Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 19, No. 1, 2026

A Dual-Task Cascade Network for Underwater Object Detection with Scene Adaptability

Haiyang Yao [1, 2], Jinhao Shi [1, 2], Yuzhang Zang [3], Xiaobo Zhao [4], Xiao Chen [1, 2], Tao Lei [1, 2],

Haiyan Wang [1, 2, 5, *]

[1] Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, Shaanxi 710021, PR China
[2] School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, Shaanxi 710021, PR China
[3] Engineering and Design Department, Western Washington University, Bellingham, WA, USA.
[4] Department of Electrical and Computer Engineering, Aarhus University, 8200 Aarhus, Denmark.
[5] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, PR China

*** Corresponding author:** Haiyan Wang

Abstract: Underwater small-object detection is often hindered by complex backgrounds and the difficulty of capturing fine-scale features. This paper addresses these challenges with three contributions: (1) A perturbation-theoretic model that rigorously quantifies background interference by decomposing loss functions into ideal and perturbation components. (2) The Dual-Task Cascade Network (DTC-Net), which classifies scenes (uniform, non-uniform, or enhanced) before activating scene-specific detection branches. (3) A Dual Attention Feature Pyramid Network (DAFPN) and a dynamic convolution Focus Module designed to recalibrate features and emphasize fine-grained details. Extensive experiments on UTDAC2020 show that DTC-Net achieves a state-of-the-art 48.1% AP, outperforming FCOS and D-FINE. Furthermore, the strategy generalizes well, improving existing detectors (e.g., GCC-Net) by up to 1.9% AP, demonstrating its effectiveness across various frameworks.

Keywords: Underwater Object Detection; Adaptive Detection Framework; Underwater Optical Image Processing.

1. Introduction

Autonomous underwater vehicles (AUVs) are critical platforms for marine exploration [1][2]. Intelligent underwater object detection is a core capability, enabling AUVs to navigate complex environments and conduct tasks like resource mapping and ecological monitoring [3][4]. However, underwater detection faces domain-specific challenges, such as light refraction, wavelength-dependent absorption, and scattering, which cause pronounced hue shifts and blurring. Furthermore, targets often exhibit extreme scale variations and subtle inter-class differences, frequently surrounded by background structures that mimic object appearance[5]. Thus, addressing both complex backgrounds and intrinsic target characteristics is essential for reliable detection.

Research has advanced through one-stage[6][7] and two-stage[8][9] paradigms, focusing on image enhancement and architecture optimization. Building on the state-of-the-art YOLOv7[10], we previously developed the lightweight YOLO_GN [11] and an online knowledge distillation framework [12] to enhance accuracy. Recently, we proposed U-TransCNN[34], a hybrid Transformer–CNN network for restoring color fidelity and fine details. Despite these advancements, small-object detection still suffers from high false-negative rates and performance instability under varying backgrounds. Motivated by this, our work makes three primary contributions:

(1) A perturbation-based theoretical model to analyze the influence of underwater backgrounds on small-object detection;

(2) A first-of-its-kind scene-adaptive architecture for underwater small-object detection;

(3) A specialized representation module to enhance discriminative feature extraction and robustness across complex underwater scenes.

2. Related Works

To address challenges regarding underwater color distortion, blurring, and small objects, research focuses on image enhancement and feature extraction.

Wang et al. [14] proposed a method based on constrained color compensation and a spatial fog line model to reduce blue-green bias. Li et al. [15] introduced Water-Net and the UIEBD benchmark to address quality degradation. Cong et al. [16] proposed PUGAN, a physical model-guided GAN, to eliminate color bias and blurring. Yeh et al. [17] proposed a lightweight network jointly learning color conversion and detection. Liu et al. [18] introduced an object-guided twin-based adversarial framework using task-aware feedback. Islam et al. [19] developed a fast enhancement method ensuring real-time performance. However, these unified frameworks often increase complexity and limit adaptability. Lin et al. [20] proposed RoIMix to fuse features from candidate regions. Chen et al. [21] introduced Swipe Net, utilizing Hyper Feature Maps to address noise. Jia et al. [22] proposed Efficient Det, integrating feature pyramids and attention mechanisms.

Song et al. [23] proposed Boosting R-CNN with uncertainty modeling. Liu et al. [24] addressed domain transfer by learning domain-invariant features, while Chen et al. [25] explored Domain Mix-up and contrast learning. Liu et al. [26] introduced TC-YOLO, incorporating Transformer self-attention. Fu et al. [27] proposed a residual feature transference module; Dai et al. [28] developed a gated cross-domain collaborative network; and Yuan et al. [29] used

interference mining to distinguish objects from backgrounds. Despite these efforts, robust generalization remains challenging. To address dataset scarcity, studies [30][31] focused on data expansion. Liu et al. [32] proposed a Poisson GAN to augment object instances and introduced Aqua Net. However, selecting effective samples from expanded datasets remains a key challenge. Despite advancements, underwater small-object detection suffers from statistical disparities between synthetic and real-world images. Existing SOTA models often exhibit inconsistent performance under varying backgrounds.

We introduce the Scene Identification and Detection Dual-Task Cascade Network (DTC-Net). The key innovations are:

1) Mathematical derivation of background influence on small-object detection

We present a framework using perturbation theory to quantify the influence of background variability. We decompose the loss function into ideal values and perturbation terms, proving that background fluctuations mathematically impact learning and limit current models.

2) Dual-Task Network with Scene Recognition and Object Detection

DTC-Net categorizes inputs into three scene types (uniform, non-uniform, and enhanced) and activates a scene-specific detection module. This cascade structure enhances adaptability and stability across diverse environments.

3) Dual Attention Feature Pyramid Network (DAFPN) and RPN Error-Based Detection Head

We design a Dual Attention Feature Pyramid Network (DAFPN) with 1D convolution for channel interaction and a global context compression mechanism. Additionally, a dynamic convolution Focus Module adaptively adjusts kernel parameters to extract fine-grained features of small objects.

Validated on specialized datasets derived from UTDAC2020 [21], DTC-Net demonstrates superior stability, adaptability, and robustness compared to SOTA models.

3. Analysis and Method

3.1. Theoretical Analysis of Background Perturbations

To address the limitations of existing detectors, we first investigate the statistical characteristics of underwater scenes. Existing datasets comprise both real-world and enhanced collections. Real-world images often exhibit pronounced color bias and non-uniform illumination, while enhanced images differ in statistical distribution. Our experiments indicate that models trained on either single or mixed datasets struggle to generalize effectively [10][11][12].

To analyze these variations, we computed the Sum of Magnitude Spectrum [13] for 1,000 samples across three distinct underwater conditions (Fig. 1). This metric quantifies frequency-domain variations, offering insight into dataset distributions. Our analysis reveals substantial disparities: non-uniform illumination images exhibit Gaussian-like distributions with unstable fluctuations, whereas uniform and enhanced images display noticeable tail-dragging. These skewed characteristics can induce large gradient variations in specific directions of the loss function, destabilizing optimization and degrading detection performance. These findings motivate the design of a scene-adaptive framework to mitigate background-induced instability.

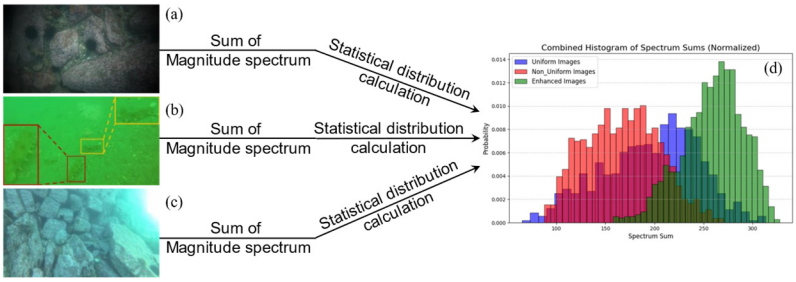


Fig 1. Challenges in underwater environments (a) non-uniform illumination image (b) uniform illumination image with green color cast (c) enhanced image (d) Magnitude Spectrum distributions

We conducted a theoretical investigation into underwater object detection under fluctuating background conditions. We selected Boosting-RCNN [23] as our analytical baseline, though the insights are applicable to other architectures. The total loss function is defined as:

$$L_{total} = \lambda_{obj-rpn}L_{obj-rpn} + \lambda_{loc-rpn}L_{loc-rpn} + \lambda_{iou-rpn}L_{iou-rpn} + \lambda_{reg}L_{reg} + \lambda_{cls}L_{cls} \quad (1)$$

where $L_{obj-rpn}$ represents object prediction loss, $L_{loc-rpn}$ is the bounding box localization loss, and $L_{iou-rpn}$ denotes the IoU prediction loss. Additionally, L_{reg} and L_{cls} correspond to the classification and regression loss, respectively. The terms $\lambda_{obj-rpn}$, $\lambda_{loc-rpn}$, $\lambda_{iou-rpn}$, λ_{reg} and λ_{cls} are the weight parameters balancing the contributions of different loss components. The classification loss L_{cls} is a softmax between the predicted class and the ground truth, given by:

$$L_{cls} = \frac{1}{k}\sum w_k \cdot \sum(-s_k^c \cdot log(\hat{s}_k^c)) \quad (2)$$

where $s_k^c = \sqrt{pr_k \times cls_k(c)}$, $pr_k = \sqrt{\widehat{g_k} \times \widehat{p_k}}$, $cls_k(c)$ represents the classification score for sample k of class c, $\widehat{g_k} = IoU$ denotes the IoU between the predicted box and ground truth, $\widehat{p_k}$ is the objectness score of anchor k. By differentiating L_{cls} with respect to the model parameter θ, we obtain:

$$\frac{\partial L_{cls}}{\partial \theta} = \frac{1}{k}\sum w_c \frac{\sum(-s_k^c \log(\hat{s}_k^c))}{\partial \theta}$$
$$= \frac{1}{k}\sum w_c \frac{\left(-\sqrt{pr_k \times cls_k(c)}\right)\log\sum\left(\sqrt{pr_k \times cls_k(c)}\right)}{\partial \theta}$$
$$= \frac{1}{k}\sum w_c \left(\sum \frac{-\sqrt{pr_k \times cls_k(c)}}{\partial \theta}\log\left(\sqrt{pr_k \times cls_k(c)}\right) + \sum\left(-\sqrt{pr_k \times cls_k(c)}\frac{1}{\sqrt{pr_k \times cls_k(c)}}\frac{\partial\sqrt{pr_k \times cls_k(c)}}{\partial \theta}\right)\right)$$
$$= \frac{1}{k}\sum w_c \left(-\sum \frac{\sqrt{pr_k \times cls_k(c)}}{\partial \theta}\left(\log\sqrt{pr_k \times cls_k(c)} + 1\right)\right) \quad (3)$$

Considering that background interference can perturb both the classification score and the objectness score, we introduce the following perturbation terms:

$$pr_k = pr_{kt} + \varepsilon_g pr_{kb}, cls_k(c) = cls_{kt}(c) + \varepsilon_c cls_{kb}(c) \quad (4)$$

where $\widehat{pr}_{kt}$ and $cls_{kt}(c)$ denote the ideal values of the objectness score and classification score of the sample k for class c, respectively. $\widehat{pr}_{kb}$ and $cls_{kb}(c)$ represent the perturbation values induced by background interference. ε_g and ε_c are the perturbation coefficients. Further differentiation yields:

$$\frac{\sqrt{pr_k \times cls_k(c)}}{\partial \theta} = \frac{\sqrt{\left(pr_{kt} + \varepsilon_g pr_{kb}\right) \times \left(cls_{kt}(c) + \varepsilon_c cls_{kb}(c)\right)}}{\partial \theta}$$
$$= \frac{1}{2\sqrt{\left(pr_{kt} + \varepsilon_g pr_{kb}\right) \times \left(cls_{kt}(c) + \varepsilon_c cls_{kb}(c)\right)}}\begin{pmatrix}\left(pr_{kt} + \varepsilon_g pr_{kb}\right) \times \frac{cls_{kt}(c) + \varepsilon_c cls_{kb}(c)}{\partial \theta} \\ +\left(cls_{kt}(c) + \varepsilon_c cls_{kb}(c)\right)\frac{pr_{kt} + \varepsilon_g pr_{kb}}{\partial \theta}\end{pmatrix}$$
$$= \frac{1}{2\sqrt{\left(pr_{kt} + \varepsilon_g pr_{kb}\right) \times \left(cls_{kt}(c) + \varepsilon_c cls_{kb}(c)\right)}}\begin{pmatrix}pr_{kt}\frac{\partial cls_{kt}(c)}{\partial \theta} + pr_{kt}\frac{\varepsilon_c \partial cls_{kb}(c)}{\partial \theta} \\ +\varepsilon_g pr_{kb}\frac{\partial cls_{kt}(c)}{\partial \theta} + \varepsilon_g pr_{kb}\frac{\varepsilon_c \partial cls_{kb}(c)}{\partial \theta} \\ +cls_{kt}(c)\frac{\partial pr_{kt}}{\partial \theta} + cls_{kt}(c)\frac{\varepsilon_g \partial pr_{kb}}{\partial \theta} \\ +\varepsilon_c cls_{kb}(c)\frac{\partial pv_k}{\partial \theta} + \varepsilon_c cls_{kb}(c)\frac{\varepsilon_g \partial pr_{kb}}{\partial \theta}\end{pmatrix} \quad (5)$$

Substituting Eq. (5) into Eq. (3) yields that:

$$\frac{\partial L_{cls}}{\partial \theta} = \frac{1}{k}\sum w_c \begin{pmatrix}-\sum\left(\log\sqrt{pcl_k(c)} + 1\right)\frac{1}{2\sqrt{pcl_k(c)}} \\ \left(cls_{kt}(c) + \varepsilon_c cls_{kb}(c)\right)\frac{\partial pr_{kt}}{\partial \theta} + \left(pr_{kt} + \varepsilon_g pr_{kb}\right)\frac{\partial cls_{kt}(c)}{\partial \theta} \\ +(\varepsilon_c pr_{kt} + \varepsilon_g \varepsilon_c pr_{kb})\frac{\partial cls_{kb}(c)}{\partial \theta} + (\varepsilon_g cls_{kt}(c) + \varepsilon_g \varepsilon_c cls_{kb}(c))\frac{\partial pr_{kb}}{\partial \theta}\end{pmatrix} \quad (6)$$

Where:

$$pcl_k(c) = \left(pr_{kt} + \varepsilon_g pr_{kb}\right) \times \left(cls_{kt}(c) + \varepsilon_c cls_{kb}(c)\right) \quad (7)$$

Under the perturbation assumption, the interaction between ideal and perturbation terms generates cross-term effects in gradient computation. Eq. (6) demonstrates that the classification score acts as a scaling coefficient modulating the gradient of the objectness score's ideal term, and vice versa. This creates a bidirectional dependency where background interference significantly influences learning dynamics.

In optimal optical conditions, the detector updates ideal feature representations with minimal background sensitivity. However, as optical quality degrades, background perturbations distort the gradient flow, forcing the model to simultaneously learn ideal feature variations and perturbation responses. This imbalance results in pronounced fluctuations in detection precision.

Crucially, when the statistical distribution of background features varies significantly (as shown in Fig. 1), the learning process becomes unstable. The resulting loss landscape is irregular, characterized by multiple local minima and steep gradients that hinder convergence. This mathematical formalization confirms that variations in background distributions fundamentally limit the ability of existing models to achieve robust generalization across heterogeneous underwater scenes.

3.2. DTC-Net: Dual-Task Cascade Network

Motivated by the statistical analysis of underwater background interference, we propose the Scene Identification and Detection Dual-Task Cascade Network (DTC-Net), an environment-adaptive architecture designed to address the instability of small-object detection. As illustrated in Fig. 2, the framework operates as a two-stage cascade system. Based

on our statistical findings, we categorize underwater images into three primary domains: uniform, non-uniform, and enhanced. A lightweight classification network first analyzes the input image to determine its specific background category. Subsequently, the image is dynamically routed to a specialized "expert" detection network corresponding to the identified scene type. Although these expert detectors share an identical architectural topology, their parameters are independently fine-tuned to specialize in their respective domains. By leveraging a pre-trained recognition module to guide the detection process, DTC-Net ensures that feature extraction and bounding box regression are performed by a model optimized for the specific optical conditions of the input. This scene-adaptive mechanism effectively mitigates performance degradation caused by background variability, offering a robust solution for heterogeneous underwater environments.

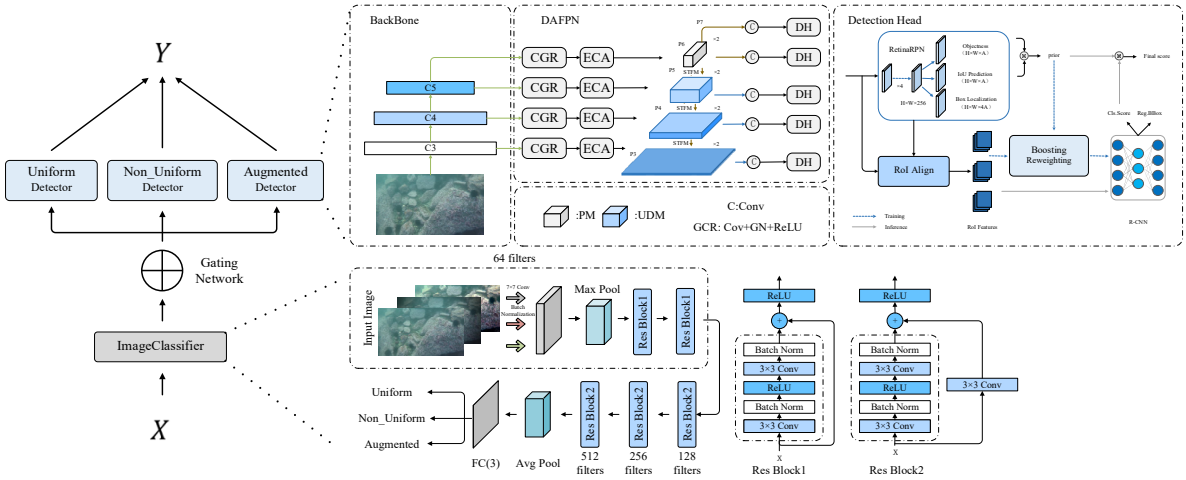


Fig 2. Architecture of the Dual-Task Cascade Network Framework

To implement the scene classification stage, we developed a lightweight recognition model based on the ResNet-18 backbone, implemented within the MMPretrain framework. As shown in Fig. 3, this module is designed to balance computational efficiency with high discrimination accuracy. The architecture begins with a stem layer that applies an initial 7×7 convolution followed by a max-pooling layer to reduce spatial resolution while preserving salient features. The core feature extractor consists of four sequential stages of residual blocks, each comprising two 3×3 convolutional layers equipped with Batch Normalization (BN) and ReLU activation. The extracted features are then aggregated via a Global Average Pooling (Avg Pool) layer, followed by a fully connected (FC) layer with three output units corresponding to the defined scene categories. This streamlined design minimizes computational overhead, allowing for efficient scene identification prior to the detection phase.

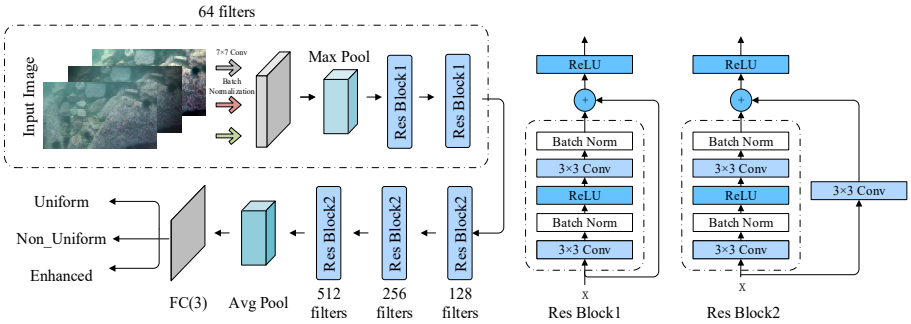


Fig 3. Architecture of the Scene Recognition Model. ResBlock1 denotes a standard ResNet block, while ResBlock2 incorporates a 1x1 convolution. FC (3) signifies a fully connected layer with three output units

3.3. The Dual Attention Feature Pyramid Network

To enhance target salience in underwater object detection through progressive multi-scale feature refinement, we propose the Dual Attention Feature Pyramid Network (DAFPN). As illustrated in Fig. 4, DAFPN adopts a four-level pyramid structure, producing five output feature maps (P3–P7) for multi-scale prediction. To ensure stable convergence in small-batch training scenarios, all normalization layers employ Group Normalization (GN) [35]. Leveraging the rich semantic context in the C5 feature map [36], DAFPN constructs P3–P5 by fusing backbone features C3–C5 through lateral and top-down connections, while P6 and P7 are generated via progressive up sampling from C5. Both the input and output branches utilize 3 × 3 convolutional layers for channel alignment, ensuring effective feature fusion and scale consistency throughout the network.

The proposed DAFPN architecture integrates two core components: the Focus Module (FM) and the Positioning Module (PM). Within this framework, the Efficient Channel Attention (ECA) mechanism and the Spatial–Temporal Fusion Module (STFM) jointly integrate channel- and spatial-domain information, thereby enhancing the extraction of critical features. The Context Exploration Block (CEB) employs multi-scale dilated convolutions to capture diverse

contextual patterns, propagating its outputs to the STFM for multi-path fusion and refinement. The FM utilizes spatial attention weights derived from the STFM to effectively distinguish foreground from background features, while simultaneously interacting with the CEB to further enhance feature quality. Concurrently, the PM fuses outputs from the ECA and spatial attention modules to deliver precise feature localization, facilitating accurate integration into subsequent pyramid layers. Through tight feature flow and coordinated inter-module interactions, DAFPN achieves an effective balance between computational efficiency and robust multi-scale representation.

A critical innovation within DAFPN is the Context Exploration Block (CEB), which is explicitly designed to expand the receptive field and mitigate the influence of background noise and distracting elements. The CEB leverages the Focus Module to distinguish between foreground-positive distractions (D_{FP}) and foreground-negative distractions (D_{FN}). Guided by a high-level feature mask, the module isolates and refines these elements to optimize the feature representation. The final refined output,(O_{refine}) is computed as:

$$O_{refine} = \alpha \cdot \mathrm{GR}(C_{current} + D_{FN}) + \beta \cdot \mathrm{GR}(C_{current} - D_{FP}) \quad (8)$$

where $C_{current}$ is the input from the current level, α and β are learnable scale parameters, and GR(·) denotes Group Normalization followed by a ReLU activation.

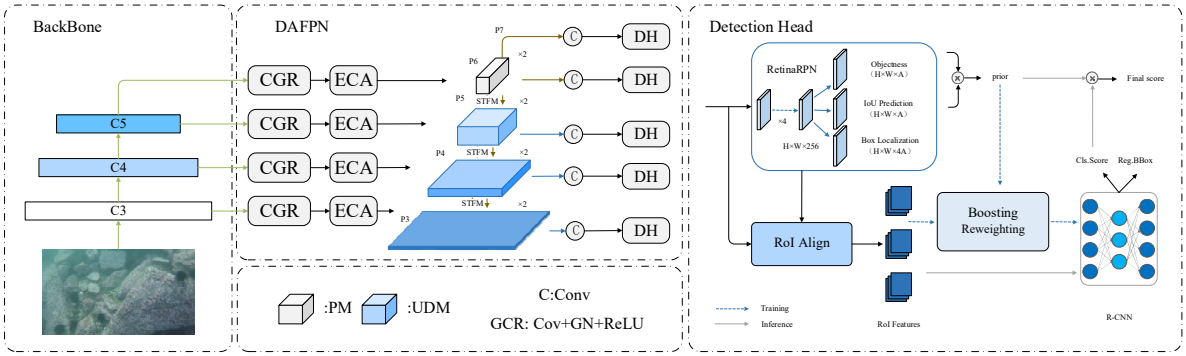


Fig 4. Overview of the architecture of DTC-Net. The network refers the pipeline of "Backbone—FPN—Heads." C3, C4, and C5 are the feature maps from the backbone, and P3 to P7 are the feature maps from the FPN, which are used for prediction tasks. The head adopts a two-stage Boosting R-CNN design, where an RPN generates proposals and the RoI head reweights samples based on the RPN's errors for final classification and regression.

3.4. Loss Function

To jointly optimize the scene classification and object detection tasks, the training objective of DTC-Net consists of two primary components: recognition loss and detection loss. The scene recognition task is designed to classify the background of underwater optical images while maintaining a lightweight computational footprint. To effectively differentiate image backgrounds with varying probability distributions, the recognition loss is formulated as a weighted cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{3} w_c y_{i,c} \log(\hat{y}_{i,c}) \quad (9)$$

where N is the number of samples, $c = 1, 2, 3$ denotes the three categories, w_c is the category weight, $y_{i,c}$ is the true label of sample i, $\hat{y}_{i,c}$ is the predicted probability of sample i for category c, and the sum of predicted probabilities is 1 for $\sum_{c=1}^{3} \hat{y}_{i,c} = 1$.

For the object detection stream, the loss function is designed to maximize detection accuracy by aggregating five distinct sub-losses: objectness loss, localization loss, binary cross-entropy loss, classification loss, and L1 loss [23]. The total detection loss is expressed as:

$$L_{total} = \lambda_{obj-rpn}L_{obj-rpn} + \lambda_{loc-rpn}L_{loc-rpn} + \lambda_{iou-rpn}L_{iou-rpn} + \lambda_{reg}L_{reg} + \lambda_{cls}L_{cls} \quad (10)$$

In this formulation,$\lambda_{obj-rpn}, \lambda_{loc-rpn}, \lambda_{iou-rpn}, \lambda_{reg}, \lambda_{cls}$ are hyper parameters that balance the contributions of each loss component. This comprehensive loss function ensures that the model effectively balances scene recognition, object classification, localization, and bounding box regression, leading to improved detection performance in challenging underwater environments.

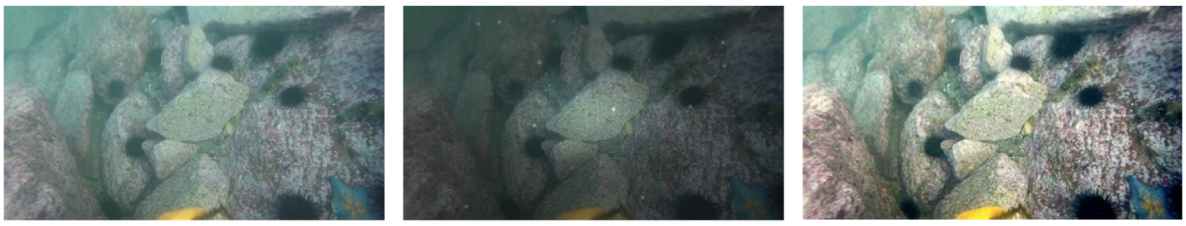
4. Experiment

4.1. Datasets

To comprehensively evaluate the effectiveness of the proposed framework, we conducted experiments on the UTDAC2020 dataset [21], a challenging benchmark for underwater object detection. This dataset comprises 5,168 training images and 1,293 validation images annotated across four categories: sea cucumbers, sea urchins, starfish, and scallops. A distinguishing feature of UTDAC2020 is its diverse resolution spectrum, which ranges from 586 × 480 to 3840 × 2160. This variability facilitates multi-scale evaluation across differing levels of detail, effectively simulating the heterogeneity of real-world underwater imaging conditions.

To further augment data diversity and simulate complex optical scenarios, we expanded the dataset using a radial masking technique combined with the Twin Adversarial Contrastive Learning (TACL) framework [18]. This process generated two distinct scene categories—non-uniform

illumination images and enhanced images—designed to mimic adverse conditions typical of deep-sea environments. Representative samples from the augmented dataset are presented in Fig. 7. These augmentation strategies yield realistic and diverse underwater scenarios, establishing a rigorous testbed for evaluating the robustness and generalization capability of the proposed method.



(a) (b) (c)

Fig 5. Examples of Augmented Underwater Images: (a) Uniformly illuminated image (b) Non-uniform image generated using the radial masking technique to simulate deep-sea artificial imaging features (c) Enhanced image generated using TACL [18] to simulate shallow-sea conditions, specifically to improve detector performance

4.2. Experimental Details

Our approach is implemented using the MMDetection framework [37]. Experiments were conducted on a workstation equipped with a single NVIDIA GeForce RTX 4090 GPU, using a batch size of 2 for both training and inference. We employed multi-scale training, resizing input images to a maximum dimension of 1300×800. The model was trained using Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of 10^{-4}, with no gradient clipping applied. We followed a standard $1 \times$ training schedule (12 epochs) with an initial learning rate of 0.005. A step decay learning rate policy was applied at epochs 8 and 11, alongside a linear warmup for the first 500 iterations with a ratio of 0.001. Unless otherwise specified, default hyperparameters and data processing pipelines from MMDetection were utilized.

Table 1. Three Scheme comparing

Configuration Dimension	Expert Detector 1 (Uniform Scenes)	Expert Detector 2 (Non-Uniform Scenes)	Expert Detector 3 (Enhanced Scenes)
Core Structure	ResNet-50 + DAFPN (ECA+STFM) + R-CNN	Identical structure	Identical structure
Training Data Subset	Uniform scene images (5168 images)	Non-uniform scene images (5168 images)	Enhanced scene images (5168 images)
Key Fine-tuned Parameters	ECA weight = 0.6; STFM fusion coefficient = 0.5	ECA weight = 0.8; STFM fusion coefficient = 0.6	ECA weight = 0.7; STFM fusion coefficient = 0.55

To ensure stable convergence and operational efficiency, DTC-Net follows a phased training protocol. Initially, the scene classification network is trained independently to establish high-precision background categorization. Subsequently, each expert detector—sharing a consistent architecture (ResNet-50 backbone, DAFPN equipped with ECA/STFM modules, and an R-CNN head)—undergoes independent optimization on its respective scene-specific dataset (5,168 images per category). As detailed in Table 1, while the structural topology remains identical, the parameters of these experts (e.g., ECA weights) are fine-tuned to capture the unique statistical characteristics of each scene. In the final inference phase, these pre-trained modules operate within a unified end-to-end framework: input images are automatically classified by the recognition network, which dynamically activates the appropriate expert detector to yield the final predictions.

Performance evaluation relies on the standard COCO-style Average Precision (AP) metrics. We report AP at Intersection over Union (IoU) thresholds of 0.5 (AP_{50}) and 0.75 (AP_{75}), as well as the mean AP (AP), which is calculated by averaging precision across ten IoU thresholds ranging from 0.50 to 0.95 in increments of 0.05.

4.3. Scene Classification Network

To demonstrate the reliability of the scene classification network, we present its performance metrics in Table 2, which reports an overall Top-1 accuracy of 98.63%, with an average Precision of 98.64% and an average Recall of 98.63%. Table 3 provides a granular breakdown by scene category: the Uniform class achieves a Precision of 98.10%, Recall of 99.92%, and F1-Score of 99.00%; the Non-Uniform class records 99.06%, 97.76%, and 98.40%, respectively; and the Enhanced class yields 98.76%, 98.22%, and 98.49%.

These quantitative results are corroborated by the confusion matrix in Fig. 6. The visualization confirms near-perfect classification for the Uniform domain (99.92% true positives) and robust identification for Non-Uniform (97.76%) and Enhanced (98.22%) scenarios, with negligible misclassification rates across boundaries. Collectively, these results validate the accuracy and robustness of the scene classification module, establishing a solid foundation for DTC-Net's adaptive detection capabilities. The specific impact of rare misclassification instances on downstream detector performance is further analyzed in Section 5 (Sensitivity Analysis).

Table 2. Three Scheme comparing

Evaluation Metric	Top-1 Accuracy	Top-3 Accuracy	Average Precision	Average Recall	Average F1-Score
Value	98.63	100.00	98.64	98.63	98.63

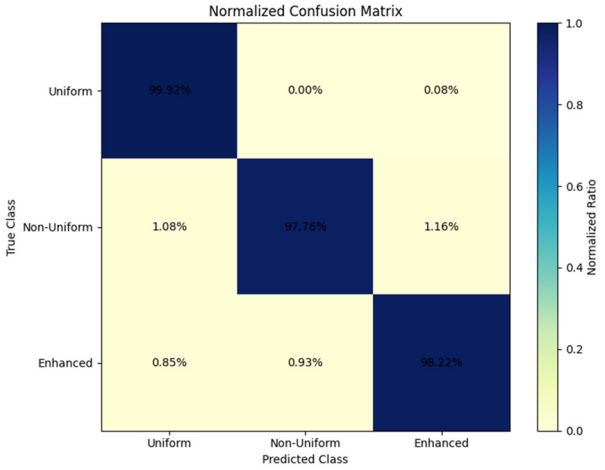


Fig 6. Normalized Confusion Matrix of Scene Classification.

4.4. Comparative Analysis

To ensure a fair comparison, all baseline models were retrained on the UTDAC2020 dataset using a unified hardware environment and a standard 12-epoch schedule, with the exception of D-FINE [45], which was trained for 160 epochs to ensure convergence. As summarized in Table V, DTC-Net demonstrates superior performance across all illumination settings. In the Uniform scenario, it achieves an AP_{50} of 83.2% and an AP of 48.1%, surpassing the competitive D-FINE by 0.6% in AP_{50} and 1.1% in AP. Furthermore, DTC-Net exhibits significant performance margins over other representative methods, outperforming Boosting-RCNN [23], Co-DETR [43], and FCOS [40] with AP gains of 1.5%, 3.6%, and 7.0%, respectively. Beyond standard conditions, the model shows robust generalization, securing the highest AP_{50} scores in Non-uniform (79.7%) and Enhanced (80.6%) settings. In terms of computational efficiency, DTC-Net strikes an optimal balance between accuracy and cost; it requires only 208.1 G FLOPs, significantly fewer than heavy architectures like Co-DETR (267.7 G) and UDMDet (234.5 G), validating its suitability for practical underwater deployment.

To further validate the robustness of our model, we present a visual comparison of detection results in Fig. 7. The figure spans three challenging underwater scenarios: low-light non-uniform backgrounds, complex textured sediments, and green-hazed waters. The first row displays the Ground Truth, while subsequent rows show predictions from representative baseline models (FCOS, Co-DETR, GCC-NET, D-FINE) and our proposed DTC-Net.

As indicated by the red arrows, baseline models exhibit distinct failure modes. FCOS [40] and Co-DETR [43] suffer from significant missed detections (False Negatives), particularly when small objects like sea urchins are camouflaged against the substrate or obscured by non-uniform lighting. GCC-NET [28] tends to produce redundant or imprecise bounding boxes in complex scenes, while D-FINE [45], despite its strong overall performance, occasionally misses subtle targets in low-contrast regions.

In contrast, DTC-Net demonstrates superior adaptability and precision. Guided by the scene identification module, it successfully retrieves small, indistinct targets that are missed by other detectors. The detection results of DTC-Net are highly consistent with the Ground Truth across all tested environments, confirming its effectiveness in handling color distortion, blurring, and scale variations in real-world underwater imagery.

Table 3. Results on the UTDAC2020 Dataset. Bold text indicates the best performance in each column.

Methods	Uniform		Non-uniform		Enhanced		Params (M)	FLOPs (G)
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}		
VFNET[33]	44.7	80.4	41.0	75.5	41.9	77.1	32.4	189.1
TOOD[6]	43.9	79.6	40.0	73.8	41.0	76.2	31.9	184.4
Boosting-RCNN[23]	46.6	81.7	42.9	76.5	43.9	78.4	45.9	160.3
UDMDet [29]	44.2	80.2	40.3	75.2	41.3	77.0	70.4	234.5
GCC_NET[28]	42.5	79.2	38.3	73.1	39.5	75.6	38.3	94.6
Co-DETR[43]	44.5	80.5	40.9	75.8	42.1	78.2	64.1	267.7
D-FINE[45]	47.0	82.6	42.7	77.2	44.1	79.5	3.7	7.1
FCOS[40]	41.1	77.8	37.6	73.0	38.5	74.2	32.2	202.1
LHDP[27]	45.0	78.6	39.0	70.6	40.4	73.7	75.5	195.8
DTC-NET(mix, our)	47.0	82.1	43.4	77.4	44.3	79.1	/	/
DTC-NET(our)	**48.1**	**83.2**	**44.8**	**79.7**	**45.5**	**80.6**	82.2	208.1

Table 4. Results on the Uniformly Illuminated Dataset. Bold text indicates the best performance in each column.

Methods	AP	AP_{50}	AP_{75}	$echinus$	$starfish$	$holothurian$	$scallop$
VFNET[33]	44.7	80.4	45.2	45.8	51.3	36.2	45.5
TOOD[6]	43.9	79.6	44.5	45.9	50.3	34.6	44.6
Boosting-RCNN[23]	46.6	81.7	48.6	47.9	52.1	38.4	47.7
UDMDet [29]	44.2	80.2	44.9	45.2	50.5	35.1	46.0
GCC_NET[28]	42.5	79.2	40.9	44.7	49.5	33.6	42.4
Co-DETR[43]	44.5	80.5	44.9	44.8	52.2	35.4	45.5
D-FINE[45]	47.0	82.6	48.3	48.1	52.7	38.2	48.5
FCOS[40]	41.1	77.8	39.1	43.3	48.1	32.0	41.0
LHDP[27]	45.0	78.6	46.6	46.1	51.4	37.3	45.2
DTC-NET(mix, our)	47.0	82.1	49.3	48.2	53.1	38.6	47.9
DTC-NET(our)	**48.1**	**83.2**	**51.2**	**49.2**	**53.9**	**39.2**	**49.6**

4.5. Ablation Experiment

1) Component 1: Class-Guided Cascade (CGC) Detector

To validate the effectiveness and generalizability of our proposed Class-Guided Cascade (CGC) strategy, we evaluated it as the core of the DTC-NET framework by integrating it with three distinct baseline detectors: GCC-Net, Boosting-RCNN, and UDMDet. As shown in the Table IX, the results demonstrate that our strategy yields consistent performance improvements across all models:

Table 5. Results on the Non-Uniformly Illuminated Dataset. Bold text indicates the best performance in each column.

Methods	*AP*	AP_{50}	AP_{75}	*echinus*	*starfish*	*holothurian*	*scallop*
VFNET[33]	41.0	75.5	40.4	43.7	47.6	30.7	42.2
TOOD[6]	40.0	73.8	39.4	43.1	46.5	28.8	41.8
Boosting-RCNN[23]	42.9	76.5	44.0	45.7	48.6	32.6	44.8
UDMDet [29]	40.3	75.2	39.2	42.8	46.4	29.9	42.2
GCC_NET[28]	38.3	73.1	35.4	41.8	45.8	27.3	38.3
Co-DETR[43]	40.9	75.8	39.3	43.1	48.8	29.0	42.7
D-FINE[45]	42.7	77.2	42.9	45.2	49.1	31.6	45.3
FCOS[40]	37.6	73.0	34.7	40.9	44.4	27.2	37.8
LHDP[27]	39.0	70.6	38.9	42.5	45.0	28.4	40.1
DTC-NET(mix, our)	43.4	77.4	44.3	46.1	49.5	32.8	45.3
DTC-NET(our)	**44.8**	**79.7**	**46.2**	**47.3**	**50.5**	**33.8**	**46.9**

Table 6. Results on the Enhanced Dataset. Bold text indicates the best performance in each column.

Methods	*AP*	AP_{50}	AP_{75}	*echinus*	*starfish*	*holothurian*	*scallop*
VFNET[33]	41.9	77.1	41.3	43.1	48.0	33.2	43.2
TOOD[6]	41.0	76.2	39.5	42.3	46.8	31.8	43.0
Boosting-RCNN[23]	43.9	78.4	44.6	44.7	49.1	35.5	46.1
UDMDet [29]	41.3	77.0	39.5	42.3	47.3	32.2	43.3
GCC_NET[28]	39.5	75.6	36.6	41.4	45.5	30.4	40.7
Co-DETR[43]	42.1	78.2	41.5	42.6	49.0	32.7	44.2
D-FINE[45]	44.1	79.5	44.3	45.4	49.4	35.8	45.9
FCOS[40]	38.5	74.2	35.3	40.6	44.5	29.8	39.2
LHDP[27]	40.4	73.7	40.4	41.7	46.5	32.1	41.5
DTC-NET(mix, our)	44.3	79.1	44.9	45.3	49.9	35.8	46.3
DTC-NET(our)	**45.5**	**80.6**	**47.1**	**46.6**	**50.8**	**36.1**	**48.6**

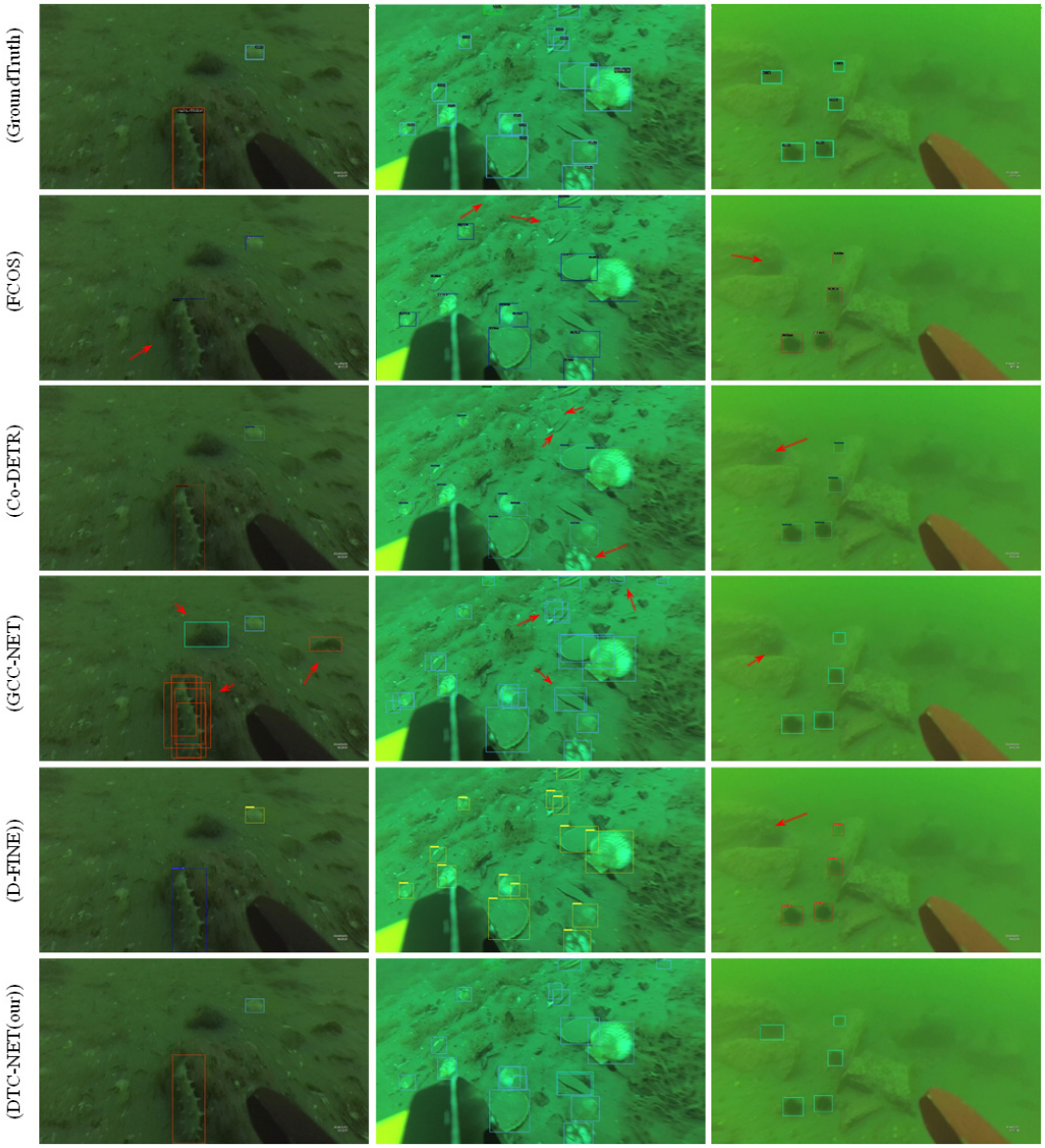


Fig 7. Visualization of detection results on challenging underwater scenes. The first row represents the Ground Truth, while the last row shows the results of our proposed DTC-NET. The intermediate rows display results from FCOS, Co-DETR, GCC-NET, and D-FINE, respectively. Red arrows indicate detection errors, such as missed detections (false negatives) or inaccurate localization. As observed, while comparison methods struggle with severe color casts, low contrast, and dense object occlusion, DTC-NET demonstrates superior robustness, generating high-quality bounding boxes that align closely with the Ground Truth

On GCC-Net: After applying our module, the AP score increased by 1.9 points (from 39.5 to 41.4), and the AP_{50} score rose by 1.6 points (from 75.6 to 77.2). On Boosting-RCNN: Our module boosted the AP by 1.3 points (from 43.9 to 45.2) and the AP_{50} by 1.8 points (from 78.4 to 80.2). On UDMDet: The integration achieved the best overall performance, reaching an AP of 45.5% and an AP_{50} of 80.6%, which represents an improvement of 0.9 and 1.0 points over the baseline, respectively.

Table 7. Ablation studies of proposed method DTC-NET combined with different detection heads. The UTDAC2020 enhanced dataset is used in this experiment.

Methods	AP	AP_{50}
GCC-Net	39.5	75.6
Boosting-RCNN	43.9	78.4
UDMDet	41.3	77.0
DTC-NET w/ GCC-Net	41.4(1.9)	77.2(1.6)
DTC-NET w/ Boosting-RCNN	**45.2(1.3)**	**80.2(1.8)**
DTC-NET w/ UDMDet	42.2(0.9)	78.0(1.0)

2) Component 2: Dual Attention Feature Pyramid Network (DAFPN)

The ablation study systematically validates the effectiveness of the proposed DTC-NET's key components—Efficient Channel Attention (ECA), Spatial-Temporal Feature Modulation (STFM)—in enhancing underwater object detection. As shown in the provided Table X, the baseline ResNet50+FPN model is progressively improved with the incremental addition of these modules; specifically, incorporating ECA refines feature discriminability, STFM significantly boosts small object recall (AR_S reaching 43.2).The complete DTC-NET (Ours), integrating all these innovations, achieves the highest performance across all metrics, with AP_{50} reaching 83.2 and strong gains in medium and large object detection (AP_M at 42.5, AP_L at 54.5), underscoring the synergistic benefits of these specialized modules for robust underwater object detection

Table 8. Ablation studies of proposed method DTC-NET. the UTDAC2020 dataset is used in this experiment.

Methods	ECA	STFM	DTC	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_S	AR_M	AR_L
ResNet50+FPN				46.6	82.0	48.6	21.3	40.6	53.0	38.2	55.3	62.6
1	✓			46.8	82.4	48.5	21.8	40.8	53.1	39.8	55.4	62.6
2	✓	✓		46.7	82.2	48.6	21.7	40.6	53.1	**43.2**	55.3	62.8
DTC-NET (Ours)	✓	✓	✓	**48.1**	**83.2**	**51.2**	**23.2**	**42.5**	**54.5**	40.6	**56.5**	**63.2**

5. Conclusion

In this work, we formulated a mathematical framework to quantify the impact of image background variations on underwater small-object detection. Through gradient-level analysis of background perturbations during loss function optimization, we demonstrated that such interference exerts a substantial influence on detection performance—an issue largely overlooked by existing approaches.

To address this challenge, we introduced the Scene Identification & Detection Dual-Task Cascade Network (DTC-Net), which integrates scene classification and object detection within a unified two-stage architecture. The framework incorporates an RPN error-based sample reweighting strategy and leverages expanded multi-scene datasets to enhance adaptability under diverse underwater conditions. Specifically, we constructed two additional datasets based on UTDAC2020: a simulated deep-sea non-uniform illumination dataset generated via the radial gradient mask technique, and an enhanced dataset produced through an object-guided dual-adversarial contrast learning method.

Extensive experimental evaluations verified that DTC-Net delivers significant improvements in detection accuracy, robustness, and generalization across challenging underwater scenarios. The proposed framework is inherently flexible, allowing for extension to additional recognition categories and substitution of the detection backbone for broader application domains. Nevertheless, the current model complexity remains relatively high; future work will focus on developing a lightweight variant of the architecture and exploring computationally efficient strategies tailored to real-time deployment in underwater environments.

References

[1] Igbinenikaro O P, Adekoya O O, Etukudoh E A. Emerging underwater survey technologies: a review and future outlook [J]. Open Access Research Journal of Science and Technology, 2024, 10(02): 071-084.

[2] Liu Z, Liu K, Chen X, et al. Deep-sea rock mechanics and mining technology: State of the art and perspectives[J]. International Journal of Mining Science and Technology, 2023, 33 (9): 1083-1115.

[3] Zhang B, Ji D, Liu S, et al. Autonomous underwater vehicle navigation: a review[J]. Ocean Engineering, 2023, 273: 113861.

[4] Chu S, Lin M, Li D, et al. Adaptive reward shaping based reinforcement learning for docking control of autonomous underwater vehicles[J]. Ocean Engineering, 2025, 318: 120139.

[5] Xu S, Zhang M, Song W, et al. A systematic review and analysis of deep learning-based underwater object detection[J]. Neurocomputing, 2023, 527: 204-232.

[6] Feng C, Zhong Y, Gao Y, et al. Tood: Task-aligned one-stage object detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society, 2021: 3490-3499.

[7] Zhao Y, Lv W, Xu S, et al. Detrs beat yolos on real-time object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 16965-16974.

[8] Sun P, Zhang R, Jiang Y, et al. Sparse r-cnn: End-to-end object detection with learnable proposals[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 14454-14463.

[9] Zhang H, Chang H, Ma B, et al. Dynamic R-CNN: Towards high quality object detection via dynamic training[C]// Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. Springer International Publishing, 2020: 260-275.

[10] Chen X, Yuan M, Yang Q, et al. Underwater-YCC: underwater object detection optimization algorithm based on YOLOv7[J]. Journal of Marine Science and Engineering, 2023, 11(5): 995.

[11] Chen, X., Fan, C., Shi, J., Wang, H., Yao, H., 2024. Underwater object detection and embedded deployment based on lightweight YOLO_GN. The Journal of Supercomputing 80, 14057–14084.

[12] X. Chen, X. Chen, F. Wu, H. Wang, H. Yao, Online_xkd: Anonline knowledge distillation model for underwater object detection, Computers and Electrical Engineering, 2024, 119:109501.

[13] Pang J, Liu W, Zhang B, et al. MCNet: Magnitude consistency network for domain adaptive object detection under inclement environments[J]. Pattern Recognition, 2024, 145: 109947.

[14] Wang J, Wan M, Xu Y, et al. Underwater Image Restoration via Constrained Color Compensation and Background Light Color Space-based Haze-Line Model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62:1-15.

[15] Li C, Guo C, Ren W, et al. An underwater image enhancement benchmark dataset and beyond[J]. IEEE transactions on image processing, 2019, 29: 4376-4389.

[16] Cong R, Yang W, Zhang W, et al. Pugan: Physical model-guided underwater image enhancement using GAN with dual-discriminators[J]. IEEE Transactions on Image Processing, 2023, 32: 4472-4485.

[17] Yeh C H, Lin C H, Kang L W, et al. Lightweight deep neural network for joint learning of underwater object detection and color conversion[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(11): 6129-6143.

[18] Liu R, Jiang Z, Yang S, et al. Twin adversarial contrastive learning for underwater image enhancement and beyond[J]. IEEE Transactions on Image Processing, 2022, 31: 4922-4936.

[19] Islam M J, Xia Y, Sattar J. Fast underwater image enhancement for improved visual perception[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 3227-3234.

[20] W. Lin, J. Zhong, S. Liu, T. Li, G. Li, Roimix: Proposal-fusion among multiple images for underwater object detection, in: ICASSP, 2020, 2588–2592.

[21] L. Chen, F. Zhou, S. Wang, J. Dong, N. Li, H. Ma, X. Wang, H. Zhou, Swipe-net: Object detection in noisy underwater scenes, Pattern Recognition, 2022, 132:108926.

[22] Jia J, Fu M, Liu X, et al. Underwater object detection based on improved Efficient-Det[J]. Remote Sensing, 2022, 14(18): 4487.

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

[24] Liu H, Song P, Ding R. Towards domain generalization in underwater object detection[C]//2020 IEEE international conference on image processing (ICIP). IEEE, 2020: 1971-1975.

[25] Chen Y, Song P, Liu H, et al. Achieving domain generalization for underwater object detection by domain mix-up and contrastive learning[J]. Neurocomputing, 2023, 528: 20-34.

[26] Liu K, Peng L, Tang S. Underwater object detection using TC-YOLO with attention mechanisms[J]. Sensors, 2023, 23(5): 2567.

[27] Fu C, Fan X, Xiao J, et al. Learning heavily-degraded prior for underwater object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(11): 6887-6896.

[28] Dai L, Liu H, Song P, et al. A gated cross-domain collaborative network for underwater object detection[J]. Pattern Recognition, 2024, 149: 110222.

[29] Yuan J, Cai Z, Cao W. A Novel Underwater Detection Method for Ambiguous Object Finding via Distraction Mining[J]. IEEE Transactions on Industrial Informatics, 2024, 20(1): 123-134.

[30] H. Liu, P. Song, et al., Towards domain generalization in underwater object detection, ICIP, 2020, 1971–1975.

[31] Y. Chen, P. Song, et al., Achieving domain generalization for underwater object detection by domain mixup and contrastive learning, Neurocomputing, 2023, 528(1):20-34.

[32] C. Liu, Z. Wang, et al., A new dataset, Poisson GAN and Aqua-net for underwater object grabbing, IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(5): 2831–2844.

[33] H. Zhang, Y. Wang, F. Dayoub and N. Sünderhauf. VarifocalNet: An IoU-aware Dense Object Detector. CVPR 2021, Nashville, TN, USA, 2021, 8510-8519.

[34] Haiyang Y, Ruige G, Zhongda Z, et al. U-TransCNN: A U-shape transformer-CNN fusion model for underwater image enhancement[J]. Displays, 2025, 88: 103047.

[35] K. He, R. Girshick, and P. Dollar, “Rethinking Imagenet pre-training,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 4917–4926.

[36] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, “You only look one-level feature,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit, 2021, pp. 13039–13048.

[37] K. Chen, J. Wang, et al., MMDetection: Open mmlab detection toolbox and benchmark, arXiv preprint arXiv:1906.07155 (2019).

[38] T.-Y. Lin, M. Maire, Microsoft coco: Common objects in context, in: ECCV, 2014, pp. 740–755.

[39] Chen X, Ge X, Yang Q, et al. A Novel Underwater Blurred Target Detection Algorithm Based on RT-DETR[J]. Concurrency and Computation: Practice and Experience, 2025, 37 (23-24): e70267.

[40] Tian Z, Shen C, Chen H, et al. FCOS: A simple and strong anchor-free object detector[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 44(4): 1922-1933.

[41] Wen J, Cui J, Zhao B, et al. EnYOLO: a real-time framework for domain-adaptive underwater object detection with image enhancement[C]//2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 12613-12619.

[42] Xu W, Wang C, Liang D, et al. NAUTILUS: A Large Multimodal Model for Underwater Scene Understanding[J]. arXiv preprint arXiv:2510.27481, 2025.

[43] Zong Z, Song G, Liu Y. Detrs with collaborative hybrid assignments training[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 6748-6758.

[44] An S, Xu L, Senior Member I, et al. HFM: A hybrid fusion method for underwater image enhancement[J]. Engineering Applications of Artificial Intelligence, 2024, 127: 107219.

[45] Peng Y, Li H, Wu P, et al. D-FINE: Redefine regression task in DETRs as fine-grained distribution refinement[J]. arXiv preprint arXiv:2410.13842, 2024.