Journal of Oceanology and Limnology

https://doi.org/10.1007/s00343-026-6192-3

Research Paper

Keypoint-guided supervision framework for underwater object detection*

Haoqiang ZHAO[1, 2, 3, 4, #], Zhantong LIANG[5, #], Wenchao MA[6], Wei WANG[1, 2, 3, 7], Xiang LI[1, 2, 3, **], Jiawei LIU[1, 2, 3, **]

[1] *Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110169, China*

[2] *Liaoning Liaohe Laboratory, Shenyang 110169, China*

[3] *Key Laboratory on Intelligent Detection and Equipment Technology of Liaoning Province, Shenyang 110169, China*

[4] *University of Chinese Academy of Sciences, Beijing 101408, China*

[5] *Dalian Maritime University, Dalian 116026, China*

[6] *Shenyang University of Technology, Shenyang 110870, China*

[7] *State Key Laboratory of Robotics and Intelligent Systems, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China*

Received Apr. 30, 2026; accepted in principle Jun. 16, 2026; accepted for publication Aug. 17, 2026

© Chinese Society for Oceanology and Limnology, Science Press and Springer-Verlag GmbH Germany, part of Springer Nature 2026

Abstract Many existing underwater object detection methods employ image enhancement to improve target visibility, but they often overlook two challenges: inconspicuous target structures and high target-background similarity caused by repetitive textures. To address these issues, we propose a keypoint-guided supervision framework, which leverages extracted structural keypoints to guide the detection model in focusing on target regions. Specifically, we design a map constrained progressive loss (MCP), which constructs coarse-to-fine structural maps to enhance global structural saliency in backbone features. To enhance foreground-background discriminability, structural prior keypoint supervision (SPK) is introduced to quantify the coverage of target-related and non-target keypoints as a supervisory signal. Extensive experiments on three public underwater datasets demonstrate the effectiveness of the proposed framework. The source code will be made publicly available at https://github.com/SIA-IDE/KG-Framework.

Keyword: deep learning; underwater object detection; scale-invariant feature transform (SIFT) keypoint; keypoint-guided supervision

1 INTRODUCTION

The ocean constitutes one of the most critical natural resource systems on Earth, not only harboring rich biodiversity but also playing an indispensable role in global climate regulation, biogeochemical cycles, and the maintenance of ecological balance (Malone et al., 2014; Danovaro et al., 2020). With the increasing intensity of marine exploration and exploitation, how to achieve long-term, continuous, and efficient observation of the marine environment has become a central issue in oceanographic research and resource management (Aguzzi et al., 2020; Mirimin et al., 2021). Particularly in complex and dynamic underwater environments, information such as the population structure of marine organisms, habitat variations, and the spatial distribution of seabed targets is essential for informed decision making in ecological assessment, resource conservation, and sustainable utilization (Er et al., 2023). The underwater environment itself constitutes a vast and intricate ecosystem, making automated perception and monitoring a highly challenging yet crucial task (Li et al., 2021a; Dai et al., 2024a). Against this backdrop, underwater object detection, as a core task in underwater visual perception, plays a pivotal

* Supported by the China Postdoctoral Science Foundation (No. 2025M781669) and the Fundamental Research Project of SIA (No. 2025JC1K05)

** Corresponding authors: lixiang@sia.cn; liujiawei@sia.cn

# Haoqiang ZHAO and Zhantong LIANG contributed equally to this work and should be regarded as co-first authors.

role in localizing and recognizing marine organisms and other underwater targets (Jian et al., 2024). Vision-based underwater detection methods provide essential technical support for applications such as organism counting (Folkman et al., 2026), growth analysis (Zhao et al., 2025), and health assessment in aquaculture, while also enabling ecological inspection, species distribution statistics, and biodiversity monitoring in marine environmental protection (Ditria et al., 2022). For camera-equipped platforms such as autonomous underwater vehicles (AUVs) and remotely operated vehicles (ROVs), object detection further serves as a fundamental capability for environmental understanding (Li et al., 2021b), autonomous decision-making (Sun et al., 2022) and trajectory tracking (Zhang et al., 2026). Zhou et al. (2024) highlighted that underwater object detection technologies play a crucial role in supporting sustainable ecological research and resource management. Meanwhile, Fu et al. (2023) emphasized that, although the acquisition of underwater imagery is challenging and data collection is costly, the growing availability of such data provides unprecedented opportunities for advancing underwater object detection.

However, compared with object detection in general scenarios, underwater object detection (UOD) faces more severe challenges arising from environmental degradation (Islam et al., 2020). Chen et al. (2026) pointed out that, due to wavelength-dependent absorption, scattering, and refraction in water, underwater images often suffer from low contrast, color distortion, structural degradation, and loss of fine details. As a result, UOD remains one of the most challenging tasks in computer vision. Furthermore, complex backgrounds, such as sandy textures, reefs, and coral structures, along with suspended particles in the water, can lead to high texture similarity between targets and backgrounds, blurred boundaries, and reduced contrast. These factors significantly increase the risks of false positives and missed detections in underwater scenarios (Guo et al., 2025; Huang et al., 2025). Therefore, achieving robust detection under the combined effects of low contrast, color shifts, and complex backgrounds remains a fundamental challenge for UOD.

To address the above challenges, a major line of UOD research improves detection performance through image enhancement, as illustrated in Fig.1a. One line of work treats underwater image enhancement as a preprocessing step, aiming to improve input quality through color correction and contrast enhancement, thereby facilitating detection. For instance, Sarkar et al. (2024) proposed the UICE-MIRNet enhancement module for detection-

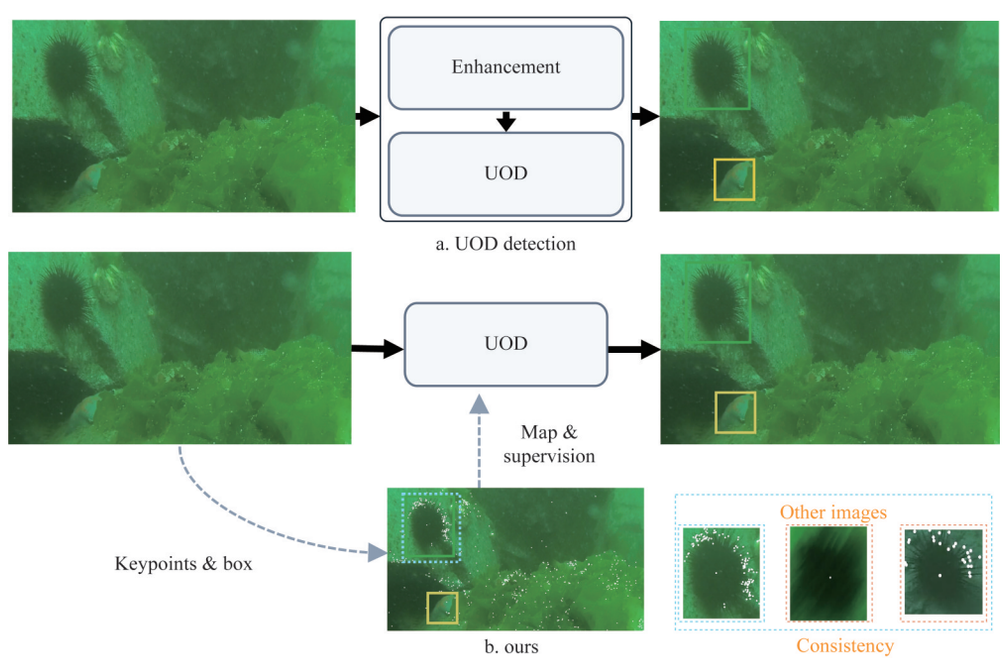


Fig.1 Comparison between image enhancement methods and the proposed approach

a. image enhancement is applied as a preprocessing step to underwater object detection (UOD); b. the proposed method extracts keypoints from the input image and leverages ground-truth (GT) bounding boxes as prior information to supervise the UOD. Similar object instances tend to exhibit comparable local structural responses.

oriented preprocessing, which extracts low-saturation regions from underwater images and performs color correction, effectively improving the visibility of small, numerous, and densely distributed objects. Liu et al. (2024b) proposed UnitModule, a plug-and-play approach that jointly optimizes image enhancement and object detection. In addition, some studies explore cross-domain information complementarity and data distribution compensation to mitigate domain shifts under varying imaging conditions. Dai et al. (2024b) argued that original images preserve natural scene characteristics and object texture details, while enhanced images improve the visibility of underwater objects. They address low visibility and low contrast challenges via cross-domain feature interaction and gated fusion. Overall, existing approaches have achieved encouraging progress in improving input quality, cross-domain fusion, and data distribution alignment. Recent multi-scale feature enhancement methods (Wang et al., 2025; Ye et al., 2025) improve underwater object representations through cross-scale feature interaction, but they mainly operate on feature-pyramid layers and pay limited attention to preserving structural cues during backbone encoding (Hua et al., 2023; Ji et al., 2024; Feng and Jin, 2025). While these methods alleviate the adverse effects of image degradation to some extent, they still exhibit two notable limitations. First, general image enhancement-based approaches (Liu et al., 2024a; Lin et al., 2026b; Xie et al., 2026) primarily focus on improving overall visual quality and target visibility, while overlooking the inherently weak structural characteristics of underwater objects. Second, repetitive textures are pervasive in underwater environments, leading to high similarity between background and target regions. Such target-background similarity reduces the model's ability to distinguish objects from the background, causing background regions to produce target-like responses, which in turn results in false positives.

To alleviate these limitations, it is desirable to introduce auxiliary cues that are more sensitive to local structural variations. Scale-invariant feature transform (SIFT) keypoints (Lowe, 2004) provide such a cue by capturing relatively stable local responses, such as corners, edges, and texture changes, which often reflect structurally informative regions in underwater targets (as shown in Fig.2a). Based on this motivation, we propose a keypoint-guided supervision framework (as shown in Fig.1b) that introduces structural priors into both the backbone and the detector. Specifically, to strengthen the weak structural characteristics commonly observed in underwater environments, map constrained progressive loss (MCP) introduces coarse-to-fine structural supervision to enhance global saliency perception for inconspicuous targets. This supervision is realized through a progressive structural map (PSM), a structural prior constructed to provide hierarchical guidance from global responses to target-focused regions. To alleviate localization ambiguity caused by target-background similarity, structural prior keypoint supervision (SPK) guides the detector to attend to target-related structural cues. The main contributions of this paper are summarized as follows:

1) We propose map-constrained progressive

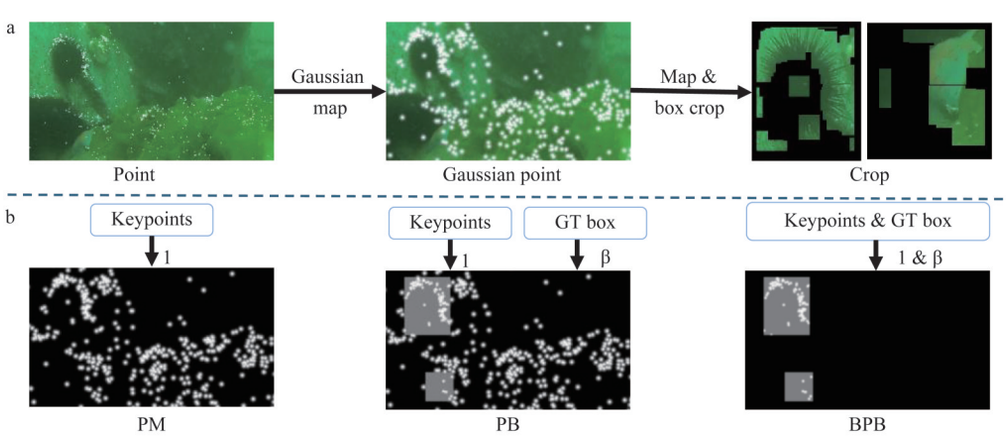


Fig.2 Illustration of SIFT keypoint extraction and map construction

a. discrete keypoints are transformed into a continuous region visualization via Gaussian diffusion; b. three progressive structural maps are constructed from keypoints and ground-truth bounding boxes. PM: gaussian points map; PB: points & box map; BPB: in-box points & box map.

loss, a coarse-to-fine map-constrained supervision mechanism that guides the model to learn structural saliency, improving the perception of visually inconspicuous underwater targets.

2) We introduce SPK with a keypoint coverage scoring mechanism to improve target-background discriminability and localization robustness.

3) Extensive experiments on public underwater datasets (DUO, URPC2020, ChinaMM2019UW) demonstrate the superiority of the proposed method.

2 METHOD

2.1 Overview

The overall framework of the proposed method is illustrated in Fig.3. First, SIFT keypoints are extracted from the input image, and a gaussian diffusion process is applied to transform these discrete keypoints into PSM as structural priors (see Section 2.2). Based on these priors, structural enhancement supervision is introduced into the detection model, including MCP and SPK. MCP incorporates coarse-to-fine progressive map loss into the multi-scale features (see Section 2.3). Meanwhile, we propose structure-aware supervision through target promotion and background suppression (see Section 2.4).

2.2 Prior generation

To capture local structural cues of underwater targets, we first extract SIFT keypoints from the original grayscale image and use only their spatial coordinates to construct structural priors. Building upon these keypoints and incorporating bounding-box region constraints, we further construct coarse-to-fine PSM:

$$P=\{(x_n,y_n)\}_{n=1}^{N}, \tag{1}$$

where N denotes the total number of extracted SIFT keypoints, (x_n, y_n) denotes the keypoint coordinates. These discrete keypoints are subsequently used as inputs to construct maps.

Map generation. Since discrete keypoints provide supervision only at a limited number of pixel locations, directly using them as supervision leads to an excessively sparse supervisory signal. By transforming discrete points into a continuous response map via gaussian diffusion, the supervision signal can be naturally propagated to the neighborhoods of keypoints, as shown in Fig.2a. This results in a smoother prior representation. Specifically, we apply gaussian diffusion to each keypoint on the image plane as follows:

$$H(x,y)=\max_{\substack{n\\ |x-x_n|\le 3\sigma\\ |y-y_n|\le 3\sigma}} e^{-\frac{(x-x_n)^2+(y-y_n)^2}{2\sigma^2}}, \tag{2}$$

where σ controls the spatial extent of the gaussian diffusion and is set to 8 in our implementation. $H(x, y)=0$ if no keypoint satisfies the above spatial constraints.

Subsequently, we incorporate ground truth bounding boxes (GT boxes) to provide additional explicit regional constraints, as illustrated in Fig.2b. Let the set of ground-truth bounding boxes B be defined as:

$$B=\{b_m\}_{m=1}^{n_b},$$
$$b_m=(c_x^m,c_y^m,w^m,h^m), \tag{3}$$

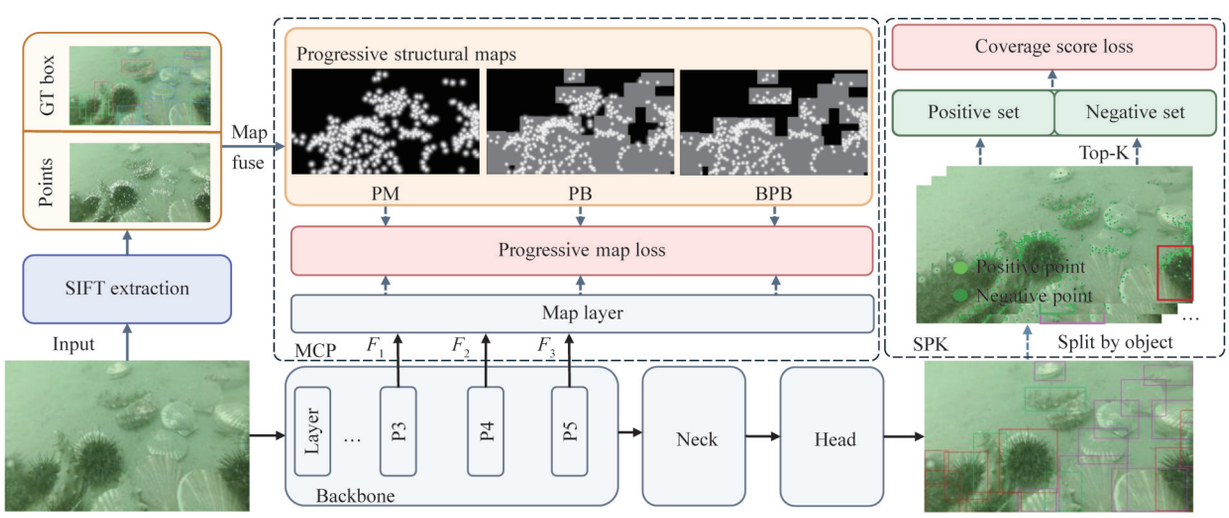


Fig.3 Illustration of the proposed method

The proposed method introduces two losses, namely MCP and SPK. PM: gaussian points map; PB: points & box map; BPB: in-box points & box map.

where c_x^m, c_y^m denote the center coordinates, w^{m}, h^{m} denote the size of the GT boxes and n_{b} denotes the total number of the GT boxes. We construct a box map:

$$M_{\mathrm{b}}(x,y)=\begin{cases}1 & \exists b_m\in B,(x,y)\in b_m\\ 0 & \text{otherwise}\end{cases}. \tag{4}$$

Progressive structural maps. We then construct three PSMs, as shown in Fig.3. The Gaussian Points Map (PM) relies solely on the gaussian priors derived from keypoints and is defined as follows:

$$\mathrm{PM}(x,y)=H(x,y). \tag{5}$$

PM can be regarded as pure point prior supervision, which captures salient structural information across the entire image, including both target and background regions.

Compared with PM, PB (Points & Box Map) further incorporates M_b introducing additional structural constraints within the target region. It can be expressed as follows:

$$\mathrm{PB}(x,y)=\max(H(x,y),\beta M_b(x,y)), \tag{6}$$

where β is a hyperparameter, which is set to 0.5 in our implementation.

The in-box points & box map (BPB) further constrains the point priors to lie within GT boxes, yielding more focused structural supervision. H_{in} denotes the map generated exclusively from keypoints located inside the target boxes, BPB is defined as follows:

$$\mathrm{BPB}(x,y)=\max(H_{\mathrm{in}}(x,y),\beta M_b(x,y)). \tag{7}$$

Compared with PB, BPB eliminates all responses outside the bounding boxes, thereby placing greater emphasis on the structural consistency within target regions.

Overall, PSM forms a progressive hierarchy of structural priors, transitioning from global structural perception to the target-level structure. PSM are then used as hierarchical supervision signals for the backbone to progressively guide multi-scale feature learning from global structural perception to fine-grained intra-object representation.

2.3 Map-constrained progressive loss

Based on the PSM, we introduce MCP on the intermediate backbone features. MCP applies structural supervision over the whole image plane and progressively shifts the supervision toward intra-object key structures, forming a coarse-to-fine learning process.

Since the backbone feature is a high-dimensional semantic representation (multiple channels) rather than a map (one channel), we employ a lightweight map layer φ_l to project it into a map space that can be aligned with the structural prior:

$$\hat{M}_l=\varphi_l(F_l),\ l\in\{1,\ 2,\ 3\}, \tag{8}$$

where F_l denotes the backbone feature. Since the PSM do not share the same spatial resolution as F_l, we resize the maps to the corresponding resolutions:

$$\begin{aligned}&M_l=\mathrm{resize}(m),\\ &m\in\{\mathrm{PM},\ \mathrm{PB},\ \mathrm{BPB}\}.\end{aligned} \tag{9}$$

We supervise the map layer using a combination of binary cross-entropy (BCE) loss and Dice loss:

$$\mathcal{L}_{\mathrm{MCP}}=\frac{1}{3}\sum_{l=1}^{3}\omega_l\begin{pmatrix}\mathcal{L}_{\mathrm{BCE}}(\hat{M}_l,M_l)\\ +\mathcal{L}_{\mathrm{Dice}}(\hat{M}_l,M_l)\end{pmatrix}, \tag{10}$$

where ω_l, $l\in\{1,\ 2,\ 3\}$ denotes the loss weight at scale l.

2.4 Structural prior keypoint supervision

While MCP provides image-level structural supervision and progressively shifts its focus toward target regions, we further introduce SPK to distinguish foreground objects from confusing background. SPK constrains predicted boxes by measuring the coverage score of target-related and non-target keypoints.

We denote the set of keypoints inside the b_m as P_m^+, and the set of keypoints outside the b_m as P_m^-. To enable differentiable optimization of keypoint coverage, we define a soft distance score of a keypoint p_n with respect to the predicted bounding box $\hat{b}_{\mathrm{m}}$ as follows:

$$\begin{aligned}&s(p_n,\hat{b}_m)=\\ &\sigma_0\left(\frac{\frac{\hat{w}_m}{2}-|x-\hat{c}_{m,x}|}{\tau}\right)\cdot\sigma_0\left(\frac{\frac{\hat{h}_m}{2}-|y-\hat{c}_{m,y}|}{\tau}\right),\end{aligned} \tag{11}$$

where $\sigma_0(\cdot)$ denotes the sigmoid function, and τ controls the smoothness of the soft boundary. We define the positive coverage as follows:

$$C_m^+=\frac{1}{N_m}\sum_{p\in P_m^+}s(p,\hat{b}_m), \tag{12}$$

where N_m denotes the number of points within the b_m. C_m^+ quantifies the average coverage of keypoints within the b_m by the predicted box $\hat{b}_{\mathrm{m}}$.

Negative coverage as follows:

$$C_m^- = \frac{1}{K} \sum_{p \in \mathcal{T}_k(P_m^-)} s\left(p, \hat{b}_m\right), \tag{13}$$

where $\mathcal{T}_k$ denotes the set of top-K. C_m^- measures the response strength to keypoints outside the b_m. To mitigate the bias caused by the large number of background keypoints, we employ a $\mathcal{T}_k$ to select a subset of negative samples with the highest responses.

Based on the above definitions, the total SPK loss can be formulated as follows:

$$\mathcal{L}_{\text{SPK}} = \frac{1}{n_{\text{b}}} \sum_{m=1}^{n_{\text{b}}} \left(1 - C_m^+ + \gamma C_m^-\right), \tag{14}$$

where γ denotes the penalty coefficient.

Finally, the overall detection loss is composed of the bounding box regression loss $\mathcal{L}_{\text{box}}$, objectness loss $\mathcal{L}_{\text{obj}}$, classification loss $\mathcal{L}_{\text{cls}}$, SPK loss $\mathcal{L}_{\text{SPK}}$ and MCP loss $\mathcal{L}_{\text{MCP}}$:

$$\mathcal{L}_{\text{det}} = \lambda_{\text{b}}\mathcal{L}_{\text{box}} + \lambda_{\text{o}}\mathcal{L}_{\text{obj}} + \lambda_{\text{c}}\mathcal{L}_{\text{cls}} + \lambda_{\text{t}}\mathcal{L}_{\text{SPK}} + \lambda_{\text{m}}\mathcal{L}_{\text{MCP}}, \tag{15}$$

where λ_{b}, λ_{o}, λ_{c}, λ_{t}, λ_{m} denote the weights of the box regression, objectness, classification, SPK, and MCP losses, respectively.

Through the joint promotion of target-related responses and suppression of background responses, SPK introduces an explicit structural constraint into detector training, encouraging predictions to be supported by target-related structural cues and thereby reducing target-background confusion.

3 EXPERIMENT AND RESULT

3.1 Dataset

We evaluate the proposed method on three public underwater object detection datasets: DUO (Liu et al., 2021), URPC2020 and ChinaMM2019UW (Liu et al., 2020). DUO is a public dataset designed for underwater object detection, consisting of 6 671 training images and 1 111 test images. It covers four common underwater categories: holothurian, echinus, scallop, and starfish. URPC2020 contains 1 973 test images and includes the same four categories (holothurian, echinus, scallop, and starfish). In this work, only the test set is used to evaluate the generalization capability of the model. ChinaMM2019UW is a real-world underwater dataset comprising 2 071 training images and 676 testing images across three categories: sea cucumber, sea urchin, and scallop. This dataset exhibits representative underwater degradations, including blue-green color casts and haze.

3.2 Experiment setting and evaluation metric

Two experimental protocols are adopted for different groups of experiments, and their base training configurations are summarized in Table 1. The first protocol is used for the experiments reported in Tables 2 & 3, while the second protocol is adopted for the experiments in Table 4.

For the experiments in Tables 2 & 3, we validate the proposed method within the YOLOv7 (Wang et al., 2023) detection framework. The model is

Table 1 Base training configurations for different experimental protocols

Configuration	Tables 2–3 (YOLOv7 and ours)	Table 4
Input resolution	640×640	640×640
Training epochs	20	20
Batch size	16	32
Optimizer	SGD	SGD
Momentum	0.937	0.937
Initial learning rate	0.01	0.01
Final learning-rate factor	0.1	0.01
Weight decay	5×10^{-4}	5×10^{-4}
MixUp probability	0.15	0.0
Mosaic probability	1.0	1.0
λ_{b}	0.05	7.5
λ_{o}	0.7	0.7
λ_{c}	0.3	0.5

The settings in Tables 2–3 apply only to YOLOv7 and the proposed method. YOLOX was evaluated using the implementation provided by the UnitModule repository, while all other methods were re-implemented based on their official code bases and configurations.

Table 2 Comparison on the DUO dataset

Method	GFLOPs	Param	mAP@0.5	mAP	Echinus	Starfish	Holothurian	Scallop	*P*	*R*
YOLOXs	26.6	8.9M	82.2	62.0	72.0	71.8	61.0	43.3	–	–
YOLOv7	103.2	37.2M	88.1	68.1	72.6	75.0	67.2	57.5	84.3	81.5
UnitModule	30.7	9.0M	82.3	61.9	71.6	72.2	61.2	42.6	–	–
GCC-Net	148.3	38.3M	84.3	64.6	71.0	72.9	63.4	51.0	–	–
YOLOv7+ours	103.2	37.2M	**88.5**	**69.1**	**73.0**	**76.5**	**68.1**	**58.8**	**85.8**	**81.8**

The per-class columns report class-wise AP@0.5:0.95. – means no data. Bold shows the best results.

Table 3 Cross-dataset comparison on URPC2020

Method	GFLOPs	Param	mAP@0.5	mAP	Echinus	Starfish	Holothurian	Scallop	*P*	*R*
YOLOXs	26.6	8.9M	73.9	38.7	39.6	46.9	39.1	29.0	–	–
YOLOv7	103.2	37.2M	79.2	43.2	41.7	49.4	43.1	38.5	74.2	78.3
UnitModule	30.7	9.0M	74.0	38.5	39.2	46.7	38.2	29.9	–	–
GCC-Net	148.3	38.3M	74.4	40.4	40.5	48.2	39.6	32.9	–	–
YOLOv7+ours	103.2	37.2M	**79.3**	**43.3**	**41.9**	**49.8**	**43.5**	**38.6**	**74.9**	**79.1**

The per-class columns report class-wise AP@0.5:0.95. – means no data. Bold shows the best results.

Table 4 Performance comparison of the proposed method across different baselines

Method	GFLOPs	Param	DUO		URPC2020	
			mAP@0.5	mAP	mAP@0.5	mAP
YOLOv8	8.2	3.0M	81.5	60.1	71.7	37.3
YOLOv8+ours	8.2	3.0M	**82.4 (+0.9)**	**61.0 (+0.9)**	**72.9 (+1.2)**	**38.5 (+1.2)**
YOLOv9	7.9	2.0M	82.1	61.7	72.8	38.5
YOLOv9+ours	7.9	2.0M	**82.9 (+0.8)**	**63.5 (+1.8)**	**73.5 (+0.7)**	**39.6 (+1.1)**
YOLOv10	8.4	2.7M	80.2	59.3	69.9	36.0
YOLOv10+ours	8.4	2.7M	**80.6 (+0.4)**	**61.3 (+2.0)**	**70.6 (+0.7)**	**37.3 (+1.3)**
YOLOv11	6.4	2.6M	81.0	59.9	70.4	36.0
YOLOv11+ours	6.4	2.6M	**82.2 (+1.2)**	**62.3 (+2.4)**	**72.6 (+2.2)**	**38.7 (+2.7)**
RT-DETR	108.3	33.0M	75.0	54.4	65.7	33.5
RT-DETR+ours	108.3	33.0M	**76.1 (+1.1)**	**55.6 (+1.2)**	**66.1 (+0.4)**	**33.8 (+0.3)**

Bold shows the better result within each baseline-ours pair. Values in parentheses denote performance improvements over the corresponding baseline.

initialized from the COCO-pretrained YOLOv7 weights. All layers are fine-tuned end-to-end without freezing any parameters. For the proposed MCP, the loss weights ω_1, ω_2, ω_3 are set to 0.6, 0.9, and 1.5. For the proposed SPK, the penalty coefficient γ is 0.5, the smoothing parameter τ is set to 0.04 and a top-K strategy with K=16 is adopted to select negative keypoints with the highest soft coverage values. The overall weights of the SPK loss λ_t and MCP loss λ_m are set to 10^{-4}, 0.05. For the experiments reported in Table 4, all compared YOLO variants (Wang et al., 2024a, b) are implemented within the YOLO-Master (Lin et al., 2026a) repository and the basic configuration is shown in Table 1. Each model is initialized with its corresponding weights pretrained on the COCO dataset before training. For the proposed MCP, ω_1, ω_2, ω_3 are set to 0.6, 0.9, and 1.5. For the SPK, γ is 0.5, τ is set to 0.04. The overall weights of the SPK loss λ_t and MCP loss λ_m are set to 0.05, 0.08. For the experiments in Table 4, all experimental results of comparison methods are re-implemented using the same configurations.

Following common practice in underwater object detection, we evaluate the proposed method using mAP@0.5, mAP, precision (*P*), and recall (*R*). mAP@0.5 is calculated at an IoU threshold of 0.5, while mAP follows the COCO-style evaluation protocol by averaging AP over IoU thresholds from 0.5 to 0.95 with a step of 0.05.

3.3 Comparison

3.3.1 Result on DUO

Table 2 presents the detection performance of different methods on the DUO dataset. Compared with YOLOX (Ge et al., 2021), UnitModule (Liu et al., 2024b), and GCC-Net (Dai et al., 2024b), the proposed method achieves superior results in both mAP@0.5 and mAP. Specifically, our method attains 88.5% mAP@0.5 and 69.1% mAP on DUO, representing improvements of 0.4% and 1.0%, respectively, over the baseline YOLOv7. In terms of per-class AP, the proposed method consistently outperforms YOLOv7 across all four categories: echinus, starfish, holothurian, and scallop. These results demonstrate that the introduced structural priors provide stable performance gains across

different types of underwater targets. Furthermore, compared with the baseline, the gains in precision and recall demonstrate the method's reliability.

3.3.2 Result on URPC2020

To further evaluate the generalization capability of the model, we train the model on DUO and directly test it on the URPC2020 dataset, shown in Table 3. The proposed method achieves 79.3% mAP@0.5 and 43.3% mAP on URPC2020, yielding improvements of 0.1% and 0.1%, respectively, over the baseline YOLOv7. Although the overall gains in cross-dataset evaluation are relatively modest, the proposed method consistently improves performance across all four categories: echinus, starfish, holothurian, and scallop. These results indicate that the keypoint-guided structural priors exhibit a certain degree of generalization capability across different underwater environments.

3.3.3 Result on different detectors

To evaluate the generalizability of the proposed method across different detectors, we integrate it into several representative models, including YOLOv8, YOLOv9 (Wang et al., 2024b), YOLOv10 (Wang et al., 2024a), YOLOv11, RT-DETR (Zhao et al., 2024) for evaluation. As shown in Table 4, the proposed method consistently improves the performance of all baseline detectors, demonstrating its effective compatibility with different detection frameworks and good general applicability.

3.3.4 Result on ChinaMM2019UW

We conduct additional experiments on the ChinaMM2019UW dataset. As shown in Table 5, the proposed method consistently improves the performance of the corresponding baseline models.

Table 5 Results of different models on ChinaMM2019UW

Method	mAP@0.5 (%)	mAP (%)
YOLOv7	96.4	68.6
YOLOv7+Ours	**96.5 (+0.1)**	**69.2 (+0.6)**
YOLOv10	95.0	70.5
YOLOv10+Ours	**95.3 (+0.3)**	**70.9 (+0.4)**

Bold shows the better result within each baseline-ours pair. Values in parentheses denote performance improvements over the corresponding baseline.

Table 6 Ablation Study of MCP weights on the DUO dataset

ω_1	ω_2	ω_3	mAP@0.5 (%)	mAP (%)	P	R
0.6	0.9	1.5	**88.1**	**68.8**	86.8	**79.3**
1.0	1.0	1.0	**88.1**	68.6	**87.5**	79.1

Bold shows the best results.

Table 7 Ablation Study of SPK Weight on the DUO dataset

τ	λ_t (10^{-3})	γ	mAP@0.5 (%)	mAP (%)	P	R
0.02	0.1	0.5	88.1	68.8	86.8	79.3
0.02	0.5	0.5	88.2	68.4	**87.8**	79.8
0.02	0.1	0.25	88.1	68.6	86.1	80.4
0.02	0.1	1.0	74.0	51.3	80.5	65.2
0.01	0.1	0.5	88.3	68.9	86.8	80.1
0.04	0.1	0.5	**88.5**	**69.1**	85.8	**81.8**

Bold shows the best results.

Table 8 Ablation study of MCP and SPK

$\mathcal{L}_{MCP}$	$\mathcal{L}_{SPK}$	mAP@0.5 (%)	mAP (%)
–	–	88.1	68.1
√	–	**88.5**	68.6
–	√	88.3	68.6
√	√	**88.5**	**69.1**

– means the corresponding component is not used. √ means the corresponding component is used. Bold shows the best results.

3.4 Ablation study

3.4.1 Parameter ablation

We further evaluate the effectiveness of the hyperparameters. The experimental results are presented in Tables 6 & 7. In Table 6, the τ=0.02, λ_t=10^{-4}, γ=0.5. In Table 7, the ω_1=0.6, ω_2=0.9, ω_3=1.5. Table 7 shows that moderate settings of λ_t, and γ lead to better overall detection performance.

3.4.2 Effectiveness of different components

To evaluate the individual contributions of MCP and SPK, we conduct component-wise ablation experiments on the DUO dataset and the results are summarized in Table 8. Both MCP and SPK bring consistent improvements over the baseline, confirming the effectiveness of each component.

3.4.3 Backbone feature heatmap

To further validate the effect of our method on intermediate backbone representations, we visualize the activation heatmaps of several backbone layers in Fig.4. Compared with the baseline, our method exhibits stronger and more concentrated responses around the target regions across multiple stages. The improvements are observed not only in shallow features but also in deeper layers, suggesting that the proposed progressive supervision helps the

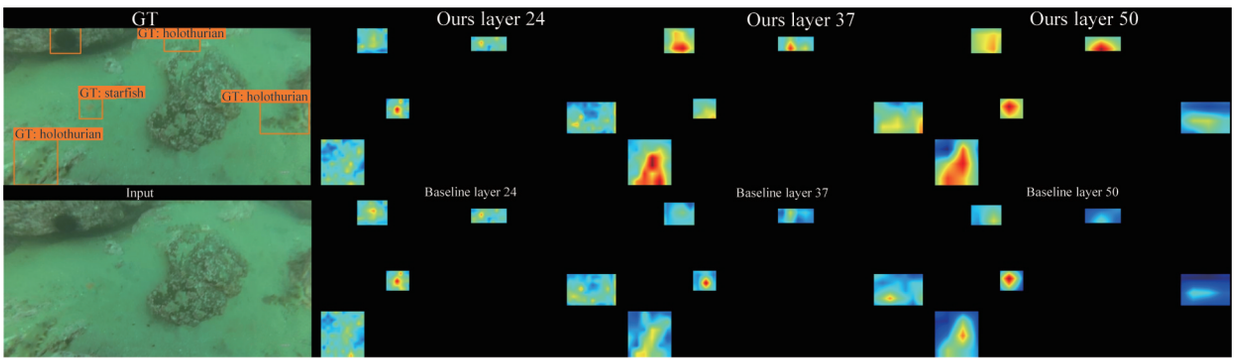


Fig.4 Visualization of backbone feature heatmaps for the baseline (YOLOv7) and the proposed method at different backbone layers

Our method yields stronger and more concentrated responses on target regions. GT: ground truth.

backbone preserve more discriminative structural cues throughout the feature extraction process.

3.4.4 Visualization

Figure 5 presents a qualitative comparison between the proposed method and several representative detectors on DUO dataset. Compared with YOLOv7, the predicted bounding boxes exhibit better alignment with the ground-truth annotations, leading to fewer missed detections and false positives. This observation is consistent with the design motivations of MCP and SPK, where MCP enhances structural saliency within target regions and SPK further improves the discriminability

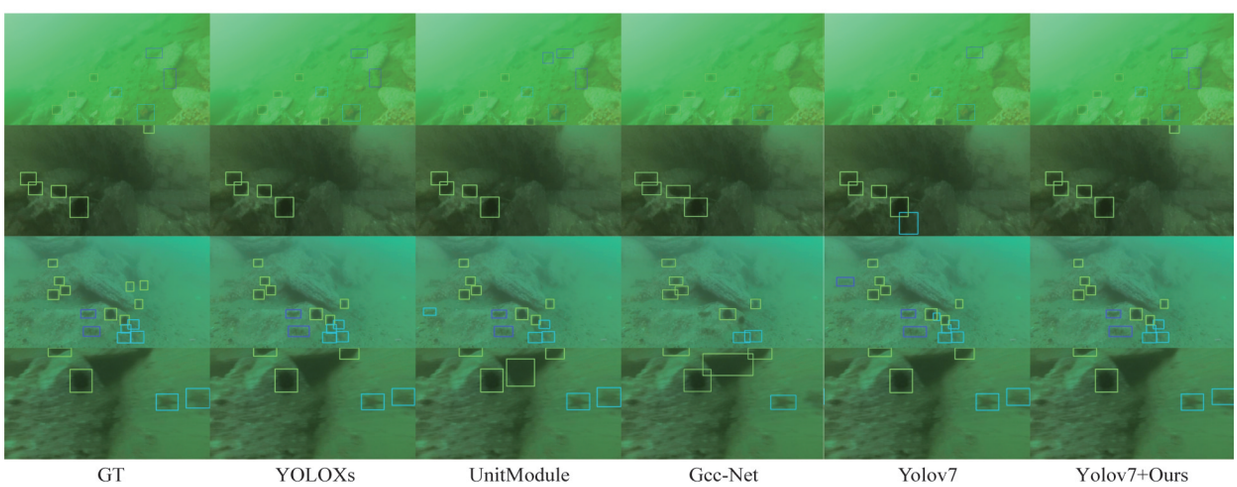


Fig.5 Qualitative comparison of different methods on the DUO dataset

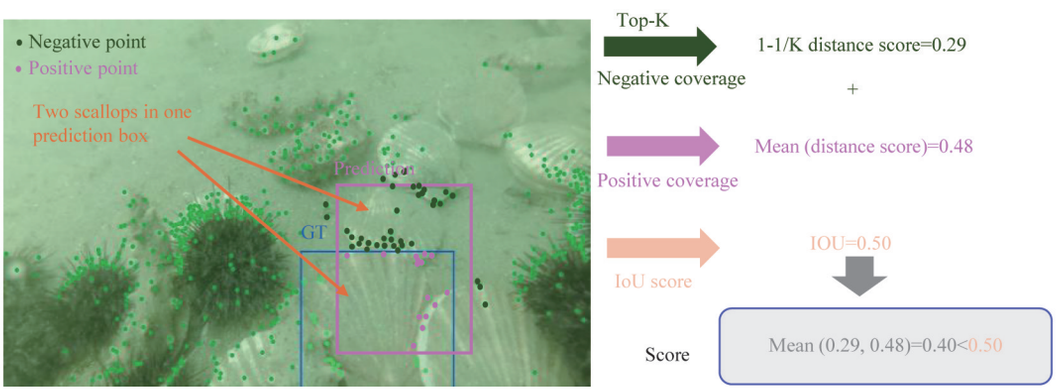


Fig.6 The visualization of SPK

SPK: structural prior keypoint supervision; IoU: intersection over union; GT: ground truth.

between targets and background. Our method achieves more reliable and consistent detection performance in complex underwater environments.

3.4.5 The visualization of SPK

As illustrated in Fig.6, SPK complements IoU-based box regression by introducing an additional structural constraint during training. Even when the IoU between a predicted box and its corresponding ground-truth box reaches 0.5, substantial local structural discrepancies may still exist, leading to structurally ambiguous or inaccurate localization. By evaluating the coverage of positive and negative keypoints by the predicted box, SPK can identify such structurally inconsistent predictions and impose additional structural penalties during training.

3.4.6 Visualization of different epochs

As shown in Fig.7, the proposed method converges rapidly during the early training stage, and most of the performance gains are obtained within the first 20 epochs. These results show that the proposed method has already delivered competitive detection performance.

4 DISCUSSION

MCP improves the structural saliency of target-related regions by progressively supervising multi-scale backbone features with coarse-to-fine structural maps. The visualization results of backbone heatmaps further support this effect, showing that the proposed method produces stronger and more concentrated activations around target regions than the baseline. This suggests that MCP helps the backbone preserve more discriminative structural cues, especially for inconspicuous underwater targets whose appearance has been weakened by low contrast, blur, and color distortion.

SPK further constrains whether a predicted box is supported by reliable local target structures. By measuring the coverage of target-related and non-target keypoints, SPK encourages candidate boxes to cover structurally meaningful foreground regions while suppressing responses to confusing background patterns. This role is particularly important in underwater scenes, where geometrically acceptable predictions may still correspond to poor detection results because they include multiple adjacent objects or extend into repetitive background regions. The visualization of SPK demonstrates that local structural evidence can effectively complement conventional IoU-based supervision and improve foreground-background discriminability at the detection stage.

5 CONCLUSION

This paper presents a keypoint-guided supervision framework for underwater object detection to address two critical challenges in underwater scenes, namely inconspicuous target

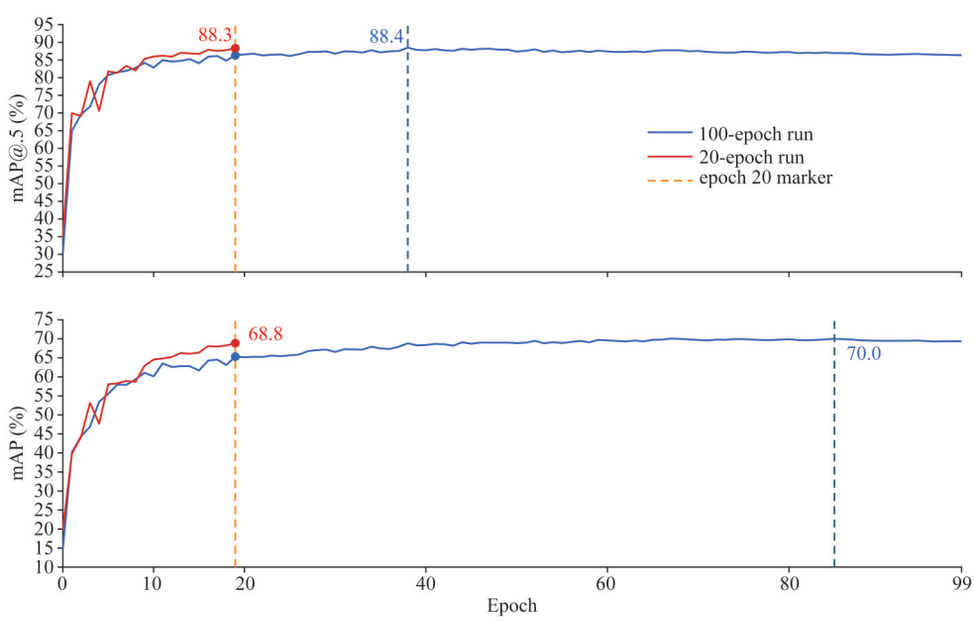


Fig.7 Comparison of training convergence between 20-epoch and 100-epoch settings on the DUO dataset

ω_1=0.8, ω_2=1.2, ω_3=1.0, τ=0.04. mAP: mean average precision.

structures and high target-background similarity caused by repetitive textures. Built upon the local structural information extracted from SIFT keypoints, the proposed approach constructs a structural supervision mechanism. Specifically, MCP enhances structural saliency through coarse-to-fine progressive maps, improving the representation of weak and inconspicuous underwater targets. In addition, SPK imposes local structural constraints on predicted bounding boxes by measuring the coverage of target-related and non-target keypoints, which helps the detector better distinguish objects from confusing background regions. By jointly exploiting global structural enhancement and local structural discrimination, the proposed method improves the model's structural awareness and detection reliability in complex underwater environments. Experimental results on public underwater datasets demonstrate that the proposed method achieves consistent improvements in detection accuracy and robustness. In future work, we will explore more lightweight strategies for constructing keypoint priors and investigate the applicability of such structural priors to a broader range of vision tasks.

6 DATA AVAILABILITY STATEMENT

The datasets used in this study are publicly available from their original sources. No new datasets were generated during this study.

References

Aguzzi J, Iveša N, Gelli M et al. 2020. Ecological video monitoring of Marine Protected Areas by underwater cabled surveillance cameras. *Marine Policy*, **119**: 104052, https://doi.org/10.1016/j.marpol.2020.104052.

Chen L, Huang Y Z, Dong J Y et al. 2026. Underwater optical object detection in the era of artificial intelligence: current, challenge, and future. *ACM Computing Surveys*, **58**(3): 62, https://doi.org/10.1145/3759243.

Dai L H, Liu H, Song P H et al. 2024a. Edge-guided representation learning for underwater object detection. *CAAI Transactions on Intelligence Technology*, **9**(5): 1078-1091, https://doi.org/10.1049/cit2.12325.

Dai L H, Liu H, Song P H et al. 2024b. A gated cross-domain collaborative network for underwater object detection. *Pattern Recognition*, **149**: 110222, https://doi.org/10.1016/j.patcog.2023.110222.

Danovaro R, Fanelli E, Aguzzi J et al. 2020. Ecological variables for developing a global deep-ocean monitoring and conservation strategy. *Nature Ecology & Evolution*, **4**(2): 181-192, https://doi.org/10.1038/s41559-019-1091-z.

Ditria E M, Buelow C A, Gonzalez-Rivero M et al. 2022. Artificial intelligence and automated monitoring for assisting conservation of marine ecosystems: a perspective. *Frontiers in Marine Science*, **9**: 918104, https://doi.org/10.3389/fmars.2022.918104.

Er M J, Chen J, Zhang Y N et al. 2023. Research challenges, recent advances, and popular datasets in deep learning-based underwater marine object detection: a review. *Sensors*, **23**(4): 1990, https://doi.org/10.3390/s23041990.

Feng J F, Jin T. 2025. LMFEN: lightweight multi-scale feature enhancement network for underwater object detection in AUVs. *Ocean Engineering*, **339**: 122109, https://doi.org/10.1016/j.oceaneng.2025.122109.

Folkman L, Vo Q L, Johnston C et al. 2026. A computer vision method to estimate ventilation rate of Atlantic salmon in sea fish farms. *Aquacultural Engineering*, **112**: 102645, https://doi.org/10.1016/j.aquaeng.2025.102645.

Fu C P, Liu R S, Fan X et al. 2023. Rethinking general underwater object detection: datasets, challenges, and solutions. *Neurocomputing*, **517**: 243-256, https://doi.org/10.1016/j.neucom.2022.10.039.

Ge Z, Liu S T, Wang F et al. 2021. YOLOX: exceeding YOLO series in 2021. arXiv preprint, arXiv: 2107.08430, https://doi.org/10.48550/arXiv.2107.08430.

Guo L J, Liu X C, Ye D S et al. 2025. Underwater object detection algorithm integrating image enhancement and deformable convolution. *Ecological Informatics*, **89**: 103185, https://doi.org/10.1016/j.ecoinf.2025.103185.

Hua X, Cui X P, Xu X H et al. 2023. Underwater object detection algorithm based on feature enhancement and progressive dynamic aggregation strategy. *Pattern Recognition*, **139**: 109511, https://doi.org/10.1016/j.patcog.2023.109511.

Huang Y Y, Huang J W, Huang M H. 2025. A lightweight YOLO network for robotic underwater biological detection. *Frontiers in Marine Science*, **12**: 1673437, https://doi.org/10.3389/fmars.2025.1673437.

Islam J, Xia Y Y, Sattar J. 2020. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters*, **5**(2): 3227-3234, https://doi.org/10.1109/LRA.2020.2974710.

Ji X, Chen S J, Hao L Y et al. 2024. FBDPN: CNN-Transformer hybrid feature boosting and differential pyramid network for underwater object detection. *Expert Systems with Applications*, **256**: 124978, https://doi.org/10.1016/j.eswa.2024.124978.

Jian M W, Yang N, Tao C et al. 2024. Underwater object detection and datasets: a survey. *Intelligent Marine Technology and Systems*, **2**(1): 9, https://doi.org/10.1007/s44295-024-00023-6.

Li C Y, Anwar S, Hou J H et al. 2021a. Underwater image enhancement via medium transmission-guided multi-color space embedding. *IEEE Transactions on Image Processing*, **30**: 4985-5000, https://doi.org/10.1109/TIP.2021.3076367.

Li J F, Xu Z, Zhu D J et al. 2021b. Bio-inspired intelligence with applications to robotics: a survey. *Intelligence & Robotics*, **1**(1): 58-83, https://doi.org/10.20517/ir.2021.08.

Lin X, Peng J L, Gan Z Y et al. 2026a. YOLO-Master: MOE-accelerated with specialized transformers for enhanced real-time detection. *In*: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. p.18440-18449.

Lin Z Y, Hou J H, Xia H Y et al. 2026b. Decoupled residual denoising diffusion models for unified and data efficient image-to-image translation. *In*: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. p.35967-35977.

Liu C W, Li H J, Wang S C et al. 2021. A dataset and benchmark of underwater object detection for robot picking. *In*: 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, Shenzhen, China. p.1-6, https://doi.org/10.1109/ICMEW53276.2021.9455997.

Liu J W, Wang Q, Fan H J et al. 2024a. Residual denoising diffusion models. *In*: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Seattle, USA. p.2773-2783, https://doi.org/10.1109/CVPR52733.2024.00268.

Liu R S, Fan X, Zhu M et al. 2020. Real-world underwater enhancement: challenges, benchmarks, and solutions under natural light. *IEEE Transactions on Circuits and Systems for Video Technology*, **30**(12): 4861-4875, https://doi.org/10.1109/TCSVT.2019.2963772.

Liu Z Y, Wang B, Li Y et al. 2024b. UnitModule: a lightweight joint image enhancement module for underwater object detection. *Pattern Recognition*, **151**: 110435, https://doi.org/10.1016/j.patcog.2024.110435.

Lowe D G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, **60**(2): 91-110, https://doi.org/10.1023/B:VISI.0000029664.99615.94.

Malone T C, DiGiacomo P M, Gonçalves E et al. 2014. A global ocean observing system framework for sustainable development. *Marine Policy*, **43**: 262-272, https://doi.org/10.1016/j.marpol.2013.06.008.

Mirimin L, Desmet S, Romero D L et al. 2021. Don't catch me if you can—using cabled observatories as multidisciplinary platforms for marine fish community monitoring: an *in situ* case study combining Underwater Video and environmental DNA data. *Science of the Total Environment*, **773**: 145351, https://doi.org/10.1016/j.scitotenv.2021.145351.

Sarkar P, De S, Gurung S et al. 2024. UICE-MIRNet guided image enhancement for underwater object detection. *Scientific Reports*, **14**(1): 22448, https://doi.org/10.1038/s41598-024-73243-9.

Sun B, Pang W, Chen M Z et al. 2022. Development and experimental verification of search and rescue ROV. *Intelligence & Robotics*, **2**(4): 355-370, https://doi.org/10.20517/ir.2022.23.

Wang A, Chen H, Liu L H et al. 2024a. YOLOv10: real-time end-to-end object detection. *In*: Proceedings of the 38th International Conference on Neural Information Processing Systems. Curran Associates Inc., Vancouver, Canada: 107984-108011, https://doi.org/10.52202/079017-3429.

Wang C Y, Bochkovskiy A, Liao H Y M. 2023. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *In*: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Vancouver, Canada, p.7464-7475, https://doi.org/10.1109/CVPR52729.2023.00721.

Wang C Y, Yeh I H, Liao H Y M. 2024b. YOLOv9: learning what you want to learn using programmable gradient information. *In*: Proceedings of the 18th European Conference on Computer Vision. Springer, Milan, Italy. p.1-21, https://doi.org/10.1007/978-3-031-72751-1_1.

Wang X, Li K, Fan F Y et al. 2025. Optimized YOLOv8s framework with deformable convolution for underwater object detection. *Scientific Reports*, **15**(1): 45446, https://doi.org/10.1038/s41598-025-28989-1.

Xie X, Liu J W, Lin Z Y et al. 2026. Unleashing the potential of large language models for text-to-image generation through autoregressive representation alignment. *In*: Proceedings of the 40th AAAI Conference on Artificial Intelligence. AAAI Press, Singapore, Singapore. p.11105-11113, https://doi.org/10.1609/aaai.v40i13.38089.

Ye Z F, Peng X, Li D K et al. 2025. Adaptive multi-scale attention networks fusion in YOLOv12-LSE: a lightweight framework for efficient and robust underwater object detection. *Applied Optics*, **64**(23): 6733-6745, https://doi.org/10.1364/AO.568505.

Zhang Y F, Zhu D Q, Chen M Z et al. 2026. Model predictive tracking control based on adaptive sliding mode constraints for unmanned underwater vehicles. *Intelligence & Robotics*, **6**(1): 101-119, https://doi.org/10.20517/ir.2026.06.

Zhao H C, Liu M, Ren Z W et al. 2025. Computer vision-based growth prediction and digestive tract assessment in pacific white shrimp (*Litopenaeus vannamei*). *Aquaculture Reports*, **45**: 103137, https://doi.org/10.1016/j.aqrep.2025.103137.

Zhao Y A, Lv W Y, Xu S L et al. 2024. DETRs beat YOLOs on real-time object detection. *In*: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Seattle, USA. p.16965-16974, https://doi.org/10.1109/CVPR52733.2024.01605.

Zhou H, Kong M W, Yuan H X et al. 2024. Real-time underwater object detection technology for complex underwater environments based on deep learning. *Ecological Informatics*, **82**: 102680, https://doi.org/10.1016/j.ecoinf.2024.102680.