ISSN 2522-3488 Vol. 9, Issue 2: 78-83, DOI: 10.25236/IJNDES.2025.090213

Research on Underwater Object Detection Method Based on Deep Learning

Yuan Liu[1], Yang Xu[1], Xingkun Li[2], Guodong Chen[1,*]

[1]School of Naval Architecture and Port Engineering, Shandong Jiao Tong University, Weihai, Shandong Province, 264209, China
[2]School of Physical Sciences, Qingdao University, Qingdao, Shandong Province, 266071, China
[*]Corresponding author: lyliuyuan2000@163.com

Abstract: *Underwater environments present unique and challenging conditions for target detection. Factors such as light attenuation, suspended particles, and water turbulence contribute to significant color distortion, low contrast, complex target morphology, and dense background clutter. To address these issues, this paper proposes UWYOLO, a specialized model based on the YOLOv8 framework that incorporates the Large Separable Kernel Attention (LSKA) module. Designed to enhance detection performance in such difficult conditions, UWYOLO improves the extraction of shape features and increases robustness, particularly for detecting small targets in cluttered backgrounds. By emphasizing shape information over texture, the model also demonstrates improved applicability in practical underwater tasks such as biodiversity monitoring and environmental surveys. Experimental results show that UWYOLO achieves a mean average precision (mAP) of 84.9% on a specialized underwater dataset, exceeding the baseline YOLOv8 by 1.0%, confirming its advanced accuracy in underwater target detection.*

Keywords: *Underwater object detection; Large separable kernel attention; Deep learning; Shape feature extraction; UWYOLO*

1. Introduction

In recent years, with the extensive exploitation and utilization of marine resources, underwater target detection has received increasing attention as a key technology in this field. From studying marine biodiversity to inspecting underwater structures and navigating autonomous underwater vehicles, accurate detection of underwater targets is essential for these applications. However, compared to terrestrial environments, underwater settings present challenges such as low visibility, low contrast, complex background noise, texture distortion, and variations in target shape and texture[1-4], all of which significantly increase the difficulty of detection tasks.

The emergence and development of artificial intelligence (AI) have made deep learning a major driver of technological innovation. In particular, convolutional neural networks (CNNs) have demonstrated powerful capabilities in image recognition and object detection. With continuous improvements in network architectures and training strategies, deep learning models have achieved substantial performance gains in underwater target detection, in some cases matching or even surpassing human-level performance. Among these models, one-stage detectors such as SSD[5] and the YOLO series[6-9] directly predict target locations and categories in a single forward pass, offering relatively low computational complexity and better real-time performance. In contrast, two-stage detectors such as Fast R-CNN[10] and Faster R-CNN[11] incorporate a Region Proposal Network (RPN) to first generate candidate regions, then classify and refine each proposal. While these tend to achieve higher accuracy by leveraging full-image convolutional features, their speed is limited by the additional processing of region proposals.

To address the various challenges in underwater environments, researchers have proposed numerous methods for underwater image object detection. Liu et al.[12] introduced TC-YOLO, a lightweight framework based on YOLOv5s, which incorporates a Transformer self-attention module (CSP-TR Block) at the end of the backbone to enhance global feature extraction, and a coordinate attention mechanism (CSP-CA Block) in the neck to improve localization of small targets. Luo et al.[13] proposed YOLO-DAFS, an improved version based on YOLOv11, which replaces standard bottleneck blocks with a lightweight DualConv structure combining group and pointwise convolutions, and introduces the C2PSF

ISSN 2522-3488 Vol. 9, Issue 2: 78-83, DOI: 10.25236/IJNDES.2025.090213

module to enhance multiscale context fusion through attention mechanisms. Hu et al.[14] developed an enhanced sea urchin detection algorithm based on SSD, incorporating a multidirectional edge detection algorithm to extract spine features and a cross-level feature fusion strategy to improve detection of small targets. Hu[15] proposed a marine organism detection model based on Faster R-CNN, which utilized both VGG16 and ResNet50 as backbones and was trained on the FathomNet 2023 dataset. The experimental results showed that ResNet50 performed better in complex seabed environments. Additionally, Liu et al.[16] improved Faster R-CNN by replacing the backbone with a Swin-Transformer structure and changing RoI pooling to RoI alignment to enhance localization accuracy.

YOLOv8 is a new target recognition network model in the current YOLO series, with high detection accuracy and fast detection speed. In this context, this paper uses the YOLOv8 model to incorporate the Large Separable Kernel Attention (LSKA)[17] into the neck of the network, which enhances the ability of the model to detect small targets and complex backgrounds, improves its robustness against challenging underwater conditions, and ultimately boosts overall detection performance.

2. Methods

2.1. The Overview of YOLOv8

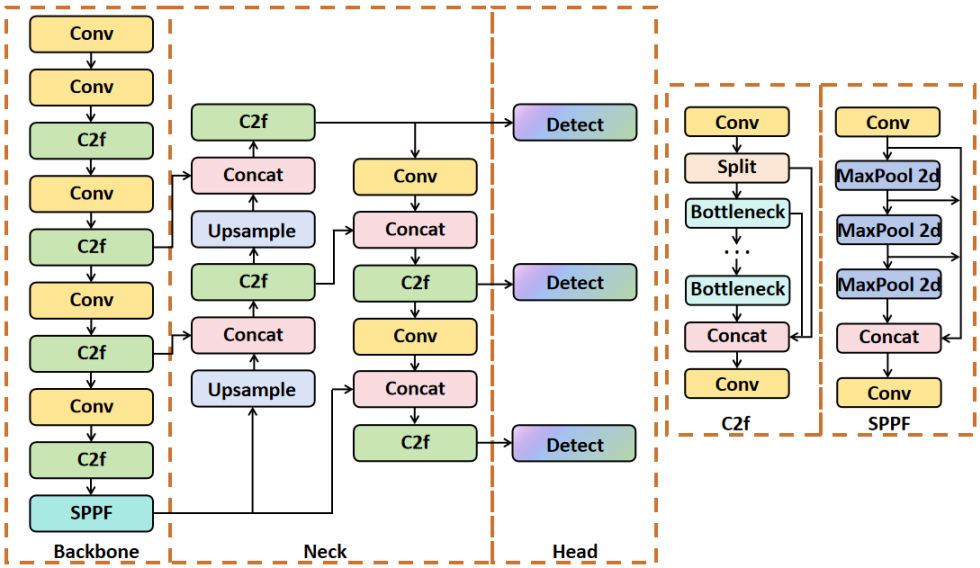

Figure 1: The structure of YOLOv8.

YOLOv8 adopts an encoder-decoder architecture composed of three core components: backbone, neck and head, as shown in Figure 1. The Backbone is constructed using an enhanced CSPDarknet structure, which introduces the C2f (CSP-to-fused) module as its fundamental building block. This module reduces parameter redundancy and strengthens gradient propagation efficiency. Furthermore, depthwise separable convolutions and dilated convolutions are incorporated to expand receptive fields and refine multi-scale feature representation. The Neck integrates a hybrid structure combining a Feature Pyramid Network (FPN) and a Path Aggregation Network (PAN), facilitating bidirectional feature fusion across different semantic levels. This design enhances the model's capacity to detect objects at various scales. For the detection Head, a decoupled architecture is employed, separating classification and regression branches. The regression component utilizes an integral representation method inspired by Distribution Focal Loss, contributing to more precise bounding box localization. The entire model operates in an anchor-free manner, which simplifies the detection pipeline and improves generalization across diverse object shapes and sizes. To support deployment under different computational constraints, YOLOv8 offers scaled variants including Nano, Small, Medium, Large, and Extra Large, each balancing accuracy and efficiency for specific application scenarios.

ISSN 2522-3488 Vol. 9, Issue 2: 78-83, DOI: 10.25236/IJNDES.2025.090213

2.2. Large Separable Kernel Attention (LSKA)

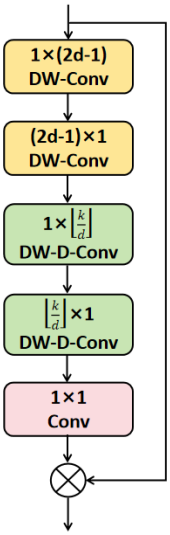

Figure 2: The structure of Large separable kernel attention(LSKA).

As illustrated in Figure 2, the LSKA module operates by decomposing the two-dimensional convolution kernel of the depthwise convolution layer into cascaded horizontal and vertical one-dimensional kernels. The calculation process is detailed in the following formulas. With this working principle, we can directly use large size convolution kernels without adding the computational amount, effectively reducing the number of parameters and computational complexity.

The local feature extraction is computed as follows:

$$\bar{Z}^C = \sum_{H,W} W^C_{(2d-1)\times 1} * (\sum_{H,W} W^C_{1\times(2d-1)} * F^C) \tag{1}$$

The global feature extraction is calculated using the following expression:

$$Z^C = \sum_{H,W} W^C_{\lfloor \frac{k}{d} \rfloor \times 1} * (\sum_{H,W} W^C_{1\times \lfloor \frac{k}{d} \rfloor} * \bar{Z}^C) \tag{2}$$

The final output is obtained through the merged BN computation:

$$A^C = W_{1\times 1} * Z^C \tag{3}$$

$$\bar{F}^C = A^C \otimes F^C \tag{4}$$

In the above formula: d is the expansion rate; * represents convolution; $\otimes$ represents the Hadamard product; $\bar{Z}^C$ represents the output of the depth convolution with two cascade kernel sizes and (2d-1)× (2d-1), which captures local spatial information and is then processed by the following two cascade kernel sizes and the depth expansion convolution of $\lfloor \frac{k}{d} \rfloor \lfloor \frac{k}{d} \rfloor$. $\lfloor \bullet \rfloor$Represents the floor operation. Dilated depthwise convolution is responsible for capturing the global spatial information output by depthwise convolution. A 1×1 convolution kernel $W_{1\times 1}$ is used to perform convolution operations on the intermediate features Z^C, thereby obtaining the output feature maps A^C. The output $\bar{F}^C$ of LSKA is obtained from the Hadamard product of the attention map A^C and the input feature map F^C.

2.3. The Overview of UWYOLO

In the architecture of YOLOv8, the Neck layer is responsible for merging multiscale features extracted by Backbone to enhance the ability of the model to detect targets of different sizes. In this paper, the LSKA module was introduced after the last C2f module in the Neck layer to further enhance the feature fusion and expression ability of the model. The improved overall structure is shown in Figure 3. This design enables the model to maintain efficient performance when processing large cores, more effectively capture and integrate features from different scales, and is more inclined to extract shape features of objects rather than texture features, so as to improve the generalization ability of the model. In addition, the introduction of LSKA module also helps to reduce the GFLOPs of the model, making the model have better real-time processing power while maintaining high performance.

ISSN 2522-3488 Vol. 9, Issue 2: 78-83, DOI: 10.25236/IJNDES.2025.090213

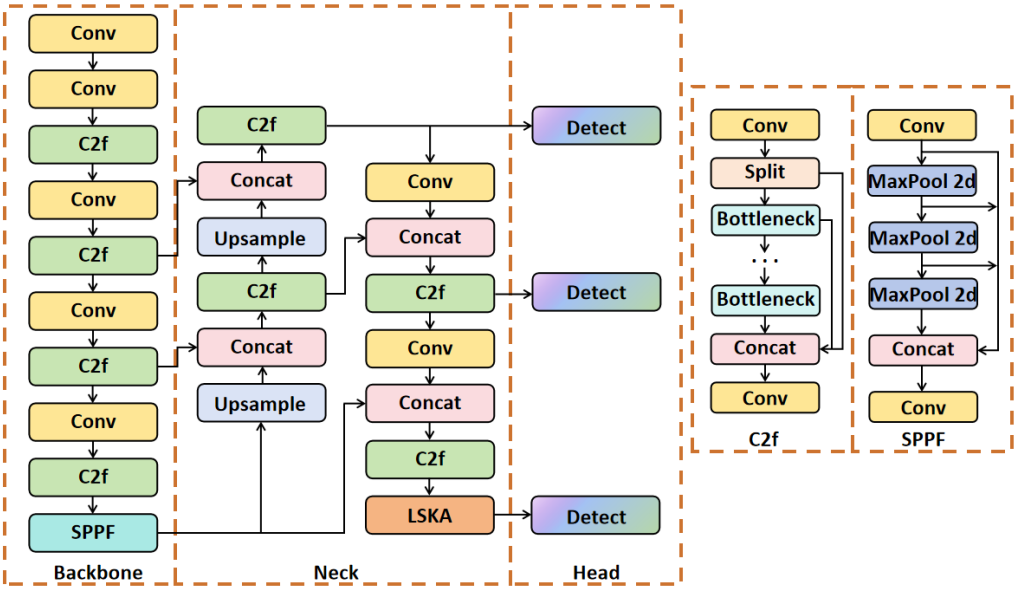

Figure 3: The structure of UWYOLO.

3. Experimental Results

3.1. Experimental Deployment Details

In this study, the model was trained using an NVIDIA GeForce RTX 2080 SUPER GPU. The network parameters were initialized from a normal distribution. Stochastic gradient descent (SGD) was employed as the optimizer, with a momentum of 0.9 and a weight decay of 0.0001. Training was conducted with a batch size of 4 and an initial learning rate of 0.01 over 300 epochs.

During both training and testing, all input images were resized to 640×640 pixels. To improve the generalization capability of the model, data augmentation techniques were applied, including random horizontal and vertical flipping with a probability of 0.5. In summary, the deep learning model was optimized through carefully tuned parameter initialization, high-performance GPU acceleration, and effective image preprocessing strategies.

3.2. Experimental Deployment Details

The dataset used in this experiment was a self-built underwater dataset, and the images were labeled by LabelImg. The dataset has four categories of 5177 images, including starfish, echinus, holothurian and scallop. The data set was divided into training set, test set and validation set in a ratio of 7:2:1.3623 images were randomly selected from the data set as training sets, 1036 images as test sets and 518 images as validation sets. The dataset sample plot is shown in Figure 4.

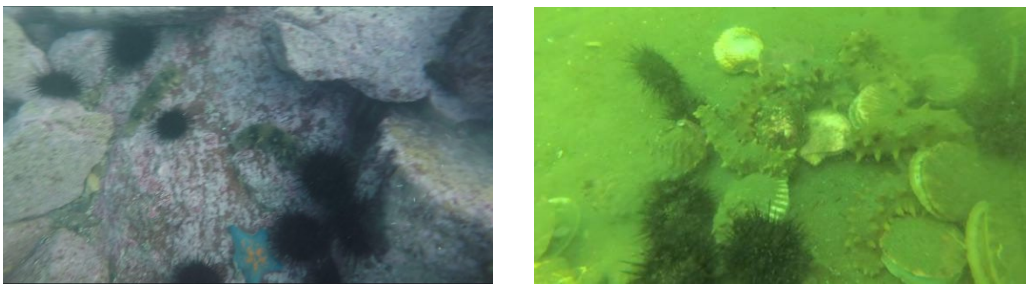

Figure 4: The images of self-build underwater dataset.

In order to comprehensively measure the performance of the model, this study mainly used the metric of Mean Precision (mAP), which is calculated based on a confidence threshold of 0.5 to obtain the Average Precision (AP) value. In addition, to evaluate the performance of the model in more detail, we also introduced recall rate, accuracy (precision), and F1 score as supplementary indicators. Through these comprehensive evaluation indicators, we can conduct in-depth analysis of the model's detection

ISSN 2522-3488 Vol. 9, Issue 2: 78-83, DOI: 10.25236/IJNDES.2025.090213

capability from multiple perspectives.

3.3. Ablation Analysis

As presented in Table 1,the proposed UWYOLO model demonstrates consistent improvements over the baseline YOLOv8 across all evaluation metrics. Specifically, UWYOLO achieves a recall of 79.6%, surpassing YOLOv8 by 2.0%. It also attains a precision of 85.5%, which is 2.3% higher than YOLOv8. Furthermore, the mAP reaches 84.9%, reflecting an improvement of 1.0 percentage points. These results indicate enhanced detection completeness, prediction accuracy, and overall performance compared to YOLOv8.

Table 1: Ablation analysis

Models	Recall	Precision	mAP
YOLOv8	77.6	83.2	83.9
UWYOLO	79.6(+2.0)	85.5(+2.3)	84.9(+1.0)

3.4. Performance Analysis of UWYOLO

According to the comparative analysis results in Table 2, UWYOLO demonstrates significant advantages in multiple detection indicators. The recall rate of this model reaches 79.6%, outperforming other comparison models and demonstrating a stronger target coverage capability. In terms of precision, UWYOLO achieved the highest value of 85.5%, indicating that its prediction results have higher accuracy. Meanwhile, this model maintains a leading position with an average accuracy of 84.9%, and its comprehensive detection performance is the most outstanding. Overall, UWYOLO comprehensively outperforms the comparison models in the three key indicators of recall rate, precision rate and mAP, demonstrating balanced and powerful detection capabilities. It can be seen from this that UWYOLO has greater potential in underwater target detection tasks with complex backgrounds, low visibility and high occlusion.

Table 2: Contrast analysis

Models	Recall	Precision	mAP
YOLOv5s	79.4	82.4	83.7
YOLOv8s	77.6	83.2	83.9
YOLOv10s	77.1	83.7	84.3
SSD	75.8	82.7	82.2
UWYOLO	79.6	85.5	84.9

4. Conclusions

To enhance the detection accuracy of underwater targets, this study proposes an improved detection model named UWYOLO. The approach incorporates the LSKA module into the Neck layer of the YOLOv8 architecture, enabling more precise capture of fine-grained target characteristics. Furthermore, the global knowledge aggregation mechanism integrates contextual information across the entire image, which assists the model in distinguishing between target and background regions in complex underwater environments, thereby improving detection precision. Comparative experiments with other state-of-the-art algorithms confirm that UWYOLO achieves higher average precision and overall detection performance, demonstrating strong adaptability and robustness in challenging underwater scenarios.

References

[1] Jia J, Fu M, Liu X, et al. Underwater object detection based on improved efficientdet[J]. Remote Sensing, 2022, 14(18): 4487.
[2] Jian M, Liu X, Luo H, et al. Underwater image processing and analysis: A review[J]. Signal Processing: Image Communication, 2021, 91: 116088.
[3] Li C, Guo J. Underwater image enhancement by dehazing and color correction[J]. Journal of Electronic Imaging, 2015, 24(3): 033023.
[4] Ancuti C O, Ancuti C, De Vleeschouwer C, et al. Color balance and fusion for underwater image enhancement[J]. IEEE Transactions on image processing, 2017, 27(1): 379-393.

ISSN 2522-3488 Vol. 9, Issue 2: 78-83, DOI: 10.25236/IJNDES.2025.090213

[5] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Cham: Springer International Publishing, 2016: 21-37.
[6] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arxiv preprint arxiv:2004.10934, 2020.
[7] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.
[8] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arxiv preprint arxiv:1804.02767, 2018.
[9] Wang C, He W, Nie Y, et al. Gold-YOLO: Efficient object detector via gather-and-distribute mechanism[J]. Advances in Neural Information Processing Systems, 2023, 36: 51094-51112.
[10] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
[11] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.
[12] Liu K, Peng L, Tang S. Underwater object detection using TC-YOLO with attention mechanisms[J]. Sensors, 2023, 23(5): 2567.
[13] Luo S, Dong C, Dong G, et al. YOLO-DAFS: A Composite-Enhanced Underwater Object Detection Algorithm[J]. Journal of Marine Science and Engineering, 2025, 13(5): 947.
[14] Hu K, Lu F, Lu M, et al. A marine object detection algorithm based on SSD and feature enhancement[J]. Complexity, 2020, 2020(1): 5476142.
[15] Hu J H. Object Detection Model for Marine Organisms Based on Faster R-CNN[J]. Advances in Engineering Technology Research, 2024, 9(1): 567-567.
[16] Liu J, Liu S, Xu S, et al. Two-stage underwater object detection network using swin transformer[J]. IEEE Access, 2022, 10: 117235-117247.
[17] Lau K W, Po L M, Rehman Y A U. Large separable kernel attention: Rethinking the large kernel attention design in cnn[J]. Expert Systems with Applications, 2024, 236: 121352.