

Article

YOLOv13-Based Two-Stage Framework for Underwater Damage Detection on Reinforced Concrete Surfaces

Xinwei Wang ¹, Muhammad Moman Shahzad ^{2,*}, Xijun Ye ³, Yinghao Zhao ⁴ and Zhihao Wang ^{1,*}

¹ School of Civil Engineering and Communication, North China University of Water Resources and Electric Power, Zhengzhou 450045, China; wangxinwei@ncwu.edu.cn

² Department of Civil Engineering, National University of Sciences and Technology, Balochistan Campus, Quetta 86000, Pakistan

³ School of Civil Engineering and Transportation, Guangzhou University, Guangzhou 510440, China; xijun_ye@gzhu.edu.cn

⁴ School of Future Transportation, Guangzhou Maritime University, Guangzhou 510725, China; zhaoyinghao@gzmtu.edu.cn

* Correspondence: m.momin@nbc.nust.edu.pk (M.M.S.); wangzhihao@ncwu.edu.cn (Z.W.)

Abstract

Prolonged underwater exposure degrades reinforced concrete (RC) structures, causing chloride-induced corrosion, spalling, cracking, and rebar exposure. Timely damage identification is critical for structural safety, but conventional non-destructive testing methods face severe limitations underwater due to restricted accessibility, image degradation, and weak-textured, irregular crack boundaries. Vision-based inspection offers a promising alternative but remains constrained by underwater optical degradation. This study proposes a two-stage detection framework for underwater RC based on YOLOv13 (YOLOv13-TSDD). First, an underwater color-detail enhancement network (UCDEN) performs color correction, detail recovery, and contour reconstruction through multi-channel color enhancement and multi-level feature refinement. Second, two detection modules are introduced: a pinwheel-shaped receptive field convolution (PRFConv), improving sensitivity to directional textures and local linear structural responses in shallow layers, and a crack-aware efficient multi-scale attention (CEMA) mechanism, enabling joint channel-spatial recalibration and multi-scale focus on crack-relevant regions. A fine-grained irregular crack IoU (FID-IoU) loss function is also developed, using auxiliary boundary boxes and piecewise weighted mapping to improve bounding-box regression for irregular cracks. Experimental results demonstrate that YOLOv13-TSDD not only achieves the best overall image enhancement performance among the evaluated methods but also delivers the highest detection performance. On the constructed underwater dataset, YOLOv13-TSDD achieves Precision, Recall, and mAP@0.50 of 93.63%, 88.97%, and 94.52%, respectively, demonstrating improved performance under complex underwater conditions.



Academic Editor: Flavio Stochino

Received: 26 July 2026

Revised: 19 August 2026

Accepted: 25 August 2026

Published: 28 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: underwater reinforced concrete; material property degradation; non-destructive testing; surface damage detection; deep learning; YOLOv13; image enhancement

1. Introduction

Reinforced concrete (RC) is widely used in marine infrastructure, including port facilities, bridge piers, breakwaters, and offshore platform substructures, due to its load-bearing capacity and cost-effectiveness [1–3]. In underwater environments, chloride penetration, wet–dry cycling, freeze–thaw effects, and hydrodynamic loading drive progressive deterioration, producing cracking, spalling, and rebar exposure, while biofouling may further

interfere with visual inspection. These deterioration phenomena reduce the effective cross-section, weaken the concrete–rebar bond, and accelerate further attack, compromising durability and load-carrying capacity. Timely detection is therefore essential for maintenance decisions and lifecycle management.

Non-destructive testing (NDT) is the standard inspection approach. Methods such as acoustic emission, ground-penetrating radar, and impact-echo are effective on land but face restricted deployment, signal attenuation, and limited resolution underwater. Vision-based detection offers a practical alternative, compatible with ROVs and AUVs, but is constrained by underwater optical degradation: blur, non-uniform color bias, low contrast, and boundary ambiguity from light scattering and absorption. Crack boundaries blend into background texture, spalling loses tonal contrast, and rebar-exposure contours become difficult to distinguish from the surrounding concrete, conditions that existing detection pipelines are not built to handle. Accurate detection therefore requires solutions addressing both image degradation and crack-specific feature representation [4,5].

Image enhancement technology (IET) has been studied extensively as a preprocessing step [6–13], shifting from general perceptual improvement toward task-oriented enhancement that preserves crack edges and geometry. Ye et al. [8] showed that image fusion-based enhancement improves Recall and mAP, though cascaded designs misalign enhancement and detection objectives. Kang et al. [9] used unsupervised learning with PatchNCE loss to recalibrate data distributions while preserving structural similarity. Talamkhani et al. [10] showed that joint, end-to-end treatment of scattering, attenuation, and low illumination is needed for meaningful gains. CNN-based inspection, optical sensing, and polarization-based recovery have also been applied to crack detection and structural inspection under challenging conditions [11–13]. Wu et al. [14] showed bandwidth and pipeline design constrain task performance, and Noda et al. [15] showed pose estimation accuracy affects image quality during close-range inspection. Collectively, these studies improve visual quality but none jointly addresses color correction, illumination compensation, and boundary preservation, leaving blurred cracks, weak spalling textures, and degraded void features that reduce detection accuracy.

Deep learning now dominates underwater crack recognition [16,17], with multi-scale fusion and edge augmentation improving detection under complex conditions [18–20]. YOLO variants are widely used for underwater targets due to their accuracy–speed trade-off on embedded ROV/AUV platforms [21–23]. Zang et al. [24] proposed MCD-YOLO for marine surface cracks using edge enhancement and dynamic heads. Li et al. [25] combined ExtraDW convolutions, SE attention, and NWD loss in YOLOv12 for crack detection. Abbas et al. [26] applied CBAM in BT-YOLO for low-illumination coastal monitoring. None of these address fully submerged RC surfaces, where color distortion, illumination imbalance, biofouling, and complex morphology combine into a harder problem than in [24–26].

Three gaps remain for underwater reinforced concrete surface structural damage (URCSSD) detection: (i) existing IET methods lack coordinated modeling of color bias, illumination, and boundary blur; (ii) detection backbones underperform at extracting orientation-sensitive, weak-boundary features with insufficient channel-spatial recalibration; and (iii) standard IoU-based losses perform poorly on small, irregular cracks, causing localization drift.

This study proposes YOLOv13-TSDD, a two-stage framework addressing these gaps. YOLOv13 is used as the baseline because it introduces HyperACE, capturing high-order dependencies across distant crack regions; FullPAD, preserving fine-grained information across stages; and DSC3k2, reducing computational cost without sacrificing resolution. These properties suit URCSSD detection, verified in Section 5. Experiments are conducted

on an underwater RC surface damage dataset covering concrete cracking, spalling, and rebar exposure.

In the first stage, the underwater color-detail enhancement network (UCDEN) applies multi-channel color enhancement (MCCE) and multi-layer feature enhancement (MLFE) to correct color bias, blur, and illumination imbalance. In the second stage, pinwheel-shaped receptive field convolution (PRFConv) replaces the first three convolution layers to improve sensitivity to directional textures, motivated by the orientation-sensitive nature of cracks and rebar exposure. The crack-aware efficient multi-scale attention (CEMA) module strengthens channel-spatial focus at mid-to-high feature levels, with a Sobel gradient-prior branch improving edge perception. The fine-grained irregular crack IoU (FID-IoU) loss improves bounding-box regression for complex, discontinuous cracks via auxiliary boxes and piecewise weighted mapping.

The main contributions are:

- (i) UCDEN, combining color correction with multi-level feature enhancement to improve visual discriminability of cracks, spalling, and rebar exposure.
- (ii) PRFConv and CEMA, addressing shallow-stage insensitivity to orientation-specific textures and weak channel-spatial recalibration, improving representation of irregular cracks.
- (iii) FID-IoU, a regression loss for small, irregular cracks, improving localization accuracy where standard IoU losses fail.
- (iv) YOLOv13-TSDD, integrating (i)–(iii) into a unified framework, validated on three representative underwater RC surface damage categories.

The paper is organized as follows: Section 2 presents the theoretical background, Section 3 details the framework, Section 4 describes the experimental setup, Section 5 presents results, and Section 6 concludes.

2. Related Work

2.1. Deep Learning-Based Image Enhancement

Deep learning-based underwater image enhancement has progressed from simple color correction toward supervision optimization, cross-domain transfer, interpretability, and multi-scale detail reconstruction. Contrastive learning has been used to extract localized benefits across candidate outputs, improving supervision stability [27]. Hybrid data- and model-driven approaches, combining deep unrolling, convolutional dictionaries, and shared-specific feature decomposition, have improved interpretability and adaptability to luminance degradation, color distortion, and texture blurring [28]. Cross-domain transfer and physics-based priors have leveraged aerial dehazing knowledge and color restoration models to compensate for the scarcity of paired underwater ground truth [29]. Other approaches combine color loss minimization, adaptive local contrast enhancement, and multi-scale fusion to improve color naturalness and detail visibility [30,31].

These methods target general perceptual quality but do not explicitly model the non-uniform color bias, local illumination imbalance, and boundary weakening characteristic of underwater damage such as spalling and rebar exposure. This limits their capacity to improve damage discriminability for downstream detection.

2.2. Feature Enhancement Networks for Object Detection

Feature enhancement research for object detection has moved from simple network stacking toward collaborative optimization, including shallow-stage detail preservation, multi-scale fusion, attention mechanisms, and local-global modeling. Conventional convolutions are widely recognized as insufficient for retaining edge and local saliency information in weak-texture, small-target, complex-background scenarios. Edge-aware branches,

shallow-deep fusion, and multi-branch contextual modeling have therefore been used to improve target discriminability [32,33]. To counter the loss of small-target information during deep semantic enhancement, dense connectivity, weighted feature reuse, and coarse-to-fine prediction have been used to strengthen feature propagation and contour representation in deeper layers [34]. In multi-scale settings, attention has been combined with feature pyramids, dynamic detection heads, or local-global modeling, using cross-scale weighting and context-dependent modeling to improve representation stability [35,36].

These advances target compact, roughly isotropic small objects, whereas cracks and spalling boundaries are the opposite case: elongated, thin, and defined by orientation rather than scale. A crack's discriminative signal lies in its directionality and continuity along an axis, not its size relative to the background, so architectures tuned for scale-invariant small-object detection do not transfer directly. This motivates a design that explicitly encodes orientation at the shallow feature stage, rather than relying on generic multi-scale or attention mechanisms to recover it implicitly.

2.3. YOLOv13

Beyond the YOLO family, other computer vision approaches offer viable alternatives for structural damage recognition and durability assessment. DeepLab, based on atrous (dilated) convolution and encoder-decoder pixel-wise segmentation, achieves strong delineation of fine, irregular boundaries in high-resolution imagery; for example, DeepLab v3 has been applied to pixel-level classification of malignant regions in gastric pathology slides, achieving robust boundary delineation across heterogeneous, high-resolution image sources [37]. Such segmentation capability is, in principle, well suited to the irregular, thin-boundary morphology of underwater cracks, though it requires dense pixel-level annotation rather than the bounding-box labels used in this study. Computer vision has also been applied to other aspects of concrete durability assessment beyond damage localization; for instance, a low-cost vision-based system has recently been developed to automate the estimation of cementitious sorptivity by tracking the water absorption front in paste, mortar, and concrete specimens, illustrating the broader applicability of vision-based monitoring to durability assessment in cementitious materials [38].

This study adopts YOLOv13 [39] as the baseline detection network over these alternatives, as its bounding-box formulation matches our annotation format (Section 4.1.4) and its anchor-free, single-stage design offers faster inference than segmentation-based pipelines, a key requirement for real-time underwater deployment. Compared with other models in the YOLO series [40–48], YOLOv13 incorporates HyperACE and FullPAD, enabling more effective modeling of high-order feature associations across spatial positions and scales while improving network-wide information flow and fine-grained feature representation. It also employs a lightweight feature extraction module based on depthwise separable convolution (DSC3k2), preserving strong feature modeling capacity while substantially reducing parameter count and computational cost. These properties make YOLOv13 a suitable baseline for URCSSD detection, supporting accurate and efficient detection of small-scale, weak-boundary, and morphologically irregular damage, including cracks, spalling, and rebar exposure, in complex underwater environments. The following subsections introduce the operating principles of HyperACE, FullPAD, and DSC3k2.

2.3.1. HyperACE

To overcome the reliance of conventional hypergraph methods on manually specified parameters [49], HyperACE introduces a learnable hyperedge generation mechanism that adaptively constructs hyperedge relationships from input visual features and mines latent high-order associations among feature vertices. Unlike prior methods constrained to

pairwise relational modeling, HyperACE enables global feature fusion and enhancement across spatial positions and scales, improving overall feature expressiveness.

Specifically, let $X = x_i \in \mathbb{C}, i = 1, 2, \dots, N$ denote the vertex feature matrix, where $\mathbb{C}$ denotes the number of feature channels. Context vectors f_{avg} and f_{max} are defined via global average pooling and global max pooling, respectively, and concatenated to form the global vertex context f_{ctx} . This process is expressed as:

$$f_{ctx} = \begin{pmatrix} f_{avg} \\ f_{max} \end{pmatrix} \in 2\mathbb{C} \quad (1)$$

Subsequently, global offsets are generated from the vertex context through a mapping layer and added to learnable global prototypes to obtain the hyperedge prototype P_m . To further promote feature diversity, a multi-head mechanism is introduced to partition the vertex query vectors z_i along feature $z_i = W_{pre}x_i$ into h subspaces $\{\hat{z}_i^\tau \in d_h\}_{\tau=1}^h$, where W_{pre} denotes the weight matrix and $d_h = \mathbb{C}/h$ denotes the corresponding transformation. Each hyperedge prototype is similarly partitioned into subspaces $\{\hat{P}_m^\tau \in d_h\}_{\tau=1}^h$. In the τ subspace, the similarity between the i -th vertex query vector and the hyperedge is then computed as:

$$s_{i,m}^\tau = \frac{\langle \hat{z}_i^\tau, \hat{P}_m^\tau \rangle}{\sqrt{d_h}} \quad (2)$$

Next, the overall similarity is defined as the average of the similarity scores across all subspaces. The similarity between each vertex query vector and the hyperedge prototype is treated as the contribution of that vertex to the corresponding hyperedge and is normalized along the vertex dimension to obtain the continuous participation matrix $A_{i,m}$.

$$A_{i,m} = \frac{\exp(\tilde{s}_{i,m})}{\sum_{j=1}^N \exp(\tilde{s}_{j,m})} \quad (3)$$

Following the generation of adaptive hyperedges, hypergraph convolution is performed to achieve feature aggregation and enhancement:

$$f_m = \sigma \left(W_e \sum_{i=1}^N A_{i,m} \cdot x_i \right) \quad (4)$$

$$\tilde{x}_i = \sigma \left(W_v \sum_{m=1}^M A_{i,m} \cdot f_m \right) \quad (5)$$

where i denotes the vertex index and m denotes the hyperedge index. i and W_v represent the projection weights of the hyperedges and vertices, respectively, and σ denotes the activation function.

2.3.2. FullPAD

FullPAD aggregates multi-scale features from the backbone and feeds them into the HyperACE module for correlation enhancement. The enhanced features are then redistributed to the corresponding network stages through separate branches, enabling fine-grained cross-layer feature interaction and cooperative representation that improves gradient propagation and overall detection performance. Specifically, after obtaining the correlation-enhanced features Y at stages B3, B4, and B5, FullPAD rescales them to the spatial resolution of the corresponding stage and performs channel alignment via 1×1 convolution. The FullPAD process is formulated as:

$$H_i = \text{Conv}_{1 \times 1}(\text{Resize}(Y \leftarrow \text{size}(B_i))), i \in \{3, 4, 5\} \quad (6)$$

For the feature map F_i at an arbitrary stage i , gated fusion is employed to regulate information flow and integration:

$$\hat{F}_i = F_i + \gamma H_i \quad (7)$$

where γ is a learnable scalar parameter that adaptively balances the contributions of the correlation-enhanced features and the original features.

2.3.3. DSC3k2

YOLOv13 adopts a lightweight design based on depthwise separable convolution, introducing DSConv, DS-Bottleneck, and DS-C3k modules to replace conventional large-kernel convolutions. Among these, the DS-C3k2 module is employed in the backbone for multi-scale feature extraction, maintaining an effective receptive field while substantially reducing parameter count and computational cost. This design strikes a favorable balance between detection accuracy and computational efficiency, enabling faster inference and lower deployment overhead, making it well-suited for real-time object detection scenarios. The architectural diagram of the DS-C3k2 module is shown in Figure 1.

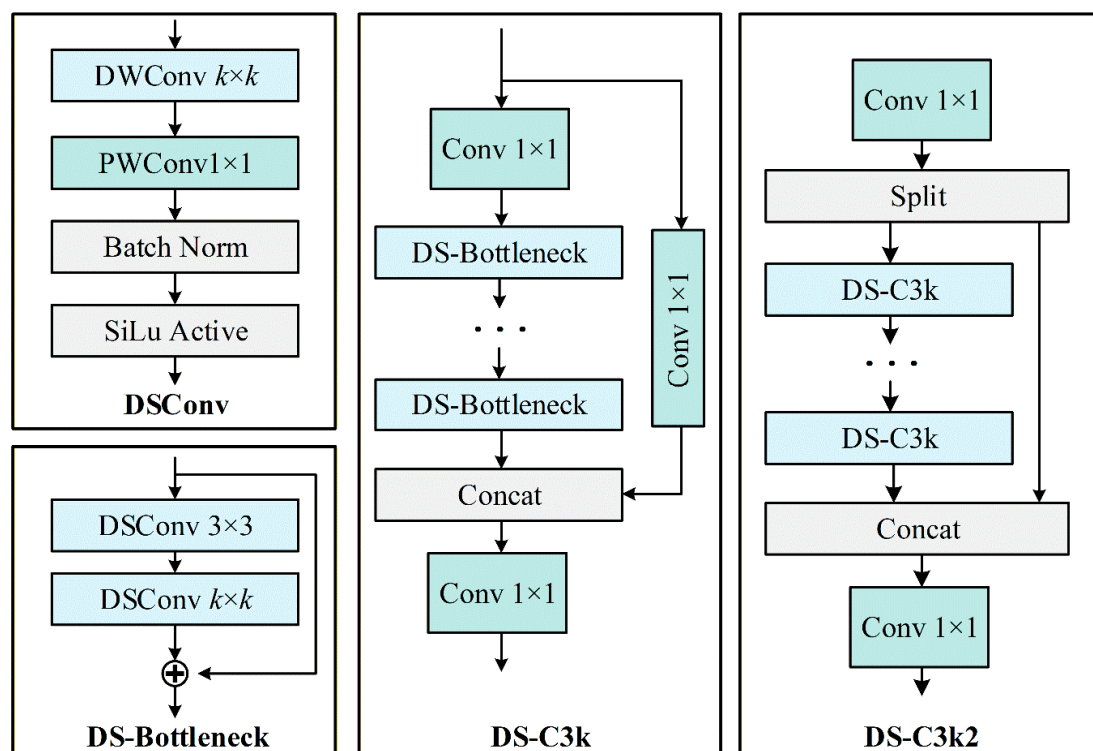


Figure 1. Architecture diagram of the DS-C3k2 module.

2.4. IoU Loss Functions

In object detection, the Intersection over Union (IoU) loss is a core method for bounding-box regression, measuring the overlap between predicted and ground truth boxes. To address the limitation of standard IoU, which provides no gradient signal when predicted and ground truth boxes do not overlap, several improved formulations have been proposed, including CIoU [50], EIoU [51], and SIoU [52]. These incorporate the minimum enclosing region, center point distance, aspect ratio consistency, and directional alignment, improving convergence speed and localization accuracy.

While these improved IoU losses perform well in general object detection, they show notable limitations for URCSSD detection. Surface damage in underwater RC structures, including cracks, spalling, and rebar exposure, is typically characterized by small scale, fine

granularity, irregular morphology, large aspect ratio variation, and boundary ambiguity, and is further affected by low contrast, scattering noise, and complex background interference underwater. As a result, existing general-purpose IoU losses remain insufficient for precise and stable bounding-box localization of such damage.

2.5. Summary of Related Work and Research Gaps

Table 1 summarizes the reviewed image enhancement and crack detection studies in terms of their scope of application, technical approach, and principal limitations. This comparison clarifies why existing methods, considered individually, do not resolve the combined challenges of underwater reinforced concrete surface structural damage (URCSSD) detection.

Table 1. Summary of related studies and their principal limitations.

Study	Scope of Application	Approach	Main Limitation
Ye et al. [8]	Underwater structural damage, two-stage pipeline	Image fusion-based enhancement + lightweight detection	Enhancement and detection stages remain cascaded and misaligned in objective
Kang et al. [9]	General underwater image enhancement	Unsupervised learning with PatchNCE loss	Not task-oriented toward crack boundary preservation
Talamkhani and Liu [10]	Underwater bridge component inspection	Joint, end-to-end treatment of scattering, attenuation, low illumination	Does not address crack-specific texture or boundary recovery
Refs. [11–13]	Micro-crack detection, turbid-water imaging	CNN, diffusion models, polarization-based recovery	Improve visual quality but do not jointly address color, illumination, and boundary preservation
Zang et al. [24]	Marine surface crack detection	MCD-YOLO with edge enhancement and dynamic heads	Not evaluated on fully submerged, color-distorted RC surfaces
Li et al. [25]	Crack detection, YOLOv12-based	ExtraDW convolution, SE attention, NWD loss	Does not address underwater-specific color bias or illumination imbalance
Abbas et al. [26]	Low-illumination coastal monitoring	BT-YOLO with CBAM attention	Targets illumination primarily; does not jointly address underwater optical degradation and complex damage morphology
Refs. [27–31]	General underwater image enhancement	Contrastive learning, deep unrolling, cross-domain transfer, color loss minimization	Target general perceptual quality; do not explicitly model crack-relevant boundary weakening
CIoU [50], EIoU [51], SIoU [52]	General object detection, bounding-box regression	Enclosing region, center distance, aspect ratio, directional alignment	Insufficient gradient signal and boundary fitting for small, irregular, low-overlap targets

Across these studies, three unresolved problems recur. First, image enhancement methods target general perceptual quality rather than the coordinated color, illumination, and boundary correction that crack discriminability requires. Second, detection backbones and attention mechanisms developed for marine or coastal targets are not evaluated under the combined color distortion, illumination imbalance, scattering degradation, and complex damage morphology of fully submerged RC surfaces. Third, standard IoU-based losses are not designed for the small scale, irregular morphology, and boundary ambiguity of

underwater cracks. YOLOv13-TSDD is proposed to address these three gaps jointly, as detailed in Section 3.

3. Proposed YOLOv13-TSDD

3.1. Underwater Color-Detail Enhancement Network

3.1.1. Overview of UC DEN

To address the prevalent challenges of non-uniform color bias, local illumination imbalance, texture degradation, and boundary blurring in underwater RC surface structural crack imagery, this study proposes an underwater color-detail enhancement network for task-oriented preprocessing of degraded images at the detection front end. Unlike conventional image enhancement methods that prioritize subjective perceptual quality improvement, UC DEN is explicitly designed around the discriminative requirements of downstream object detection, specifically the need for reliable crack texture, edge contour, and luminance distribution information. The network aims to improve the visual discriminability and structural integrity of crack regions, including cracks, spalling, and rebar exposure, directly at the input stage.

In terms of overall architecture, UC DEN consists of two principal components: a multi-channel color enhancement strategy and a multi-layer feature enhancement mechanism, which together constitute a progressive enhancement process proceeding from color restoration to fine-grained detail reinforcement. MCCE addresses color distortion, insufficient local contrast, and non-uniform color bias inherent in underwater imaging by performing separate modeling and adaptive correction of the three RGB channel degradation characteristics, thereby restoring color naturalness and recovering fundamental structural information. MLFE further integrates multi-resolution feature modeling and cross-scale information fusion to cooperatively reinforce local texture, edge contour, and contextual information in crack regions, compensating for the limitations of color correction alone in recovering fine-grained structural detail.

In summary, UC DEN is not designed merely to improve visual appearance, but functions as an enhancement network optimized for downstream detection task performance. Through coordinated modeling of color correction, detail recovery, and contour reconstruction, UC DEN effectively mitigates the adverse effects of water scattering, low illumination, and color attenuation on crack representation, providing the downstream detection network with cleaner, more stable, and more discriminative input features. This establishes a stronger representational foundation for crack recognition and localization in complex underwater environments. The architecture of the proposed UC DEN image enhancement mechanism is illustrated in Figure 2.

3.1.2. MCCE

Non-uniform absorption and scattering of light across different wavelength bands by the aquatic medium is a primary cause of image quality degradation in underwater RC surface structural crack imagery. This degradation manifests as reduced visibility, color distortion, and weakening of crack texture and boundary information, directly compromising detection accuracy for targets such as cracks, spalling, and rebar exposure.

To address the above challenges, UC DEN incorporates a multi-channel color enhancement strategy that performs joint modeling of the three-channel RGB degradation characteristics, combined with multi-scale feature extraction and channel-adaptive weighting, to achieve coordinated optimization of color correction and structural enhancement, thereby improving input image quality for downstream URCSSD detection.

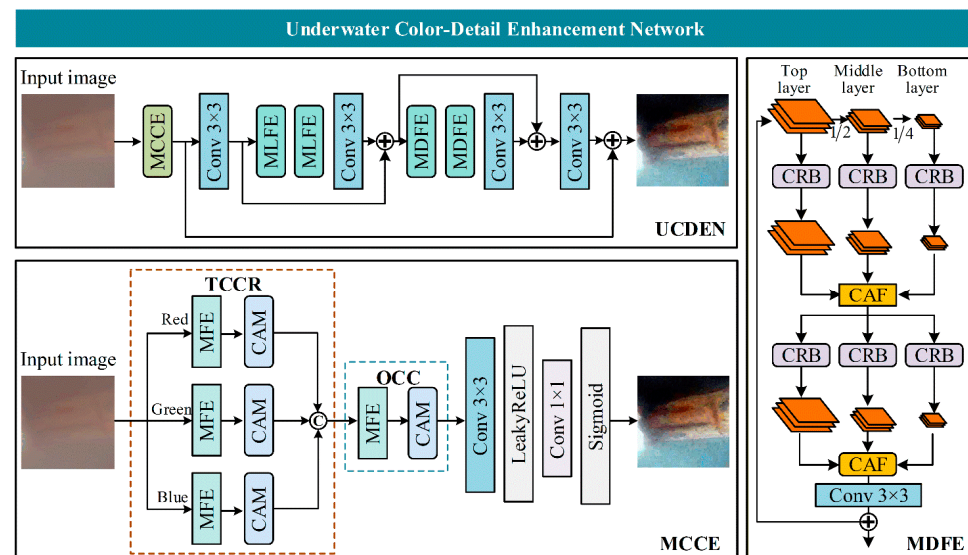


Figure 2. Architecture of the proposed UC DEN image enhancement mechanism.

Specifically, MCCE comprises two stages: three-channel color restoration (TCCR) and overall color correction (OCC). In the TCCR stage, the three RGB channels of the input image are first separated and processed independently, with multi-scale feature extraction (MFE) and a channel attention mechanism (CAM) applied to each channel.

As illustrated in Figure 3, the internal architecture of MFE employs parallel convolutional branches with feature concatenation to simultaneously model local detail, mid-scale structure, and global contextual information. The 1×1 convolutional branch first performs channel compression, reducing the channel dimension from 256 to 64 to lower subsequent computational cost. Combined with LeakyReLU (LReLU) nonlinear activation, this branch enhances the network's sensitivity to fine-grained foreground textures and subtle crack responses. The 3×3 convolutional branch extracts local structural features at a moderate receptive field, facilitating the representation of surface corrosion patches, crack contours, and mid-scale texture variations. The 5×5 convolutional branch focuses on capturing broader contextual information, including non-uniform illumination distributions, global color shifts, and feature attenuation in distant regions, thereby alleviating information loss caused by illumination decay. Concurrently, the max-pooling branch preserves the maximum local response to reinforce edge and contour feature expression, preventing excessive smoothing during the enhancement process. A final 1×1 convolution adjusts the channel dimensionality to ensure consistent channel structure across all branches prior to concatenation.

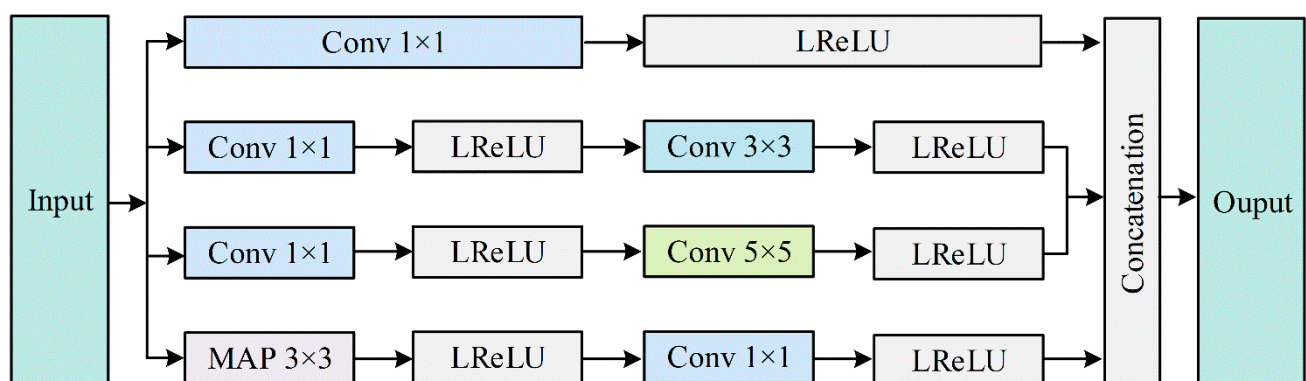


Figure 3. Internal architecture of the MFE module.

Building upon these multi-scale features, CAM performs adaptive recalibration of the fused representations. This process begins with global average pooling applied to each channel to capture its overall statistical contribution to the current feature response, computed as:

$$GAP_m = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W F_m(i, j), m = \{1, 2, \dots, M\} \quad (8)$$

where F_m denotes the m -th channel of the input feature map, GAP_m denotes the spatial dimension of the feature map, and i and j are the pixel coordinate indices, with the result representing the global average pooling output for the m -th channel. Based on this global descriptor, the weight coefficient for the m -th channel is subsequently generated as:

$$Coeff_m = \psi(C_2 \psi(C_1 GAP_m)) \quad (9)$$

where $Coeff_m$ denotes the weight coefficient of the m -th channel, C_1 and C_2 are the weight matrices of the fully connected layers, and ψ denotes the activation function. Finally, the weight coefficients are multiplied element-wise with the input feature map channel by channel to dynamically adjust the response intensity of each channel, and the adjusted result is reassigned to F_m , expressed as:

$$F_m = F_m \otimes Coeff_m, m = \{1, 2, \dots, M\} \quad (10)$$

The roles of MFE and CAM are channel-specific by design. In the red channel, MFE compensates for low contrast caused by red light attenuation, enhancing crack texture and edge information. In the green channel, adaptive modulation suppresses scattering-induced mid-scale blur, preventing over- or under-enhancement of corrosion contours and structural features. In the blue channel, redundant blue-light responses are constrained to reduce the global blue cast and contour distortion common in distant regions, sharpening crack edges and outlines. This channel-specific processing selectively restores the texture, edge, and contour information most relevant to crack detection.

After independent per-channel enhancement, the outputs are concatenated and passed to the OCC stage. Since independent restoration can introduce local oversaturation or abrupt contrast transitions, cross-channel correlation modeling and global constraint optimization are applied to the fused features to smooth color transitions while reinforcing local contrast and edge texture. Adaptive reweighting moderately constrains strongly responding channels and compensates weak ones, suppressing spurious color bias and restoring a spatially consistent color distribution.

Finally, a 3×3 convolution reduces the fused feature dimensionality while retaining useful information, a 1×1 convolution maps the result to a three-channel RGB output, and a Sigmoid activation constrains values to a physically meaningful range. Through this design, MCCE preserves color naturalness while improving detail, edge, and illumination representation, producing higher-quality inputs for the downstream URCSSD detection task.

3.1.3. MLFE

Following multi-channel color enhancement, color distortion and global bias are largely corrected. However, color correction alone cannot fully recover the fine-grained structural features of cracks, spalling, and rebar exposure.

To address this, UCDEN incorporates a multi-layer feature enhancement (MLFE) strategy, which jointly models feature maps at multiple resolutions using a coordinate

residual block (CRB) and a cross-scale adaptive fusion (CAF) module, improving the suitability of the enhanced representation for downstream detection.

MLFE uses three parallel branches at full, half, and quarter resolution, enabling enhancement from local detail to global structure. The top branch extracts fine-grained detail such as micro-cracks, spalling textures, and local edges. The middle branch, with a larger receptive field, models the spatial distribution of rebar exposure, crack contour extension, and local illumination variation. The bottom branch, at the lowest resolution but largest receptive field, represents background illumination, global contour organization, and large-scale degradation patterns. This hierarchical design lets the network capture both high-resolution detail and low-resolution global context, improving crack discriminability against complex backgrounds.

Within each branch, CRB with coordinate attention first refines and reweights features, strengthening responses at critical spatial positions while suppressing irrelevant interference (Figure 4). A convolution performs initial feature refinement, followed by decomposing two-dimensional global pooling into separate one-dimensional encodings along the horizontal and vertical directions, preserving spatial position information along each axis. Specifically, average pooling along the horizontal direction is applied to each channel to generate a feature vector retaining height-dimension spatial information, computed as:

$$Z_{c,h}^{XAP} = \frac{1}{W} \sum_{w=1}^W X_{c,h,w} \quad (11)$$

where $Z_{c,h}^{XAP}$ denotes the feature map compressed along the horizontal direction, W denotes the width dimension of the input feature map, and $X_{c,h,w}$ denotes the input feature map. Here, c , h , and w represent the channel index, height index, and width index, respectively. Similarly, average pooling is applied along the vertical direction for each channel to retain spatial information in the width dimension, expressed as:

$$Z_{c,w}^{YAP} = \frac{1}{H} \sum_{h=1}^H X_{c,h,w} \quad (12)$$

where $Z_{c,w}^{YAP}$ denotes the feature map compressed along the vertical direction, and H denotes the height dimension of the input feature map.

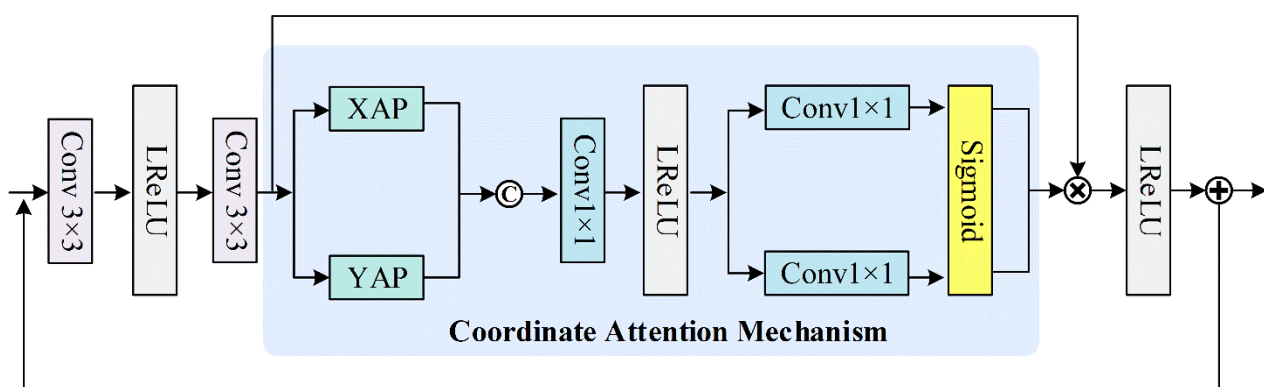


Figure 4. Internal architecture of the CRB module.

The horizontal and vertical feature vectors are concatenated along the spatial dimension and passed through a 1×1 convolution for channel compression, yielding a low-dimensional joint representation. Separate convolutional operations then generate attention weight maps for the horizontal and vertical directions, which are mapped to the range $[0, 1]$ via Sigmoid activation. Finally, the weight maps are multiplied element-wise with the input features, and the result is added to the original input through a residual

connection, simultaneously preserving original detail while reinforcing direction-sensitive global information. This design improves the network's ability to capture orientation-specific crack features such as elongated cracks and strip-like spalling patterns, while alleviating the disruption of pixel neighborhood relationships caused by conventional downsampling operations.

After obtaining enhanced features at each resolution level, cross-scale information interaction is further achieved through CAF. The internal architecture of CAF is illustrated in Figure 5. Defining the high, medium, and low-resolution feature maps output by the three branches as F_1 , F_2 , and F_3 , respectively, a preliminary multi-scale feature fusion is performed via element-wise addition, expressed as:

$$F_{in} = F_1 \oplus F_2 \oplus F_3, F \in H \times W \times C \quad (13)$$

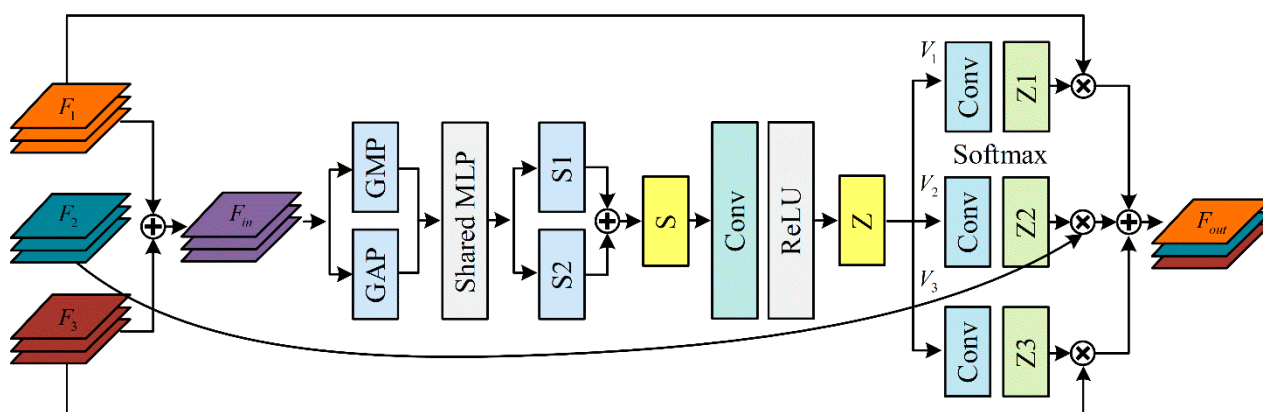


Figure 5. Internal architecture of the CAF module.

Subsequently, global average pooling and max pooling are applied simultaneously to the input feature F_m to extract global statistical information and locally salient responses. Max pooling retains the maximum value $m \in \mathbb{R}^{1 \times 1 \times C}$ across each channel to highlight salient regions and critical structural features in the image, while average pooling produces the channel-wise mean $a \in \mathbb{R}^{1 \times 1 \times C}$ to preserve overall luminance and global tonal information. The pooled outputs m and a are then independently passed through a shared multi-layer perceptron to generate intermediate features s_1 and s_2 , expressed as:

$$s_1 = MLP(a), s_2 = MLP(m) \quad (14)$$

The combined feature is then expressed as:

$$s = s_1 + s_2, s \in \mathbb{R}^{1 \times 1 \times C} \quad (15)$$

A 1×1 convolution is subsequently applied to s for dimensionality reduction, producing a compact feature representation $z \in \mathbb{R}^{1 \times 1 \times c/8}$. Three parallel 1×1 convolutional layers then generate three feature descriptors v_1 , v_2 , and v_3 , from which corresponding attention weights z_1 , z_2 , and z_3 are derived via a softmax function. The original feature maps are then combined through an attention-weighted summation to obtain the adaptively fused output feature map, expressed as:

$$F_{out} = z_1 \cdot F_1 + z_2 \cdot F_2 + z_3 \cdot F_3, F_{out} \in \mathbb{R}^{H \times W \times C} \quad (16)$$

where F_{out} denotes the adaptively fused output feature map.

For the URCSSD detection task, relying exclusively on high-resolution features preserves local texture but introduces sensitivity to noise, specular reflections, and suspended particulate interference. Conversely, relying exclusively on low-resolution features risks losing critical information for elongated cracks and weak-boundary corrosion regions. MLFE addresses this trade-off through coordinated direction-sensitive enhancement and cross-layer adaptive fusion: the high-resolution branch emphasizes local detail and edge information, while the mid- and low-resolution branches supplement contour, illumination, and contextual features. This cooperative design achieves a more favorable balance among detail clarity, contour integrity, and luminance consistency. Furthermore, cascading multiple MLFE modules progressively strengthens detail, contour, and illumination representation, improving the separability of crack regions from complex backgrounds and providing more reliable input features for downstream object detection.

3.2. Collaborative Feature Enhancement Network for Underwater Surface Damage Detection

3.2.1. PRFConv

In underwater RC surface structural crack detection, targets such as cracks, spalling, and rebar exposure commonly exhibit small scale, weak boundaries, strong directionality, and high susceptibility to complex background interference, resulting in weak responses in shallow feature representations. Conventional convolutions are constrained by fixed receptive fields and limited directional modeling capacity, making them insufficient for extracting the fine-grained structural information characteristic of these cracks.

To address this, a pinwheel-shaped receptive field convolution is developed in this study. PRFConv employs asymmetric convolutional paths with directional extension characteristics to enhance the representational capacity of shallow-stage features for texture, edge, and directional structural information, thereby improving the quality of foundational features for downstream URCSSD detection.

The structure of PRFConv is illustrated in Figure 6. Unlike standard convolution, PRFConv constructs horizontal and vertical convolutional kernels targeting different spatial regions of the input through asymmetric padding, allowing the convolutional response to propagate outward along multiple directions. Let the input tensor be $X^{(h_1, w_1, c_1)}$, where h_1 , w_1 , and c_1 denote the height, width, and channel number of the input feature map, respectively. To improve training stability and convergence speed, batch normalization and Sigmoid Linear Unit (SiLU) activation are applied after each convolutional operation. The first-layer parallel convolutions of PRFConv are expressed as:

$$X_1^{(h', w', c')} = \text{SiLU} \left(\text{BN} \left(X_{P(1,0,0,3)}^{(h_1, w_1, c_1)} \otimes W_1^{(1,3,c')} \right) \right) \quad (17)$$

$$X_2^{(h', w', c')} = \text{SiLU} \left(\text{BN} \left(X_{P(0,3,0,1)}^{(h_1, w_1, c_1)} \otimes W_2^{(3,1,c')} \right) \right) \quad (18)$$

$$X_3^{(h', w', c')} = \text{SiLU} \left(\text{BN} \left(X_{P(0,1,3,0)}^{(h_1, w_1, c_1)} \otimes W_3^{(1,3,c')} \right) \right) \quad (19)$$

$$X_4^{(h', w', c')} = \text{SiLU} \left(\text{BN} \left(X_{P(3,0,1,0)}^{(h_1, w_1, c_1)} \otimes W_4^{(3,1,c')} \right) \right) \quad (20)$$

where $\otimes$ denotes the convolution operation and $W_1^{(1,3,c')}$ denotes the 1×3 convolutional kernel with c' output channels. The padding parameter $P(1,0,0,3)$ specifies the number of padding pixels applied in the left, right, top, and bottom directions, respectively. After the first-layer interleaved convolutions, the relationships between the height h' , width w' , and channel c' dimensions of the output and input feature maps satisfy:

$$h' = \frac{h_1}{s} + 1, \quad w' = \frac{w_1}{s} + 1, \quad c' = \frac{c_1}{4} \quad (21)$$

where c_2 denotes the channel number of the final output feature map of the PRFConv module, and s denotes the convolution stride.

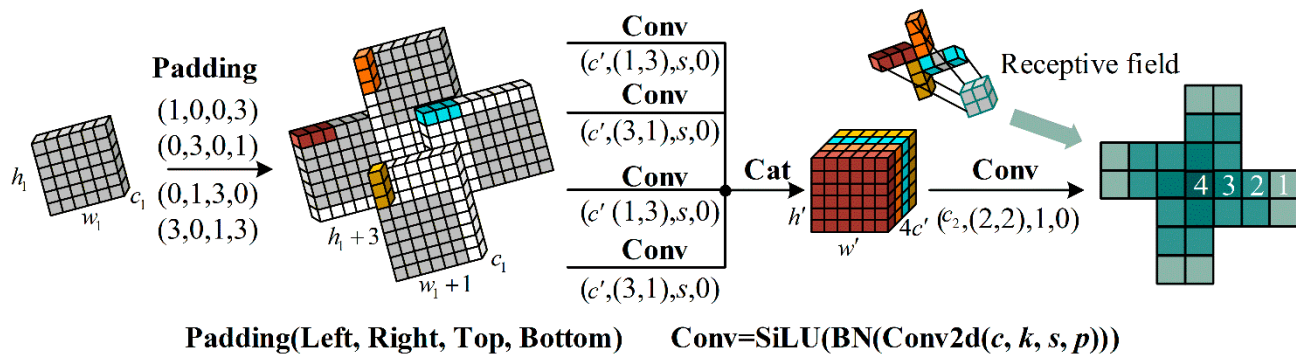


Figure 6. Architecture of PRFConv.

The outputs of the four directional branches in the first layer are then concatenated along the channel dimension to yield:

$$X^{(h', w', 4c')} = \text{Cat}(X_1^{(h', w', c')}, \dots, X_4^{(h', w', c')}) \quad (22)$$

Finally, a padding-free convolutional kernel $W^{(2,2,c_2)}$ is applied to the concatenated tensor for normalization and fusion, while simultaneously adjusting the spatial dimensions of the output feature map to the predefined values h_2 and w_2 , enabling PRFConv to serve as a direct replacement for standard convolutional layers. This process can also be interpreted as adaptive integration of the contributions from different convolutional directions. The final output $Y^{(h_2, w_2, c_2)}$ is computed as:

$$h_2 = h' - 1 = \frac{h_1}{s}, \quad w_2 = w' - 1 = \frac{w_1}{s} \quad (23)$$

$$Y^{(h_2, w_2, c_2)} = \text{SiLU}\left(\text{BN}\left(X^{(h', w', 4c')} \otimes W^{(2,2,c_2)}\right)\right) \quad (24)$$

From a mechanistic perspective, the effective receptive field decays progressively from the center outward, exhibiting a distribution profile analogous to a Gaussian response. For small-scale targets, whose informative content is typically more spatially concentrated, central region features are particularly critical. PRFConv constructs a pinwheel-shaped receptive field structure through four-directional interleaved convolutions, enabling the convolutional influence to extend outward from the center with a monotonically decreasing response pattern. This structure substantially expands the effective receptive field with a minimal increase in parameter count, while simultaneously strengthening shallow-stage feature extraction capacity. For the URCSSD detection task, elongated cracks, micro-scale rebar exposure, and small spalling regions typically manifest as low-contrast, small-scale, and strongly directional local responses. Standard convolutions in shallow stages are generally insufficient for capturing such orientation-sensitive features, whereas the asymmetric directional modeling of PRFConv more effectively enhances the representation of crack propagation directions, spalling boundary morphology, and local texture variation.

In terms of parameter count, the parameter count of a standard convolution is expressed as:

$$\text{Conv}_{\text{params}} = c_2 \times c_1 \times k, \quad (\text{bias} = \text{False}) \quad (25)$$

where k denotes the kernel size. When the output channel number c_2 equals the input channel number c_1 , the parameter count of a standard 3×3 convolution is given by $9c_1^2$. The parameter count of PRFConv is expressed as:

$$PRFConv_{params} = 4 \times ((c_2/4) \times c_1 \times 3 \times 1) + 4c_2c_1 = 7c_2c_1 = 7c_1^2 \quad (26)$$

It can be seen that PRFConv reduces the parameter count relative to standard 3×3 convolution while substantially expanding the effective receptive field. In shallow-stage feature extraction, this design provides a larger spatial receptive range at a lower parameter cost, thereby improving the contrast expressiveness between target and background regions.

3.2.2. CEMA Attention Mechanism

Following PRFConv's enhancement of shallow-stage orientation-sensitive features, the network better perceives crack propagation direction and local texture at the front end. However, underwater RC cracks still suffer from boundary ambiguity, specular reflection, and background noise, and convolution alone cannot sufficiently highlight crack regions.

To address this, this study proposes a crack-aware efficient multi-scale attention (CEMA) mechanism, consisting of a dual-pooling channel attention module and a coordinate-driven spatial attention module connected in series. The channel attention reinforces discriminative responses along the channel dimension, while the spatial attention highlights edge, orientation, and position information along the spatial dimension. Together, they address both the semantic saliency and structural integrity of crack features (Figure 7).

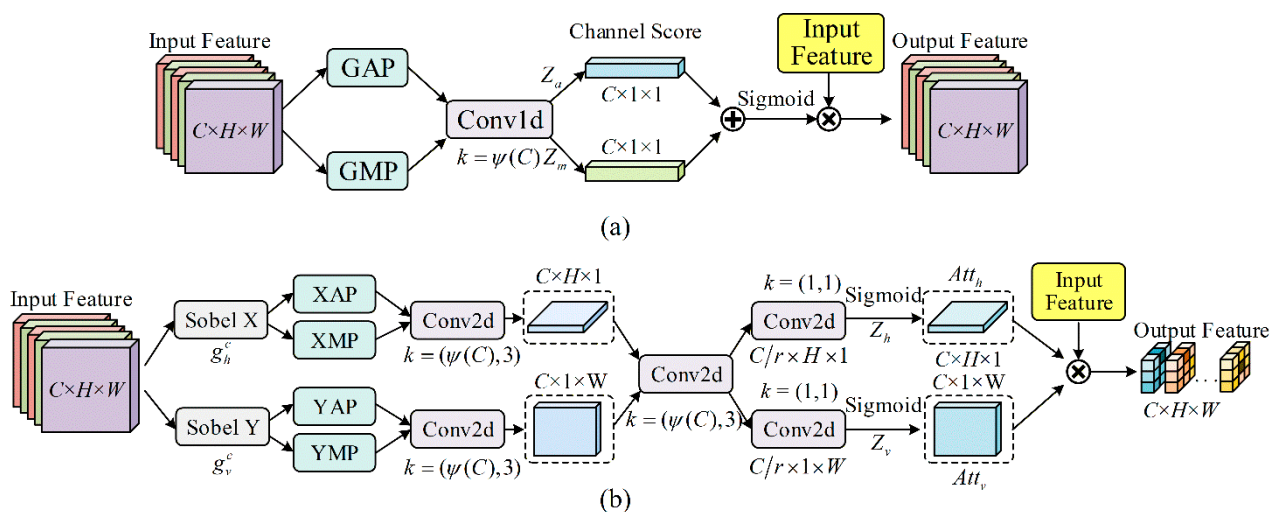


Figure 7. Architecture of the proposed CEMA attention mechanism. (a) Dual-pooling channel attention. (b) Coordinate-driven spatial attention.

In the channel stage, global average pooling and global max pooling are applied in parallel to capture overall feature distribution and locally salient activations. The two descriptors are fused and passed through a lightweight convolution to model inter-channel dependencies, producing channel attention weights that reweight the original feature map.

In the spatial stage, the channel-enhanced features are modeled along the horizontal and vertical directions separately, combined with multi-scale convolution to extract local contour and context at different receptive fields. This spatial attention draws on the Sobel

operator, a standard discrete differential operator for capturing gradient variation along the horizontal and vertical directions. The two directional 3×3 kernels are defined as:

$$S_h = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \quad (27)$$

$$S_v = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (28)$$

where S_h is used to extract gradient variation along the horizontal direction and S_v is used to extract gradient variation along the vertical direction.

Given the input feature map $X_o \in \mathbb{R}^{C \times H \times W}$, the coordinate-driven spatial attention first applies Sobel convolution separately along the horizontal and vertical directions to obtain edge-aware features. For the c^{th} channel, the feature maps in the horizontal and vertical directions are expressed as:

$$g_h^c = x_o^c * S_h \quad (29)$$

$$g_v^c = x_o^c * S_v \quad (30)$$

Building on this, average pooling and max pooling are applied independently to the feature maps of each direction to retain global statistical information and locally salient responses. For the horizontal direction feature map g_h^c , the dual-pooling results are expressed as:

$$z_{h,avg}^c = \frac{1}{W} \sum_{j=1}^W g_h^c(i, j), \quad j = 1, 2, \dots, H \quad (31)$$

$$z_{h,max}^c = \max_i g_h^c(i, j), \quad j = 1, 2, \dots, H \quad (32)$$

Similarly, for the vertical direction feature map g_v^c , the dual-pooling results are expressed as:

$$z_{v,avg}^c = \frac{1}{H} \sum_{i=1}^H g_v^c(i, j), \quad j = 1, 2, \dots, W \quad (33)$$

$$z_{v,max}^c = \max_i g_v^c(i, j), \quad j = 1, 2, \dots, W \quad (34)$$

Through these operations, four parallel directional feature branches are obtained, denoted $z_{h,avg} \in \mathbb{R}^{C \times H \times 1}$, $z_{h,max} \in \mathbb{R}^{C \times H \times 1}$, $z_{v,avg} \in \mathbb{R}^{C \times 1 \times W}$, and $z_{v,max} \in \mathbb{R}^{C \times 1 \times W}$. The average-pooled and max-pooled features within the same direction are then concatenated and passed through a convolutional layer for directional feature refinement, with convolutional parameters shared between branches of the same direction. This is expressed as:

$$Z_h = f_{conv}^{1 \times 1} \left(f_{conv}^{k \times 3} (Z_{h,avg} \parallel Z_{h,max}) \right) \quad (35)$$

$$Z_v = f_{conv}^{1 \times 1} \left(f_{conv}^{k \times 3} (Z_{v,avg} \parallel Z_{v,max}) \right) \quad (36)$$

where $Z_h \in \mathbb{R}^{C/r \times H \times 1}$ and $Z_v \in \mathbb{R}^{C/r \times 1 \times W}$ denote the horizontally and vertically refined feature maps after convolution, respectively, and $k = \left\lfloor \frac{(\log 2(C)+1)}{2} \right\rfloor$ denotes the convolutional kernel size.

Two separate 1×1 convolutional layers followed by Sigmoid activation are then applied to generate attention weights for the two directions, which are subsequently

applied to the original input feature map to produce direction-aware output features, expressed as:

$$Att_h = \sigma(f_{\text{conv}}^{1 \times 1}(Z_h)) \quad (37)$$

$$Att_v = \sigma(f_{\text{conv}}^{1 \times 1}(Z_v)) \quad (38)$$

$$Y = X_o \cdot Att_h \cdot Att_v \quad (39)$$

where Att_h and Att_v denote the attention weight maps in the horizontal and vertical directions, respectively, and Y denotes the final output feature map.

Compared with conventional spatial attention mechanisms, CEMA retains the long-range dependency modeling capability of coordinate attention while incorporating a Sobel gradient prior, making the module more sensitive to edge and texture information. This design more effectively highlights boundary-type targets such as crack propagation paths, spalling boundaries, and rebar exposure contours, while simultaneously leveraging the attention mechanism to suppress the noise amplification that may arise from direct edge enhancement.

3.3. Fine-Grained Irregular Crack IoU Loss Function

Bounding-box regression loss is critical for underwater crack detection accuracy. Cracks, spalling, and rebar exposure typically have small scale, fine granularity, large aspect ratio variation, and ambiguous boundaries. Combined with underwater image degradation and background interference, conventional IoU-based losses often produce weak gradient signals, unstable localization, and inaccurate bounding-box localization. To address this, this study proposes a fine-grained irregular crack IoU (FID-IoU) loss, which strengthens regression for hard samples and weak-boundary targets through an auxiliary bounding-box mechanism and a piecewise weighted mapping strategy.

FID-IoU modifies the conventional IoU loss to improve regression across different sample types. Beyond measuring overlap between predicted and ground truth boxes, it applies differentiated optimization across IoU intervals using auxiliary boxes and interval-based weighting, improving learning for hard-to-detect targets in complex scenes (Figure 8).

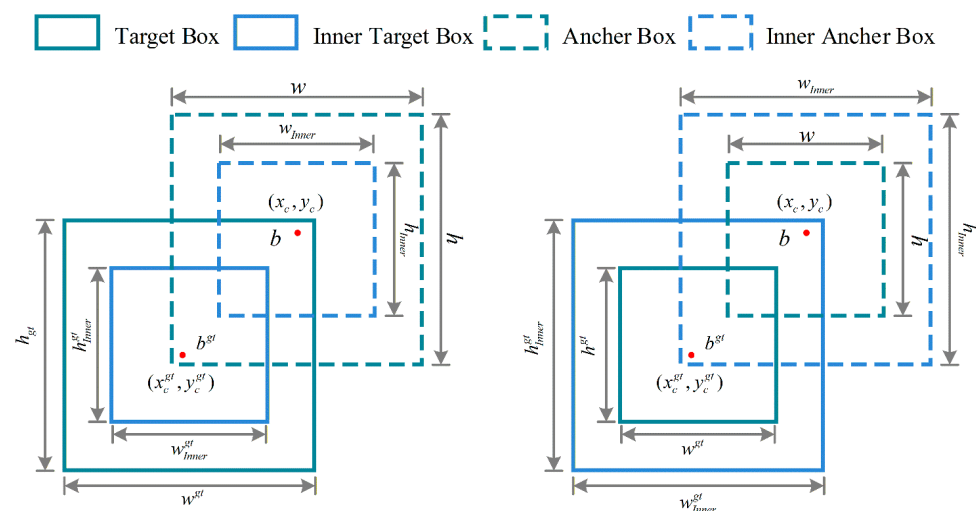


Figure 8. Auxiliary bounding-box mechanism of the proposed FID-IoU.

First, FID-IoU introduces an auxiliary bounding-box mechanism to extend the loss to both high-IoU and low-IoU samples. Using a scaling factor, auxiliary boxes are constructed

for both the ground truth and predicted box. The boundary coordinates of the ground truth auxiliary box are computed as:

$$\begin{aligned} b_{gt}^l &= x_c^{gt} - \frac{w_{gt} \cdot ratio}{2}, & b_{gt}^r &= x_c^{gt} + \frac{w_{gt} \cdot ratio}{2} \\ b_{gt}^t &= y_c^{gt} - \frac{h_{gt} \cdot ratio}{2}, & b_{gt}^b &= y_c^{gt} + \frac{h_{gt} \cdot ratio}{2} \end{aligned} \quad (40)$$

The boundary coordinates of the predicted auxiliary bounding box are computed as:

$$\begin{aligned} b^l &= x_c - \frac{w \cdot ratio}{2}, & b^r &= x_c + \frac{w \cdot ratio}{2} \\ b^t &= y_c - \frac{h \cdot ratio}{2}, & b^b &= y_c + \frac{h \cdot ratio}{2} \end{aligned} \quad (41)$$

where x_c^{gt} and y_c^{gt} denote the center coordinates of the ground truth box, w_{gt} and h_{gt} denote its width and height, and x_c , y_c denote the center coordinates of the predicted box, with w and h representing the predicted box width and height, respectively. Based on the auxiliary bounding boxes defined above, the intersection area and union area are further computed as:

$$inter_{inner} = \left(\min(b_{gt}^r, b^r) - \max(b_{gt}^l, b^l) \right) \times \left(\min(b_{gt}^b, b^b) - \max(b_{gt}^t, b^t) \right) \quad (42)$$

$$union_{inner} = \left(w_{gt} \cdot h_{gt} \cdot (ratio)^2 + w \cdot h \cdot (ratio)^2 - inter_{inner} \right) \quad (43)$$

The expression for IoU^{Inner} is thereby obtained as:

$$IoU^{Inner} = \frac{inter_{inner}}{union_{inner}} \quad (44)$$

The advantage of this auxiliary bounding-box mechanism is that even when the original overlap between the predicted and ground truth boxes is small or near zero, the auxiliary boxes still provide a more stable overlap measurement, alleviating the insufficient optimization signal that conventional IoU-based losses exhibit on low-overlap samples.

Building on this, FID-IoU further incorporates a piecewise interval mapping strategy to assign differentiated weights to samples across different IoU intervals. Let $[d, u] \subset [0, 1]$ be defined as above. When IoU^{Inner} falls below the lower bound d , the loss value is set to zero to reduce the interference of uninformative samples on the training process. When IoU^{Inner} falls within the intermediate interval $[0, 1]$, it is linearly mapped to the range $[0, 1]$. This process is expressed as:

$$FID - IoU = \frac{inter_{inner} - union_{inner} \cdot d}{union_{inner} \cdot (u - d)} \quad (45)$$

This linear mapping substantially increases the gradient weight of samples in the intermediate IoU interval, directing the model's attention toward these hard samples. When $IoU^{Inner} > u$ is satisfied, the value is mapped to 1 to reduce the dominance of already well-localized samples over the overall loss. Integrating the piecewise processing strategies described above, FID-IoU is ultimately defined as:

$$FID - IoU = \begin{cases} 0, & IoU^{Inner} < d \\ \frac{inter_{inner} - union_{inner} \cdot d}{union_{inner} \cdot (u - d)}, & d \ll IoU^{Inner} \ll u \\ 1, & IoU^{Inner} > u \end{cases} \quad (46)$$

Overall, FID-IoU enhances the adaptability of bounding-box regression to small-scale, fine-grained, and complex-shape targets through its auxiliary bounding-box mechanism, hard sample weighting strategy, and piecewise interval mapping. This loss function more

effectively improves localization stability and bounding-box localization accuracy for weak-boundary targets such as cracks, spalling, and rebar exposure, reducing missed detections and localization drift under complex background conditions, and thereby providing more reliable regression constraints for downstream detection.

3.4. Overview of the Proposed YOLOv13-TSDD Framework

To address the difficulty of simultaneously restoring image quality, enhancing shallow-detail perception, and accurately localizing complex-boundary cracks in underwater scenes, this study proposes a two-stage detection framework, YOLOv13-TSDD.

Figure 9 shows the baseline YOLOv13 pipeline, where extraction, multi-scale feature interaction, and prediction are handled by the backbone, neck, and head. For underwater RC crack detection, however, performance is limited not only by network architecture but also by image degradation, weak crack textures, and boundary regression difficulty. Figure 10 shows the task-adaptive modifications. UC DEN is added before the detection network to enhance degraded images. PRFConv is introduced at the shallow extraction stage to strengthen directional texture features such as cracks and spalling boundaries, while CEMA highlights critical crack regions. FID-IoU improves bounding-box regression for irregular cracks. The core YOLOv13 structures, HyperACE, FullPAD, and DSC3k2, are retained, preserving high-order feature association, multi-scale information flow, and lightweight extraction.

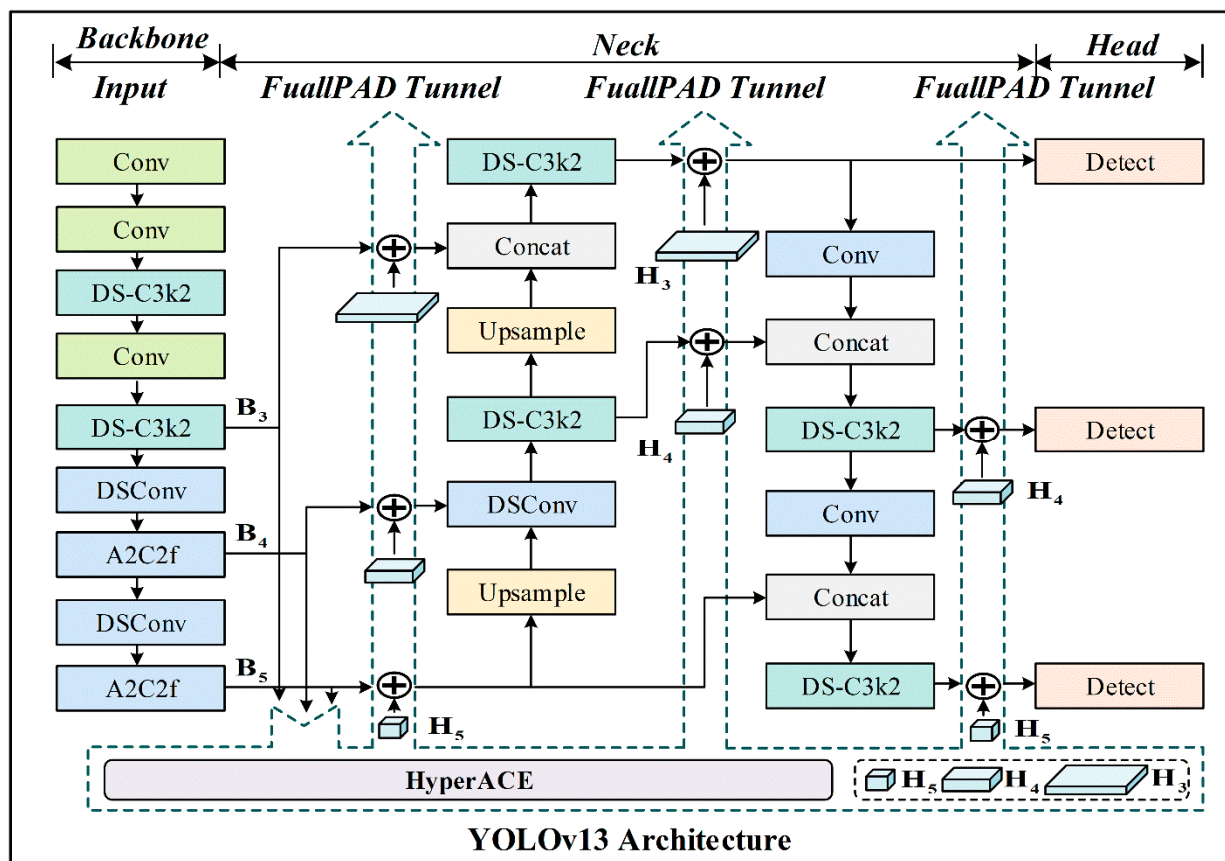


Figure 9. Architecture of YOLOv13.

The effectiveness of YOLOv13-TSDD comes from the complementarity between stages and modules, not from any single component. This is detailed below.

UCDEN as the representational foundation. UC DEN performs task-oriented enhancement calibrated to the texture, edge, and luminance CEMAnds of the detection

network, not general perceptual improvement. It corrects color bias and recovers texture and contour information in crack, spalling, and rebar exposure regions, giving the detection network cleaner, higher-contrast inputs that better reflect true crack structure. Without this step, shallow-stage feature extraction is vulnerable to scattering, low illumination, and background noise, causing downstream attention and boundary regression to operate on incomplete or distorted features.

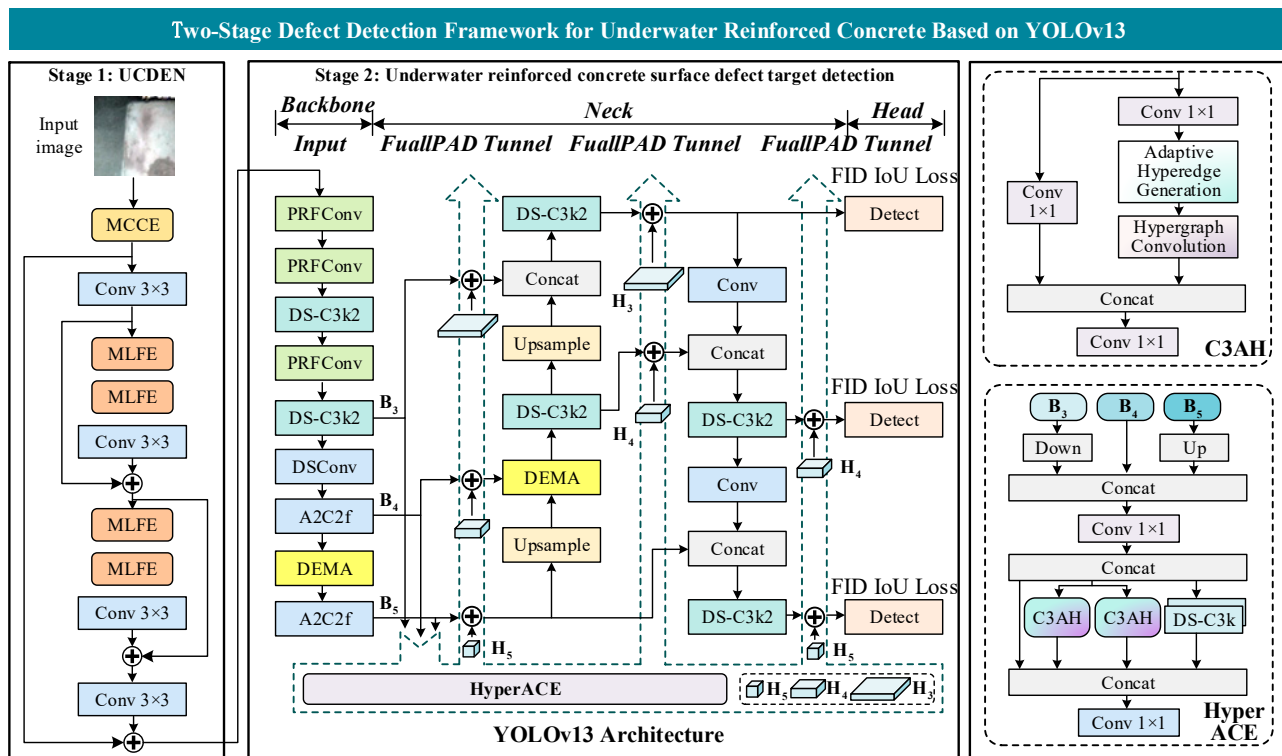


Figure 10. Architecture of the proposed YOLOv13-TSDD.

PRFConv as early structural activation. Using UC DEN's enhanced inputs, PRFConv amplifies orientation-sensitive and locally linear structural responses at the front of the detection network, activating crack propagation directions, spalling edge orientations, and rebar exposure texture variations early in the feature hierarchy. UC DEN addresses whether cracks can be seen clearly; PRFConv addresses whether their key structural signatures can be captured early. Together, they improve the fidelity of the mapping from image to feature domain.

CEMA as semantic focusing and edge localization. CEMA refines the shallow-stage responses activated by PRFConv. Channel attention preserves semantically discriminative crack responses, and Sobel gradient-prior spatial attention then sharpens edge and positional sensitivity, allowing shallow texture cues to transition into mid-to-high-level semantic representations. PRFConv and CEMA are therefore not simply additive, but a complementary pairing of front-end structural perception and back-end saliency focusing.

FID-IoU as regression consolidation. With UC DEN improving input appearance, PRFConv strengthening early structural response, and CEMA improving semantic focus and edge localization, FID-IoU works from a more reliable feature representation. Its auxiliary bounding-box mechanism and piecewise interval mapping convert this into more stable and precise bounding-box localization for irregular, complex-boundary damage. Its advantage on hard samples and weak boundaries is fully realized only when the upstream stages are working in coordination.

3.5. Summary of Proposed Modules

To systematically summarize the major modifications of YOLOv13-TSDD and further distinguish established technical foundations from the proposed innovations, Table 2 summarizes the underlying mechanisms, task-specific innovations, and functional roles of UCDEN, PRFConv, CEMA, and FID-IoU within the overall framework.

Table 2. Function and role of the proposed modules in YOLOv13-TSDD.

Module	Established Technical Basis	Proposed Innovation	Role in Framework
UCDEN	Multi-scale convolution, channel/coordinate attention, cross-scale fusion	Progressive color-detail collaborative enhancement for underwater degradation	Stage-I input enhancement
PRFConv	Asymmetric convolution, multi-branch feature extraction	Four-direction interleaved convolution with a pinwheel-shaped expanded receptive field	Shallow directional structure perception
CEMA	Channel/spatial attention, Sobel gradient operator	Dual-pooling channel attention with gradient-prior-guided coordinate spatial attention	Critical-region and edge enhancement
FID-IoU	IoU-based bounding-box regression	Joint auxiliary-box and piecewise weighted mapping optimization	Precise irregular-target regression

As summarized in Table 2, UCDEN, PRFConv, CEMA, and FID-IoU respectively address input enhancement, directional structure perception, critical-feature focusing, and bounding-box regression, forming a continuous collaborative optimization process. Stage I suppresses underwater color distortion, detail degradation, and contour blurring through UCDEN, providing more discriminative inputs for Stage II. Stage II then uses PRFConv to strengthen directional structural responses, CEMA to emphasize critical regions and edge features, and FID-IoU to optimize regression for irregular targets. Thus, Stage I improves the basis for subsequent feature learning, while Stage II further converts the enhanced visual information into structural representation and localization gains, enabling collaborative optimization from degradation suppression and feature enhancement to precise regression.

It should be clarified that all four modules are task-specific methods proposed in this study for underwater RC surface damage detection, and their overall structural designs, functional organization, and integration constitute the principal methodological innovations of this work. The convolution, attention mechanisms, Sobel gradient operator, and IoU measurement employed within these modules are established techniques and are not claimed as individual innovations.

4. Experimental Setup

4.1. Dataset Construction and Partitioning

4.1.1. Specimen Preparation and Damage Type Design

The dataset was constructed for underwater RC surface damage detection. Rectangular RC specimens with dimensions of $70 \times 20 \times 12$ cm were cast in the laboratory and cured for 28 days, after which three representative surface damage types were manually introduced: concrete spalling (CS), exposed rebar (ER), and concrete cracking (CC). Specifically, CS was produced by locally removing surface concrete to form irregular spalled regions; ER was generated by removing local concrete cover around the reinforcement to expose portions of the steel bars; and CC was produced through controlled mechanical treatment to form linear surface cracks with different propagation directions and lengths. The artificial damage

preparation was intended to provide representative visual characteristics with different scales, morphologies, and boundary patterns rather than to reproduce a prescribed level of structural damage under a specific loading condition.

4.1.2. Underwater Image Acquisition and Preprocessing

The specimens were submerged in water tanks and recorded using a smartphone-based RGB camera at a fixed distance of 15 cm. The fixed imaging device and distance were adopted to reduce non-target variations in sensor response, target scale, and imaging geometry. After acquisition, image frames were extracted according to their source video sequences, and frames without valid damage information, with severe blur, or with excessive visual redundancy were first excluded. For temporally adjacent frames exhibiting highly similar visual content, only representative images were retained to reduce redundancy caused by video continuity. The retained images were subsequently cropped and normalized to 512×512 pixels.

4.1.3. Imaging Conditions Under Different Turbidity Levels

Two turbidity levels were used: Condition 1 (≤ 10 NTU, relatively clear) and Condition 2 (20–30 NTU, more degraded), representing two controlled levels of underwater degradation within the investigated range. Turbidity was controlled by adding measured sediment. Representative images are shown in Figure 11.

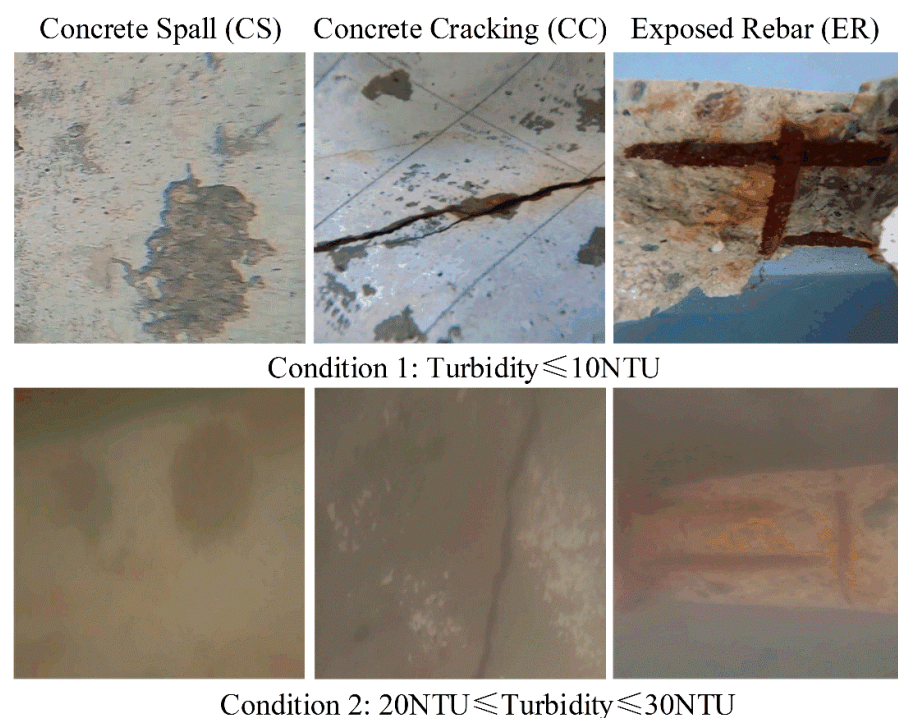


Figure 11. RC surface damage images acquired under different underwater turbidity conditions.

It should be noted that although a completely nighttime underwater imaging condition was not separately established, the collected data include low-illumination samples affected by underwater absorption, scattering, and illumination attenuation. Combined with the two turbidity ranges, the dataset therefore incorporates multiple levels of optical degradation involving turbidity variation, scattering, color attenuation, contrast reduction, blur, and low illumination, rather than representing only a single clear-water tank condition.

4.1.4. Dataset Composition and Partitioning Strategy

All images used for detector training and evaluation (Datasets C and D) were manually annotated using Labellmg (version 1.8.6), with bounding boxes and category labels stored in Pascal VOC format. Each bounding box was drawn to enclose the visible damage region as completely as possible and assigned to CS, ER, or CC. Images for which the damage category or visible boundary could not be reliably determined were excluded during data screening. Representative annotation examples for the three damage categories are shown in Figure 12.

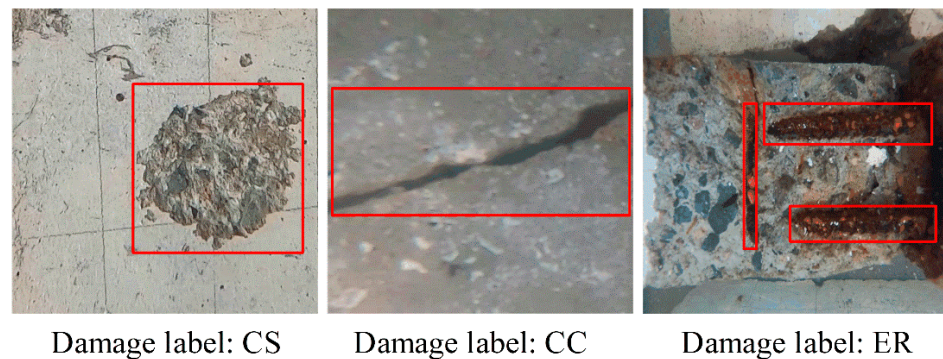


Figure 12. Annotation examples for each damage category.

To prevent training–test leakage caused by temporally correlated video frames, the source video sequence was used as the grouping unit for separating the development pool from the final test set. All retained frames from the same continuous video sequence were assigned either to the development pool or to Dataset D, ensuring that adjacent or visually similar frames from the same sequence could not cross the development–test boundary. Dataset D was first reserved at the source-sequence level as the final independent test set. The remaining development pool was then used to construct Datasets A–C. Datasets A and B are task-specific Stage-I training subsets and may overlap with Dataset C within the development pool; however, none of these training subsets overlap with Dataset D. The composition and task-specific partitioning strategy of Datasets A–D are summarized in Table 3.

Table 3. Partitioning strategy of the underwater RC surface damage dataset.

Dataset	Condition	Number of Images	CS	ER	CC	Purpose
Dataset A	Condition 1	2000	--	--	--	Training UCDEN
Dataset B	Condition 2	2000	--	--	--	Training UCDEN
Dataset C	Condition 1 + 2	5126	1712	1640	1774	Training detection framework
Dataset D	Condition 1 + 2	1241	473	412	356	Final test

In addition, after group-wise partitioning, a cross-subset near-duplicate screening was performed between Dataset D and the development pool used to construct Datasets A–C. Candidate images with high visual similarity were further manually reviewed to ensure that duplicate or near-duplicate samples originating from adjacent video frames were not retained across the training and test sets. Dataset D was therefore kept independent from model training at both the source-sequence and image levels.

4.2. Metrics Evaluation

To comprehensively assess the performance of the proposed method on both underwater image enhancement and object detection tasks, distinct evaluation protocols are adopted for each.

For image enhancement, five underwater-specific image quality metrics are employed: UCIQE, UISM, UICM, UIConM, and UIQM. UCIQE evaluates non-uniform color bias, blurriness, and image contrast, with higher values indicating better quality. UISM measures detail sharpness. UICM quantifies color fidelity. UIConM measures local contrast. UIQM is a linear combination of UICM, UISM, and UIConM, providing an integrated quality score. For detailed formulations of these metrics, the reader is referred to [53,54].

For object detection, four standard metrics are adopted: Precision, Recall, Average Precision (AP), and mean Average Precision (mAP). Precision quantifies the proportion of predicted positive samples that are true positives, reflecting detection exactness. Recall quantifies the proportion of actual crack instances successfully detected by the model, reflecting detection completeness. AP characterizes single-class detection performance across different recall levels as the area under the Precision–Recall curve. mAP averages AP across all categories to provide an overall measure of multi-class detection performance. In addition, mAP@0.50–0.95 is introduced to further evaluate bounding-box regression performance under stricter localization requirements. This metric averages AP over IoU thresholds ranging from 0.50 to 0.95 with an interval of 0.05. Compared with mAP@0.50, it imposes progressively stricter overlap requirements between predicted and ground-truth boxes and therefore provides a more comprehensive assessment of localization accuracy and bounding-box regression quality.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (47)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (48)$$

$$\text{mAP} = \frac{1}{N} \sum_{c=1}^N \int_0^1 P(R) dR \quad (49)$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, FN denotes false negatives, and N denotes the number of categories.

4.3. Experimental Environment and Training Configuration

The experiments were conducted on a workstation equipped with an Intel Core i7-14650HX CPU (Intel Corporation, Santa Clara, CA, USA) and an NVIDIA GeForce RTX 4060 GPU (16 GB VRAM, NVIDIA Corporation, Santa Clara, CA, USA), running Windows 11 with Python 3.11, PyTorch 2.2.2, and CUDA 12.1. For detection model training, the initial learning rate was set to 0.01, with 200 training epochs, a batch size of 32, a momentum of 0.825, and a weight decay coefficient of 0.002. Model parameters were optimized using stochastic gradient descent (SGD). The internal structural parameter configurations of the MFE and CRB modules within the proposed UCDEN are provided in Table 4.

Table 4. Internal structural parameter configuration of the MFE module.

Layer Name	Kernel Size	Output Shape
1×1 Conv, LReLU	[1, 1, 1]	$64 \times H \times W$
3×3 Conv, LReLU	[3, 3, 1]	$64 \times H \times W$
5×5 Conv, LReLU	[5, 5, 2]	$64 \times H \times W$
MAP	--	$64 \times H \times W$
Concatenation	--	$256 \times H \times W$

The internal structure of the CRB module follows the coordinate residual block design of Lin et al. [55]. Briefly, the input feature map passes through two successive 3×3 convolutions with LReLU activation, followed by coordinate attention: horizontal and vertical

average pooling (XAP, YAP) generate direction-aware descriptors, which are concatenated and reduced via a 1×1 convolution to eight channels before being split into two branches. Each branch is restored to the original channel dimension through a separate 1×1 convolution with Sigmoid activation, producing horizontal and vertical attention weights. These weights are applied multiplicatively to the input features, and the result is combined with the original input through a residual addition. Full layer-wise kernel sizes and output dimensions are given in the original reference.

A single training schedule (200 epochs, batch size 32, initial learning rate 0.01, momentum 0.825, weight decay 0.002, SGD optimizer) was applied uniformly to all six compared methods to isolate the effect of architectural and module-level differences from confounding variation in optimizer tuning. No architecture-specific hyperparameter search, pretrained-weight initialization control, or parameter-count matching was performed across methods. This protocol therefore evaluates relative performance under a common, fixed training budget rather than each method's individually optimized performance ceiling.

In addition, to reduce the influence of random initialization and training variability, each method was independently trained five times using the same data partition and training configuration. Precision, Recall, AP, and mAP are reported as the mean $\pm$ standard deviation over five independent runs.

5. Experimental Results and Analysis

5.1. Baseline and Comparison Methods

To systematically evaluate the effectiveness of YOLOv13-TSDD, the comparison methods were selected from two complementary categories: underwater-specific detection frameworks and general object detection variants, as summarized in Table 5.

Table 5. Module descriptions and improvement configurations of comparison methods.

Methods	Image Enhancement Method	Improvement Mechanism	IoU Loss	Unified Training Settings
YOLOv5s-CycleGAN [8]	CycleGAN	Ghost lightweight, SimAM, SiLU	C-IoU [50]	epoch = 200, image size = 640, batch size = 32, SGD, initial learning rate = 0.01, momentum = 0.825, weight decay = 0.002
YOLOv5s-MSRCP [56]	MSRCP	Retinex pre-enhancement	G-IoU	
YOLOv13-Baseline	--	--	Default	
YOLOv12-CBAM	--	CBAM	Default	
YOLOv8-ECA	--	ECA	Default	
YOLOv13-TSDD	UCDEN	PRFConv, CEMA, FID-IoU	FID-IoU	

Among them, YOLOv5s-CycleGAN [8] first enhances underwater images degraded by scattering, absorption, and turbidity through an improved CycleGAN and image fusion mechanism, and then employs an improved YOLOv5s detector for concrete spalling, exposed rebar, and cracking. YOLOv5s-MSRCP [56], targeting apparent defects in underwater concrete structures, combines MSRCP-based underwater image enhancement with YOLOv5s detection to improve low-quality underwater images and achieve localization and recognition of cracks, spalling, and exposed rebar.

To further examine the influence of different detection backbones and attention mechanisms, three single-stage comparison methods are included: YOLOv13-Baseline, YOLOv12-CBAM, and YOLOv8-ECA. YOLOv13-Baseline is the unmodified YOLOv13 model and

serves as the primary reference for quantifying the contribution of the proposed modifications. YOLOv12-CBAM integrates CBAM attention into the detection network to enhance channel and spatial representation of critical crack regions. YOLOv8-ECA incorporates ECA attention to assess the effect of lightweight channel attention on crack feature enhancement. Together, these comparisons enable performance evaluation across multiple dimensions, including baseline architecture, classical attention design, and lightweight feature enhancement. All methods were evaluated using the same data partition, input size, number of training epochs, batch size, optimizer, and major training hyperparameters, together with the same test set and evaluation metrics, to ensure comparable experimental conditions as far as possible.

5.2. Experimental Design

To systematically evaluate YOLOv13-TSDD on the URCSSD detection task, the experimental protocol is organized around three complementary levels of analysis: comparative validation, visualization analysis, and ablation evaluation. First, comprehensive performance comparisons are conducted across methods representing different image enhancement strategies, feature enhancement mechanisms, and bounding-box regression approaches. Second, a systematic analysis of training convergence behavior, image enhancement quality, and crack detection performance is presented for each method on the URCSSD dataset. Third, ablation experiments are conducted to assess the independent contribution and cooperative effect of each key module: UCDEN, PRFConv, CEMA, and FID-IoU, covering image enhancement strategy, feature enhancement mechanism, and loss and activation function combinations. This three-level design provides a comprehensive evaluation of the validity and internal consistency of the proposed two-stage framework.

5.3. Comparative Experimental Results and Analysis

5.3.1. Training Convergence Analysis

To evaluate the optimization capacity and training stability of each method on the URCSSD detection task, a comprehensive convergence analysis is performed by integrating the training curves in Figure 13, the final performance metrics in Table 6, and the confusion matrices in Figure 14.

As shown in Figure 13, all methods exhibit rapid performance improvement and loss reduction in the early training phase, indicating that each model can extract meaningful crack discriminative information from the training data. However, as training progresses, the methods diverge noticeably in convergence speed, curve smoothness, and final plateau values. The proposed YOLOv13-TSDD consistently achieves the highest mAP@0.50 curve throughout training and enters a stable convergence stage earlier than competing methods, demonstrating a higher detection ceiling and superior optimization stability. In contrast, YOLOv5s-MSRCP converges most slowly and reaches the lowest final plateau, indicating that conventional MSRCP pre-enhancement alone is insufficient to sustain improvements in downstream detection discriminability. YOLOv13-Baseline, YOLOv12-CBAM, and YOLOv8-ECA maintain relatively stable convergence trajectories but achieve final accuracy values noticeably below YOLOv13-TSDD, confirming that generic backbone architectures and standard attention mechanisms provide limited adaptation to the weak-texture, low-contrast, and complex-boundary characteristics of underwater cracks. YOLOv5s-CycleGAN outperforms most single-stage methods as a two-stage approach, validating the general effectiveness of the enhancement-detection pipeline for this task, though its final convergence level remains below that of the proposed method.

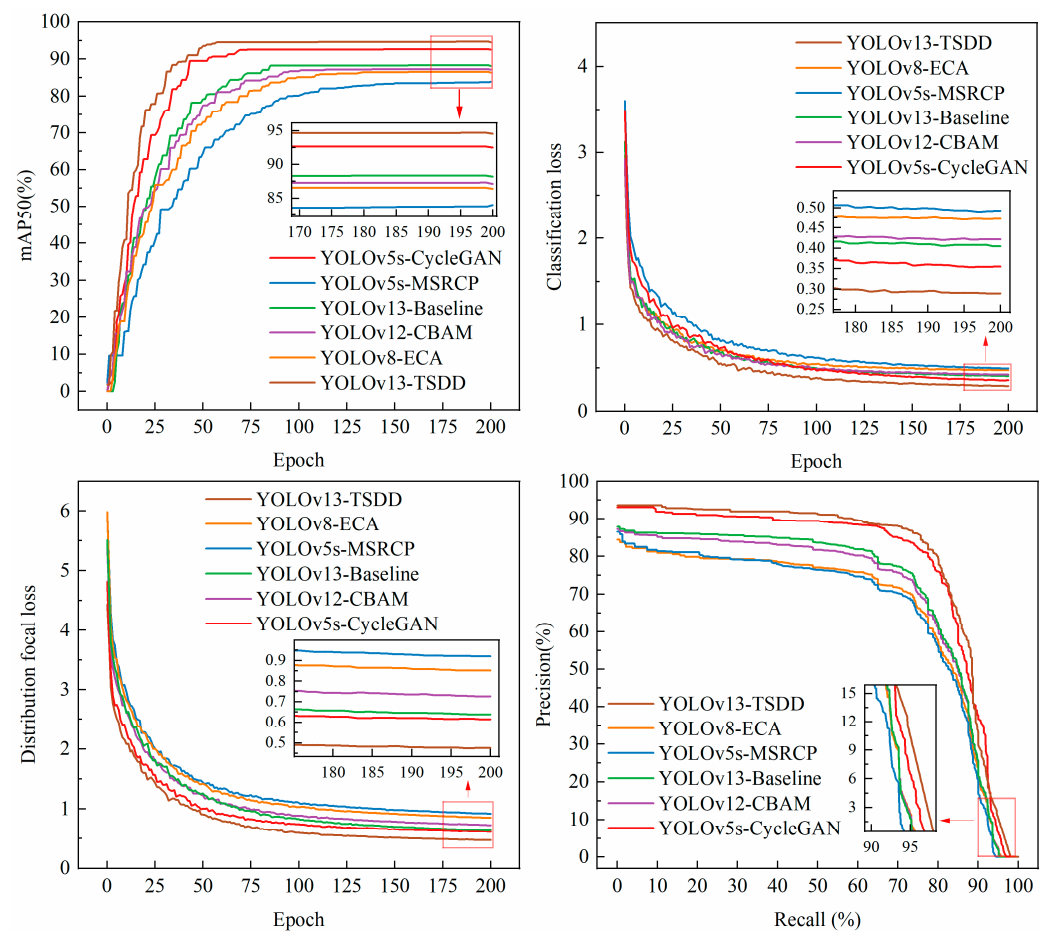


Figure 13. Training convergence performance of different object detection methods.

Table 6. Final detection performance of different methods on the URCSSD task.

Methods	Precision (%)	Recall (%)	mAP@0.50 (%)
YOLOv5s-CycleGAN	93.10 ± 0.38	86.70 ± 0.52	92.51 ± 0.31
YOLOv5s-MSRCP	83.70 ± 0.61	75.20 ± 0.74	84.00 ± 0.55
YOLOv13-Baseline	87.15 ± 0.46	82.73 ± 0.59	88.20 ± 0.42
YOLOv12-CBAM	86.56 ± 0.49	81.58 ± 0.63	87.15 ± 0.45
YOLOv8-ECA	85.82 ± 0.53	80.65 ± 0.68	86.41 ± 0.48
YOLOv13-TSDD	93.63 ± 0.27	88.97 ± 0.36	94.52 ± 0.24

The classification loss (Cfs loss) and distribution focal loss (Dfl loss) curves further corroborate these observations: YOLOv13-TSDD exhibits faster descent rates and lower final stable values for both loss terms, indicating stronger optimization capacity in both category discrimination and bounding-box regression. This behavior is attributable to the cooperative design of the proposed framework: UCDEN mitigates color bias, blur, and local detail degradation at the input stage, providing the detection network with more discriminative visual representations; PRFConv enhances shallow-stage sensitivity to orientation-sensitive texture and local linear structures; CEMA reinforces joint channel-spatial focusing on critical crack regions; and FID-IoU improves regression stability for fine-grained, irregular, and weak-boundary targets. The convergence advantages of the proposed method therefore reflect the cumulative and synergistic gains of input enhancement, feature extraction, attention focusing, and regression optimization operating in coordination.

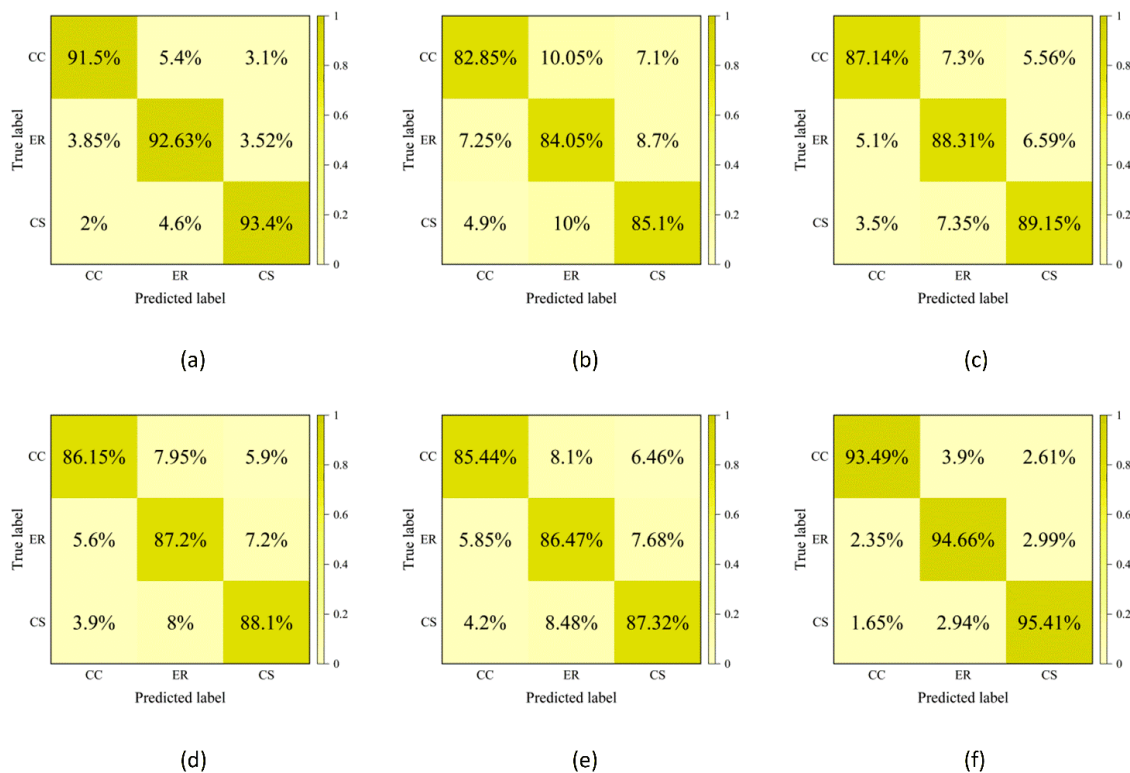


Figure 14. Confusion matrices of different detection methods. (a) YOLOv5s-CycleGAN. (b) YOLOv5s-MSRCP. (c) YOLOv13-Baseline. (d) YOLOv12-CBAM. (e) YOLOv8-ECA. (f) YOLOv13-TSDD.

As shown in Table 6, YOLOv13-TSDD achieves the best results across all comparison methods, attaining Precision, Recall, and mAP@0.50 of 93.63%, 88.97%, and 94.52%, respectively. Furthermore, the standard deviations of Precision, Recall, and mAP@0.50 for YOLOv13-TSDD across five independent runs are only 0.27, 0.36, and 0.24 percentage points, respectively, indicating low performance variability and good statistical stability across independent training runs. Relative to YOLOv13-Baseline, the proposed method improves Precision, Recall, and mAP@0.50 by 6.48, 6.24, and 6.32 percentage points, respectively, demonstrating that the two-stage collaborative modifications substantially enhance detection performance over the unmodified baseline. Compared with YOLOv12-CBAM and YOLOv8-ECA, YOLOv13-TSDD again achieves higher detection accuracy and lower loss values, confirming that the task-specific designs of PRFConv and CEMA are more effective for underwater crack feature representation than general-purpose CBAM or ECA attention mechanisms. Further comparison with the two underwater-specific methods shows that YOLOv13-TSDD improves Precision, Recall, and mAP@0.50 by 0.53, 2.27, and 2.01 percentage points over YOLOv5s-CycleGAN, respectively, and by 9.93, 13.77, and 10.52 percentage points over YOLOv5s-MSRCP. These results indicate that the advantage of the proposed method is not limited to general YOLO and attention-based variants, but is also maintained against existing detection frameworks specifically designed for underwater degradation conditions.

The confusion matrices in Figure 14 provide additional evidence: YOLOv13-TSDD achieves the highest diagonal occupancy and lowest off-diagonal misclassification rates across all three crack categories, demonstrating superior discriminative capacity for CS, ER, and CC targets. The remaining comparison methods exhibit varying degrees of inter-class confusion, particularly for samples with similar textures, blurred boundaries, or strong underwater scattering interference.

5.3.2. Image Enhancement Performance Analysis

To evaluate the input quality improvement capability of each comparison method for the URCSSD detection task, this section presents a comprehensive analysis of image enhancement effectiveness based on the visual results in Figures 15–17, the enhancement performance comparison in Figure 18, and the quantitative evaluation in Table 7.

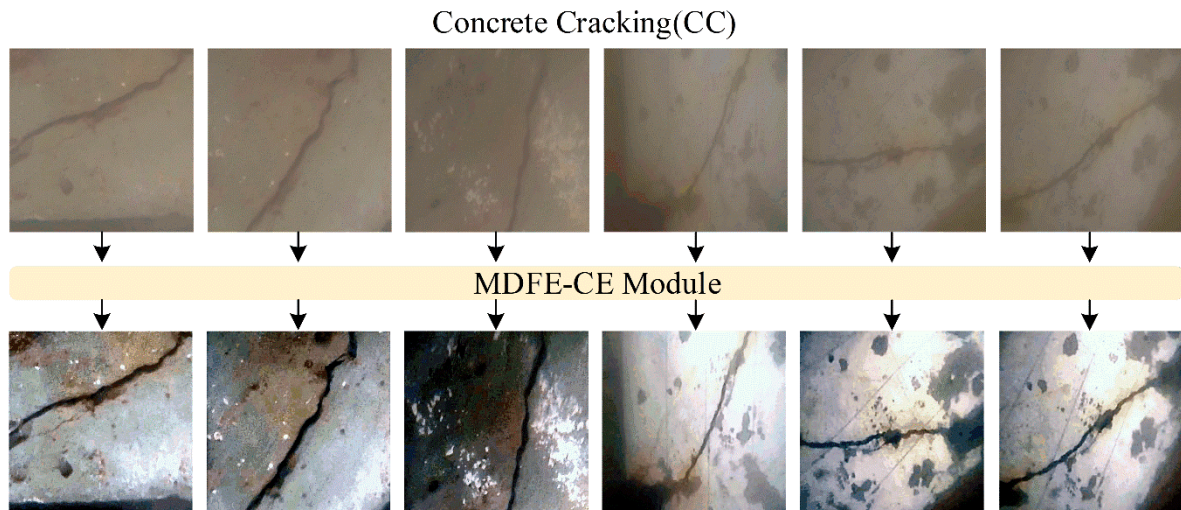


Figure 15. Visual enhancement results of underwater concrete crack images processed by UC DEN.

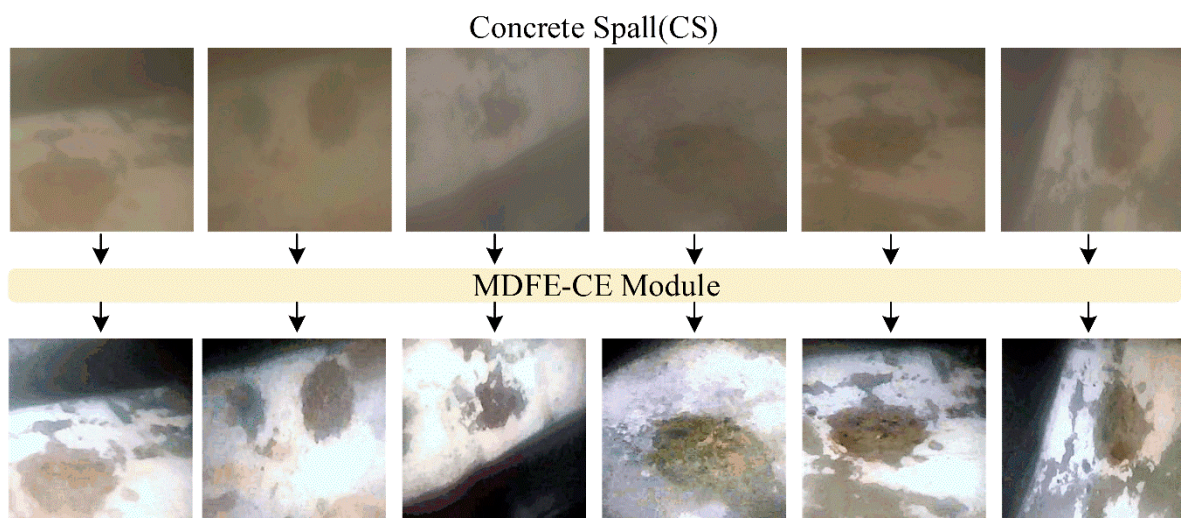


Figure 16. Visual enhancement results of underwater concrete spalling images processed by UC DEN.

It should be emphasized that the role of image enhancement in this task extends beyond subjective perceptual improvement. The more critical consideration is whether enhancement can effectively strengthen texture detail, edge continuity, local contrast, and color stability in crack regions, including cracks, spalling, and rebar exposure, thereby providing more discriminative inputs for the downstream detection network.

As shown in Figures 15–17, the proposed YOLOv13-TSDD consistently demonstrates effective task-oriented enhancement across all three crack categories. For crack images, UC DEN substantially improves the continuity of elongated linear structures, producing clearer grayscale transitions between crack edges and surrounding backgrounds. For spalling images, the enhanced outputs preserve the irregular boundaries and internal texture hierarchy of crack regions, resulting in more complete crack contours. For rebar exposure images, the boundary between steel and concrete substrate becomes more distinct

after enhancement, with rust coloration contrast and strip-like structural features effectively highlighted. These results confirm that the proposed method does not rely on global brightness amplification or contrast stretching alone, but instead performs targeted enhancement of the edge, texture, and structural information that crack detection depends upon.

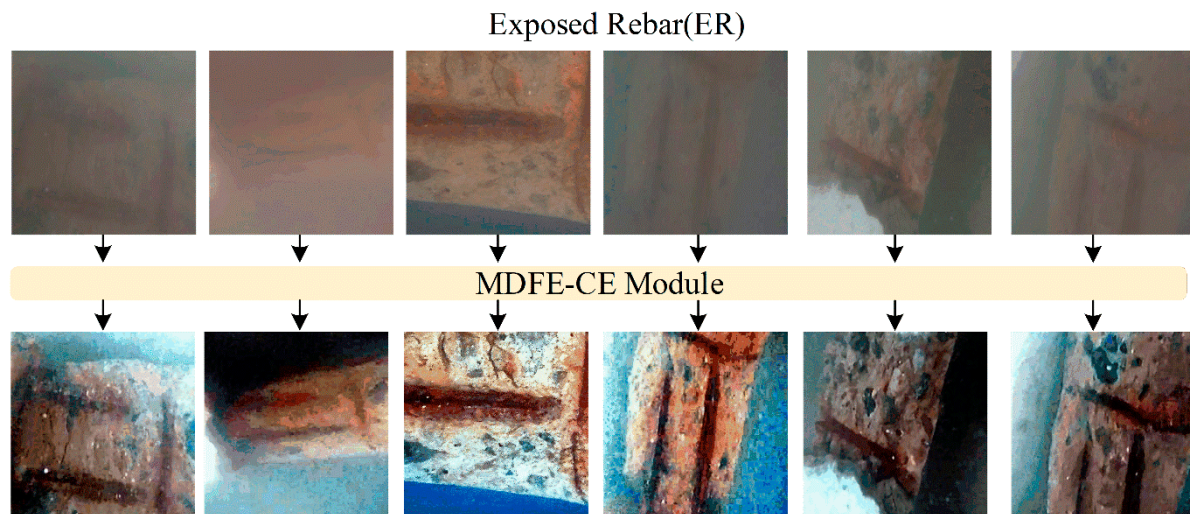


Figure 17. Visual enhancement results of underwater exposed rebar images processed by UCDEN.

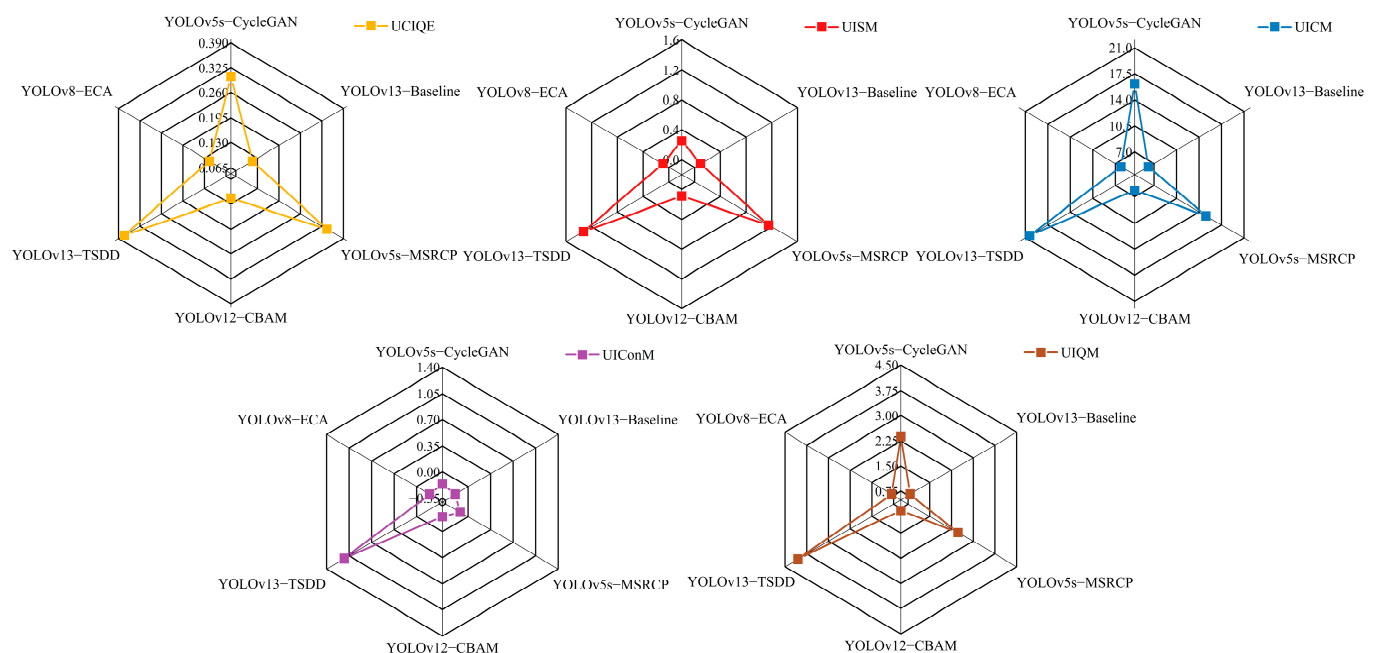


Figure 18. Image enhancement performance of different object detection methods.

Table 7. Quantitative image enhancement evaluation results for different object detection methods.

Methods	Image Enhancement Method	UCIQE	UISM	UICM	UICoM	UIQM
YOLOv5s-CycleGAN	CycleGAN	0.302	0.249	16.114	−0.168	2.378
YOLOv13-Baseline	-	0.115	0.092	6.127	−0.201	0.821
YOLOv5s-MSRCP	MSRCP	0.34	1.15	15.065	−0.123	2.473
YOLOv12-CBAM	-	0.115	0.092	6.127	−0.201	0.821
YOLOv8-ECA	-	0.115	0.092	6.127	−0.201	0.821
YOLOv13-TSDD	UCDEN	0.322	1.327	17.325	−0.125	2.855

As shown in Figure 18, the differences in enhancement effectiveness among the comparison methods fundamentally reflect their varying degrees of adaptation to underwater imaging degradation characteristics and detection task requirements. YOLOv5s-CycleGAN exhibits advantages in UCIQE and UICM, indicating a degree of improvement in overall color restoration and visual naturalness. However, its UISM of only 0.249 reveals limited recovery of detail sharpness and high-frequency structural information, suggesting that CycleGAN-based enhancement is predominantly oriented toward global visual distribution transfer, improving general image clarity while insufficiently reinforcing the local structural detail required for fine-grained weak-texture targets such as cracks. YOLOv5s-MSRCP achieves stronger performance on UCIQE, UISM, and UIQM, indicating effective brightening and sharpening of low-illumination and low-contrast images. However, its enhancement is fundamentally based on global luminance compensation and dynamic range compression, which tends to introduce localized over-enhancement and luminance imbalance, limiting its ability to stably preserve true crack boundaries and texture structures.

By contrast, YOLOv13-Baseline, YOLOv12-CBAM, and YOLOv8-ECA, which incorporate no front-end image enhancement module, produce nearly identical image quality scores. This confirms that improvements confined to the detection backbone or attention mechanism alone cannot directly address the input quality degradation inherent in underwater concrete crack imagery. For this task, front-end image enhancement is not an optional preprocessing step but a foundational component that determines the quality of subsequent feature learning. Without effective mitigation of color bias, blur, and low contrast at the input stage, downstream attention mechanisms operate on already-degraded feature representations, fundamentally limiting their ability to improve crack discriminability. The superior overall enhancement performance of YOLOv13-TSDD is attributable to the coordinated optimization of color restoration, detail enhancement, and structural fidelity achieved through UC DEN. MCCE performs multi-channel joint correction of non-uniform color bias and color distortion in underwater imagery, while MLFE reinforces local texture, edge contour, and luminance hierarchy through multi-level feature enhancement and cross-scale fusion. The advantage of the proposed method therefore does not reside in the isolated improvement of any single metric, but in balanced optimization across multiple key indicators.

As shown in Table 7, YOLOv13-TSDD achieves the best results in terms of UISM, UICM, and UIQM, with values of 1.327, 17.325, and 2.855, respectively. For UIConM, YOLOv13-TSDD attains -0.125 , marginally below YOLOv5s-MSRCP's -0.123 , indicating comparable local contrast performance between the two methods rather than a clear advantage for either. Meanwhile, its UCIQE reaches 0.322, remaining within a favorable and reasonable range without evident over-enhancement. These results indicate that the proposed method achieves the best overall balance among color naturalness, detail sharpness, and image quality, while remaining competitive rather than uniformly superior across every individual metric.

5.3.3. Damage Recognition Performance Analysis

To further evaluate the practical recognition capability of different methods for underwater reinforced concrete surface structural crack detection, this study conducts a comprehensive analysis of crack recognition performance based on visual detection results, per-category AP comparisons, and quantitative metrics. Since damage types such as CS, ER, and CC exhibit significant differences in scale distribution, morphological structure, and texture characteristics, the results in this section not only reflect the overall detection performance of each model but also reveal the adaptability of different methods to complex underwater crack features.

As shown in Figure 19 and Table 8, the proposed YOLOv13-TSDD achieves the best detection performance across all three crack categories, with AP@0.50 of 95.41%, 94.66%, and 93.49% for CS, ER, and CC, respectively, and an overall mAP@0.50 of 94.52%. Notably, YOLOv13-TSDD maintains consistently high AP@0.50 across all three damage categories, indicating that its performance advantage is not dominated by a single category but is consistently maintained across different damage morphologies. Meanwhile, the standard deviations of AP@0.50 for CS, ER, and CC are 0.22, 0.26, and 0.31 percentage points, respectively, indicating that the performance advantage of YOLOv13-TSDD remains stable across different damage categories and independent training runs. Compared with the underwater-specific YOLOv5s-CycleGAN, the proposed method improves AP@0.50 by 2.01, 2.03, and 1.99 percentage points for CS, ER, and CC, respectively, indicating that its performance gain is consistently maintained across different types of underwater RC surface damage rather than being dominated by a single category.

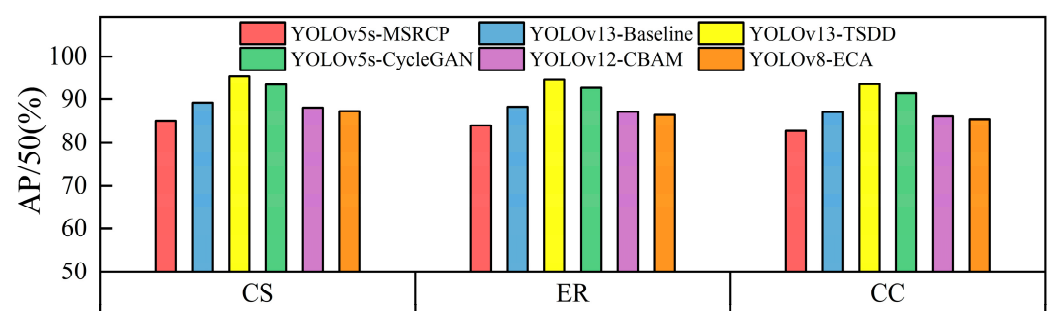


Figure 19. Detection performance of different methods across damage categories.

Table 8. Per-category detection performance of different methods.

Methods	AP@0.50(%)			mAP@0.50(%)
	CS	ER	CC	
YOLOv5s-CycleGAN	93.40 ± 0.32	92.63 ± 0.38	91.50 ± 0.45	92.51 ± 0.31
YOLOv5s-MSRCP	85.10 ± 0.57	84.05 ± 0.63	82.85 ± 0.71	84.00 ± 0.55
YOLOv13-Baseline	89.15 ± 0.41	88.31 ± 0.47	87.14 ± 0.54	88.20 ± 0.42
YOLOv12-CBAM	88.10 ± 0.44	87.20 ± 0.50	86.15 ± 0.57	87.15 ± 0.45
YOLOv8-ECA	87.32 ± 0.47	86.47 ± 0.54	85.44 ± 0.61	86.41 ± 0.48
YOLOv13-TSDD	95.41 ± 0.22	94.66 ± 0.26	93.49 ± 0.31	94.52 ± 0.24

Across all methods, a consistent difficulty ranking emerges: CS detection is relatively straightforward, CC presents the greatest challenge, and ER falls between the two. This pattern reflects the morphological characteristics of each crack type: spalling regions are typically large and visually salient, making them easier for models to detect. Crack targets, characterized by weak boundaries, small scale, and fragile continuity under background interference, are most susceptible to missed detection and false alarms. Rebar exposure, while possessing distinctive structural features, remains difficult to recognize under underwater blur and color bias.

As shown in Figure 20, differences among methods in underwater crack recognition primarily manifest in target response completeness, localization precision, and category discrimination stability. Overall, YOLOv13-TSDD achieves the most complete target coverage, highest detection confidence, and more accurate bounding-box localization across all three damage categories. This advantage is particularly evident in weakly contrasted cracks, locally blurred spalling regions, and multi-instance rebar exposure scenarios, where the proposed method maintains strong detection continuity and robustness. YOLOv5s-CycleGAN and YOLOv5s-MSRCP achieve better overall detection performance than most

single-stage methods, owing to the input quality improvement provided by their respective front-end enhancement modules. However, both methods remain inferior to YOLOv13-TSDD in bounding-box localization accuracy and complete local region response. Notably, YOLOv13-Baseline, YOLOv12-CBAM, and YOLOv8-ECA, despite incorporating attention-based feature enhancement in the detection stage, exhibit substantially degraded detection performance under highly turbid conditions due to the absence of front-end image enhancement, further confirming the indispensable role of input-stage preprocessing in this task.

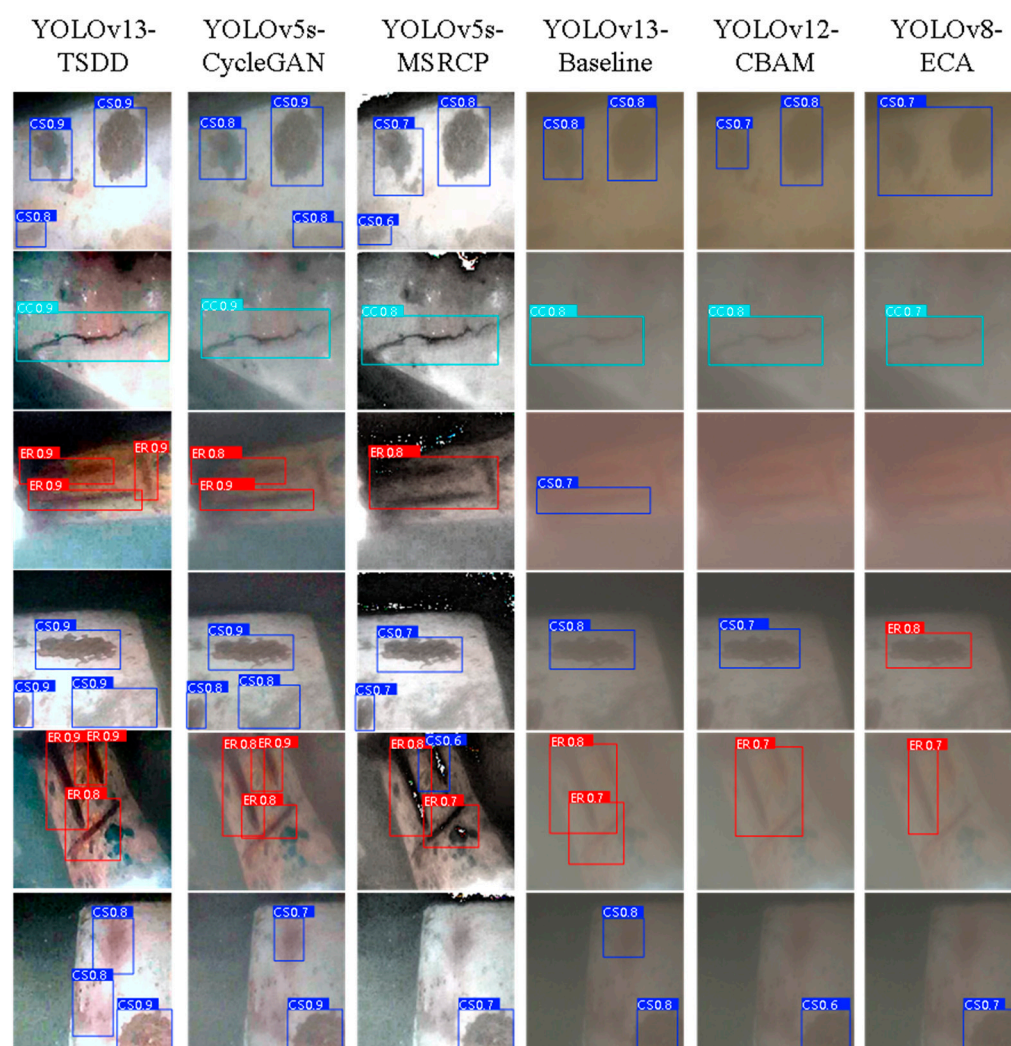


Figure 20. Visual detection results of different detection methods.

5.4. Ablation Study and Module Contribution Analysis

5.4.1. Ablation Study on Image Enhancement Strategy

To assess the effect of front-end image enhancement strategy on the overall detection performance of the proposed framework, four enhancement configurations are evaluated while keeping the back-end feature enhancement structure fixed: no enhancement, MSRCP enhancement, CycleGAN enhancement, and the proposed UC DEN. Results are analyzed with reference to Figures 21 and 22 and Table 9.

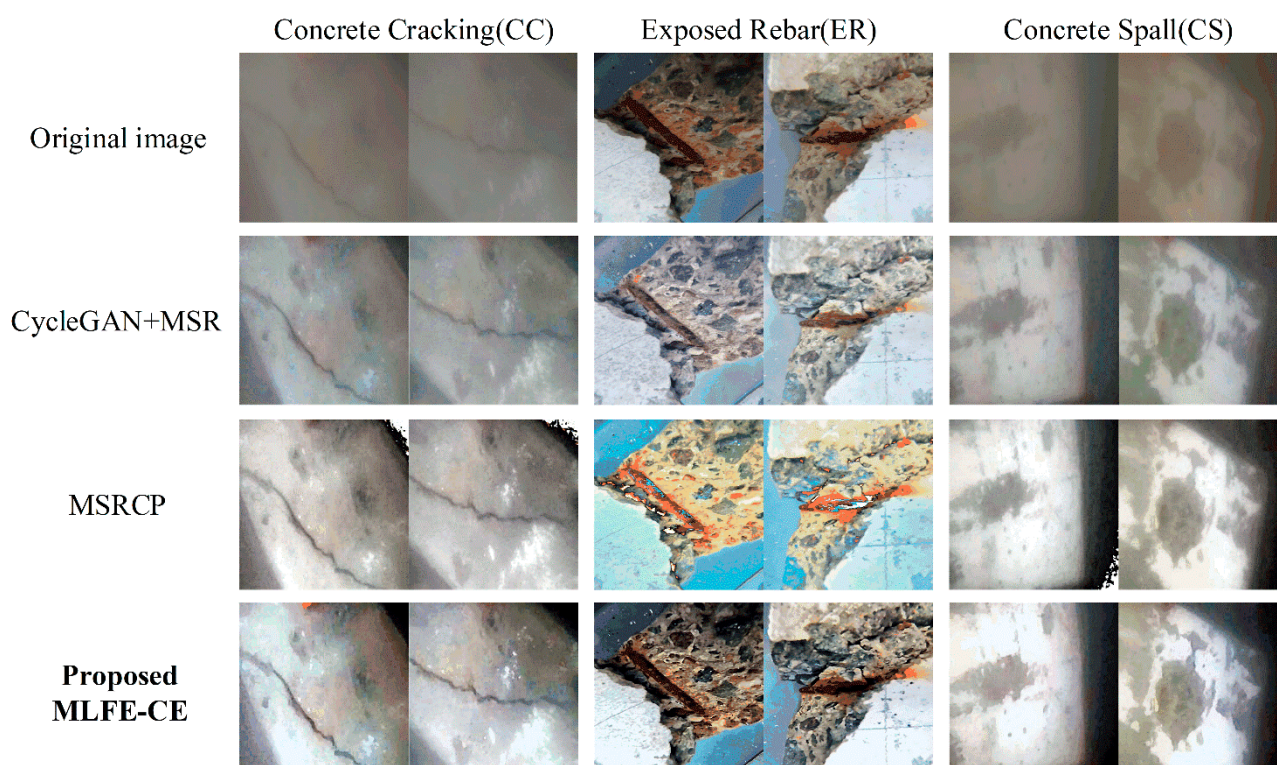


Figure 21. Visual comparison of different image enhancement strategies.

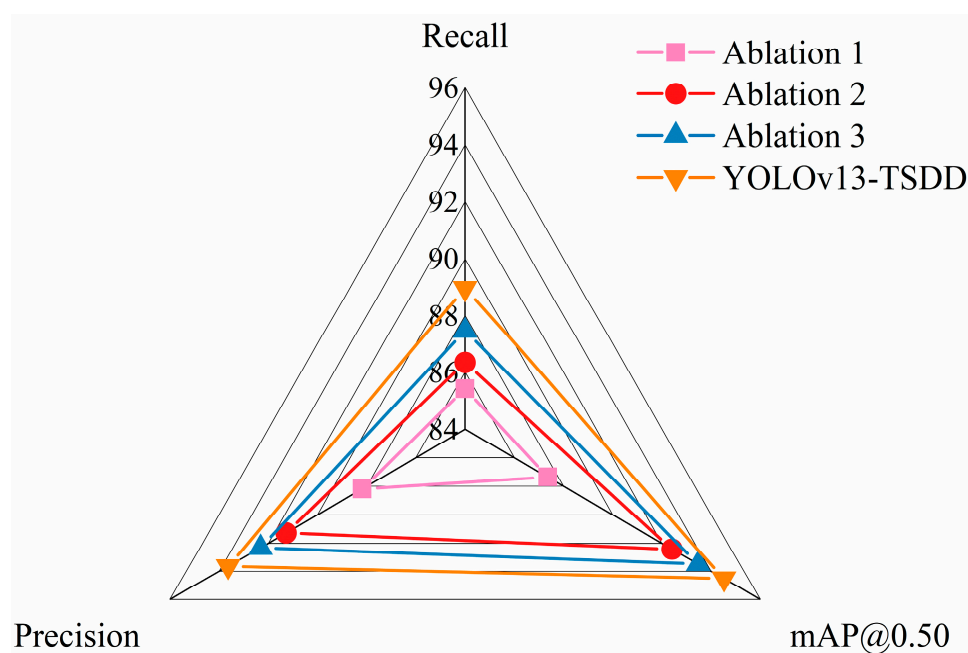


Figure 22. Graphical comparison of precision, recall, and mAP@0.50 for YOLOv13-TSDD under different image enhancement strategies.

Table 9. Detection performance of YOLOv13-TSDD under different image enhancement strategies.

Ablation Method	Feature Enhancement	Image Enhancement	Precision (%)	Recall (%)	mAP@0.50 (%)
Ablation 1	PRFConv + CEMA	--	88.18	85.42	87.36
Ablation 2		MSRCP	91.27	86.34	92.41
Ablation 3		CycleGAN	92.31	87.46	93.47
YOLOv13-TSDD		UCDEN	93.63	88.97	94.52

As shown in Figure 21, raw images consistently exhibit low contrast, color bias, texture blur, and boundary ambiguity, all of which weaken the separability of crack, spalling, and rebar exposure regions from the background. MSRCP effectively improves overall brightness and local contrast, making crack regions more visually prominent, but tends to introduce localized over-enhancement and luminance imbalance. CycleGAN achieves superior visual naturalness and color restoration relative to MSRCP, more effectively mitigating distortion caused by underwater degradation, but remains limited in reinforcing the continuity of fine-grained linear structures such as cracks. In contrast, UC DEN demonstrates more balanced enhancement across all three damage categories: it strengthens the elongated continuous texture of cracks and the strip-like structural characteristics of rebar exposure while more completely preserving the irregular contours and internal texture hierarchy of spalling regions, confirming its superior capacity to simultaneously achieve color restoration, detail reinforcement, and structural preservation.

As shown in Figure 22, without any image enhancement, Precision, Recall, and mAP@0.50 are limited to 88.18%, 85.42%, and 87.36%, respectively, demonstrating that even with PRFConv and CEMA incorporated in the back end, unmitigated input degradation substantially constrains detection performance.

As can be further seen from Table 9, applying MSRCP enhancement raises these metrics to 91.27%, 86.34%, and 92.41%, confirming that traditional Retinex-based enhancement can partially improve image visibility and target saliency. CycleGAN further improves performance to 92.31%, 87.46%, and 93.47%, indicating that deep generative enhancement provides greater benefit to downstream detection discriminability than traditional methods. However, the proposed UC DEN achieves the best results, with Precision, Recall, and mAP@0.50 of 93.63%, 88.97%, and 94.52%, respectively, corresponding to improvements of 5.45, 3.55, and 7.16 percentage points over the no-enhancement baseline.

5.4.2. Ablation Study on Feature Enhancement Mechanism

To quantify the specific contribution of each feature enhancement component to the detection performance of YOLOv13-TSDD, ablation experiments are conducted at two levels: convolutional design and attention mechanism, with the front-end enhancement strategy fixed as UC DEN throughout. Results are presented in Figure 23 and Table 10. It should be noted that the feature enhancement modules do not simply augment the detection network externally; they directly determine the model's capacity to model orientation-sensitive texture, local linear structures, and critical region responses for underwater cracks, and therefore exert a systematic influence on overall crack representation quality.

Table 10. Detection performance of YOLOv13-TSDD under different feature enhancement methods.

Ablation Method	Image Enhancement	Convolution Method	Attention Mechanism	Precision (%)	Recall (%)	mAP@0.50 (%)
Ablation 4	UC DEN	Default	--	88.76	85.88	87.94
Ablation 5		Default	CEMA	90.08	87.24	90.18
Ablation 6		PRFConv	--	89.54	87.57	89.56
Ablation 7		PRFConv	CBAM	91.74	88.03	92.86
YOLOv13-TSDD		PRFConv	CEMA	93.63	88.97	94.52

As shown in Figure 23, the progressive introduction of convolutional and attention enhancement mechanisms produces a consistent simultaneous improvement in Precision, Recall, and mAP@0.50, with the complete model consistently occupying the outermost position in the radar plot, confirming its optimal balance across false alarm suppression, missed detection reduction, and overall detection performance.

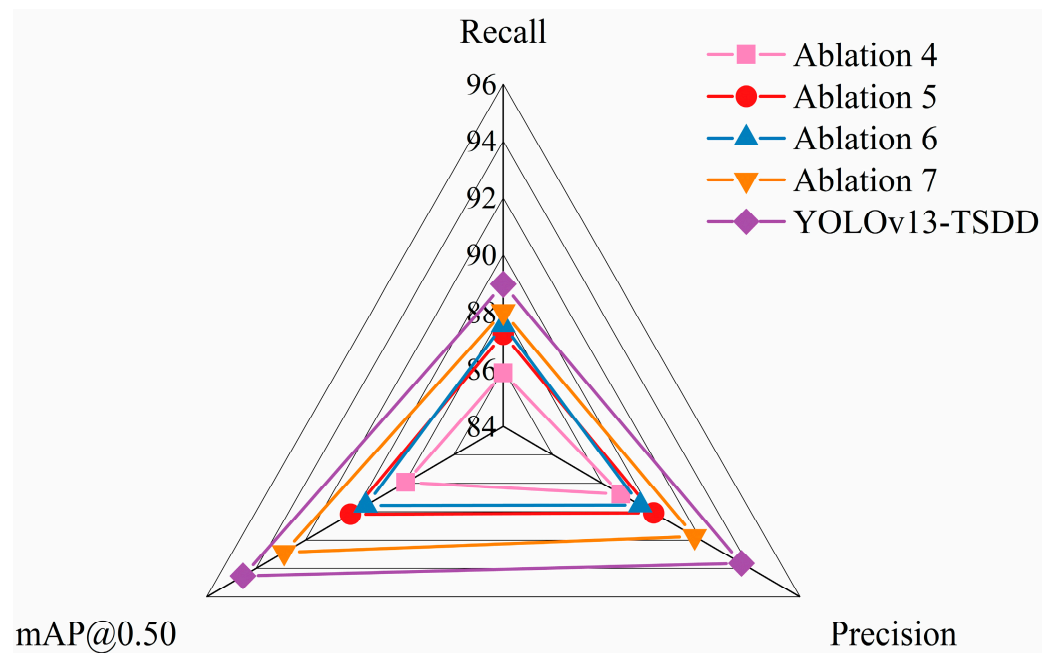


Figure 23. Graphical comparison of precision, recall, and mAP@0.50 for YOLOv13-TSDD under different feature enhancement configurations.

As shown in Table 10, when only UC DEN is applied without any convolutional or attention improvement (Ablation 4), Precision, Recall, and mAP@0.50 are 88.76%, 85.88%, and 87.94%, respectively. This result indicates that even with improved input quality, retaining the default convolutional structure without targeted feature recalibration still limits the network's representational capacity for weak-texture, blurred-boundary, and irregularly shaped cracks. Introducing CEMA alone (Ablation 5) raises the three metrics to 90.08%, 87.24%, and 90.18%, demonstrating that attention-based region focusing improves feature selection under complex backgrounds. However, since CEMA still operates on default convolution outputs, the quality of shallow-stage initial representations remains constrained, limiting the overall performance gain. In contrast, replacing default convolution with PRFConv alone (Ablation 6) achieves 89.54%, 87.57%, and 89.56%, confirming that the convolutional design improvement also contributes substantially by enhancing the model's initial sensitivity to orientation-sensitive texture and local linear structures. In functional terms, convolutional improvement addresses whether crack features can be perceived early, while the attention mechanism addresses whether already-perceived features can be further reinforced and selectively focused.

The complete YOLOv13-TSDD achieves Precision, Recall, and mAP@0.50 of 93.63%, 88.97%, and 94.52%, representing improvements of 1.89, 0.94, and 1.66 percentage points over Ablation 7. This margin confirms that the performance advantage of the proposed method does not derive simply from combining convolutional and attention modules, but from their task-specific internal designs. PRFConv strengthens initial shallow-stage responses to crack propagation directions, spalling boundaries, and rebar strip textures, while CEMA's crack-aware joint channel-spatial attention further highlights critical structural and edge information. Together, they provide more effective adaptation to the feature representation CEMAnds of underwater crack detection than generic convolution combined with general-purpose attention mechanisms.

The collaborative effect between the two stages can be further verified by jointly considering Tables 10 and 11. With PRFConv and CEMA fixed in Stage II, introducing UC DEN increases mAP@0.50 from 87.36% to 94.52%. Conversely, with UC DEN fixed in Stage I, the complete PRFConv-CEMA design increases mAP@0.50 from 87.94% to 94.52%.

These results indicate that Stage I improves the reliability of the input representation for subsequent feature learning, while Stage II further converts the enhanced color, texture, and contour information into directional structural and edge features, ultimately producing detection gains. The two stages therefore form a continuous collaborative optimization process from input enhancement to feature representation and target localization.

Table 11. Detection performance of YOLOv13-TSDD under different losses and activation functions.

Combination	Enhancement Method	Loss Function	Activation Function	Precision (%)	Recall (%)	mAP@0.50 (%)	mAP@0.50–0.95 (%)
1	Stage 1: UC DEN Stage 2: PRFConv + CEMA	C-IoU [50]	ReLU	90.92	86.04	92.01	61.72
2		S-IoU [52]	ReLU	91.36	86.48	92.43	62.48
3		E-IoU [51]	ReLU	92.04	87.16	93.11	63.91
4		FID-IoU	ReLU	92.58	87.79	93.67	65.37
5		C-IoU	SiLU	91.51	86.71	92.58	62.83
6		S-IoU	SiLU	92.03	87.25	93.06	63.66
7		E-IoU	SiLU	92.74	88.01	93.82	65.12
8		FID-IoU	SiLU	93.63	88.97	94.52	67.08

5.4.3. Ablation Study on Loss Function and Activation Function Combinations

To assess the influence of loss function and activation function configuration on detection performance, a systematic ablation is conducted across different IoU loss and activation function combinations while keeping the front-end enhancement and feature enhancement structures fixed. Results are presented in Figure 24 and Table 11. Unlike image enhancement and feature enhancement, which operate on the input and feature domains respectively, the loss function and activation function directly govern gradient propagation quality, parameter update stability, and bounding-box regression optimization direction during training, making them critical training-level determinants of final detection performance.

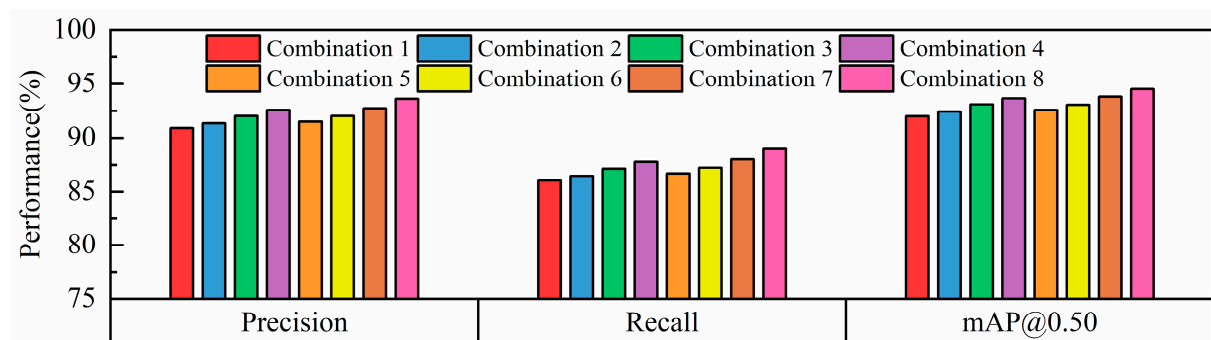


Figure 24. Graphical comparison of precision, recall, and mAP@0.50 for YOLOv13-TSDD under different loss and activation function combinations.

As shown in Figure 24, the effects of the loss function and activation function are not independent but exhibit a clear coupling relationship. The loss function defines the regression optimization objective, while the activation function determines whether the model can stably retain and learn the weak-response features associated with that objective. If the loss function is poorly adapted to complex-boundary cracks, improvements in the activation function are insufficient to compensate. Conversely, if the activation function provides limited weak-feature retention, the advantages of a more targeted loss function are similarly diminished. The best results are therefore achieved not by FID-IoU or SiLU in isolation, but by their combination, confirming that training optimization for underwater

crack detection requires simultaneous attention to boundary modeling capacity and feature expression continuity.

As shown in Table 11, under ReLU activation, progressively replacing C-IoU with S-IoU, E-IoU, and FID-IoU increases mAP@0.50 from 92.01% to 93.67%, while mAP@0.50–0.95 increases from 61.72% to 65.37%. This trend becomes more pronounced under SiLU activation, where FID-IoU achieves mAP@0.50 and mAP@0.50–0.95 of 94.52% and 67.08%, respectively, corresponding to improvements of 1.94 and 4.25 percentage points over C-IoU. Notably, the larger gain in mAP@0.50–0.95 indicates that the advantage of FID-IoU is maintained and becomes more evident under stricter overlap requirements. This suggests that the auxiliary bounding-box mechanism and piecewise weighted mapping provide more effective regression constraints for low-overlap and difficult samples, thereby improving bounding-box localization quality for small-scale, irregular, and weak-boundary damage.

6. Conclusions

This study proposed YOLOv13-TSDD, a two-stage surface damage detection framework based on YOLOv13, to address the persistent challenges of color bias, severe local detail degradation, boundary ambiguity, and unstable localization of small-scale irregular targets in underwater RC surface structural crack detection. The framework adopts a coordinated image enhancement and object detection pipeline: UC DEN is constructed at the front end to perform task-oriented enhancement of degraded underwater imagery, while PRFConv, CEMA, and FID-IoU are integrated in the detection stage to provide targeted optimization of shallow-stage orientation-sensitive feature extraction, critical crack region focusing, and complex boundary regression. The framework achieves high-accuracy detection of typical surface cracks, including concrete spalling, rebar exposure, and concrete cracking. It is worth noting that a current limitation of the proposed method is that UC DEN and the detection network are connected in a sequential cascaded manner. As a result, the image enhancement results produced by UC DEN are not yet constrained and optimized in an end-to-end manner through the loss of the dFietection network, indicating that there remains room for further improvement in the collaborative optimization between the two stages. A further limitation concerns the experimental protocol: all comparison methods were trained using an identical schedule and hyperparameter configuration, without architecture-specific tuning, pretrained-weight matching, or parameter-count normalization. The reported comparisons should therefore be interpreted as performance under a common training budget rather than each method's best achievable result.

The scope of experimental validation remains subject to certain limitations. Although the current dataset includes low-illumination samples caused by underwater absorption, scattering, and light attenuation, completely nighttime underwater inspection scenarios were not separately established. In addition, the current imaging conditions were mainly limited to two turbidity ranges, a single camera, and a fixed imaging distance, without systematic consideration of biofouling, water type, or different imaging devices.

The principal conclusions are as follows:

- (i) The proposed UC DEN effectively mitigates non-uniform color bias, luminance attenuation, detail blurring, and boundary weakening in underwater imagery, substantially improving texture, contour, and local contrast representation in crack regions, including cracks, spalling, and rebar exposure. Experimental results demonstrate that UC DEN achieves the best performance on UISM, UICM, and UIQM, and remains competitive on UIConM, confirming a well-balanced optimization among color restoration, detail enhancement, and overall image quality preservation. This provides the downstream detection network with more discriminative input representations.

- (ii) To address weak shallow-stage crack responses, insufficient directional feature sensitivity, and susceptibility to background interference at mid-to-high feature levels, a feature enhancement network integrating PRFConv and CEMA was developed. PRFConv strengthens initial perception of crack propagation directions, spalling boundary morphology, and rebar strip structures, while CEMA further improves saliency and edge expressiveness in critical crack regions. Ablation experiments confirm strong cooperative gains between the two modules, demonstrating their collective capacity to improve model representation of weak-texture, morphologically diverse, and complex-background cracks.
- (iii) To overcome the insufficient adaptation of conventional IoU losses to small-scale, complex-boundary, and geometrically irregular cracks, the FID-IoU loss function was proposed. Through its auxiliary bounding-box mechanism and piecewise weighted mapping strategy, FID-IoU provides more stable optimization constraints for hard samples, improving bounding-box localization accuracy and regression stability for complex-boundary targets. Experimental results confirm that the combination of FID-IoU and SiLU activation achieves the best overall detection performance, validating the effectiveness of a task-oriented bounding-box regression strategy.
- (iv) Comprehensive comparative experiments demonstrate that YOLOv13-TSDD achieves the best overall performance on the URCSSD detection task, with Precision, Recall, and mAP@0.50 reaching 93.63%, 88.97%, and 94.52%, respectively, indicating high prediction accuracy and a reduced risk of missed detections for weak-texture damage. Despite the sample imbalance among CS, ER, and CC, the AP@0.50 of the three categories reaches 95.41%, 94.66%, and 93.49%, respectively, demonstrating stable multi-class detection capability under imbalanced class distributions.

Future work will extend the current dataset and conduct external validation under broader field conditions, including different RGB imaging devices, imaging distances, illumination conditions, natural water types, biofouling interference, and turbidity levels above 30 NTU. More extreme low-light and nighttime scenarios will also be considered to further evaluate the cross-condition and cross-device generalization capability of YOLOv13-TSDD. Meanwhile, model deployment will be optimized for autonomous underwater vehicles (AUVs), remotely operated vehicles (ROVs), and fixed underwater monitoring cameras.

Author Contributions: Conceptualization, X.W. and M.M.S.; methodology, X.W. and Z.W.; software, X.Y. and Y.Z.; validation, X.W., Y.Z., X.Y. and Z.W.; formal analysis, Y.Z., X.Y. and Z.W.; investigation, X.W., M.M.S., Y.Z., X.Y. and Z.W.; resources, X.W.; data curation, X.W., M.M.S., Y.Z., X.Y. and Z.W.; writing—original draft preparation, X.W., M.M.S., Y.Z., X.Y. and Z.W.; writing—review and editing, X.W., M.M.S., Y.Z., X.Y. and Z.W.; visualization, X.W., M.M.S., Y.Z., X.Y. and Z.W.; supervision, X.W. and Z.W.; project administration, X.W.; funding acquisition, X.W. and Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant no. 52378300), Key Scientific Research Projects of Colleges and Universities in Henan province (Grant no.26A560013), Key R&D and Promotion Projects in Henan Province (Grant no. 262102320052), Henan Out-standing Youth Science Foundation (Grant no.242300421014), and the Science and Technology Research and Development Project of China Construction Sixth Engineering Bureau Corp., Ltd. (CSCEC6B-2024-Z-4).

Data Availability Statement: The datasets presented in this article are not readily available due to technical and time limitations. Requests to access the datasets should be directed to wangxinwei@ncwu.edu.cn.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fan, Z.; Chen, J.; Li, Y.; Huang, Q. Few-shot crack image generation for surface detection on precast concrete components. *Adv. Eng. Inform.* **2026**, *70*, 104210. [\[CrossRef\]](#)
2. Han, S.; Xiao, G.; Wang, P.; Xu, C.; Zhou, A.; Yu, J.; Ou, J. Axial-bending behavior of hybrid fiber-reinforced polymer-steel reinforced concrete columns with novel closed ties: Confinement mechanism and equivalent design. *Eng. Struct.* **2026**, *353*, 122236. [\[CrossRef\]](#)
3. Guo, H.; Jafari-Asl, J.; Dai, J.G.; Dong, Y.; Fang, J.; Cui, H. Life-cycle performance evaluation framework for GFRP-reinforced concrete structures subject to marine environment. *Eng. Struct.* **2026**, *348*, 121762. [\[CrossRef\]](#)
4. Veruz, E.G.; Silva, A.J.; Michalski, M.A.C.; da Silva, R.F.; de Souza, G.F.M.; Oshiro, A.T. A framework for planning underwater hull inspections based on computer vision and degradation assessment. *Ocean Eng.* **2025**, *317*, 120053. [\[CrossRef\]](#)
5. Elmezain, M.; Saad Saoud, L.; Sultan, A.; Heshmat, M.; Seneviratne, L.; Hussain, I. Advancing Underwater Vision: A Survey of Deep Learning Models for Underwater Object Recognition and Tracking. *IEEE Access* **2025**, *13*, 17830–17866. [\[CrossRef\]](#)
6. Jia, Z.; Li, Z.; Wang, K.; Zhi, G.; Liu, Z.; Vong, C. Graph convolutional contrastive learning for few-shot open-set fault diagnosis in electro-hydrostatic actuators. *IEEE Trans. Aerosp. Electron. Syst.* **2026**, *62*, 3076–3092. [\[CrossRef\]](#)
7. Jia, Z.; Liu, Z.; Li, Z.; Wang, K.; Vong, C.-M. Lightweight Fault Diagnosis via Siamese Network for Few-Shot EHA Circuit Analysis. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 15585–15596. [\[CrossRef\]](#)
8. Ye, X.; Luo, K.; Wang, H.; Zhao, Y.; Zhang, J.; Liu, A. An advanced AI-based lightweight two-stage underwater structural damage detection model. *Adv. Eng. Inform.* **2024**, *62*, 102553. [\[CrossRef\]](#)
9. Kang, F.; Tang, P.; Chen, J.; Geng, Z.; Zhou, D. Synthesizing sonar image of underwater dam cracks with conditional generative adversarial networks. *Eng. Struct.* **2026**, *354*, 122378. [\[CrossRef\]](#)
10. Talamkhani, S.; Liu, K. Underwater vision-enhanced image segmentation for supporting automated inspection of underwater bridge components. *Autom. Constr.* **2025**, *175*, 106230. [\[CrossRef\]](#)
11. Liu, H.; Yuan, J.; Ren, Q.; Li, M.; Qi, Z.; Deng, X. Remotely operated vehicle (ROV) underwater vision-based micro-crack inspection for concrete dams using a customizable CNN framework. *Autom. Constr.* **2025**, *173*, 106102. [\[CrossRef\]](#)
12. Liu, P.; Li, X.; Zhang, A.; Zhang, J.; Zhang, J. Slope Surface Crack Sensor Based on a Single-Mode Fiber Conjugate Helical Structure. *Opt. Laser Technol.* **2025**, *192*, 114128. [\[CrossRef\]](#)
13. Jiao, Y.; Wang, H.; Du, H.; Shan, Y.; Pan, Z.; Zhang, C.; Huang, H. Robust polarimetric image restoration for underwater concrete crack inspection in turbid environments. *Opt. Lasers Eng.* **2026**, *201*, 109669. [\[CrossRef\]](#)
14. Wu, T.; Hou, S.; Wu, Z.; Xiong, W.; He, X. Optical measurement method for large-scale surface cracks on underwater bridge piers. *Eng. Struct.* **2026**, *350*, 121966. [\[CrossRef\]](#)
15. Noda, T.; Kondo, H.; Nagaokaya, K.; Someya, T. Real-Time Underwater Localization in Three-Dimensional Environment. *IEEE J. Ocean. Eng.* **2025**, *50*, 1585–1595. [\[CrossRef\]](#)
16. Sun, W.; Hou, S.; Wu, G.; Zhang, Y.; Zhao, L. Two-step rapid inspection of underwater concrete bridge structures combining sonar, camera, and deep learning. *Comput.-Aided Civ. Infrastruct. Eng.* **2025**, *40*, 2650–2670. [\[CrossRef\]](#)
17. Liu, M.; Lin, H.; Li, Y.; Wang, T.; Li, Y.; Luo, J. High-precision stripe center extraction method using laser line scanning for underwater concrete 3D reconstruction. *Constr. Build. Mater.* **2026**, *517*, 145778. [\[CrossRef\]](#)
18. Wang, X.; Zhao, P.; Yue, S.; Gu, Y. An Improved Cross Algorithm Edge Detection Method and Its Application in Coal Gangue Identification. *Int. J. Coal Prep. Util.* **2026**, *46*, 835–847. [\[CrossRef\]](#)
19. Huang, W.; Pang, R.; Zhou, Y.; Zhong, F.; Chen, J.; Kang, F. Multi-condition dam crack segmentation based on a novel lightweight CEE-YOLO model. *Adv. Eng. Inform.* **2025**, *68*, 103697. [\[CrossRef\]](#)
20. Wei, B.; Zheng, Z.; Hu, Y.; Yuan, D.; Chi, Y.; Qiu, C. Underwater concrete crack detection of dams via CycleGAN-based data enhancement and optimized multi-scale YOLO11. *Constr. Build. Mater.* **2026**, *514*, 145497. [\[CrossRef\]](#)
21. Chen, Q.; Liu, L. Unveiling dynamic shrinkage behaviors of bamboo: Automatic tracking of vascular bundles during drying with YOLO and ByteTrack. *Constr. Build. Mater.* **2026**, *534*, 146863. [\[CrossRef\]](#)
22. Ge, K.; Han, Q.B.; Xia, J.Y.; Wang, J.H.; Sun, L.J. AWC-ECRA-YOLO: A lightweight sonar crack detection network with adaptive wavelet convolution and enhanced cluster routing attention. *Results Eng.* **2026**, *29*, 109847. [\[CrossRef\]](#)
23. Ji, C.; Li, Y. UWLIGHT-YOLO: An adaptive lightweight attention network for underwater sonar image object detection. *Cogn. Robot.* **2026**, *6*, 78–91. [\[CrossRef\]](#)
24. Zang, S.; Hong, G.; Liu, Y.; Gan, H.; Sun, F.; Lu, D. MCD-YOLO: An improved YOLOv8-based approach for surface crack detection of marine machinery components. *Ocean Eng.* **2025**, *341*, 122624. [\[CrossRef\]](#)
25. Li, H.; Kang, F.; Zhu, X.; Chen, J.; Li, H. A crack detection system for underwater marine structures based on an improved YOLOv12 model. *Ocean Eng.* **2026**, *343*, 123495. [\[CrossRef\]](#)
26. Abbas, N.; Zheng, K.; Li, Z.; Chi, W.; Chen, C.; Lin, H.; Jiang, D.; Gu, H.; Li, J. BT-YOLO: An attention-enhanced framework for monitoring bioluminescent algal blooms in coastal tourism zones. *Ecol. Inform.* **2026**, *93*, 103595. [\[CrossRef\]](#)

27. Li, K.; Wu, L.; Qi, Q.; Liu, W.; Gao, X.; Zhou, L.; Song, D. Beyond Single Reference for Training: Underwater Image Enhancement via Comparative Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 2561–2576. [CrossRef]
28. Chang, B.; Yuan, G.; Li, J. CDDU: Convolutional Dictionary Deep Unfolding Network for Underwater Image Enhancement. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 2540718. [CrossRef]
29. Zhang, Y.; Jiang, Q.; Liu, P.; Gao, S.; Pan, X.; Zhang, C. Underwater Image Enhancement Using Deep Transfer Learning Based on a Color Restoration Model. *IEEE J. Ocean. Eng.* **2023**, *48*, 489–506. [CrossRef]
30. Zhang, W.; Zhuang, P.; Sun, H.H.; Li, G.; Kwong, S.; Li, C. Underwater Image Enhancement via Minimal Color Loss and Locally Adaptive Contrast Enhancement. *IEEE Trans. Image Process.* **2022**, *31*, 3997–4010. [CrossRef] [PubMed]
31. Liang, Z.; Huang, H.; Zhang, W.; Song, H.; Wan, X.; Wang, C.; Huang, L.; Zhuang, P. Underwater Image Enhancement via Piecewise Colour Balancing and Multiscale Enhancement Fusion. *IEEE J. Ocean. Eng.* **2025**, *50*, 1960–1977. [CrossRef]
32. Jin, X.; Liu, B.; Tang, L.; Yao, S.; Wang, P.; Jiang, Q. ECMSFNet: Real-Time Infrared Small Target Detection Network With Efficient Convolution and Multiscale Feature Fusion. *IEEE Internet Things J.* **2026**, *13*, 11537–11552. [CrossRef]
33. Chu, C.; Shen, J.; Xiao, Q. Graph-Regularized Geometric Deep Learning for Motor Imagery EEG Decoding via Multi-Domain Fusion. *Pattern Recognit.* **2026**, *174*, 112976. [CrossRef]
34. Li, C.; Zhang, Y.; Shi, Z.; Zhang, Y.; Zhang, Y. Moderately Dense Adaptive Feature Fusion Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5616712. [CrossRef]
35. Liu, W.; Liu, S. Bearing Remaining Useful Life Prediction Based on Optimized VMD and BiLSTM-CBAM. *PLoS ONE* **2025**, *20*, e0326399. [CrossRef] [PubMed]
36. Wang, X.; Zou, X.; Wang, T.; Chen, J.; He, Y.; Pan, Z. GAENet: A Global Attention-Aware Enhanced Network for Remote Sensing Target Detection. *IEEE Geosci. Remote Sens. Lett.* **2026**, *23*, 4004305. [CrossRef]
37. Song, Z.; Zou, S.; Zhou, W.; Huang, Y.; Shao, L.; Yuan, J.; Gou, X.; Jin, W.; Wang, Z.; Chen, X.; et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nat. Commun.* **2020**, *11*, 4294. [CrossRef] [PubMed]
38. Kabir, H.; Wu, J.; Dahal, S.; Joo, T.; Garg, N. Automated estimation of cementitious sorptivity via computer vision. *Nat. Commun.* **2024**, *15*, 9935. [CrossRef] [PubMed]
39. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733.
40. Zhou, K.; Lei, D.; Chun, P.; She, Z.; He, J.; Du, W.; Hong, M. Evaluation of BFRP strengthening and repairing effects on concrete beams using DIC and YOLO-v5 object detection algorithm. *Constr. Build. Mater.* **2024**, *411*, 134594. [CrossRef]
41. Jocher, G. Ultralytics YOLOv5. 2020. Available online: <https://github.com/ultralytics/yolov5> (accessed on 1 March 2026).
42. Zhou, Z.; Zhou, S.; Li, S.; Li, H.; Yang, H. Tunnel lining quality detection based on the YOLO-LD algorithm. *Constr. Build. Mater.* **2024**, *449*, 138240. [CrossRef]
43. Wang, C.Y.; Bochkovskiy, A.; Liao, H. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. *arXiv* **2022**, arXiv:2207.02696.
44. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 1 March 2026).
45. Wang, C.Y.; Liao, H. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.
46. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *arXiv* **2024**. [CrossRef]
47. Jocher, G.; Qiu, J. Ultralytics YOLO11. 2024. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 1 March 2026).
48. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Realtime Object Detectors. *arXiv* **2025**, arXiv:2502.12524.
49. Wang, H.; Zhang, S.; Leng, B. HGFormer: Topology-Aware Vision Transformer With Hypergraph Learning. *IEEE Trans. Multimed.* **2025**, *27*, 5746–5757. [CrossRef]
50. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12993–13000. [CrossRef]
51. Zhang, Y.F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and Efficient IoU Loss for Accurate Bounding Box Regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
52. Bochkovskiy, A. Siou Loss: More Powerful Learning for Bounding Box Regression. *arXiv* **2022**, arXiv:2205.12740.
53. Yang, M.; Sowmya, A. An Underwater Color Image Quality Evaluation Metric. *IEEE Trans. Image Process.* **2015**, *24*, 6062–6071. [CrossRef] [PubMed]
54. Liu, R.; Fan, X.; Zhu, M.; Hou, M.; Luo, Z. Real-World Underwater Enhancement: Challenges, Benchmarks, and Solutions Under Natural Light. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4861–4875. [CrossRef]

55. Lin, P.; Fan, Z.; Wang, Y.; Sun, X.; Yáng, Y.R.; Fu, X. Msffnet: Multi-stream feature fusion network for underwater image enhancement. *Displays* **2025**, *88*, 103023. [[CrossRef](#)]
56. Zhang, D.; Ye, X.; Chen, D.; Luo, K. Intelligent Identification of Apparent Defects in Underwater Structures Based on Deep Learning. *Struct. Eng.* **2025**, *41*, 25–30. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.