T network [7]. This network processes the image through four stages, generating four sets of feature maps with varying scales. To ensure accurate spatial alignment, the feature maps from the second to fourth stages undergo Group Normalization (GN) [8], a technique that normalizes the features to produce three fixed-dimensional encodings. These encodings are then combined with hierarchical encodings to derive a Reference Point, denoted as . The Reference Point is subsequently fed into the Encoder-Decoder module, where it undergoes further feature enhancement. This module is designed to refine the features and improve the detection accuracy. Finally, bounding boxes and categories are predicted through the head module to output the final detection result.

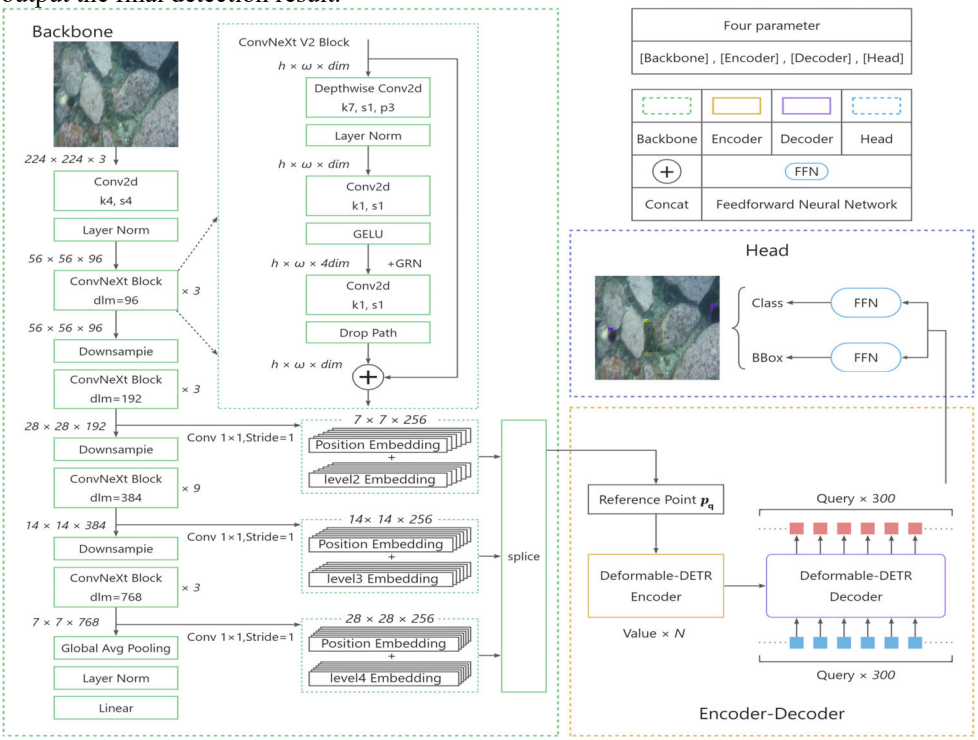

Fig. 1. Structure of the Efficient-Deformable-DETR Model.

2.1 Backbone enhancement

The Deformable-DETR model's reliance on ResNet50 as its backbone poses challenges in feature extraction within complex underwater environments due to its fixed receptive field and spatial invariance [9]. To address this, we replace ResNet50 with ConvNeXt V2-Tiny (ConvNeXt V2-T), which is more suited for feature extraction in dynamic underwater settings. ConvNeXt V2-T leverages self-supervised learning through the Fully Convolutional Masked Autoencoder (FCMAE), enabling it to capture both global and local features, and incorporates Global Response Normalization (GRN) techniques to optimize the detection architecture. The lightweight ConvNeXt V2-T network is chosen as the backbone, utilizing down-sampling layers and depth-wise separable convolution to reduce spatial dimensions and model parameters, respectively [10]. The integration of layer normalization, regularization, and GRN techniques enhances feature extraction capability. In this paper, targets smaller than 32 × 32 pixels will be considered as small targets. Table 1 compares the performance of the original ResNet50 backbone and ConvNeXt V2-T as the backbone of the Deformable-DETR model. The results demonstrate that ConvNeXt V2-T achieves higher mean Average Precision (mAP) and mAP for small objects (mAP_s) values, indicating its superior performance in detecting small underwater targets.

Table 1. Comparing ResNet50 and ConvNeXt V2-T Backbones for Deformable-DETR Model.

Model Name	**Input Size**	**mAP(%)**	**mAP_s(%)**
Deformable-DETR (ResNet50)	1×224×224×3	82.8	67.0
Deformable-DETR (ConvNeXt V2-T)	1×224×224×3	**84.1**	**70.2**

2.2 Loss function improvement

To tackle class imbalance in datasets, Efficient-Deformable-DETR utilizes the Gradient Harmonizing Mechanism-based Classification Loss (GHM-C) and Regression Loss (GHM-R), alongside the Generalized Intersection over Union Loss (GIoULoss) for IoU loss [11]. In contrast, Deformable-DETR employs Focal Loss (FL) for bounding box loss, which focuses on hard-to-classify samples to enhance performance on such cases [12].

However, FL overlooks the gradient distribution, risking misclassification of outliers post-convergence. The Gradient Harmonizing Mechanism (GHM) introduces gradient density, defined as the count of examples with similar gradient norms within a unit area. The formula for Gradient Density is as follows:

$$GD(g) = \frac{1}{l \in N} \sum_{k=1}^{N} \delta_{\epsilon}(g_k, g) \quad (1)$$

where g is the gradient norm, l is the number of unit regions, $\in$ is the length of each unit region, N is the number of examples, and $\delta_{\epsilon}(g_k, g)$is an indicator function that equals 1 if $y - \frac{\epsilon}{2} \leq x < y + \frac{\epsilon}{2}$ and 0 otherwise.

The equation for GHM-C Loss is as follows:

$$L_{GHM-C} = \frac{1}{N} \sum_{i=1}^{N} \frac{L_{CE}(p_i, p_i^*)}{GD(g_i)} \quad (2)$$

where L_{CE} is the cross-entropy loss, p_i is the predicted probability, p_i^* is the ground truth, and $g_i = |p_i - p_i^*|$ is the gradient norm.

The equation for GHM-R Loss is as follows:

$$L_{GHM-R} = \frac{1}{N} \sum_{i=1}^{N} \frac{ASL1(x_i, x_i^*)}{GD(g_i)} \quad (3)$$

Where ASL1 is the asymmetric smooth L1 loss, $x_i$is the predicted box coordinate, x_i^* is the ground truth, and g_i is the gradient of ASL1 concerning x_i. As shown in Table 2,

employing GHMLoss as the loss function for Deformable-DETR enhances performance both in detecting small objects and overall object detection.

Table 2. Comparing FL and GHMLoss Loss Functions for Deformable-DETR Model.

Model Name	Input Size	mAP(%)	mAP_s(%)
Deformable-DETR (FL)	1×224×224×3	82.8	67.0
Deformable-DETR (GHMLoss)	1×224×224×3	**83.1**	**69.1**

3 Experiments and analysis

3.1 Dataset and data pre-processing

The Underwater Robot Perception Challenge 2020 (URPC2020) dataset, created by the Dalian Municipal Government and Pengcheng Laboratory [13], contains real-world underwater images featuring four marine organisms: echinus, holothurian, scallop, and starfish. Due to the small size of these targets, URPC2020 was chosen for this study. The dataset includes 4800 training images, 1080 test images, and 1661 validation images. Data preprocessing involves resizing images to 224×224 pixels, applying random Gaussian blur, and adding random noise to improve model robustness and generalization.

3.2 Experimental results and analysis

This paper compares the Efficient-Deformable-DETR model with mainstream detection models on the URPC2020 dataset, using the input image sizes listed in Table 3, and evaluates performance with mAP and mAP_s metrics. Fig. 2 shows detection outcomes, with objects enclosed in color-coded bounding boxes: echinus (blue), holothurian (yellow), scallop (red), and starfish (green). The numbers within the boxes indicate the model's confidence level, with higher percentages reflecting greater certainty in detection and classification. The analysis reveals that the Efficient-Deformable-DETR model (e) excels in detecting small targets in complex underwater environments, offering a lower false negative rate and higher precision compared to other models.

Table 3. Performance parameter comparison of different models on the URPC2020 dataset.

Model Name	Backbone	Loss	Input Size	epochs	mAP(%)	mAP_s(%)
Faster R-CNN	VGG16 +RoIMiX	CEL	224×224×3	150	60.1	55.8
YOLO v5s	Darknet-53	CEL	224×224×3	150	75.2	61.9
DETR	ResNet50	FL	224×224×3	500	74.6	62.5
Deformable-DETR	ResNet50	FL	224×224×3	150	82.8	67.0
Efficient-Deformable-DETR	ConvNeXt V2-T	GHMLoss	224×224×3	150	**84.5**	**72.1**
Efficient-Deformable-DETR	ResNet50	GHMLoss	224×224×3	150	**83.0**	**68.8**
Efficient-Deformable-DETR	ConvNeXt V2-T	FL	224×224×3	150	**83.7**	**69.1**

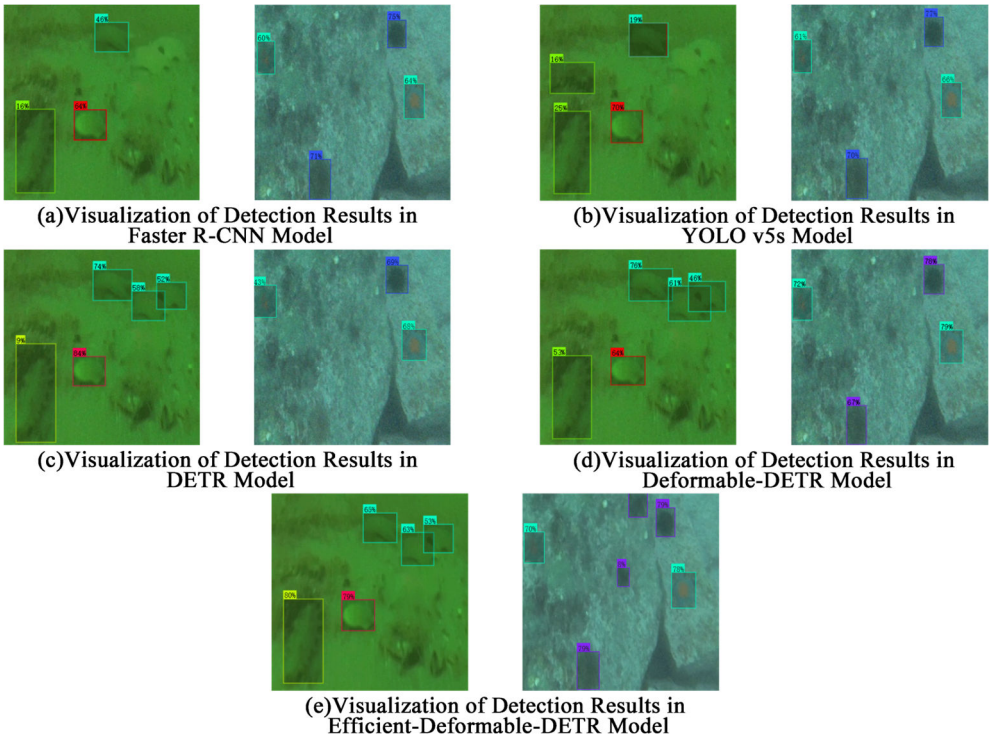

Fig. 2. Visual results of various comparative methods.

4 Conclusion

This paper proposes Efficient-Deformable-DETR, a model designed to enhance the detection capability of small underwater targets in complex environments. By introducing ConvNeXt V2-T and a loss function with a gradient coordination mechanism, the model's detection ability is improved. Comparative experiments and ablation studies on the URPC2020 dataset demonstrate the effectiveness of the proposed design improvements, and visual results further validate this finding. The experimental results show that the improved model achieves a mAP value of 84.5% on the URPC2020 dataset, a 1.7% increase over the original model. Future research directions include combining image enhancement preprocessing methods, such as using efficient super-resolution reconstruction technology to enhance image resolution, to improve the robustness and accuracy of target detection in complex underwater environments. The proposed method has broad application potential in ocean target detection research, contributing to more efficient and accurate ocean environment protection and management.

References

1. M. Zhou, B. Li, J. Wang, K. Fu, A lightweight object detection framework for underwater imagery with joint image restoration and color transformation. J. King Saud Univ. - Comput. Inf. Sci. **35**, 101749 (2023). https://doi.org/10.1016/j.jksuci.2023.101749
2. G. Xin, H. Xie, S. Kang, Design and development of an intelligent connected access integrated training and assessment platform. J. Electr. Syst. **20**, 900-908 (2024).

https://doi.org/10.52783/jes.1250

3. G. Xin, H. Xie, S. Kang, Y. Chen, Y. Jiang, Improved research on coral bleaching detection model based on FCOS model. Mar. Environ. Res. **200**, 106644 (2024). https://doi.org/10.1016/j.marenvres.2024.106644
4. L. Zeng, B. Sun, D. Zhu, Underwater target detection based on Faster R-CNN and adversarial occlusion network. Eng. Appl. Artif. Intell. **100**, 104190 (2021). https://doi.org/10.1016/j.engappai.2021.104190
5. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in Proceedings of Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28 (2020), 213. https://doi.org/10.1007/978-3-030-58452-8_13
6. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: deformable transformers for end-to-end object detection, in Proceedings of 9th International Conference on Learning Representations, Vienna, Austria, May 3-7 (2021), 1. https://doi.org/10.48550/arXiv.2010.04159
7. S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I.S. Kweon, S. Xie, ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders, in Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, June 17-24 (2023), 16133. https://doi.org/10.48550/arXiv.2301.00808
8. Y. Wu, K. He, Group normalization, in Proceedings of the European conference on computer vision (ECCV 2018), Munich, Germany, September 8-14 (2018), 3. https://doi.org/10.48550/arXiv.1803.08494
9. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, June 27-30 (2016), 770. https://doi.org/10.1109/CVPR.2016.90
10. F. Chollet, Xception, Deep learning with depthwise separable convolutions, in Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, July 21-26 (2017), 1251. https://doi.org/10.48550/arXiv.1610.02357
11. B. Li, Y. Liu, X. Wang, Gradient harmonized single-stage detector, in Proceedings of the AAAI conference on artificial intelligence, Honolulu, USA, January 27-February 1 (2019), 8577. https://doi.org/10.1609/aaai.v33i01.33018577
12. T.Y. Lin, P. Goyal, R. Girshick, K. Hef, P. Dollár, Focal loss for dense object detection, in Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October 22-29 (2017), 2980. https://doi.org/10.48550/arXiv.1708.02002
13. C. Liu, H. Li, S. Wang, M. Zhu, D. Wang, X. Fan, Z. Wang, A dataset and benchmark of underwater object detection for robot picking, in Proceedings of 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shenzhen, China, July 5-9 (2021), 1. https://doi.org/10.1109/ICMEW53276.2021.9455997