International Core Journal of Engineering Volume 11 Issue 5, 2025
ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0040

YOLOv8-CDD: A Salient Target Detection Model for Underwater Cultural Heritage in Complex Environments

Tiancheng Liu*

School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 100044, P.R. China

*Corresponding author: liutiancheng1998@163.com

Abstract

Existing object detection methods often experience a decline in accuracy when directly applied to underwater scenarios, making it difficult to meet the precision and reliability requirements of practical cultural heritage detection tasks. To address common challenges in underwater heritage detection-such as low image contrast, severe target occlusion, and insufficient detection accuracy-this paper proposes a novel detection model, YOLOv8-CDD. The model is designed to improve the recognition of salient target regions in complex underwater conditions, thereby enhancing the automation and efficiency of archaeological detection. YOLOv8-CDD is developed by deeply optimizing the backbone and detection head of the original YOLOv8 architecture. It integrates a C2f-DCNv4 structure, the CBAM attention mechanism, and the Dynamic4 detection head to significantly enhance feature extraction and the perception of small or occluded objects, while maintaining high computational efficiency. The model is systematically evaluated on both the public DUO dataset and a self-constructed underwater archaeological image dataset. Experimental results show that YOLOv8-CDD achieves an mAP@0.5 of 84.9% and an mAP@0.5:0.95 of 63.8% on the DUO dataset, representing improvements of 4.0% and 3.4%, respectively, over the baseline. On the self-constructed dataset, it attains an mAP@0.5 of 70.4%, with a gain of 4.1%.

Keywords

Underwater Object Detection; Cultural Heritage; Salient Region Recognition.

1. Introduction

Object detection technology simulates the human visual attention mechanism by automatically identifying and localizing key targets within images. In underwater archaeology, its application holds significant value in improving both the efficiency and accuracy of cultural heritage detection. However, underwater environments are often characterized by insufficient lighting, dense suspended particles, and turbid water, all of which severely degrade image quality and pose substantial challenges to target recognition. Existing detection methods tend to perform poorly under conditions of low contrast, heavy occlusion, and small-scale targets, making them inadequate for meeting the accuracy and robustness requirements of underwater archaeological tasks. Therefore, object detection methods tailored for underwater environments must be designed to enhance recognition performance under complex conditions. Such advancements not only support the development of intelligent underwater archaeological equipment but also provide critical technical support for the preservation of marine cultural heritage.

Chenping Fu[1] proposes a Residual Feature Transfer Module (RFTM), which leverages the Heavy Degradation Prior (HDP) to optimize underwater object detection and improve the performance of

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0040

CNN-based detectors. Ranjith Dinakaran[2] first employs a DCGAN-based image enhancement approach to improve the detectability of low-quality underwater images using SSD. Then, Particle Swarm Optimization (PSO) is introduced to fine-tune the learning rate, momentum, and weight decay of the DCGAN+SSD framework, thereby enhancing the stability and generalization ability of the detection model. Long Chen[3] proposes the SWIPENET-CMA model, which combines dilated convolutions and skip connections to address noise interference in small-object detection in underwater environments and improve detection accuracy.In subsequent research, two algorithms—Noise Removal (NR) and Factor-Agnostic Gradient Reweighting (FAGR)—are introduced[4] to eliminate unreliable annotations from the dataset and reduce the impact of label noise on the model. These methods lead to enhanced object detection performance on underwater datasets with imbalanced noise distributions.Jiahao Qi[5] proposes UTD-Net, which combines hyperspectral unmixing (HU) techniques with a depth estimation model to separate mixed pixels of objects and water bodies. This approach reduces background interference from the underwater environment and improves object detection accuracy.Bowen Wang[6] proposes a Dual-Branch Joint Learning Network (DJL-Net), which integrates edge enhancement and adaptive feature fusion to strengthen edge features of small objects, thereby improving detection stability.Jian Zhang[7] enhances the capture of both global and local information by introducing RTMDet as the backbone and integrating the BoT3 module to improve feature extraction capability.Fei Lei[8] introduces the Swin Transformer as the backbone in YOLOv5, optimizes PANet for multi-scale feature fusion, and improves the confidence loss function, resulting in enhanced detection accuracy and recall.Sixian Cai[9] introduces a dual-YOLOv5 training framework combined with a batch filtering module, leveraging a co-teaching mechanism to optimize object detection and enhance robustness against noisy samples.Fubin Zhang[10] enhances the accuracy and robustness of underwater object detection by integrating the CIB module into YOLOv8 to enlarge the receptive field, employing the PSA module to optimize computational efficiency, and incorporating the Neck structure from Gold-YOLO to improve feature fusion.Alsuwaylimi[11] improves YOLOv8-Seg by integrating RepBlock and SimConv modules to enhance feature extraction and optimize parameter efficiency. These improvements lead to better instance segmentation accuracy and real-time performance in underwater debris detection.Shaobin Cai[12] proposes AGW-YOLOv8 by integrating attention mechanisms, GSConv, and WIoU to achieve a balance between model lightweighting and accuracy optimization, thereby improving detection stability.Jie Chen[13] proposes Dynamic YOLO by incorporating a lightweight DCNv3-based backbone, a dynamic feature fusion framework, and an extended decoupled detection head into the YOLOv8 architecture. This model enhances feature extraction for small objects while reducing computational cost.

This paper proposes an improved object detection model, YOLOv8-CDD, designed for complex underwater environments. Built upon the YOLOv8 architecture, the model addresses challenges such as severe occlusion of small targets and weak feature representation in underwater imagery by optimizing both the backbone network and detection head. Specifically, the C2f-DCNv4 module is introduced to enhance feature extraction, the CBAM attention mechanism is integrated to strengthen the focus on salient regions, and a Dynamic4 detection head is designed to improve multi-scale target perception. This enhanced framework aims to balance detection accuracy and computational efficiency, thereby improving the practicality and robustness of the model in real-world underwater cultural heritage detection tasks.

2. Methods

2.1 YOLOv8-CDD Model Architecture

YOLOv8-CDD is optimized through the following steps. First, the original C2f structure is enhanced by integrating DCNv4, which enables adaptive receptive field adjustment. This allows the model to more effectively handle deformed objects and low-contrast regions commonly found in underwater environments. Compared to DCNv3, DCNv4 also improves computational efficiency, significantly reducing the deployment cost of adaptive receptive fields. Second, the CBAM attention mechanism

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0040

is incorporated to strengthen the network's focus on critical features, thereby improving the accuracy of salient object detection underwater. Finally, the detection head is replaced with a novel Dynamic4 structure, which leverages multi-dimensional adaptive attention mechanisms to optimize feature representation and significantly boost detection performance.

Through the above improvements and optimizations, a YOLOv8-CDD network more suitable for underwater object detection is developed. The architecture of the improved network is illustrated in Figure 1.

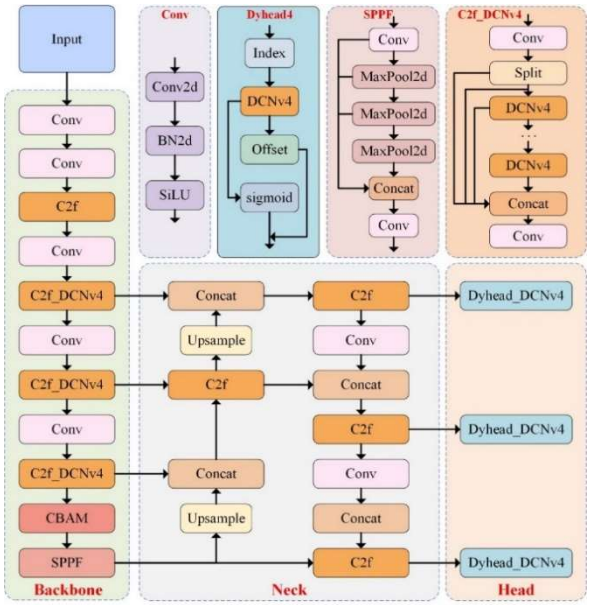

Figure 1. Network Structure of YOLOv8-CDD

2.2 Cross-Stage Partial Feature Fusion based on Deformable Convolutions

Convolutional Neural Networks (CNNs) have achieved remarkable progress in computer vision tasks. However, their fixed receptive fields limit their ability to adapt to object deformation and complex backgrounds. To address this limitation, Deformable Convolution (DCN) was introduced. After multiple iterations, DCN has evolved to its latest version, DCNv4[14]. The previous version, DCNv3 , improved computational efficiency through a grouping mechanism, but still faced limitations in both computational cost and feature learning capability, which motivated the development of DCNv4.

DCNv4 introduces improvements in two key aspects. First, it removes the Softmax normalization constraint used in DCNv3, allowing the convolutional kernel to assign unconstrained weights to different sampling points. This enhances the model's learning capacity and enables more flexible feature distribution. Second, DCNv4 significantly improves computational efficiency by optimizing CUDA thread scheduling and reducing redundant memory access (Memory Access Cost, MAC), resulting in a 2–3× increase in inference speed compared to DCNv3.

These improvements make DCNv4 particularly advantageous for small object detection and target recognition in complex environments. When applied to challenging conditions such as underwater imagery, DCNv4 demonstrates enhanced adaptability.

Given an input feature map X, the grid sampling offset is denoted as Δp_k, and p_k represents the relative position of the k-th sampling point in the convolutional kernel. The convolution kernel weights are denoted as W_k, with an additional bias term b and attention weight Δm_k. K indicates the total number of sampling points in the kernel, p represents the pixel location on the feature map, and the output feature map Y at location p is computed as follows:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0040

$$Y(\mathrm{p}) = \sum_{\mathrm{k}=1}^{K} W_k \cdot X(p + p_k + \Delta p_k) \cdot \Delta m_k + b \tag{1}$$

In this study, DCNv4 is integrated into the C2f structure of YOLOv8 (as illustrated in figure 2) to enhance performance in underwater cultural heritage detection. The original C2f structure promotes feature flow by passing a subset of channels through successive layers; however, it relies on fixed 3×3 standard convolutions, which limits its ability to cope with challenges in underwater imagery-such as object damage, occlusion, and shape variation-due to its restricted receptive field and lack of adaptability.

By incorporating DCNv4, the model gains the capability to dynamically adjust convolutional sampling locations based on image content, enabling more flexible feature extraction. This enhances its adaptability to low-contrast conditions, small-scale targets, and complex backgrounds. Furthermore, the sparse attention mechanism embedded in DCNv4 helps the model focus more precisely on critical heritage regions while effectively suppressing background noise, thereby further improving detection accuracy and robustness.

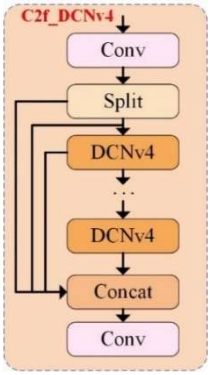

Figure 2. Structure of C2f_DCNv4

2.3 Channel and Spatial Attention Integration Module

To enhance the model's ability to perceive salient regions in complex underwater environments, this study integrates the Convolutional Block Attention Module (CBAM) into the backbone network. By strengthening the network's focus on critical features, CBAM significantly improves the accuracy of underwater cultural heritage detection [15].

The channel attention mechanism aims to evaluate the relative importance of each channel in feature representation. In CBAM, for each channel of the input feature map, global average pooling (GAP) and global max pooling (GMP) are applied to extract channel-wise statistical information. These pooled features are then passed through a shared multilayer perceptron (MLP) to generate the channel attention weights. The computation of this branch is defined as follows:

$$M_c(F) = \sigma\begin{pmatrix} MLP\big(AvgPool(F)\big) \\ +MLP\big(MaxPool(F)\big) \end{pmatrix} \tag{2}$$

In the above equation, $M_c(F)$ denotes the generated channel attention map, σ is the Sigmoid activation function, and AvgPool and MaxPool represent global average pooling and global max pooling operations, respectively.

The spatial attention mechanism is designed to identify the most informative spatial regions within the feature map. In CBAM, global max pooling and global average pooling are first applied along the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0040

channel axis to extract spatial-level statistical features. These two pooled feature maps are then concatenated and passed through a 7×7 convolution layer to generate the spatial attention map. The computation is formulated as follows:

$$M_s(F) = \sigma\left(f^{7\times7}([AvgPool(F_c); MaxPool(F_c)])\right) \tag{3}$$

In the above equation, $M_s(F)$ denotes the spatial attention map, $f^{7\times7}$ represents the 7×7 convolution operation, and $[;]$ indicates concatenation along the channel dimension.

Finally, the feature representation enhanced by the combined channel and spatial attention is computed as:

$$F' = M_s(M_c(F) \odot F) \odot F \tag{4}$$

In the above equation, $\odot$denotes element-wise multiplication.

In the YOLOv8 architecture, the CBAM module is integrated before the Spatial Pyramid Pooling Fast (SPPF) module within the backbone network. This design fully leverages the advantages of CBAM by enhancing feature attention before the multi-scale pooling operation performed by SPPF. The rationale behind this integration is that SPPF is responsible for optimizing multi-scale feature representation, while CBAM improves the focus of the features beforehand, ensuring that more discriminative information is passed into the SPPF module.

This structural design proves particularly effective in underwater cultural heritage detection, as it enhances the recognizability of target artifacts in low-contrast and high-background-noise environments, thereby improving overall detection accuracy and robustness.

2.4 Adaptive Object Detection Head based on DCNv4

The object detection head serves as the decision-making module of a detection model, responsible for receiving feature maps from the backbone and feature pyramid modules, and performing core tasks such as object classification, bounding box regression, and confidence score prediction. Traditional detection heads typically employ fixed convolutional structures, which limits their adaptability to multi-scale targets, complex background interference, and task-specific variations. This rigidity hinders dynamic adjustment of feature representation and ultimately degrades overall detection performance and robustness.

To address these limitations, the Dynamic Head is proposed. It leverages multi-dimensional adaptive attention mechanisms to optimize feature hierarchy, spatial information, and task-level representations, thereby achieving a better balance between detection accuracy and computational efficiency [16].

The dynamic detection head employs spatially-aware attention to adaptively transform features, enhancing focus on critical regions while suppressing background noise. The computation is formulated as follows:

$$\pi_S(F) = \frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K} w_{l,k} \cdot F(l; p_k + \Delta p_k; c) \cdot \Delta m_k \tag{5}$$

In the above formulation, $\pi_S(F)$ denotes the spatial attention weights, which are used to adjust the response values of the feature map at different spatial locations. L represents the number of feature hierarchy levels, and K denotes the number of sampling points in deformable convolution, i.e., the number of neighborhood points sampled per pixel. Δp_k is the dynamically learned offset, $w_{l,k}$ is

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0040

the fusion weight, Δm_k is the normalized modulation factor, and c is the channel index, indicating the specific feature channel being considered during attention computation.

Finally, object detection typically involves two subtasks: object classification and bounding box regression. These tasks exhibit significant differences in their feature representation requirements. However, conventional detection heads often rely on shared features for joint modeling, which limits their ability to simultaneously capture the semantic richness needed for classification and the spatial precision required for regression, thus constraining overall detection performance.

The dynamic detection head introduces task-aware attention, which adaptively adjusts features according to task-specific demands. This enables the classification branch to focus on global semantic information, while the regression branch emphasizes precise boundary localization, thereby improving the accuracy of bounding box prediction. The computation is expressed as follows:

$$\pi_C(F) = max\begin{pmatrix}\alpha_1(F)\cdot F_c + \beta_1(F),\\ \alpha_2(F)\cdot F_c + \beta_2(F)\end{pmatrix} \tag{6}$$

In the above equation, $\pi_C(F)$ represents the task-aware attention weights, which are used to optimize feature channels to better suit different tasks. The parameters α, β are learnable scalars used to adjust the weighting.

Although the dynamic detection head effectively improves detection accuracy through multi-dimensional attention mechanisms, its original implementation employs DCNv2 for spatial feature transformation, which still suffers from high computational cost, limited expressive capacity, and substantial memory usage. To further enhance the modeling efficiency and feature representation capability of the detection head, this study proposes Dynamic4, which replaces DCNv2 with the more efficient and expressive DCNv4. This substitution enables more effective spatial modeling and more precise object localization. The structure of Dynamic4 is illustrated in Figure 3.

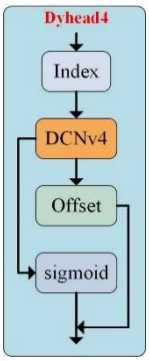

Figure 3. Structure of Dynamic4

Within the dynamic detection head framework, the spatially-aware attention module plays a key role in enhancing the feature representation capability of the detection head. By replacing DCNv2 with DCNv4, the detection head can capture spatial information more precisely, significantly improving detection performance for deformed objects. In conjunction with the task-aware attention mechanism, DCNv4 not only adapts to varying object shapes but also dynamically adjusts feature representations based on task-specific requirements, thereby optimizing bounding box regression accuracy.

Moreover, to ensure the stability of multi-scale feature fusion, a hierarchical cascading strategy is introduced into the detection head. This strategy strengthens the interaction between high-level semantic features and low-level spatial features, further improving the robustness of object detection.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0040

3. Experimental Results and Analysis

3.1 Experimental Setup and Design

This study first conducts extensive experiments on the publicly available DUO (Detecting Underwater Objects) dataset to evaluate the effectiveness of the proposed model in underwater object detection tasks. Beyond conventional detection experiments, a salient region detection task is further designed for underwater cultural heritage imagery, aiming to automatically localize potential areas of archaeological significance within the images.

The underwater archaeological image dataset is constructed from multi-source data, comprising two primary sources: manually annotated images from publicly available datasets such as HURLA and SQUID, focusing on shipwreck scenes and their salient regions; web-crawled images of underwater archaeological sites, covering various artifact types including ship hulls, ceramics, and metal objects.

After initial image collection, several data augmentation techniques-such as random rotation, cropping, color perturbation, and noise injection-are applied to expand the sample set and enhance the model's robustness to varying environmental conditions and object appearances.

3.2 Experimental Environment and Evaluation Metrics

Model training and testing in this study are conducted on a local desktop computer. The hardware and software configurations are as follows: the CPU is an Intel® Core™ i9-13900K @ 5.80GHz, the GPU is an NVIDIA GeForce RTX 4060, and the system has 8 GB of RAM. The operating system is Windows 11, the programming environment is Python 3.9.7, and PyTorch 1.12.1 is used as the deep learning framework.

During training, all input images are uniformly resized to 640×640 pixels. The total number of training epochs is set to 100, with a batch size of 8 per epoch. Stochastic Gradient Descent (SGD) is used as the optimizer, with an initial learning rate set to 0.01.

To comprehensively evaluate the performance of the proposed salient region detection model, several widely used metrics in object detection tasks are adopted, including Precision, Recall, Average Precision (AP), and mean Average Precision (mAP).

In addition, to assess the model's deployment efficiency in practical applications, the following auxiliary metrics are also introduced: frames per second (FPS), number of parameters (Params), and floating point operations per second (FLOPs).

The corresponding evaluation formulas are defined as follows:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN} \tag{7}$$

$$AP = \int_0^1 P(R)dR \tag{8}$$

$$mAP = \frac{1}{f}\sum_{i=1}^{f} AP_i \tag{9}$$

3.3 Ablation Study

To validate the efficiency and effectiveness of each proposed module in the salient region detection task for underwater cultural heritage, a series of ablation experiments are conducted. The original YOLOv8 model is used as the baseline, and improvements-including the DCNv4 module, CBAM attention mechanism, and the Dynamic4 detection head-are incrementally incorporated to construct

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0040

multiple model variants. Each variant is then trained and evaluated to assess performance gains. The specific model configurations are summarized in Table 1.

Table 1. Results of ablation experiments on the DOU dataset

Group	DCNv4	CBAM	Dyhead4	Param (M)	FLOPs (G)	Precision (%)	Recall (%)	mAP@ 0.5 (%)	mAP@ 0.5:0.95 (%)
1	×	×	×	3.01	8.20	81.8	72.8	80.9	60.4
2	√	×	×	2.86	7.90	82.0	73.6	82.6	61.6
3	×	√	×	3.08	8.20	81.3	73.4	81.3	60.8
4	×	×	√	5.19	17.70	83.4	74.2	83.2	62.3
5	√	√	×	5.69	21.00	83.0	74.1	83.5	61.9
6	√	×	√	5.05	17.40	83.9	74.8	84.4	63.4
7	×	√	√	5.18	17.50	83.6	74.5	84.3	62.6
8	√	√	√	5.47	22.00	84.3	75.3	84.9	63.8

In the ablation experiments conducted on the DUO dataset, the baseline YOLOv8 model-without any enhancement modules-achieves an mAP@0.5 of 80.9% and an mAP@0.5:0.95 of 60.4%. With the step-by-step introduction of the three key modules DCNv4, CBAM, and Dyhead each contributes measurable performance improvements.

In terms of detection accuracy, the proposed YOLOv8-CDD model demonstrates a clear advantage, achieving an mAP@0.5 of 84.9% and an mAP@0.5:0.95 of 63.8%, which represent absolute gains of 4.0% and 3.4%, respectively, over the baseline. These results validate the effectiveness of the proposed collaborative module optimization strategy.

Figure 4. Visual comparison of detection results between YOLOv8 and YOLOv8-CDD

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0040

The detection results on the DUO dataset are visualized in Figure 4. The left three images display the output of the original YOLOv8 model, while the right three images show the results from YOLOv8-CDD. As illustrated, the baseline model suffers from false positives (yellow boxes) and missed detections (blue boxes).

In summary, the proposed model demonstrates superior performance on the DUO underwater dataset and offers robust support for salient object detection tasks in underwater archaeology.

3.4 Salient Object Recognition Results

To further verify the overall performance advantages of the proposed model in salient region detection for underwater cultural heritage, this study conducts comparative experiments with several mainstream object detection models, including Faster R-CNN, YOLOv5, YOLOv7, and YOLOv8. The comparison results on the underwater archaeology dataset are presented in Table 2.

Table 2. Performance of different models on underwater archaeological datasets

Model	Param (M)	FLOPs (G)	mAP@ 0.5(%)	mAP@ 0.5:0.95 (%)	FPS
Faster R-CNN	41.1	91.2	64.2	47.6	15
YOLOv5	7.2	8.3	66.1	49.8	36
YOLOv7	6.21	6.79	65.4	48.5	37
YOLOv8	3.03	8.16	66.3	51.2	32
YOLOv8-CDD	5.5	21	70.4	53.6	28

In terms of detection accuracy, the proposed YOLOv8-CDD achieves an mAP@0.5 of 70.4% and an mAP@0.5:0.95 of 53.6%, which represent improvements of 4.1 and 2.4 percentage points, respectively, over the baseline YOLOv8 model. Among all compared models, YOLOv8-CDD delivers the best performance.

Regarding inference speed, YOLOv8-CDD reaches 28 FPS, outperforming Faster R-CNN and only slightly behind the original YOLOv8 (32 FPS). Overall, it meets the real-time processing requirements of most underwater archaeological applications.

Therefore, YOLOv8-CDD achieves a well-balanced trade-off between detection accuracy, model size, and inference speed. This makes it highly practical for real-world deployment in underwater cultural heritage detection tasks.

To provide a more intuitive demonstration of the improved model's ability to detect salient regions, Figure 5 presents a visual comparison between the detection results of YOLOv8 and YOLOv8-CDD. The first row displays four sample images with detection results from the original YOLOv8 model, while the second row shows the corresponding results produced by YOLOv8-CDD.

In the YOLOv8 results, yellow bounding boxes highlight issues such as false positives, missed detections, and redundant boxes. In contrast, the proposed YOLOv8-CDD model eliminates these issues, with no false detections, missed targets, or redundant bounding boxes observed in the visualized examples.

This improvement reflects the model’s enhanced capability in feature representation and spatial discrimination, further validating its robustness in complex underwater environments.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0040

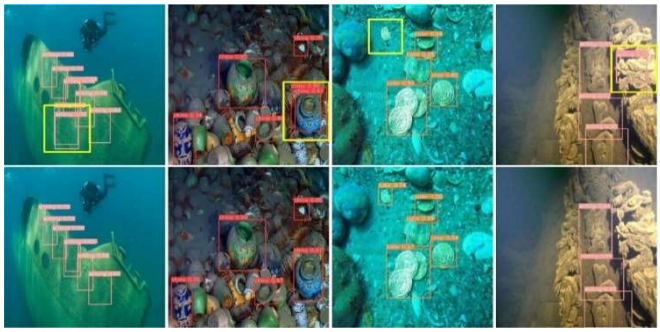

Figure 5. Visual comparison of the detection results of YOLOv8 and YOLOv8-CDD

4. Conclusion

To address the limitations of existing object detection models in underwater environments, this study presents YOLOv8-CDD, a model designed to improve salient region detection in underwater cultural heritage imagery. Based on the YOLOv8 framework, the model integrates C2f-DCNv4, CBAM, and the Dynamic4 detection head to enhance feature extraction and reduce background interference while maintaining computational efficiency.Experiments on the public DUO dataset and a custom underwater archaeological dataset indicate that YOLOv8-CDD achieves an mAP@0.5 of 84.9% and mAP@0.5:0.95 of 63.8% on DUO, with improvements of 4.0% and 3.4% over the baseline. On the custom dataset, mAP@0.5 reaches 70.4%, showing a 4.1% increase.

In summary, YOLOv8-CDD shows improved accuracy and stability over the YOLOv8, indicating its potential for underwater artifact detection in practical scenarios.

Acknowledgments

The authors wish to acknowledge the support of the graduate Student Innovation Projects of Beijing University of Civil Engineering and Architecture (No. PG2024121).

References

[1] Fu C, Fan X, Xiao J, et al. Learning Heavily-Degraded Prior for Underwater Object Detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(11): 6887–6896.

[2] Dinakaran R, Zhang L, Li C T, et al. Robust and Fair Undersea Target Detection with Automated Underwater Vehicles for Biodiversity Data Collection[J]. Remote Sensing, 2022, 14(15): 3680.

[3] Chen L, Zhou F, Wang S, et al. SWIPENET: Object detection in noisy underwater scenes[J]. Pattern Recognition, 2022, 132: 108926.

[4] Chen L, Li T, Zhou A, et al. Underwater object detection in noisy imbalanced datasets[J]. Pattern Recognition, 2024, 155: 110649.

[5] Qi J, Gong Z, Xue W, et al. An Unmixing-Based Network for Underwater Target Detection From Hyperspectral Imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 5470–5487.

[6] Wang B, Wang Z, Guo W, et al. A dual-branch joint learning network for underwater object detection[J]. Knowledge-Based Systems, 2024, 293: 111672.

[7] Zhang J, Zhang J, Zhou K, et al. An Improved YOLOv5-Based Underwater Object-Detection Framework[J]. Sensors, 2023, 23(7): 3693.

[8] Lei F, Tang F, Li S. Underwater Target Detection Algorithm Based on Improved YOLOv5[J]. Journal of Marine Science and Engineering, 2022, 10(3): 310.

[9] Cai S, Li G, Shan Y. Underwater object detection using collaborative weakly supervision[J]. Computers and Electrical Engineering, 2022, 102: 108159.

[10] Zhang F, Cao W, Gao J, et al. Underwater Object Detection Algorithm Based on an Improved YOLOv8[J]. Journal of Marine Science and Engineering, 2024, 12(11): 1991.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202505_11(5).0040

[11] Alsuwaylimi A A. Enhanced YOLOv8-Seg Instance Segmentation for Real-Time Submerged Debris Detection[J]. IEEE Access, 2024, 12: 117833–117849.

[12] Cai S, Zhang X, Mo Y. A Lightweight underwater detector enhanced by Attention mechanism, GSConv and WIoU on YOLOv8[J]. Scientific Reports, 2024, 14(1): 25797.

[13] Chen J, Er M J. Dynamic YOLO for small underwater object detection[J]. Artificial Intelligence Review, 2024, 57(7): 165.

[14] Xiong Y, Li Z, Chen Y, et al. Efficient Deformable ConvNets: Rethinking Dynamic and Sparse Operator for Vision Applications[A/OL]. arXiv, 2024[2025-03-14]. https://arxiv.org/abs/2401.06197.

[15] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module[M]// In: Ferrari V, Hebert M, Sminchisescu C, et al. Computer Vision – ECCV 2018. Cham: Springer, 2018, 11211: 3–19.

[16] Dai X, Chen Y, Xiao B, et al. Dynamic Head: Unifying Object Detection Heads with Attentions[A/OL]. arXiv, 2021[2025-03-17]. https://arxiv.org/abs/2106.08322.