Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimodal Detection for Underwater Cultural Heritage: A Physics-Guided Color-Bias-Aware Network and Dataset

Zigeng Song
Xiamen University of Technology
Ruifeng Wei
Xiamen University of Technology
Wang Man
manwang@xmut.edu.cn
Xiamen University of Technology
Qin Nie
Xiamen University of Technology
Zongmei Li
Xiamen University of Technology
Xiaofeng Du
Xiamen University of Technology
Xinchang Zhang
Xiamen University of Technology

Article

Keywords:

Posted Date: October 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11214605/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multimodal Detection for Underwater Cultural Heritage: A Physics-Guided Color-Bias-Aware Network and Dataset

Zigeng Song[1,†], Ruifeng Wei[1,†], Wang Man[1*], Qin Nie[1], Zongmei Li[1], Xiaofeng Du[1], Xinchang Zhang[1,2]

[1] Institute of Spatiotemporal Intelligence Application and Innovation, College of Computer Science and Technology, Xiamen University of Technology, Xiamen 361021, China

[2] School of Geographical Sciences, Guangzhou University, Guangzhou 510006, China

Corresponding author: Wang Man (manwang@xmut.edu.cn)

[†] These authors contributed equally to this work.

Abstract

Reliable object detection is essential for underwater cultural heritage research, yet its development remains limited by the scarcity of domain-specific annotations, underwater color distortion, and attenuated object boundaries. To address these challenges, we construct the Wrecks, Ruins, and Fragmentary heritage (WRF) dataset, comprising 4,732 RGB images collected from multiple underwater archaeological sites, corresponding estimated depth maps, 17,059 annotated instances, and 21 categories of artifacts and ship components. We further develop UCH-Net, a dual-stream RGB-depth detector that integrates appearance and geometric cues through physics-guided color-bias-aware fusion, multi-scale feature aggregation, and decoupled localization. On the WRF, UCH-Net achieves an mAP of 77.5% and an mAP75 of 88.0%, surpassing the representative Deformable-DETR baseline by 12.8 and 14.9 percentage points, respectively, while exceeding the strongest competing results by 1.2% in mAP75 and 1.5% in recall. UCH-Net further achieves an mAP of 66.0% on RUOD. These results demonstrate the potential of the proposed dataset and detector for automated screening and object localization in underwater archaeological imagery.

Introduction

Underwater cultural heritage, including ancient shipwrecks, submerged settlements, and underwater cities, provides important material evidence of human history and maritime civilizations, with significant historical, artistic, and scientific value [1-2]. These remains provide direct insights into ancient maritime transportation, human migration, and the development of shipbuilding and navigation technologies, while also constituting important targets for heritage conservation and scientific investigation [3]. Prolonged exposure to complex underwater environments can progressively compromise structural integrity and obscure surface details, potentially resulting in the irreversible loss of archaeological information. Therefore, systematic documentation and monitoring through object detection and related computer vision techniques are essential for supporting the investigation, interpretation, and conservation of underwater cultural heritage.

Reliable visual datasets provide the foundation for automated object detection in underwater archaeology and heritage monitoring. Early underwater visual benchmarks have predominantly been developed for specific ecological or environmental applications. For example, Underwater Robot Professional Contest (URPC2020) [4], Detecting Underwater Objects (DUO) [5], and Real-world Underwater Object Detection (RUOD) [6] mainly contain benthic organisms such as sea cucumbers and starfish. The UIIS dataset [7] introduces a multiclass benchmark for underwater instance segmentation, covering seven categories that include fish and coral reefs, whereas the TrashCAN dataset [8] focuses on marine debris and plastic waste. However, underwater cultural heritage encompasses a substantially different range of targets, including shipwreck structures, pottery, porcelain, anchors, and rudders, whose visual appearance, physical scale, and spatial distribution differ from those of conventional marine ecological objects. Consequently, existing benchmarks provide limited support for heritage-oriented detection tasks, particularly in terms of category definitions, scene composition, and annotation requirements [9]. Some shipwreck datasets have been developed for broader underwater applications [10], but their limited sampling coverage and coarse object taxonomies restrict their use for fine-grained object detection in complex archaeological sites. Dedicated multimodal benchmark datasets for underwater cultural heritage remain scarce, limiting the development and systematic evaluation of detection methods tailored to these environments [11-12].

Beyond the scarcity of domain-specific data, underwater imaging conditions introduce additional sources of uncertainty for object detection. Earlier approaches relied on handcrafted features and conventional classifiers [13,14], whose performance was often sensitive to illumination attenuation, visual clutter, and background interference. Deep learning detectors, including Faster R-CNN [15], the YOLO series [16], and the DETR series [17],

have since become widely adopted in general vision tasks because of their end-to-end detection capability. Recent studies have further addressed scale variation and the detection of small marine targets through CNN- and Mamba-based multiscale feature aggregation [18] and context-aware attention mechanisms [19]. Nevertheless, RGB-based representations remain vulnerable to the degradation of color, texture, and boundary cues under complex underwater imaging conditions, particularly in environments characterized by weak illumination, scattering and substantial optical degradation.

Compared with RGB images, depth information provides geometric contours and spatial-structure cues that can partially compensate for visual information obscured by underwater color distortion and texture degradation. UMIS-YOLO exploits the complementary properties of RGB and depth modalities, with RGB capturing color and texture features and depth providing additional boundary and geometric information [20]. Underwater visual datasets have also increasingly incorporated multimodal annotation [21-22]. These studies highlight the potential of depth-derived geometric features, which are less sensitive to certain appearance and texture variations, for improving detection performance in complex underwater scenes.

Underwater optical degradation further limits detection performance. Selective light absorption and scattering produce spatially heterogeneous color shifts and contrast degradation [23]. Some object detection methods adopt enhancement-before-detection pipelines to explicitly correct color distortion. For example, YOLO-DBS [24] uses green-channel statistics as a reference to derive red- and blue-channel compensation factors from their means and variances before the image enters the YOLO backbone. U-DECN [25] treats color shift as underwater noise and improves detector robustness through a dedicated denoising training mechanism. However, these methods generally overlook the relationship between color distortion and imaging distance, despite the nonuniform variation in distortion with propagation distance. Robustness learned solely through parameters may therefore be insufficient for adapting to spatially heterogeneous color shifts associated with variations in propagation distance.

To address these limitations, we construct a multimodal dataset for underwater cultural heritage object detection and propose an object detection method that integrates color-bias awareness with depth-derived geometric information.

1. A multimodal benchmark dataset for underwater cultural heritage. We collect and screen RGB images from complex underwater scenes containing multiple cultural heritage categories and perform object-level annotation. Because the original images lack depth information, Depth Anything V2 [26] is used to generate depth maps spatially aligned with the RGB images. These data constitute the Wrecks, Ruins, and Fragmentary Heritage multimodal detection dataset, providing a foundation for underwater cultural heritage detection research.

2. A multimodal object detection architecture for underwater scenes. We develop a framework comprising a Dual-Stream Backbone for parallel extraction of RGB texture and depth-derived geometric features, a Dual-Feature Fusion Pyramid for cross-scale feature interaction, and a Shared Decoupled Detection Head for separate classification and bounding-box regression.

3. A Color-Bias-Aware Fusion Module (CBFM). CBFM uses early RGB reference features to estimate local color bias, regional brightness, and global degradation context, generating bounded scale and bias parameters for adaptive modulation of current-layer RGB features. The modulated RGB features are then fused with depth-derived geometric features through spatial, channel-wise, and pixel-level gating, improving feature discriminability and structural representation for cultural heritage objects in complex underwater environments.

Methods

Dataset acquisition and annotation

To address the limited availability of data for underwater cultural heritage detection, we construct a dedicated multi-site dataset. To enhance geographic diversity and capture a broad range of underwater optical conditions, the original images are collected from multiple deep-sea archaeological sites in the Gulf of Mexico, the Black Sea, and the Mediterranean Sea (Fig.1).

1. The Gulf of Mexico subset includes several deep-sea shipwreck sites. Wreck 15377 (28°28′N, 90°02′W) is located near the edge of a deep-sea basin east of the Mississippi River mouth [27], with images primarily depicting hull structures and scattered cargo. The subset also includes Wreck 15577 [28] and Wreck 15563 (25°41′N, 94°13′W) [29], where remains of wooden hulls from early nineteenth-century merchant vessels are visible.

2. The Black Sea subset contains ancient archaeological remains from the Eregli site (40°56′N, 27°57′E) [30], which contains numerous wooden components, and the Sinop I shipwreck site (42°01′N, 35°09′E) [31], where Roman-period artifacts, such as amphorae, are together with wooden materials.

3. The Mediterranean subset includes the Knidos shipwreck site in the Aegean Sea (36°42′N, 27°22′E) [32], where images depict Byzantine-period shipwreck pottery and partially buried or fragmented artifacts. It includes the Thonis underwater site in Egypt (31°25′N, 30°09′E), which contains ancient Egyptian granite statues and small figurines.

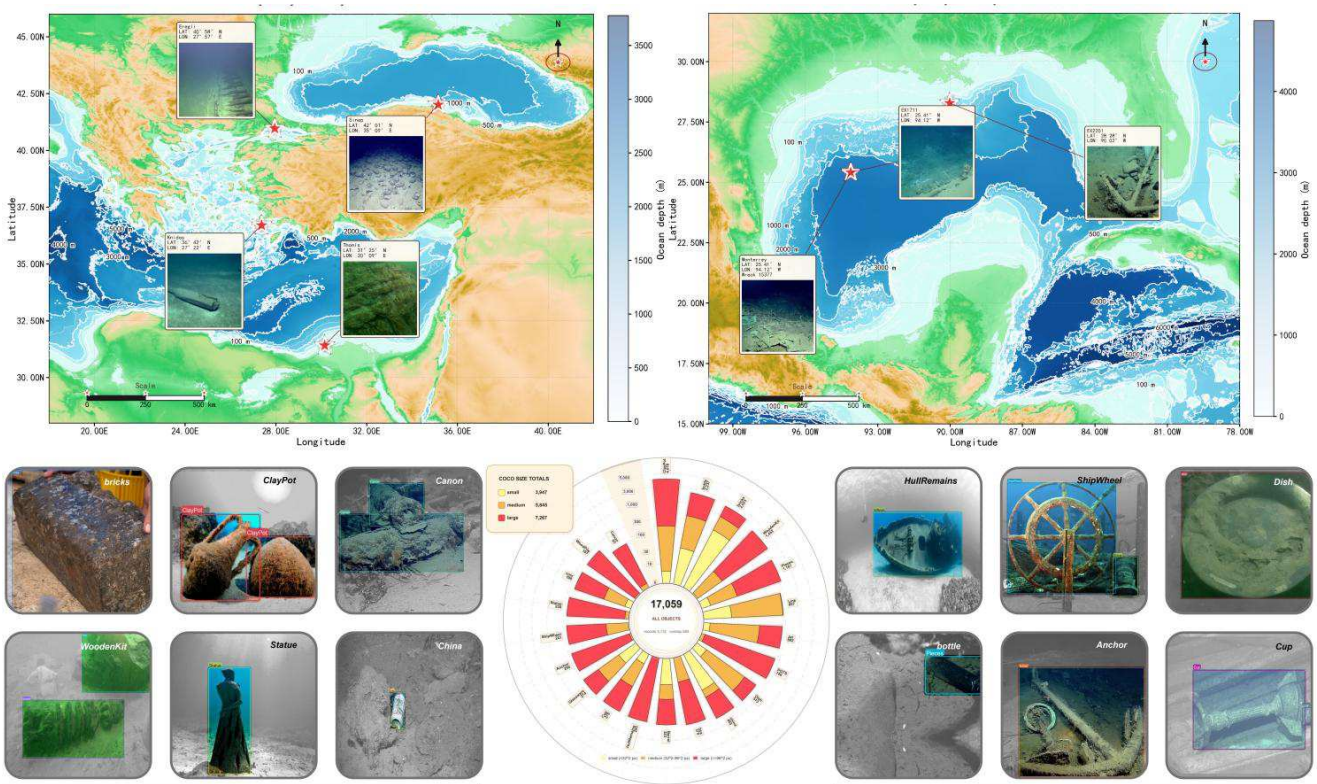


Fig.1. Geographic coverage, representative underwater archaeological scenes and annotation characteristics of the Wrecks, Ruins and Fragmentary Heritage (WRF) dataset. The maps show the spatial distribution of the archaeological sites across the Gulf of Mexico, the Black Sea and the Mediterranean Sea, with representative underwater images from the surveyed locations. The lower panels illustrate representative cultural heritage categories and their object-level bounding-box annotations, together with the distribution of instances across the 21 object categories.

After obtaining the original image sequences, we constructed the multimodal dataset through the following procedure. The images were manually deduplicated and screened for quality [33], with severely blurred images and images with an excessively limited effective field of view removed. The retained images were annotated according to unified guidelines, with each visually identifiable object instance assigned to a predefined category and enclosed by a tight bounding box. Partially occluded, truncated, or moderately blurred objects were retained when their category and approximate spatial extent can still be identified with sufficient confidence, whereas severely ambiguous instances are excluded. Finally, Depth Anything V2 was applied to the filtered RGB images for pixel-wise depth estimation, generating a corresponding depth map for each image.

The resulting WRF Dataset contains 4,732 multimodal images and 17,059 cultural heritage object instances across 21 categories of cultural artifacts and ship components, including wooden components, pottery jars, sculptures, coins, and glass bottles. The category distribution is long-tailed, with pottery jars and glass bottles represented by relatively more instances. This imbalance increases detection difficulty and enables model evaluation under class-imbalanced conditions. The dataset was partitioned into training and test sets at an 8:2 ratio using a temporally ordered split, reducing the possibility that highly similar frames acquired at adjacent time points were distributed across both sets.

In addition to WRF dataset, we evaluate the proposed model on the public RUOD dataset as an additional benchmark for general underwater object detection. RUOD contains 14,000 high-resolution images acquired in real underwater environments, with 74,903 object instances across 10 categories of underwater organisms and objects with substantial morphological variation. The dataset includes densely distributed targets, scale variation, and background

interference caused by light scattering and blue-green color casts, making it suitable for assessing color-bias-aware fusion module under severe underwater color distortion.

UCH-Net architecture

To address visual degradation and color shifts caused by light attenuation and underwater scattering, we developed UCH-Net, a multimodal object detection network for underwater cultural heritage detection. As shown in Fig. 2, UCH-Net follows a dual-stream feature extraction, physical-prior-guided fusion, and bidirectional feature decoding pipeline, comprising four main components.

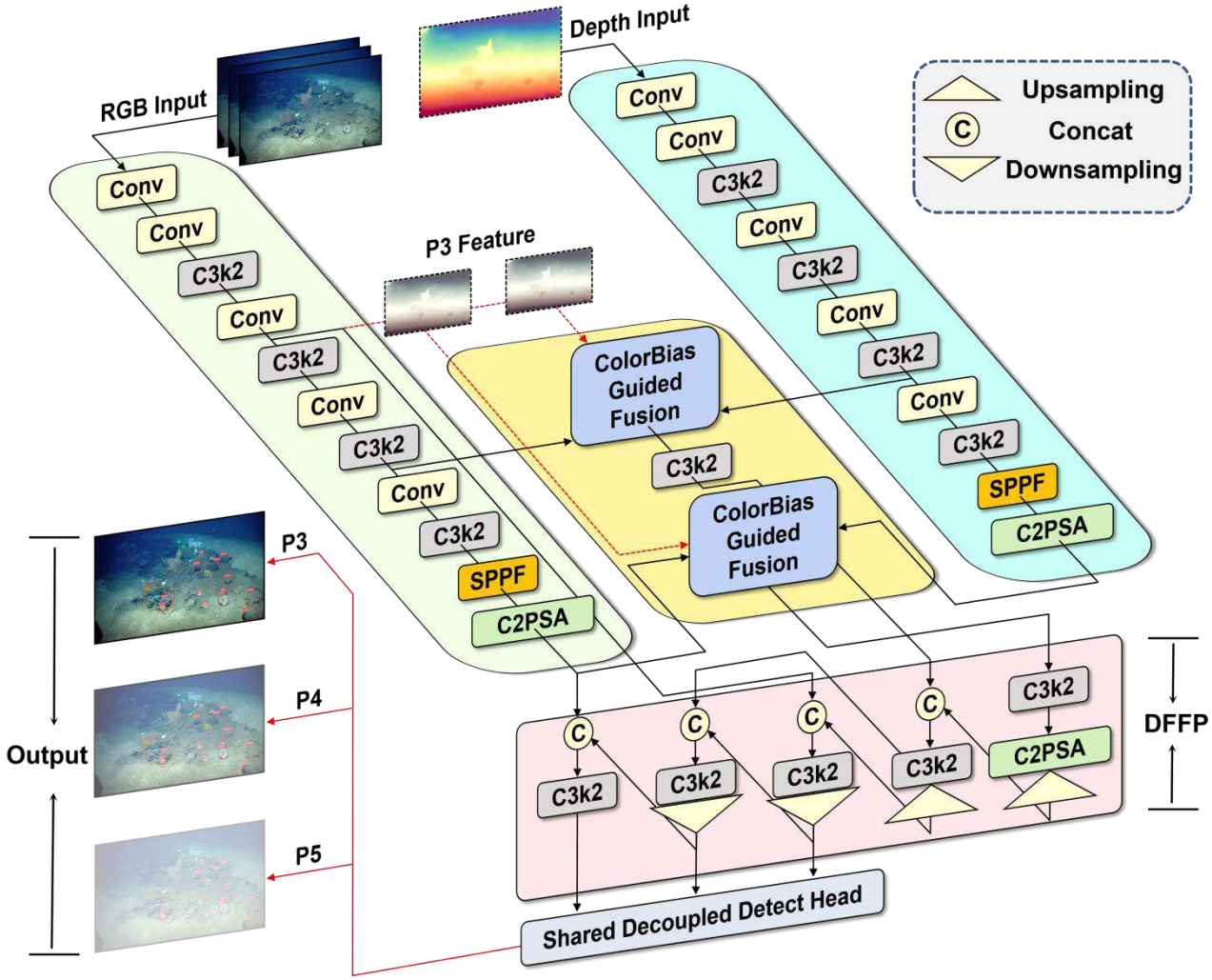


Fig.2. Architecture of the UCH-Net for underwater cultural heritage detection. The framework integrates dual-stream RGB and depth feature extraction, Color-Bias-Guided Fusion for physical-prior-guided multimodal feature integration, and a Shared Decoupled Detect Head for multiscale object prediction.

Dual-Stream Backbone

The first component consists of an RGB branch and a depth branch that extract multiscale hierarchical features from the two modalities. The branches maintain the same spatial resolution and channel dimension at each corresponding stage to facilitate subsequent feature fusion.

The RGB branch alternates C3k2 modules (Cross Stage Partial bottleneck modules with 2 convolutions) with convolutional downsampling layers to extract color and texture features. C3k2 is an improved module based on CSPNet (Cross Stage Partial Network) [34], designed to preserve spatial details and color variations while facilitating stable gradient propagation in deeper layers. At deeper stages, the network further introduces a Spatial Pyramid Pooling-Fast (SPPF) module and a C2PSA attention module. SPPF applies consecutive 5×5 max-pooling operations to enlarge the receptive field without altering the spatial dimensions of feature maps, thereby aggregating multiscale contextual information for targets ranging from hull remains to artifact fragments. C2PSA generates channel- and

spatial-wise attention weights to suppress irrelevant background responses and enhance target-related representations, reducing interference from suspended particles and other background noise.

The depth branch processes depth maps in parallel with the RGB branch using a hierarchical structure aligned with its RGB counterpart. It primarily provides geometric contours and spatial structural information. Independent yet spatially aligned feature extraction helps reduce modal noise interference and semantic misalignment associated with premature fusion.

Dual-Feature Fusion Pyramid

The second component is the Dual-Feature Fusion Pyramid (DFFP), whose topology is adapted from PANet [35]. PANet extends FPN [36] with a bottom-up feature pathway, shortening the information path from low-level, high-resolution localization features to higher-level representations and benefiting boundary regression. This multiscale design is particularly relevant to underwater cultural heritage scenes, where large shipwreck structures coexist with small artifacts and fragments.

In multimodal settings, however, sequential feature propagation may progressively weaken geometric priors and color-texture information or introduce semantic shifts through successive convolutional operations. To preserve the PANet's spatial information propagation while mitigating these effects, we modify its topology to construct DFFP for multiscale feature encoding and fusion. DFFP introduces cross-layer skip connections between encoding input nodes and decoding output nodes at the same scale, allowing original multimodal features to participate directly in subsequent fusion and reducing information loss from deep convolutional smoothing without adding separate computational pathways. For the medium-scale feature Output_{P_4}, the multi-path adaptive weighted fusion is defined as:

$$\text{Output}_{P_4} = \frac{w_1 F_{\text{Fused}}^{P_4} + w_2 P_{4,\text{in}} + w_3 \,\text{Conv}\left(P_{3,\text{out}}\right)}{w_1 + w_2 + w_3 + \varepsilon} \quad (1)$$

where Output_{P_4} denotes the fused output feature, $P_{4,\text{in}}$ denotes the semantic feature propagated through the top-down pathway, $F_{\text{Fused}}^{P_4}$ denotes the multimodal intermediate-level fused feature, and $P_{3,\text{out}}$ is the low-level high-resolution feature. Because these features have different spatial resolutions, $\text{Conv}(\cdot)$ is applied for scale alignment. w_i denotes a learnable scalar weight, with all weights initialized to 1.0, and ε is a small constant that prevents division by zero. These weights allow the network to dynamically adjust the contributions of different feature pathways under varying underwater optical conditions and modality quality.

Shared Decoupled Detection Head

The third component is the Shared Decoupled Detection Head (SDDH). To accommodate scenes in which large-scale shipwreck remains coexist with small objects such as glass cups, the neck network outputs three feature maps, P3, P4, and P5, corresponding to small-, medium-, and large-scale targets, respectively. Conventional three-scale YOLO detection heads assign independent classification and regression convolutional layers to each scale after channel alignment, increasing computational requirements on resource-constrained underwater deployment platforms.

In SDDH, each scale first undergoes an independent 3×3 convolution for channel alignment, after which P3, P4, and P5 are processed by a shared depthwise separable convolutions:

$$F_i = \mathcal{F}_{\text{shared}}\left(\sigma\left(\text{GN}\left(\text{Conv}_{3\times3}(X_i;\theta_i)\right)\right), \Theta_{\text{shared}}\right), \quad i \in \{3,4,5\} \quad (2)$$

where X_i denotes the input feature at the P3, P4, or P5 level, and θ_i denotes the parameters of the scale-specific alignment layer. $\text{GN}(\cdot)$ denotes Group Normalization. $\mathcal{F}_{\text{shared}}$ is the depthwise separable convolution operator, whose parameters Θ_{shared} are shared across all three scales. Cross-scale weight sharing encourages consistent target representations across spatial scales while reducing the tendency of scale-specific branches to overfit particular underwater noise patterns.

Following the YOLO detection head design, the classification and bounding-box regression outputs at each scale are defined as:

$$C_i = \text{Conv}_{1\times1}(F_i; W_{\text{cls}}) \quad (3)$$

$$B_i = \alpha_i \,\text{Conv}_{1\times1}\left(F_i; W_{\text{reg}}\right) \quad (4)$$

where C_i and B_i denote the class prediction and bounding-box regression map, respectively. W_{cls} and W_{reg} denote the 1×1 projection parameters shared across scales for classification and regression, respectively. The adaptive parameter

α_i adjusts the response magnitude of the regression branch to accommodate geometric size variations among underwater targets.

Color-Bias-Aware Fusion Module

The fourth component is the Color-Bias-Aware Fusion Module (CBFM), which extracts color-bias-related physical priors from early RGB reference feature by jointly modeling local, regional, global, and positional information. These priors guide the adaptive modulation of current RGB features and, together with spatial and channel attention, regulate pixel-level gating between the modulated RGB and depth representations. This design preserves informative RGB appearance cues while adaptively exploiting depth-derived geometric structure in regions affected by color degradation. CBFM is deployed at intermediate and high feature levels in UCH-Net, with further details provided in Supplementary Note 1.

Experimental settings and evaluation metrics

All experiments were implemented in PyTorch and trained using two NVIDIA RTX 3090 GPUs. The model was trained for 200 epochs with a batch size of 8, using AdamW as the optimizer with a weight decay of 1×10^{-4}. The learning rate decays from 1×10^{-3} to 1×10^{-5}. Data augmentation includes random horizontal and vertical flipping, 90° rotation, and color augmentation.

For underwater object detection, predicted bounding boxes are matched with ground-truth bounding boxes at a specified Intersection over Union (IoU) [37] threshold to determine True Positives (TP), False Positives (FP), and False Negatives (FN). Precision (P), Recall (R), and F1-score are calculated as follows:

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \tag{5}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{6}$$

$$\text{F1-score}=\frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision+Recall}} \tag{7}$$

Precision and Recall at a single threshold cannot fully characterize detection performance. Therefore, Average Precision (AP) and mean Average Precision (mAP) are further adopted for evaluation. AP is defined as the area under the Precision-Recall (P-R) curve, while mAP is the mean AP over all object categories:

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

$$\text{mAP}=\frac{1}{\text{N}_\text{c}}\sum_{\text{i}=1}^{\text{N}_\text{c}}\text{A P}_\text{i} \tag{9}$$

where N_c denotes the number of object categories. Three mAP metrics are reported. mAP_{50} is calculated at an IoU threshold of 0.50, mAP_{75} at 0.75, and $\text{mAP}_{50:95}$ at the mean of the ten mAP values obtained at IoU thresholds from 0.50 to 0.95 in increments of 0.05. For simplicity, $\text{mAP}_{50:95}$ is referred to as mAP throughout this paper. Model complexity is evaluated using the number of parameters (Params) and floating-point operations (GFLOPs).

Comparison methods and experimental protocol

To evaluate UCH-Net, we compare it with 16 representative methods, including eight DETR-based end-to-end detectors (Deformable-DETR [38], DINO [39], UDINO [40], RT-DETR, BSR5-DETR [41], EAMSF-DETR [42], DAB-DETR [43], and U-DECN), six YOLO-based one-stage detectors (RG-YOLO [44], YOLOv12 [45], YOLOv26 [46], LFN-YOLO, BSR5-YOLO, and UWA-PRB [47]), and two two-stage detectors (Boosting R-CNN [48] and FBDPN [49]). These methods encompass general-purpose, recent, and underwater-specific detectors, enabling comparisons of detection accuracy, multiscale feature fusion, and computational cost.

For consistency, all methods used the same dataset splits, with RUOD following its official split and WRF dataset divided into training and testing sets at an 8:2 ratio. All methods were trained for 200 epochs, while other implementation details, including input preprocessing, data augmentation, optimization parameters, and inference thresholds, follow the corresponding public repositories and original papers. Because implementations and default configurations vary across methods, reproduced results may differ slightly from those reported in the original papers. All results are obtained under the same dataset splits and training epochs to facilitate consistent comparative evaluation.

Results

Comparison on WRF dataset

Table 1 and Fig. 3 summarize the detection performance of UCH-Net on WRF dataset. UCH-Net achieved a Precision of 94.9%, Recall of 89.9%, F1-score of 92.1%, mAP50 of 94.4%, mAP75 of 88.0% and mAP of 77.5%, attaining the highest values across all reported accuracy metrics except Precision. LFN-YOLO achieved the highest Precision at 95.2% exceeding UCH-Net by 0.3 percentage points. Among the compared methods, BSR5-YOLO yielded the closest overall mAP at 77.0%, 0.5 percentage points below UCH-Net, while UCH-Net achieved a 1.2 percentage-points advantage in mAP75. EAMSF-DETR and BSR5-DETR achieved mAP values of 71.7% and 73.0%, respectively, both lower than the 77.5% obtained by UCH-Net. Compared with LFN-YOLO, UCH-Net exhibited slightly lower Precision but higher Recall and F1-score, with a larger performance difference at mAP75 than at mAP50. These results indicate consistent detection

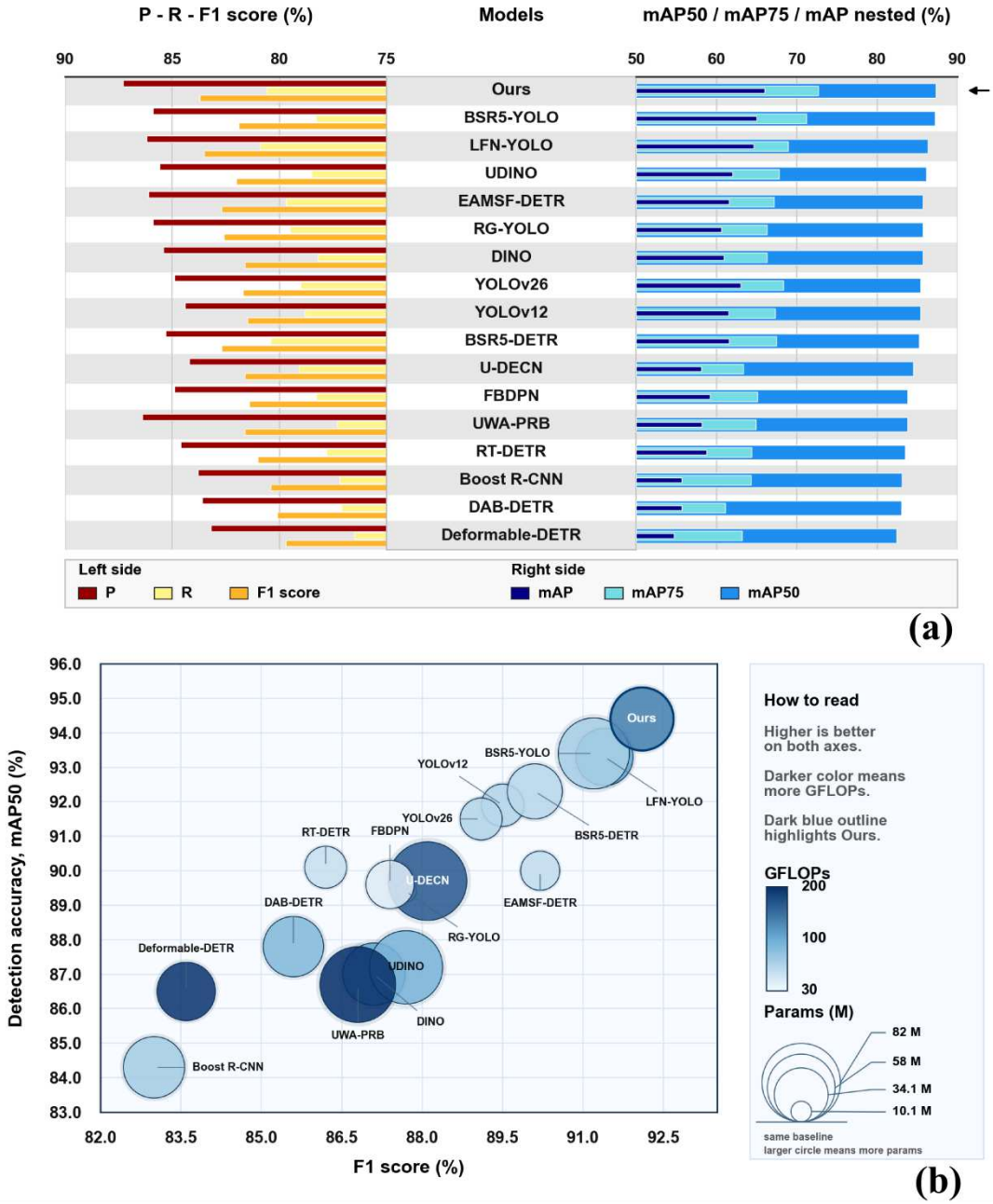


Fig.3. Comparative detection performance and accuracy–efficiency trade-offs of UCH-Net and competing methods on the WRF dataset. The upper panel (a) compares Precision, Recall, F1-score, mAP50, mAP75 and mAP across the evaluated detectors. The lower panel (b) illustrates the relationship between F1-score and detection accuracy, with bubble size representing the number of model parameters and color intensity indicating computational complexity in GFLOPs. The figure provides a comparative overview of detection accuracy and model efficiency across different architectures.

performance across different IoU thresholds.

Comparison on RUOD dataset

Table 2 and Fig.4. summarize the detection performance of UCH-Net on the RUOD dataset. UCH-Net achieved a Precision of 87.3%, Recall of 80.6%, F1-score of 83.7%, mAP50 of 87.3%, mAP75 of 72.7% and mAP of 66.0%. It obtained the highest Precision, F1-score, mAP50, mAP75, and mAP among the evaluated methods, whereas LFN-YOLO achieved the highest Recall at 80.9%. BSR5-YOLO yielded the closest overall mAP at 65.0%, 1.0 percentage point below UCH-Net, while UCH-Net achieved a 1.5 percentage-point advantage in mAP75. Compared with LFN-YOLO, UCH-Net showed slightly lower Recall but higher values across the remaining reported accuracy metrics, with the largest difference observed at mAP75. The consistent performance across the evaluated metrics suggests that UCH-Net maintains effective detection accuracy under the diverse object appearances and optical degradation conditions represented in RUOD, while preserving advantages in stricter localization evaluation.

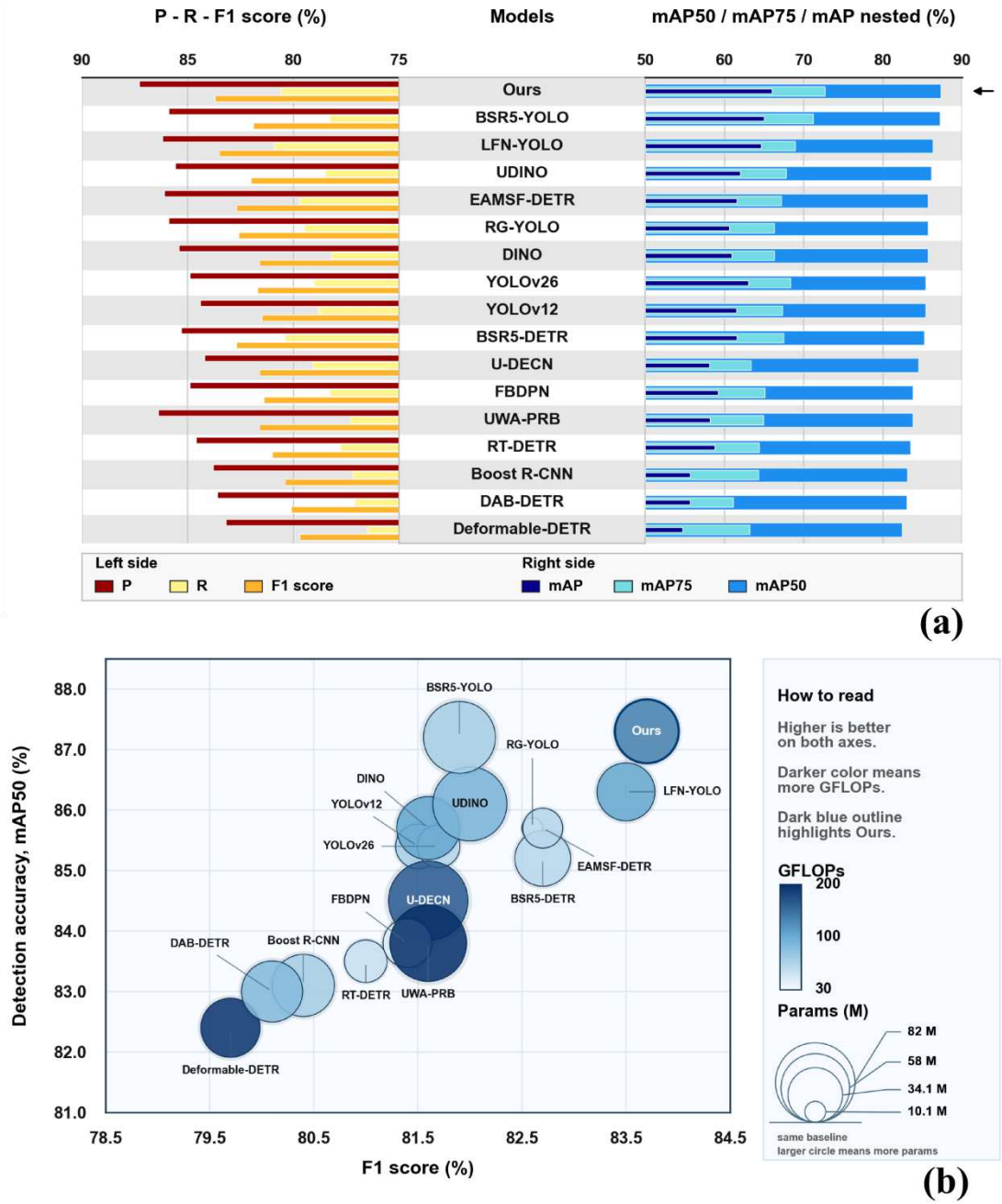


Fig.4. Comparative detection performance and accuracy–efficiency trade-offs of UCH-Net and competing methods on the RUOD dataset. The upper panel (a) compares Precision, Recall, F1-score, mAP50, mAP75 and mAP across the evaluated detectors. The lower panel (b) illustrates the relationship between detection accuracy and model complexity, with bubble size representing the number of parameters and color intensity indicating computational cost in GFLOPs.

Beyond detection accuracy, Tables 1 and 2 compare model complexity in terms of parameter count and computational cost. UCH-Net contains 48.8 M parameters and requires 143.8 GFLOPs per forward pass. Its parameter count is lower than those of BSR5-YOLO, UDINO, U-DECN, and UWA-PRB. In particular, relative to BSR5-YOLO, UCH-Net reduces the parameter count from 64.8 M to 48.8 M, corresponding to a reduction of ~ 24.7%. However, this reduction is accompanied by a higher computational cost than that of most lightweight YOLO-based detectors. BSR5-YOLO requires 76.8–76.9 GFLOPs, LFN-YOLO ~ 118 GFLOPs, and several lightweight detectors less than 70 GFLOPs. These results characterize UCH-Net as a model with a moderate parameter count but relatively high computational demand, reflecting a trade-off between detection performance and computational efficiency. The computational overhead should therefore be considered when deploying UCH-Net on resource-constrained underwater platforms.

Table 1. Quantitative evaluation results on the *WRF* dataset. The best results are highlighted in bold (%).

Method	*WRF Dataset*							
	P	*R*	*F1*	*mAP50*	*mAP75*	*mAP*	*Params*	*GFLOPs*
Boost R-CNN	88.7	78.8	83.0	84.3	59.7	53.7	45.7	77.5
Deformable-DETR	88.4	79.2	83.6	86.5	73.1	64.7	41.0	196.0
RG-YOLO	93.0	83.7	87.7	89.4	78.3	68.1	**10.1**	**31.1**
YOLOv12	92.9	86.4	89.5	91.9	85.1	73.1	20.4	67.9
YOLOv26	92.7	85.7	89.1	91.5	82.1	71.7	20.0	68.1
LFN-YOLO	**95.2**	88.1	91.4	93.3	82.9	74.6	39.0	117.9
DINO	92.9	82.5	87.1	87.0	74.2	65.4	47.7	119.6
UDINO	94.1	82.4	87.7	87.2	74.8	65.9	69.5	104.8
RT-DETR	89.0	84.3	86.2	90.1	78.4	67.2	19.9	57.0
BSR5-DETR	92.7	87.6	90.1	92.3	80.3	73.0	35.6	66.8
BSR5-YOLO	94.7	88.4	91.2	93.4	86.8	77.0	64.8	76.9
U-DECN	91.3	85.4	88.1	89.7	68.7	59.7	82.0	170.0
FBDPN	91.7	84.2	87.4	89.6	77.4	67.4	26.3	42.9
UWA-PRB	91.6	82.8	86.8	86.7	74.0	65.6	75.4	226.8
EAMSF-DETR	92.9	87.9	90.2	90.4	83.5	71.7	17.5	60.2
DAB-DETR	90.7	81.8	85.6	87.8	73.0	64.4	44.1	98.1
Ours	94.9	**89.9**	**92.1**	**94.4**	**88.0**	**77.5**	48.8	143.8

Table 2. Quantitative evaluation results on the RUOD dataset. The best results are highlighted in bold(%).

Method	*RUOD Dataset*							
	P	*R*	*F1*	*mAP50*	*mAP75*	*mAP*	*Params*	*GFLOPs*
Boost R-CNN	83.8	77.2	80.4	83.1	64.3	55.7	45.9	77.6
Deformable-DETR	83.2	76.5	79.7	82.4	63.2	54.7	41.0	196.0
RG-YOLO	85.9	79.5	82.6	85.7	66.3	60.6	**10.1**	**31.1**
YOLOv12	85.3	78.6	81.8	84.9	66.4	60.7	21.8	74.9
YOLOv26	84.9	79.0	81.7	85.4	68.3	63.0	20.0	68.0
LFN-YOLO	86.2	**80.9**	83.5	86.3	68.9	64.6	39.0	118.1
DINO	85.4	78.2	81.6	85.7	66.3	60.9	47.5	119.5
UDINO	85.6	78.5	82.0	86.1	67.8	62.0	69.5	104.8
RT-DETR	84.6	77.8	81.0	83.5	64.4	58.8	19.9	57.1
BSR5-DETR	85.3	80.4	82.7	85.2	67.5	61.6	35.6	66.7
BSR5-YOLO	85.9	78.3	81.9	87.2	71.2	65.0	64.8	76.8
U-DECN	84.2	79.1	81.6	84.5	63.4	58.1	82.0	170.0
FBDPN	84.9	78.3	81.4	83.8	65.1	59.2	26.3	42.9
UWA-PRB	86.4	77.3	81.6	83.8	64.9	58.2	75.4	226.8
EAMSF-DETR	86.1	79.7	82.7	85.7	67.2	61.6	17.7	67.6
DAB-DETR	83.6	77.1	80.1	83.0	61.1	55.7	44.0	98.0
Ours	**87.3**	80.6	**83.7**	**87.3**	**72.7**	**66.0**	48.8	143.8

Ablation study of UCH-Net

Table 3 presents the cumulative ablation study of UCH-Net on RUOD and WRF datasets. A single RGB branch serves as the baseline, and the remaining components are introduced sequentially under otherwise identical training settings. Accordingly, differences between adjacent configurations reflect the marginal contribution of each newly added component under the specified order of integration, rather than its isolated effect.

Introducing DSB increased mAP from 63.0% to 64.0% on RUOD and from 71.7% to 72.5% on WRF dataset, accompanied by improvements in F1-score on both datasets. Adding CBFM resulted in a marginal mAP increase on RUOD, from 64.0% to 64.2%, but a larger improvement on WRF dataset, from 72.5% to 75.0%. The subsequent integration of DFFP further increased mAP on RUOD to 65.0%, whereas mAP on WRF dataset remained at 75.0%. Finally, incorporating SDDH yielded the complete UCH-Net, achieving mAP values of 66.0% on RUOD and 77.5% on WRF dataset. These results correspond to overall improvements of 3.0 and 5.8 percentage points over the single-

RGB baseline, respectively, and demonstrate performance gains across the cumulative integration process, with the magnitude of individual improvements varying between datasets.

Table 3. Ablation results on the WRF dataset and RUOD. The best results are highlighted in bold (%).

Method	*RUOD*					*WRF Dataset*				
	P	*R*	*F1*	*mAP50*	*mAP*	*P*	*R*	*F1*	*mAP50*	*mAP*
None	84.9	79.0	81.8	85.4	63.0	92.7	85.7	89.1	91.5	71.7
+DSB	86.2	80.1	82.9	86.7	64.0	92.7	88.1	90.1	92.6	72.5
+CBFM	85.3	80.4	82.7	86.5	64.2	92.7	**89.9**	91.1	93.8	75.0
+DFFP	86.4	**81.4**	83.7	87.1	65.0	93.8	89.0	91.1	93.4	75.0
+SDDH(Final)	**87.3**	80.6	**83.7**	**87.3**	**66.0**	**94.9**	**89.9**	**92.1**	**94.4**	**77.5**

Component analysis of the CBFM

Table 4 evaluates the internal components of CBFM on WRF dataset. Introducing PPE improved all six reported metrics relative to the baseline, increasing mAP from 75.9% to 76.8% and F1-score from 90.1% to 91.4%. This provides the reference configuration for evaluating the subsequent attention modules. Among the individual attention configurations added to PPE, PA delivered the most consistent improvement, outperforming the PPE configuration across all six metrics. In particular, Recall increased from 88.7% to 90.0%, while mAP75 and F1-score increased from 87.0% to 87.6% and from 91.4% to 92.0%, respectively. In contrast, CA and SA produced mixed changes across the reported metrics.

The combined configurations showed different performance profiles. Among the two-attention combinations, SA+PA achieved the highest mAP75 and mAP, reaching 87.9% and 77.1%, respectively. With PPE, CA, SA, and PA enabled, the complete CBFM achieved the highest Precision (94.9%), mAP50 (94.4%), mAP75 (88.0%), mAP (77.5%) and F1-score (92.1%) among the evaluated configurations. Its Recall reached 89.9%, only 0.1 percentage points below the maximum value obtained by the PA configuration.

Table 4. Ablation results of CBFM on the WRF Dataset. The best results are highlighted in bold (%).

Setting	***Module activation***				***WRF Dataset (%)***					
	PPE	*CA*	*SA*	*PA*	*Precision*	*Recall*	*mAP50*	*mAP75*	*mAP*	*F1-score*
Baseline	–	–	–	–	92.6	87.8	92.4	86.4	75.9	90.1
PPE	●	–	–	–	94.3	88.7	93.6	87.0	76.8	91.4
CA	●	●	–	–	93.5	88.9	93.3	87.4	76.9	91.1
SA	●	–	●	–	93.6	88.8	93.4	87.4	76.9	91.0
PA	●	–	–	●	94.5	**90.0**	93.9	87.6	77.0	92.0
CA + PA	●	●	–	●	94.7	89.6	93.8	87.4	76.9	91.9
SA + PA	●	–	●	●	94.4	89.8	93.9	87.9	77.1	91.9
CA + SA	●	●	●	–	94.5	89.6	93.9	87.2	76.4	91.8
CBFM	●	●	●	●	**94.9**	89.9	**94.4**	**88.0**	**77.5**	**92.1**

Attention heatmap and detection analysis

To complement the quantitative evaluation, we examined the spatial response patterns and detection behavior of different models on representative underwater cultural heritage images. Fig.5 shows Grad-CAM++ [50] visualizations, where warmer colors indicate stronger model responses. All compared methods exhibit activations around target regions, whereas several general-purpose detectors, including RT-DETR and DAB-DETR, produce more dispersed responses extending into surrounding background areas. In contrast, UCH-Net generates more concentrated high-response regions around the annotated objects, particularly pottery jars and glass bottles.

Fig.5 compares the detection results of UCH-Net with those of LFN-YOLO, UDINO, and EAMSF-DETR on challenging WRF Dataset. The examples include heavily overlapping coins, densely distributed and partially buried artifacts, and distant small objects affected by suspended particles and color distortion. The compared methods show varying degrees of missed detections, duplicate predictions, category errors, or localization deviations in these cases. UCH-Net also exhibits occasional missed detections, but produces comparatively fewer visible errors in several selected examples involving occlusion, densely distributed objects, and small objects.

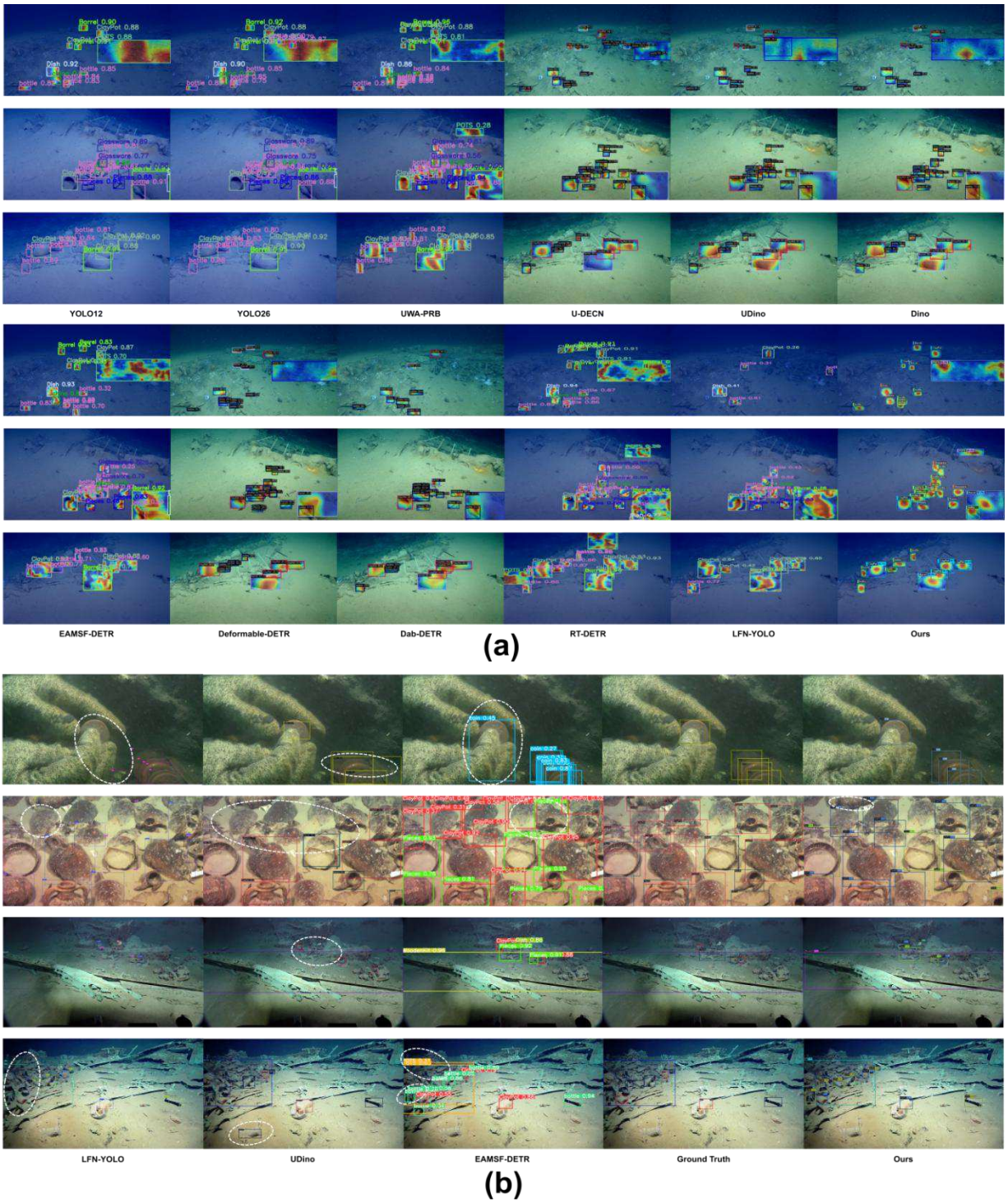


Fig.5. Grad-CAM++ response patterns and qualitative detection performance of UCH-Net and competing methods on challenging underwater cultural heritage scenes from the WRF dataset. The upper panels (a) compare UCH-Net with twelve representative detectors, with predicted boxes labelled by class and confidence, and inset Grad-CAM++ heatmaps showing spatial attention (warmer colors indicate stronger activation). On crowded and low-visibility scenes, competing methods diffuse attention over the background, emit spurious boxes or miss faint targets, while UCH-Net concentrates on the artifacts. The lower panels (b) close up on heavily overlapping, partially buried, or small distant targets under suspended particles and color distortion. White dashed ellipses mark the difficult instances that LFN-YOLO, UDINO and EAMSF-DETR miss or mislocalise, whereas UCH-Net recovers them consistently and matches the ground truth.

Discussion

This study addresses key challenges in underwater cultural heritage imagery, including limited task-specific annotations, severe color distortion, weakened object boundaries, and substantial scale variation. We developed the WRF dataset and proposed UCH-Net, a detection network that incorporates depth-modality input. UCH-Net achieved

the highest overall mAP and mAP75 among the compared methods on both datasets. Although individual metrics, such as Recall, were not consistently optimal, the overall results indicate that UCH-Net provides a favorable balance between detection performance and localization accuracy.

The improvements under stricter localization criteria are particularly relevant to cultural heritage applications. Fig.3. summarize the results on WRF dataset, where UCH-Net shows only a modest mAP50 improvement over the strongest baseline but a more evident advantage at IoU = 0.75. A similar pattern is observed on RUOD. In underwater archaeological image analysis, identifying object presence is only one aspect of the task. Accurate delineation of object extent is also important for candidate-object cropping, archaeological site inventory construction, repeated-survey comparisons, and subsequent expert review. The contribution of UCH-Net therefore extends beyond detection accuracy to improving the spatial consistency of automatically generated candidate regions.

The ablation experiments in Table 3 show that introducing the depth branch through the DSB improves the principal detection metrics on RUOD and WRF datasets. This finding suggests that relative geometric information provides complementary structural cues when RGB features are degraded by color distortion, weakened textures, or partial occlusion. The performance gain associated with CBFM is greater on the WRF dataset than on RUOD, indicating that adaptive appearance modulation may be particularly beneficial in archaeological scenes characterized by complex color distortion and background interference. The result supports the utility of incorporating a physical prior to characterize underwater imaging degradation. Ablation experiments of CBFM internal components indicate that CA, SA, and PA make complementary contributions to the complete module.

The qualitative results are consistent with the quantitative findings. Fig.5. shows Grad-CAM++ [50] visualizations in which UCH-Net generally produces more concentrated high-response regions around manually annotated cultural heritage objects. The detection examples further suggest relatively stable performance in challenging scenarios involving overlapping coins, densely distributed or partially buried artifacts, distant small objects, and complex background interference. These observations provide complementary visual evidence of model behavior under difficult conditions. Nevertheless, attention heatmaps should be interpreted as qualitative supporting evidence rather than direct causal evidence of the underlying feature-fusion mechanism.

This study still has several limitations. First, the depth maps are estimated from RGB images by Depth Anything V2 [26], rather than acquired independently with calibrated depth sensors. They may be affected by underwater color distortion, scattering, weak textures, and domain differences between underwater scenes and the depth-estimation model's training data. These depth maps are consequently better understood as auxiliary geometric priors rather than direct measurements of physical distance. Future work could incorporate measured depth information and confidence-aware fusion mechanisms to improve the reliability of geometric cues [23]. Second, WRF dataset currently adopts a temporally ordered 8:2 image-level split. Site-independent splits would provide a more rigorous evaluation of cross-site generalization. The long-tailed class distribution of the dataset [51] and the current reliance mainly on single-run experiments also warrant further investigation. In addition, although UCH-Net contains fewer parameters than BSR5-YOLO [41], its computational cost remains relatively high at 143.8 GFLOPs, potentially limiting real-time deployment on resource-constrained underwater platforms [44]. Therefore, model lightweighting and deployment-oriented optimization remain important priorities.

Within practical underwater archaeological workflows, UCH-Net could support the automatic screening and localization of underwater cultural heritage objects. For large-scale imagery acquired by remotely operated vehicles (ROVs) [52], autonomous underwater vehicles (AUVs) [53], or divers, the model could identify potential artifacts and ship components for subsequent verification. Confirmed detections could then contribute to archaeological site inventories, digital archive management, and repeated-survey documentation. The proposed framework provides a potential technical basis for improving the efficiency and consistency of underwater archaeological image analysis and documentation [54]. Overall, the results support the value of combining degradation-aware appearance modeling with auxiliary geometric information for underwater cultural heritage detection and indicate that the UCH-Net has practical potential for automated documentation and monitoring of underwater cultural heritage.

Data availability

The WRF dataset generated and analyzed during the current study is publicly available at https://doi.org/10.57760/sciencedb.0148c..The RUOD dataset used in this study is publicly available from its original source.

Code availability

The source code for UCH-Net and the code used to reproduce the experiments in this study are publicly available at https://github.com/duyanxiong/UCH-Net.

Acknowledgments

The authors gratefully acknowledge the National Oceanic and Atmospheric Administration (NOAA) for making the underwater video footage used in this study publicly available. The authors also acknowledge the Copernicus Data Space Ecosystem (CDSE) for providing access to the marine bathymetric data used in this work. This work was supported by the Research Startup Fund of Xiamen University of Technology (Grants #YKJ25049R), the Natural Science Foundation of Fujian Province (Grants #2025J011276, #2024J011194, Grants #2021H0026, and #2023J011427), the Research Project of Fujian Provincial Department of Industry and Information Technology (Grants #50102250012), the Natural Resources Science and Technology Innovation Project of Fujian Province (Grants #KY-030000-04-2025-022), the Xiamen Industry University Research Cooperation Project (Grants #2024FCD012025010117), and the Special Project of the Academician Expert Station of Xiamen University of Technology (Grants #50199250001).

Author contributions

Z.S. and R.W.: methodology, data processing and analyses, and writing. W.M.: conceptualization, funding acquisition, supervision and writing. Q.N., Z.L., X.D., and X.Z. participated in interpreting the results and refining the manuscript.

Competing Interests

The authors declare no competing financial or non-financial interests.

References

[1] Raveendran,S., S., Patil, M. D. & Birajdar, G. K.“Underwater image enhancement: a comprehensive review, recent trends, challenges and applications.” *Artificial Intelligence Review* **54,** 5413–5467 (2021).

[2] O'keefe,P.J.*Shipwrecked Heritage:A Commentary on the UNESCO Convention on Underwater Cultural Heritage* (Institute of Art and Law, Builth Wells, 2014).

[3] Perez-Alvaro, E.Climate change and underwater cultural heritage: Impacts and challenges. *Journal of Cultural Heritage*, **21**, 842-848 (2016).

[4] Wang, J. et al. B-YOLOX-S: A lightweight method for underwater object detection based on data augmentation and multiscale feature fusion. *Journal of Marine Science and Engineering* **10**, 1764 (2022).

[5] Liu,C. et al. A dataset and benchmark of underwater object detection for robot picking. In *Proc. 2021 IEEE international conference on multimedia&expo workshops (ICMEW)* (pp.1-6) (IEEE, 2021).

[6] Fu,C. et al. Rethinking general underwater object detection: Datasets, challenges, and solutions. *Neurocomputing*, **517**, 243-256 (2023).

[7] Lian,S. et al.Watermask: Instance segmentation for underwater imagery. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* 1305-1315 (2023).

[8] Hong,J., Fulton,M.S., & Sattar,J. TrashCan 1.0 an instance-segmentation labeled dataset of trash observations. *Data Repository for the University of Minnesota (DRUM).* Preprint at https://arxiv.org/abs/2007.08097 (2020).

[9] Chen,Y. et al. Achieving domain generalization for underwater object detection by domain mixup and contrastive learning. *Neurocomputing*, **528**, 20-34 (2023).

[10] Sethuraman,A.V., Sheppard,A., Bagoren,O., Pinnow,C., Anderson,J., Havens,T.C., & Skinner,K.A. Machine learning for shipwreck segmentation from side scan sonar imagery:Dataset and benchmark. The International Journal of Robotics Research, **44(3)**, 341-354 (2025).

[11] Yang,Y. et al. Object Detection for Underwater Cultural Artifacts Based on Deep Aggregation Network with Deformation Convolution. *Journal of Marine Science and Engineering*, **11(12)**, 2228 (2023).

[12] Papageorgiou,M. Underwater cultural heritage facing maritime spatial planning:Legislative and technical issues. *Ocean&Coastal Management*, **165**, 195-202 (2018).

[13] Lowe,D.G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, **60(2)**, 91-110 (2004).

[14] Dalal,N., & Triggs,B. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR),* **1**, pp.886-893 (IEEE, 2005).

[15] Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149, (2017).

[16] Wang, A. et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, **37**, 107984-108011 (2024).

[17] Zhao, Y. et al. DETRs Beat YOLOs on Real-Time Object Detection,. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16965–16974 (2024).

[18] Gao,G., Li,C., Zhang,X., Yao,B., & Chen,Z. Mamba-CNN hybrid Multi-scale ship detection Network driven by a Dual-perception feature of Doppler and Scattering. *ISPRS Journal of Photogrammetry and Remote Sensing*, **232**, 656-674 (2026).

[19] Liu, Mingxin, et al. LFN-YOLO: precision underwater small object detection via a lightweight reparameterized approach. *Frontiers in Marine Science,* **11**, 1513740 (2025).

[20] Chen, Y. et al. Implicit multi-spectral transformer:An lightweight and effective visible to infrared image translation model, in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 1–8 (IEEE, 2024).

[21] Quang-Trung Truong. et al. Marine Video Kit: A New Marine Video Dataset for Content-based Analysis and Retrieval.In *Proceedings of the 29th International Conference on Multimedia Modeling(MMM)*, 539-550 (Springer, 2023).

[22] Y.Yang. et al. UMIS-YOLO: Underwater Multimodal Images Instance Segmentation With YOLO. in *IEEE Transactions on Geoscience and Remote Sensing*, .**63**, pp.1-14, 4212414 (2025).

[23] Cai,Z., & Metzler,C.Underwater Monocular Metric Depth Estimation: Real-World Benchmarks and Synthetic Fine-Tuning with Vision Foundation Models. Preprint at https://arxiv.org/abs/2507.02148 (2025).

[24] Wang,X., Song,X., Li,Z., & Wang,H. YOLO-DBS:efficient target detection in complex underwater scene images based on improved YOLOv8. *Journal of Ocean University of China*, **24(4)**, 979-992 (2025).

[25] Liu,Z., Wang,B., Wang,B., & Li,Y. U-decn:End-to-end underwater object detection convnet with improved denoising training. *IEEE Transactions on Geoscience and Remote Sensing*. **63**, 1-9 (2025).

[26] Yang, L. et al. Depth Anything V2. *Advances in Neural Information Processing Systems*. 37, 21875 – 21911 (2024).

[27] Delgado, J.P., Irion, J. & Brennan, M.L. Wreck 15377: A Probable Immigrant and Cargo Packet of the Mid-Nineteenth Century in the Gulf of Mexico. *J. Marit. Archaeol*. **18**, 521 – 552 (2023).

[28] Brennan, M. et al. Telepresence-Enabled Maritime Archaeological Exploration in the Deep. *Journal of Maritime Archaeology* **13(2)**: 97-121 (2018).

[29] Delgado, J.P., Brennan, M.L. Wreck 15563: An Early Nineteenth-Century Merchant Vessel in the Gulf of Mexico. *Journal of Maritime Archaeology.* **18**, 477–520 (2023).

[30] Kocabaş, U. et al. Following the Karaburun Anchors: 2024 Karaburun-Philea Underwater Survey (Black Sea, Türkiye). *International Journal of Nautical Archaeology*, 1-17 (2026).

[31] Opaiţ, A., Davis, D. & Brennan, M. L. The Sinop I shipwreck: a Black Sea merchant ship from the Roman Imperial era. In *Connecting the Ancient West and East: Studies Presented to Prof. Gocha R. Tsetskhladze* **1**, 633 – 646 (2022).

[32] Türkmenoğlu, E., & Davis, D. Knidos F, L and N: Three Deep Sea Shipwrecks from the Byzantine Period. *Heritage*, **9(6)**, 216 (2026).

[33] Li, B. et al. MARIS: Marine Open-Vocabulary Instance Segmentation with Geometric Enhancement and Semantic Alignment. Preprint at arXiv: 2510.15398 (2025).

[34] Wang, C.-Y. et al. CSPNet: A New Backbone That Can Enhance Learning Capability of CNN. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 390 – 391 (2020).

[35] Liu, S., Qi, L., Qin, H., Shi, J. & Jia, J. Path aggregation network for instance segmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 8759 – 8768 (2018).

[36] Lin, T.-Y. et al. Feature Pyramid Networks for Object Detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125 (2017).

[37] Zheng, Z. et al. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *Proc. AAAI. Artificial Intelligence.* **34**, no. 7, pp. 12993–13000, (2020).

[38] Zhu, X. et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Proc. International Conference on Learning Representations (ICLR)* (2021).

[39] Zhang, H. et al.. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In *Proc. The Eleventh International Conference on Learning Representations (ICLR)* (2023).

[40] Wang, W., Yu, Z. & Huang, M. Refining features for underwater object detection at the frequency level. *Frontiers in Marine Science*. **12**, 1544839 (2025)

[41] Zhou, J. et al. Spatial residual for underwater object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **47(6),** 4996–5013 (2025).

[42] Li, J. & Yan. EAMSF-DETR: Edge-aware multi-scale feature fusion network based on DETR for underwater object detection. *Optics & Laser Technology*, **194**, 114357 (2026).

[43] Liu, S. et al. DAB-DETR: dynamic anchor boxes are better queries for DETR. In *Proc. International Conference on Learning Representations (ICLR)* (2022).

[44] Zheng, Z. & Yu, W. RG-YOLO: multi-scale feature learning for underwater target detection. *Multimedia Systems*. **31**, 26 (2025).

[45] Tian, Y., Ye, Q. & Doermann, D. YOLOv12: Attention-centric real-time object detectors. *Advances in Neural Information Processing Systems*, **38**, pp. 78433–78457, (2025).

[46] Jocher, G. et al. Ultralytics YOLO26: Unified real-time end-to-end vision models. Preprint arXiv:2606.03748, (2026).

[47] Tsai, Y.-S., Tsai, C.-T. & Huang, J.-H. Multi-scale detection of underwater objects using attention mechanisms and normalized Wasserstein distance loss. *The Journal of Supercomputing*, **81**, 785, (2025).

[48] Song, P., Li, P., Dai, L., Wang, T. & Chen, Z. Boosting R-CNN: Reweighting R-CNN samples by RPN's error for underwater object detection. *Neurocomputing*, **530**, pp. 150–164, (2023).

[49] Ji, X., Chen, S., Hao, L.-Y., Zhou, J. & Chen, L. FBDPN: CNN-Transformer hybrid feature boosting and differential pyramid network for underwater object detection. *Expert Systems with Applications*, **256**, 124978, (2024).

[50] Chattopadhyay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In *Proc. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* 839 – 847 (IEEE, 2018).

[51] Alexandridis, K. P., Elezi, I., Deng, J., Nguyen, A., & Luo, S. Fractal Calibration for Long-tailed Object Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15139–15150 (2025). DOI: 10.1109/CVPR52734.2025.01410.

[52] Huang, Z. & Liu, T. The application of marine intelligent platforms in underwater archaeology. *Measurement*, **275**: 121373 (2026). DOI: 10.1016/j.measurement.2026.121373.

[53] Benetatos, C. et al. Exploring the Mediterranean: AUV high-resolution mapping of the Roman wreck offshore of Santo Stefano al Mare (Italy). *Journal of Marine Science and Engineering*, **13(10**), 1921 (2025). DOI: 10.3390/jmse13101921.

[54] Diamanti, E. & Ødegård, Ø. Visual sensing on marine robotics for the 3D documentation of underwater cultural heritage: A review. *Journal of Archaeological Science*, **166,** 105985 (2024). DOI: 10.1016/j.jas.2024.105985.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- npjHeritageScienceSupplementaryInformation.docx