frontiers | Frontiers in Marine Science

TYPE Technology and Code
PUBLISHED 20 November 2025
DOI 10.3389/fmars.2025.1697267

OPEN ACCESS

EDITED BY
David Alberto Salas de León,
National Autonomous University of Mexico, Mexico

REVIEWED BY
Hao Wang,
Laoshan National Laboratory, China
Yize Wang,
Waseda University, Japan

*CORRESPONDENCE
Xianpeng Shi
xpsh@ndsc.org.cn

RECEIVED 04 September 2025
REVISED 31 October 2025
ACCEPTED 04 November 2025
PUBLISHED 20 November 2025

CITATION
Chen D, Shi X, Liu M, Qiu S and Zhou Z (2025) Deep-sea organism detection method based on the SDA-HTransYOLOv8n model. *Front. Mar. Sci.* 12:1697267. doi: 10.3389/fmars.2025.1697267

Deep-sea organism detection method based on the SDA-HTransYOLOv8n model

Dali Chen[1,2], Xianpeng Shi[1,2]*, Meng Liu[1,2], Shaojian Qiu[3] and Zihan Zhou[3]

[1]National Deep Sea Center, Qingdao, China, [2]The College of Ocean Science and Engineering, Shandong University of Science and Technology, Qingdao, China, [3]College of Mathematics and Informatics, South China Agricultural University, Guangzhou, China

Deep-sea organism detection is one of the key technologies in deep-sea resource research and conservation. However, challenges such as low recognition accuracy and insufficient robustness arise due to issues like dim lighting, severe water scattering, and blurred target features in the deep-sea environment. To address these issues, this study proposes a deep-sea organism recognition method based on an improved SDA-HTransYOLOv8n model. The model introduces significant improvements to the neck network structure of YOLOv8n. First, it replaces the traditional upsampler with an improved point sampling dynamic sampler, which adaptively adjusts the sampling rate based on the target size, reducing redundant information interference and enhancing the efficiency of image feature extraction. Second, a Semantics and Detail Infusion module (SDI) is designed to adaptively fuse feature map information across different scales, addressing the issue of small deep-sea organisms being easily overlooked while enhancing the edge and detail features of deep-sea organisms. Third, a HyperTransformer-based HT_C2f module is designed to dynamically adjust attention weights, enhancing the model's ability to capture target organism features in complex deep-sea environments and improving sensitivity to blurry and low-contrast targets. Fourth, an improved downsampling convolution module (ADown) is introduced to reduce the dimension of feature maps while retaining more key feature information, avoiding feature loss in deep-sea organism images caused by information compression during sampling. Experimental results demonstrate that, on the deep-sea organism dataset obtained by the Jiaolong manned submersible in the western Pacific Ocean, the SDA-HTransYOLOv8n model developed in this study achieves a precision of 87.6%, a mAP50 of 67.7%, and a mAP50–95 of 51.6%, respectively, representing improvements of 8.9%, 2.8%, and 1.8% compared to the original YOLOv8n model, significantly enhancing the accuracy of deep-sea organism recognition. This study effectively meets the target detection requirements in complex deep-sea environments, providing technical support for deep-sea exploration and underwater operations. Code and models are available at https://github.com/Riokuli/SDA-HTransYOLOv8n-Model.

KEYWORDS

deep-sea organisms, SDA-HTransYOLOv8n, deep learning, image recognition, Jiaolong manned submersible

1 Introduction

The deep sea, as both a frontier for earth science research and a treasure trove of strategic resources, has elevated the capability to explore and develop it into one of the core indicators for gauging a country's scientific and technological prowess (Costa et al., 2020). As a pivotal technology for researching and conserving deep-sea biodiversity, underwater object detection faces numerous technical challenges. The unique characteristics of the deep-sea environment, including low illumination, light attenuation, color distortion, and low contrast, severely compromise the quality of underwater images and thereby impair the accuracy and robustness of existing detection algorithms (Xie et al., 2022). Furthermore, the scarcity of deep-sea image data and the exorbitant cost of image acquisition make data collection and annotation more arduous, further constraining the generalization capacity of models. Meanwhile, many existing detection models typically pursue higher accuracy at the expense of increased computational complexity, rendering them difficult to deploy effectively in resource-constrained deep-sea exploration systems. Therefore, how to boost detection accuracy while reducing model computational complexity and ensuring real-time performance and efficiency in deep-sea detection tasks has emerged as a critical issue demanding urgent resolution in this field.

In recent years, scholars have conducted extensive research on underwater biological detection algorithms. Current deep learning algorithms have revealed significant limitations when addressing the challenges of complex deep-sea environments: In traditional CNN models, the fixed receptive field design fails to dynamically adapt to targets of varying sizes and background distractors, leading to insufficient feature discrimination capability for deep-sea images (Han et al., 2020). YOLOv5 relies on the Focus structure and traditional convolution-dominated feature extraction, a reliance that cannot effectively address the blurred features of small targets caused by low-light conditions in deep seas (Kim et al., 2022). Although YOLOv8 is equipped with the C2f module and a dynamic detection head, it exhibits insufficient ability to distinguish low-contrast targets and incomplete capture of features of crustaceans with significant morphological variations (Wu and Dong, 2023). YOLOv10 centers on a decoupled detection head and Layer-wise Feature Aggregation, yet it struggles to adapt to the blurred feature hierarchy issue in highly turbid deep-sea environments (Hu et al., 2025). While YOLOv11 enhances detailed feature extraction through the Spatial Pyramid Pooling-Feature Pyramid Network enhancement module, it still adopts a local convolution-based core architecture, which prevents the complete capture of the global morphology of soft-bodied organisms (Cheng et al., 2025). Anchor-free models such as CenterNet rely on feature pyramids for key point regression; they show poor adaptability to inter-layer feature confusion induced by low light in deep seas, lack a mechanism for dynamically adjusting feature hierarchies, and are prone to key point localization deviations (Duan et al., 2019). R-CNN depends on preset anchor boxes, making it difficult to adapt to the diverse morphologies of deep-sea organisms and liable to misjudgment in dense scenarios (Bharati and Pramanik, 2019). The Region Proposal Network of Faster R-CNN generates candidate regions based on preset anchor boxes; it lacks an effective recognition mechanism for densely distributed similar features and tends to misclassify dense background distractors as target clusters (Sisodiya and Bhoite, 2025). Although Swin Transformer and Vision Transformer models possess global perception advantages, Swin Transformer uses a fixed 9×9 window for attention computation—this window size cannot be adaptively adjusted according to target size. Additionally, its static weight update mechanism fails to respond in real time to dynamic scene changes, reducing recognition stability (Wei et al., 2024). The self-attention mechanism of ViT models has not been optimized for special deep-sea environments (e.g., low light), resulting in insufficient perception capability for regions with weak features (Li et al., 2025).

In the latest underwater recognition research, scholars have overcome environmental constraints from the perspectives of data augmentation and cross-modal fusion. WaterCycleDiffusion is an underwater image enhancement method driven by vision-text fusion; it guides the diffusion model to generate enhanced images consistent with real scenarios via text descriptions, effectively mitigating the loss of image details under low light. However, this method does not undergo end-to-end joint optimization with downstream detection tasks, leading to a mismatch between the enhanced image features and the feature requirements of the detection network (Wang et al., 2025). The enhancement algorithm that combines histogram similarity-based color compensation with multi-attribute adjustment dynamically corrects color distortion and contrast attenuation of underwater images, improving the discriminability between targets and the background. Nevertheless, this method has insufficient adaptability to dynamic changes in water scattering coefficients, limiting its enhancement effect in highly turbid deep-sea regions (Wang et al., 2023). In research on integrated detection and tracking, the integrated detection and tracking paradigm for Compact High-Frequency Surface Wave Radar based on reinforcement learning optimizes the fusion strategy of radar data and visual data through reinforcement learning, enhancing the continuous tracking capability of dynamic targets. However, this paradigm is more suitable for long-range monitoring of medium-to-large marine organisms; it suffers from insufficient precision in close-range fine detection of small deep-sea organisms. Moreover, the modal differences between radar and visual features cause information redundancy and loss during the fusion process (Li et al., 2025) (Li et al., 2024).

To address these limitations, this paper proposes an enhanced detection model for deep-sea scenarios, SDA-HTransYOLOv8n (S: Semantics and Detail Infusion module; D: Dynamic Sampling Module; A: Adaptive Downsampling Module; HTrans: HyperTransformer Module; YOLOv8n: You Only Look Once 8n), with innovative breakthroughs in three dimensions:

1. Designing a cross-domain adaptable Transformer module that uses an environment-aware dynamic attention mechanism to achieve precise focusing on target features under low signal-to-noise ratio conditions;
2. Constructing a SDI multi-level feature fusion architecture, which enhances the consistency of multi-scale feature

information through dynamic scale alignment and cross-level feature product interaction, thereby improving the feature consistency of multi-scale target recognition;

3. Innovating the DySample dynamic sampling and ADown enhanced downsampling mechanisms, which enhance the retention rate of small target features while reducing the loss of critical information during the dimensionality reduction process.

Experimental results on the deep-sea biological dataset obtained by the Jiaolong manned submersible in the western Pacific Ocean indicate that the model achieves improvements of 8.9%, 2.8%, and 1.8% in precision and mean average precision (mAP50, mAP50-95) compared to the original YOLOv8n model, providing technical support for deep-sea resource exploration and ecological monitoring.

2 Method

2.1 YOLOv8n network structure

The YOLO series is celebrated for its exceptional efficiency and accuracy in object detection (Wang et al., 2022). The YOLOv8n model (Chen et al., 2025), building on the achievements of YOLOv5n, introduces significant improvements. Specifically, it replaces the conventional C3 module with the more sophisticated C2f module, thereby refining residual learning and facilitating improved gradient propagation via an optimized bottleneck module. Moreover, the model incorporates a novel image segmentation algorithm that synergistically combines deep learning with an adaptive threshold function (Deng et al., 2023), resulting in a lightweight framework that effectively captures gradient stream data (Shi and Wang, 2023). The input image is sequentially processed through multiple convolutional layers and C2f modules to extract feature maps at varying scales, which are then refined by an SPPF module prior to being forwarded to the detection head. This detection head seamlessly integrates anchor-free and decoupled-head strategies, while the loss function (Hu et al., 2024) leverages binary cross-entropy for classification alongside regression losses based on the CIOU and VFL. Additionally, the frame matching process has been improved with the Task-Aligned Assigner, further enhancing detection accuracy.

2.2 Improved YOLOv8n model— SDA-HTransYOLOv8n model

The SDA-HTransYOLOv8n model structure proposed in this paper is shown in Figure 1. Its core lies in a completely new improvement to the neck network of YOLOv8n, achieving a breakthrough in performance through the collaborative design of four key modules: First, the traditional sampler is replaced with an improved point sampling dynamic sampler (Liu et al., 2023). This module adaptively adjusts the sampling rate based on the size characteristics of the target, effectively filtering out redundant information interference while significantly enhancing the efficiency of image feature extraction; Second, an innovative multi-level feature fusion module (SDI) is constructed (Yang

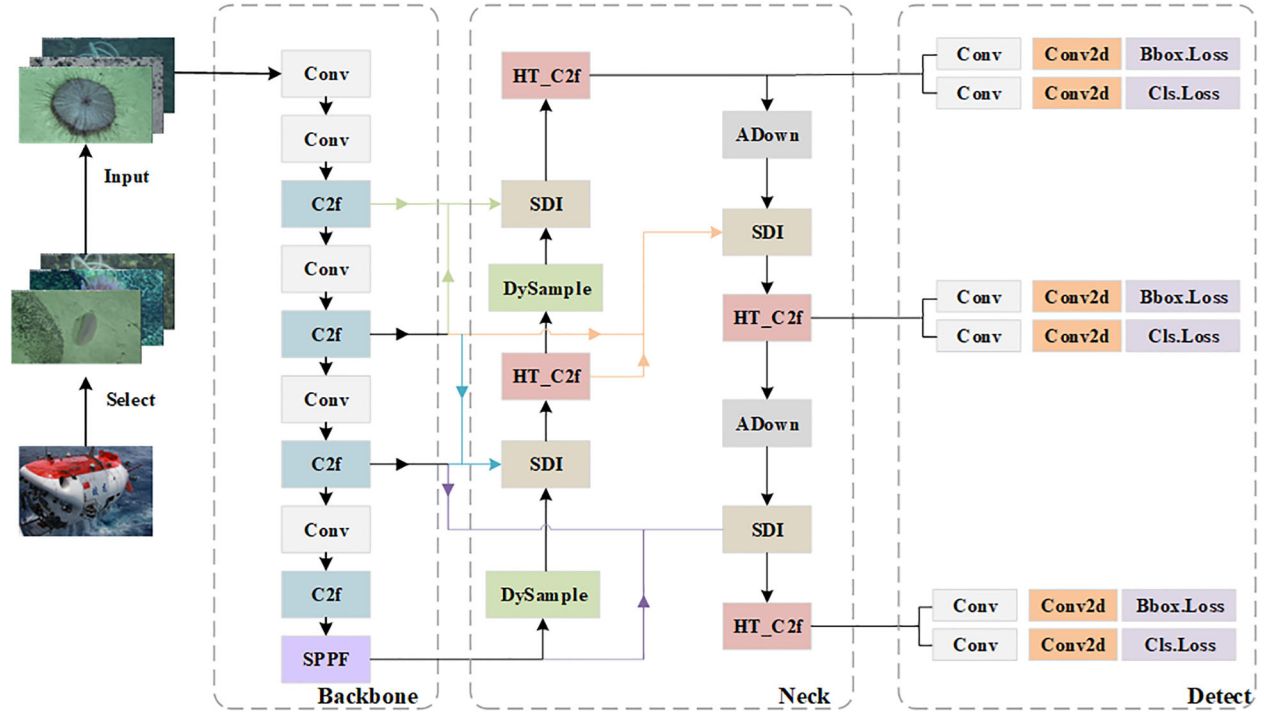


FIGURE 1
Overall structure diagram and basic modules of SDA-HTransYOLOv8n (All modules within the neck were designed by the authors).

et al., 2021), which introduces an adaptive fusion mechanism to enable deep interaction between feature maps of different scales, thereby enhancing the representation of target edges and detailed features; Third, a HyperTransformer-based HT_C2f module is designed (Dong et al., 2024). This module dynamically adjusts attention weight distributions to enhance the model's ability to capture target biological features in complex deep-sea environments, particularly improving sensitivity to blurry and low-contrast targets; Finally, an improved downsampling convolution module (ADown) is introduced (Wen et al., 2025). This module compresses feature map dimensions while retaining more critical feature information, effectively avoiding feature loss in deep-sea biological images caused by information compression during sampling. The organic integration of these improved modules significantly enhances the model's adaptability to extreme deep-sea environments, providing a reliable foundation for the efficient identification of diverse deep-sea organisms.

2.3 Dynamic sampling module

To address issues such as blurred features, diverse morphologies (e.g., posture distortion of soft-bodied organisms and blurred edges of transparent organisms) in deep-sea biological images, as well as low discriminability of local features caused by low illumination, this paper designs a dynamically adaptive upsampling module. This module achieves dynamic alignment and enhancement of features by learning pixel-level offsets, thereby effectively capturing the key features of deep-sea organisms.

The core of the DySample module is to predict offsets through 1×1 convolution, generate dynamic sampling coordinates in combination with the initial reference grid, and finally complete feature resampling via bilinear interpolation. Its overall structure is shown in Figure 2, and the specific working principles are as follows:

First, the offset layer predicts the base offsets. For the lp mode (from low resolution to high resolution), which is applicable to super-resolution feature enhancement of deep-sea images, the number of output channels is $2 \times groups \times scale^2$ (where 2 corresponds to offsets in x/y directions, *groups* is the number of groups, and *scale* is the sampling scaling factor). For the pl mode (from high resolution to low resolution), which is suitable for retaining key information during feature dimension reduction, the number of output channels is $2 \times groups$. Meanwhile, to avoid excessive offsets caused by deep-sea noise, an optional scope layer is set. When scope=True, an offset scaling factor is generated through sigmoid activation to dynamically control the offset amplitude. The offset formulas are shown in Equations (1) and (2).

$$p_0 = \{(\frac{i}{scale}, \frac{j}{scale}) | i,j \in [-\frac{scale-1}{2}, \frac{scale-1}{2}]\} \tag{1}$$

$$\Delta p = \begin{cases} (O(x) \cdot \delta(S(x))) \cdot 0.5 + p_0, (scope = True) \\ O(x) \cdot 0.25 + p_0, (scope = False) \end{cases} \tag{2}$$

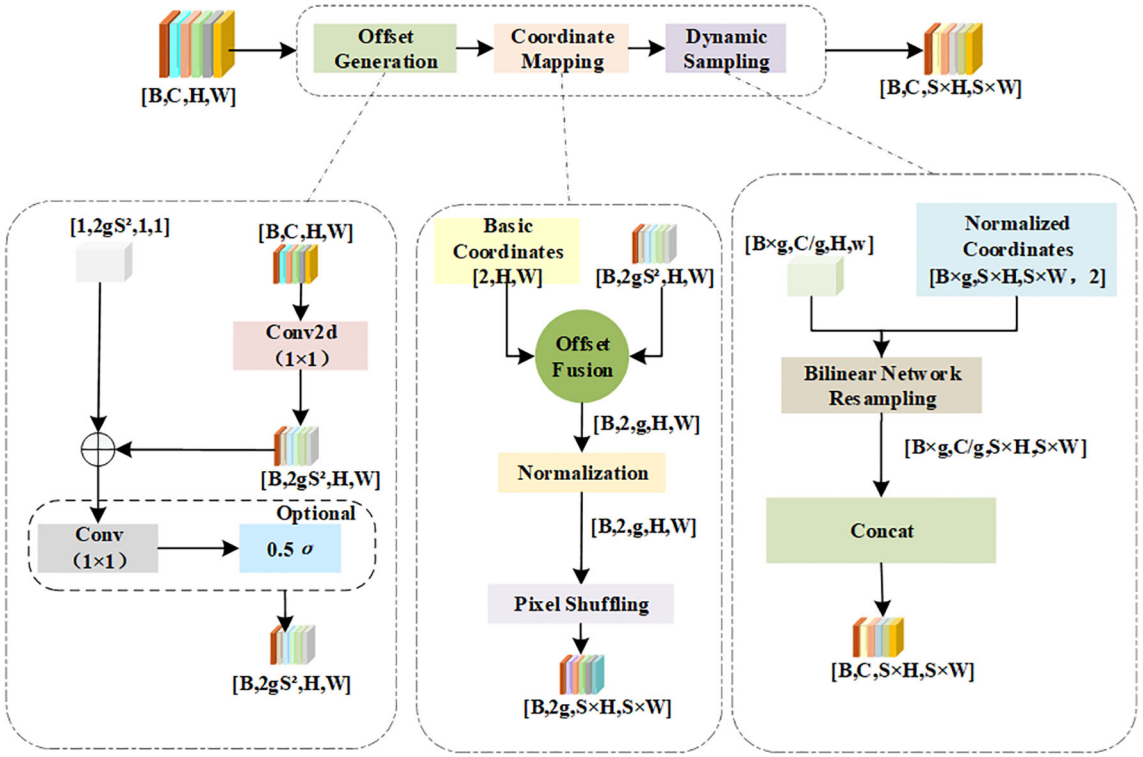


FIGURE 2
Structure diagram of dynamic sampling module.

Among them, p_0 is the initial reference grid; $O(x)$ is the output of the offset layer; $S(x)$ is the output of the scope layer; δ is the sigmoid function.

The dynamic sampling process is implemented through the sample function. Its core lies in fuse the predicted offsets with the original coordinates, generate normalized sampling coordinates, and then complete feature resampling via bilinear interpolation. The specific formulas are shown in Equations (3) and (4).

$$coords_{raw} = (x + 0.5, y + 0.5) \tag{3}$$

$$coords_{norm} = 2 \cdot \frac{coords_{raw} + \Delta p}{(W, H)} - 1 \tag{4}$$

Among them, $coords_{raw}$ is the generated original pixel center coordinate; (x, y) is the pixel index; (W, H) is the size of the input feature map.

The DySample module adaptively handles feature variations in deep-sea biological images through dynamic sampling by selectively employing two sampling modes. This module not only reduces computational load but also enhances feature diversity. It breaks through the limitations of traditional fixed-grid sampling, enabling sampling points to actively converge in high-information regions and improving the discriminability of feature representation.

2.4 SDI multi-level feature fusion module

As a spatial dimension interaction module, the SDI module can effectively address challenges such as low illumination and high noise in deep-sea biological image recognition through its unique multi-scale feature fusion mechanism, thus demonstrating significant advantages in deep-sea biological recognition tasks. The SDI module enhances the semantic and detailed information in images by integrating the hierarchical feature maps generated by the encoder. It specifically consists of three parts: feature extraction and integration of deep-sea biological images, fusion of high-level and low-level features at different levels, and feature transmission and segmentation. Its structure is shown in Figure 3.

To address the issue of low signal-to-noise ratio in deep-sea image features, the SDI module achieves effective integration through multi-scale feature extraction and noise suppression. The formula for multi-scale basic feature extraction are shown in Equations (5)–(7).

$$F_l = Backbone(I) \in R^{C_l \times H_l \times W_l}, (l \in 1, 2, \ldots, L) \tag{5}$$

$$H_l = \frac{H}{2^{l-1}} \tag{6}$$

$$W_l = \frac{W}{2^{l-1}} \tag{7}$$

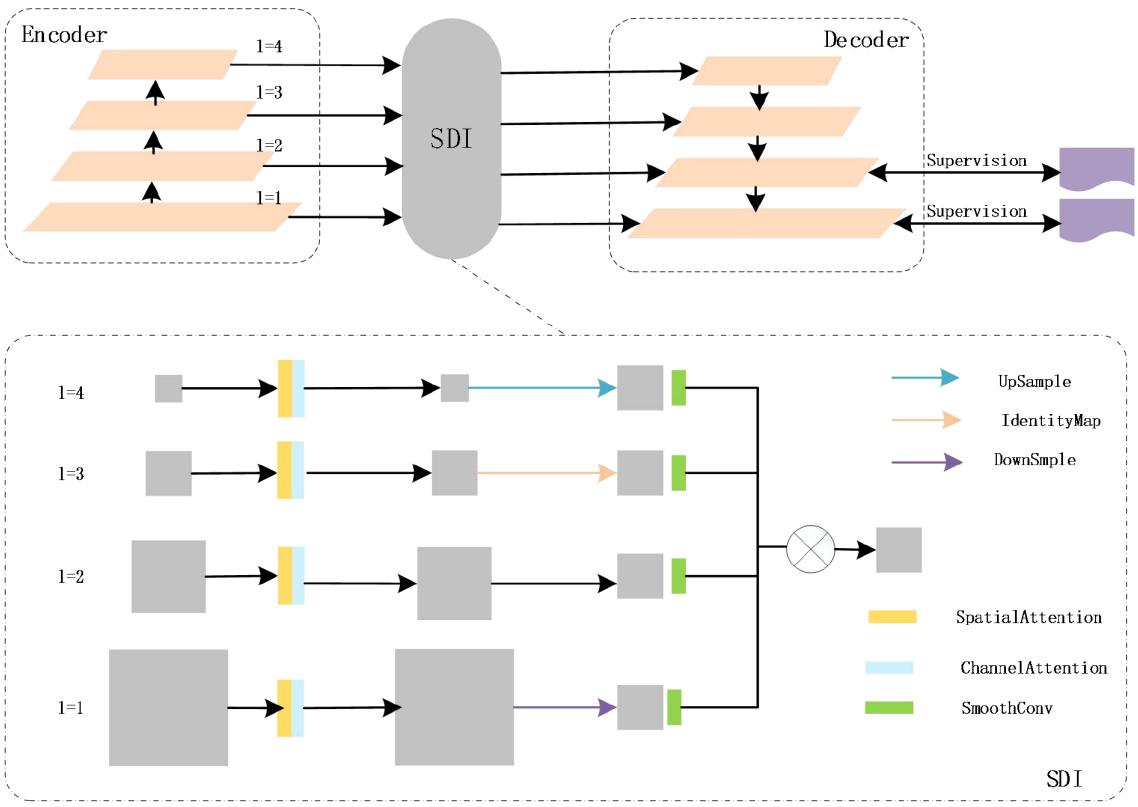


FIGURE 3
The structural diagram of the SDI multi - level feature fusion module.

Among them, $I \in R^{3\times H\times W}$ represents the original deep-sea image; C_l denotes the number of channels of the feature map at the l-th layer; F_1 stands for low-level features (such as edges and textures); F_L refers to high-level features (semantics and robustness).

The SDI module performs scale alignment and convolution purification on the extracted features. The integrated formulas are shown in Equations (8) and (9).

$$F_{int} = \prod_{l=1}^{L}(W_l * Align(F_l, H_l, W_l) + b_l) \tag{8}$$

$$Align(F_l, H_l, W_l) = \begin{cases} BI(F_l, H_l, W_l), & (l > 1, \text{upsampling}) \\ F_l, & (l = 1, \text{ baseline size}) \end{cases} \tag{9}$$

Among them, $Align(\cdot)$ is the scale alignment function; $W_l * (\cdot) + b_l$ refers to the noise filteringtemplate learned by the convolutional layer for noise filtering; $BI(\cdot)$ denotes bilinear interpolation.

Deep-sea biological recognition requires simultaneous consideration of low-level features and high-level features. SDI achieves the fusion of high-level and low-level features at different levels through hierarchical interaction, as shown in Equation (10).

$$F_{fusion} = \prod_{l=1}^{L}(\alpha_l \cdot (W_l * Align(F_l, H_l, W_l) + b_l)) \tag{10}$$

Among them, α_l satisfies $\sum \alpha_l = 1$. For small deep-sea organisms α_1 and α_2, it is increased (to enhance details); for large organisms α_{L-1} and α_L, it is increased (to enhance semantics).

In the deep-sea biological segmentation task, the features fused by SDI module need to guide pixel-level classification through a transmission mechanism. The fused features are transmitted to the original image size through F_{fusion} upsampling, as shown in Equation (11).

$$F_{seg} = BI(F_{fusion}, H, W) \in R^{3\times H\times W} \tag{11}$$

Then, for each pixel (i, j), its biological category k is predicted by the classifier, as shown in Equations (12) and (13).

$$P(k|i, j) = Soft\max(W_{seg}^{\star} F_{seg}[:, i, j] + b_{seg})_k \tag{12}$$

$$M(i, j) = \arg\max P(k|i, j) \tag{13}$$

Among them, W_{seg} is the convolution weight of the segmentation head; F_{seg} is the fused feature map transmitted to the segmentation head; b_{seg} is the bias term of the segmentation head; $M(i, j)$ represents the value of the segmentation mask at pixel (i, j); and arg maxis the index function corresponding to taking the maximum value.

Meanwhile, to improve the segmentation accuracy of low-illumination regions, a confidence-weighted loss based on SDI module features is introduced, as shown in Equation (14).

$$\varsigma = -\sum_{i,j} \log P(M^{\star}(i, j)|i, j) \cdot \frac{1}{Z}\sum_{l=1}^{L} \|Align(F_l, H, W)[:, i, j]\|_2 \tag{14}$$

Among them, $M^{\star}$ is the annotation mask; Z is the normalization constant.

The SDI module adapts to the low illumination and scale differences of deep-sea images through multi-scale product fusion. Its core formulas suppress noise during the feature extraction stage, balance details and semantics in the hierarchical fusion stage, and achieve accurate pixel classification through feature transmission in the segmentation stage, providing a mathematically rigorous solution for deep-sea biological recognition.

2.5 HT_C2f module based on HyperTransformer

Biometric recognition in deep-sea environments faces significant challenges such as low light levels, high noise, low contrast between targets and backgrounds, and complex spatial relationships. Traditional convolutional neural networks are limited by their local receptive fields and struggle to effectively capture global context dependencies, while pure Transformer architectures suffer from high computational costs and loss of local details. To address these issues, this paper introduces an HT_C2f module based on a hybrid architecture, inspired by the C2f structure in YOLOv8 (as shown in Figure 4). By integrating the local feature extraction capabilities of convolutional neural networks with the global modelling capabilities of Transformers, the module achieves efficient enhancement and modelling of target features in deep-sea images.

The core improvement of the HT_C2f module lies in the use of an "alternating replacement" strategy to reconstruct the feature processing chain, forming a hybrid feature interaction mode of "Conv-Bottleneck-Transformer." Among these, the HyperTransformer serves as the core enhancement unit, consisting of three key components: the Hyper Edge feature extraction submodule (Wazirali and Chaczko, 2016), the Transformer global modelling submodule (Wang et al., 2023), and feature fusion and residual connection (Fu et al., 2025).

The Hyper Edge feature extraction submodule captures local spatial features through 3×3 convolutions, introduces nonlinear transformations via the GELU activation function (Lee, 2023), and then compresses the feature dimensions to hyper_dim through 1×1 convolutions, thereby retaining key details while reducing the computational complexity of the Transformer. Its feature transformation process is shown in Equation (15).

$$hyper\_feat = Conv_{1\times1}(GELU(Conv_{3\times3}(x))) \tag{15}$$

Among them, $x \in R^{B\times C\times H\times W}$ represents the input feature; $hyper\_feat \in R^{B\times K\times H\times W}$ (K is $hyper_$dim) denotes the output low-dimensional local feature.

The Transformer global modeling submodule flattens the feature map output by the Hyper Edge into a sequence form (with the dimension converted to $HW \times B \times K$), and captures the global context dependencies through the Transformer encoder layer. The multi-head self-attention mechanism of Transformer is shown in Equations (16)–(18).

$$MultiHead(Q, K, V) = Concat(head_1, \cdots head_h)W^O \tag{16}$$

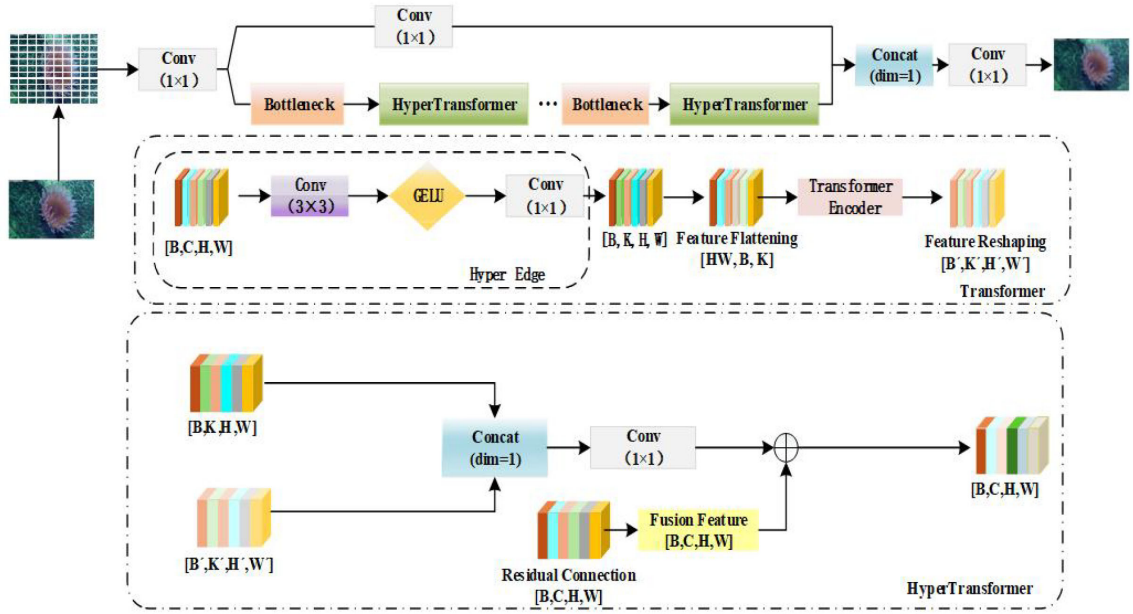


FIGURE 4
Structure diagram of the HT_C2f module based on HyperTransformer.

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{17}$$

$$Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{18}$$

Among them, $head_i$ denotes the output of the i-th attention head; W^O is the output projection matrix; W_i^Q, W_i^K, and W_i^V are the projection matrices of the i-th head, which project Q, K, and V into the low-dimensional subspace, respectively.

The feature sequence processed by the Transformer is reshaped back to the original spatial dimension, resulting in the globally enhanced feature $trans\_feat \in R^{B\times K\times H\times W}$.

The feature fusion and residual connection concatenate local feature $hyper\_feat$ and global feature $trans\_feat$ along the channel dimension (with the dimension being $B \times 2K \times H \times W$), compress them to the input dimension C through convolution 1×1, and perform residual fusion with the original input x, as shown in Equation (19).

$$output = x + Conv_{1\times1}(Concat(hyper\_feat, trans\_feat)) \tag{19}$$

In the HT_C2f module, the aforementioned HyperTransformer units and original Bottlenecks are arranged alternately, with each HyperTransformer processing only half of the channels ($C_2//2$). This design ensures global modeling capability while maintaining computational efficiency. In deep-sea biological recognition, this design can not only effectively extract detailed features of blurred targets but also model the correlation between targets and complex backgrounds through global attention, significantly enhancing the feature expression ability for complex deep-sea scenes.

2.6 Adaptive downsampling-ADown

The ADown module is a lightweight downsampling module designed for the characteristics of deep-sea environments such as low illumination, high noise, and blurred targets. This module adopts a dual-path feature parallel processing mechanism, which achieves 2x spatial downsampling while effectively preserving subtle features and edge information in deep-sea biological images. ADown module enhances the representation ability for weak textures and small targets through multi-scale feature fusion, and at the same time controls computational complexity to meet the requirements of practical deep-sea biological image recognition.

The structural design of ADown follows the logic of "feature divide-and-conquer -parallel enhancement - fusion output", which is specifically divided into three parts: input preprocessing, dual-path feature transformation, and feature fusion. Its structure is illustrated in Figure 5.

In the input preprocessing stage, first, average pooling with no size change is performed on the input feature map to suppress high-frequency noise in deep-sea images, as shown in Equation (20).

$$X' = AvgPool2d(X; k = 1, s = 1, p = 0) \tag{20}$$

Among them, the pooling kernel k = 1; the stride s = 1; the padding p = 0; and the output is $X' \in R^{b\times c_1\times h\times w}$.

Then, the input channels are split: X' is evenly divided into two branches along the channel dimension (dim=1) to achieve differentiated feature processing, as shown in Equation (21)

$$X_1, X_2 = torch.chunk(X', \text{chunks} = 2, \text{dim} = 1) \tag{21}$$

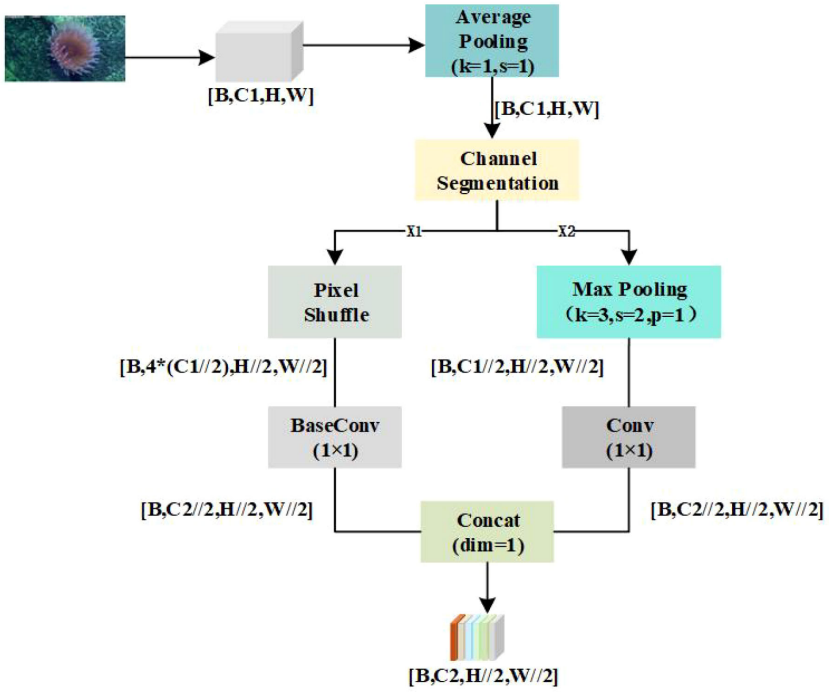


FIGURE 5
Structural diagram of adaptive downsampling.

Among them, $X_1, X_2 \in R^{b\times(c_1/2)\times h\times w}$ are respectively routed to two distinct feature transformation branches.

In the dual-path feature transformation stage, the two paths are designed for different feature types of deep-sea images, forming complementary feature sets. First, for small targets and weak texture features in deep-sea images, fine-grained information is preserved through spatial partitioning-channel concatenation operations, as shown in Equations (22) and (23).

$$X_1 \rightarrow \begin{cases} X_{L1} \\ X_{L2} \\ X_{L3} \\ X_{L4} \end{cases} \tag{22}$$

$$X_1^{sea} = torch.sea([X_{L1}, X_{L2}, X_{L3}, X_{L4}], \mathrm{dim} = 4) \tag{23}$$

Among them, X_{L*} refers to 4 regional sizes $h/2 \times w/2$; the number of channels is $c_1/2$; $X_1^{sea} \in R^{b\times 2c_1\times h/2\times w/2}$ with the number of channels being $2c_1 = 4 \times (c_1/2)$.

Then, the BaseConv module is used to compress and activate the channels, as shown in Equation (24).

$$Y_1 = \delta_{SiLU}(BN(X_1^{sea} * W_1)) \tag{24}$$

Among them, $W_1 \in R^{(c_2/2)\times 2c_1\times 1\times 1}$ is the convolution kernel; $\delta_{SiLU}(x) = x \cdot \delta(x)$, δ are Sigmoid functions; $Y_1 \in R^{b\times(c_2/2)\times h/2\times w/2}$ is the output.

The second stage is the salient feature enhancement path with max pooling. For relatively clear targets in deep-sea environments, local salient features are strengthened through max pooling, as shown in Equation (25)

$$X_2^{pool} = MaxPool2d(X_2; k = 3, \mathrm{s} = 2, \mathrm{p} = 1) \tag{25}$$

Then, the Conv module is used to compress and activate the channels, as shown in Equation (26)

$$Y_2 = \delta_{SiLU}(BN(X_2^{pool} * W_2)) \tag{26}$$

Among them, $W_2 \in R^{(c_2/2)\times(c_1/2)\times 1\times 1}$ is the convolution kernel; $Y_2 \in R^{b\times(c_2/2)\times h/2\times w/2}$ is the output.

Finally, the outputs of the two paths are concatenated along the channel dimension to fuse detailed features and salient features, as shown in Equation (27)

$$Y = torch.sea([Y_1, Y_2], \mathrm{dim} = 1) \tag{27}$$

Among them, $Y \in R^{b\times c_2\times h/2\times w/2}$ is the final output.

The ADown module achieves a balance between detail preservation and salient feature enhancement in the downsampling task of deep-sea images through its dual-path parallel design and refined feature processing (He et al., 2025). With its mathematically rigorous dimension transformation and scenario-adaptive design, it serves as an effective component for improving feature representation capabilities in deep-sea image recognition models. This module can be embedded into backbone networks or feature pyramid structures to optimize the performance of multi-scale feature extraction.

3 Experiments and results

3.1 Dataset establishment and processing

This experimental dataset was derived from real video data captured by Jiaolong (Shi et al., 2019), a manned submersible of the 7000 m underwater class in the western Pacific Ocean, and the process shown in Figure 6 below was used to construct the deep-sea biological dataset. Firstly, the video key frame images were intercepted to obtain the deep-sea organism images, then the images were expanded by rotating, inverting, noise addition, and other data enhancement methods (Cheung and Yeung, 2023), and finally, the organism images were annotated using the traditional labeling tool to obtain 5002 images and the corresponding labels.

During the annotation process, all image organisms were categorized according to their biological phyla, totaling eight phyla, as illustrated in Figure 7, which include Coral Polyp, Crinoid, Starfish, Crustacean, Ray-Finned Fish, Sea Urchin, Sea Cucumber, and Hexactinellida.

3.2 Experimental environment and parameter configuration

The experiments in this paper use Ubuntu 20.04 as the operating system, PyTorch as the deep learning framework, and the experimental platform uses Python 3.8.19 and torch2.0.1 +cuda11.8. The graphics card model is (NVIDIA GeForce RTX4090, 24GB). The detailed production parameters of the experiment are shown in Table 1.

3.3 Evaluation criteria

In this study, the performance of the improved YOLOv8n model was evaluated using the precision (Hestness et al., 2019) and mean average precision (*mAP*) as the evaluation metrics. For details, see Equations (28)–(31).

$$P = \frac{T_P}{T_P + F_N} \tag{28}$$

$$R = \frac{T_P}{T_P + F_P} \tag{29}$$

$$AP = \int_0^1 P \cdot \mathrm{Rd}_R \tag{30}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \times 100\,\% \tag{31}$$

where P represents the precision of the described model, i.e., what percentage of instances predicted by the model to be positive instances are actually positive instances. R represents the ratio of the instances of correctly identified deep-sea organisms to all the annotated instances of deep-sea organisms. T_P denotes the number of accurate identifications of deep-sea organism detections made by the YOLOv8n network model. F_P denotes the number of inaccurate identifications of deep-sea organism detections made by the YOLOv8n network model. *AP* is the mean average precision. AP50 is the mean average precision for this category of samples when the threshold value of the IoU of the confusion matrix is taken to be 0.5. *mAP* is the precision of the samples of all the categories averaged, which reflects the trend of the model's precision with the recall rate; the higher the value, the easier it is for the model to maintain a high precision at a high recall rate. mAP50–95 represents the average *mAP* value over different IoU thresholds (from 0.5 to 0.95 in steps of 0.05). N represents the number of categories.

In model evaluation metrics, the precision is a crucial indicator for assessing the accuracy of the model recognition, while the mean average precision serves as a comprehensive performance metric that aggregates multiple precision values across different recall rates. As a key evaluation criterion, the *mAP* holds even greater significance. It not only reflects the model's precision in recognizing positive samples but also provides a comprehensive evaluation of all the object detections. The *mAP* plays a pivotal role in assessing the model effectiveness and selecting the optimal model.

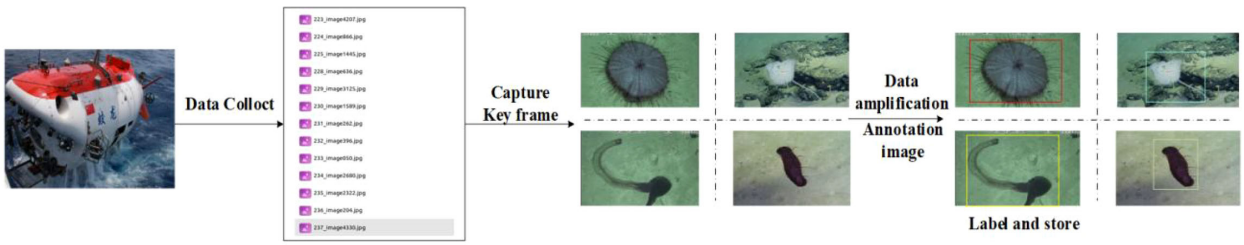


FIGURE 6
Flowchart of dataset production.

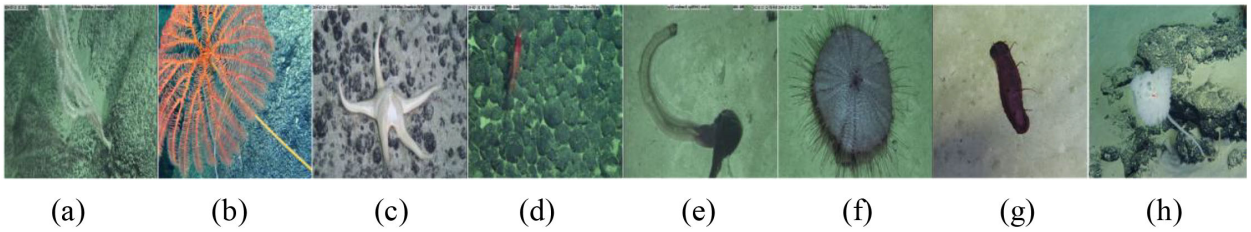


FIGURE 7
The Jiaolong dataset contains eight phyla of organisms: **(a)** Coral Polyp; **(b)** Crinoid; **(c)** Starfish; **(d)** Crustacean; **(e)** Ray-Finned Fish; **(f)** Sea Urchin; **(g)** Sea Cucumber; **(h)** Hexactinellida.

TABLE 1 Detailed hyperparameters of the experiment.

Parameters	Value or type
Epochs	100
Batch size	8
Optimizer	SGD
Image size	640 × 640
Initial learning rate	0.01
Optimizer momentum	0.937
Weight decay	5×10^{-4}

3.4 Experimental results and analysis

3.4.1 Experimental comparison before and after model improvement

Figure 8 shows the precision, mean average precision at IoU=0.5, and mean average precision at IoU=0.5–0.95 of the original YOLOv8n model and the SDA-HTransYOLOv8n model after 100 training iterations. Specifically, the SDA-HTransYOLOv8n model exhibits better precision when trained on the deep-sea organism dataset. In terms of the mAP50 and mAP50–95 metrics, the SDA-HTransYOLOv8n model has maintained a leading position after iteration, and it shows a significant advantage especially in mAP50. This indicates that the SDA-HTransYOLOv8n model not only has improved detection accuracy but also demonstrates stronger robustness under different IoU thresholds. To more intuitively compare the detection effects of YOLOv8n and SDA-HTransYOLOv8n on samples not involved in training, Figure 9 presents some detection results covering different deep-sea organisms. Compared with the YOLOv8n model, the SDA-HTransYOLOv8n model proposed in this paper performs better in terms of accuracy and recognition error rate. Especially in the complex deep-sea environment, the SDA-HTransYOLOv8n model has higher confidence and more stable detection performance. Although YOLOv8n can also accurately identify deep-sea organisms, SDA-HTransYOLOv8n shows stronger robustness under different deep-sea backgrounds, reduces interference from the complex deep-sea environment, and exhibits higher accuracy.

3.4.2 Ablation experiments

To comprehensively evaluate and verify the effectiveness of the improved model, ablation experiments were conducted. Under the

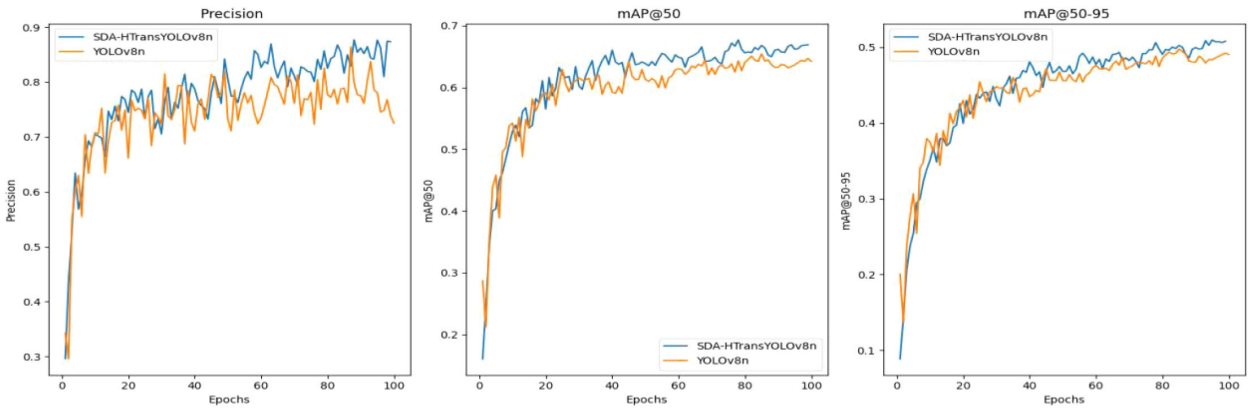
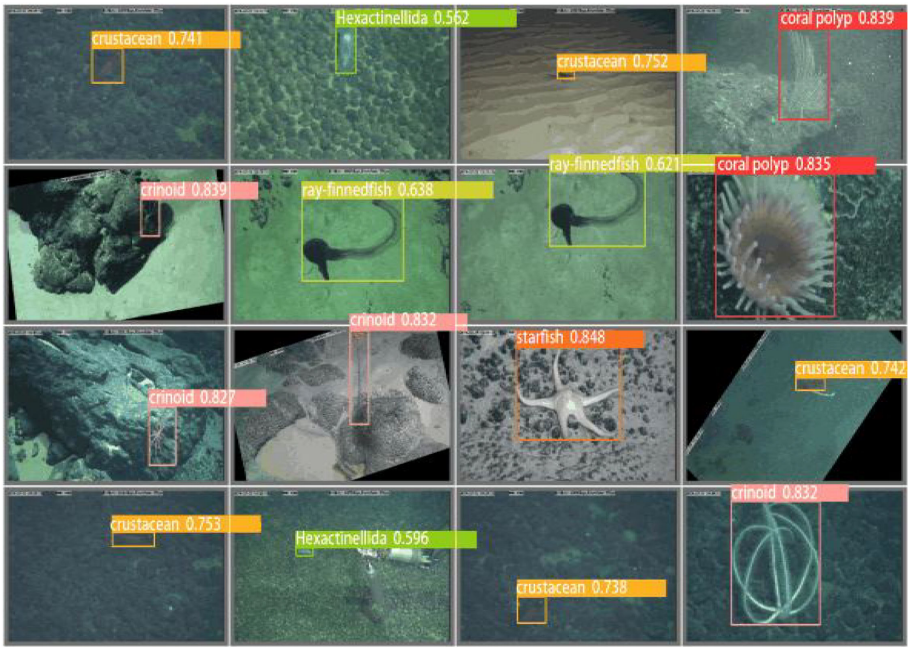
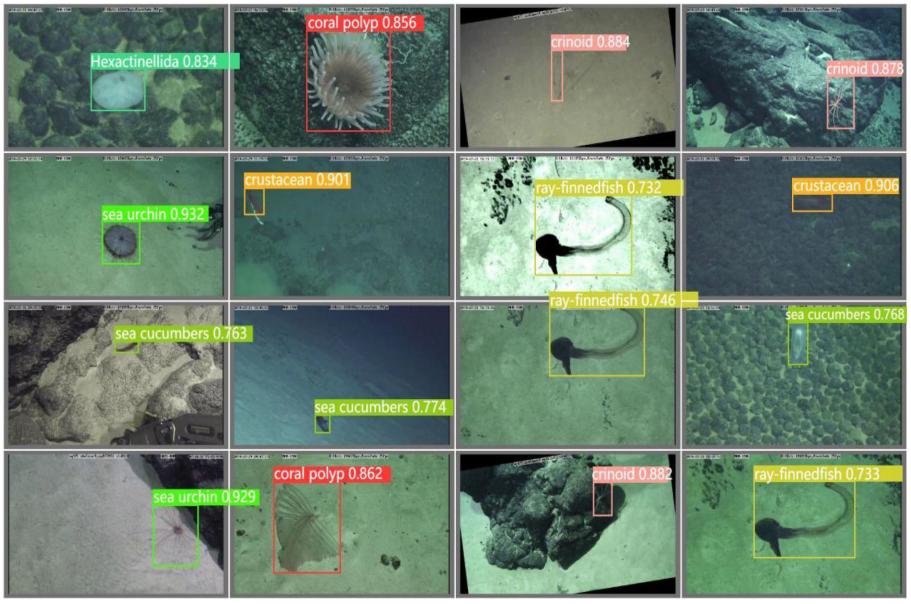


FIGURE 8
Comparison of indicators before and after model improvement.



(a)



(b)

FIGURE 9
YOLOv8n **(a)** vs. SDA-HTransYOLOv8n **(b)** detection results for different organisms.

premise of keeping the environment and training parameters consistent, the optimization effects of adding different modules of YOLO or different combinations on deep-sea biological detection were analyzed. The results are shown in Table 2 below.

3.4.3 Comparative tests of different models

In order to fully evaluate the performance of the SDA-HTransYOLOv8n model, the same dataset is used as a sample and it is analyzed in comparison with a series of target detection models, which include SSD, YOLOv3, YOLOv5, YOLOv7, YOLOv8n, and YOLOv11 five models. The results are shown in Table 3, and SDA-HTransYOLOv8n excels in all performance metrics. Its model reaches 87.6%, 67.7%, and 51.6% in Precision, and mean average precision (mAP50,mAP50-95), respectively, which fully proves its strong ability in accurately identifying and localizing targets.

TABLE 2 Ablation experiments.

YOLOv8n	HT_C2f	DySample	SDI	ADown	Precision(%)	Average precision mean	
						mAP50/(%)	mAP50-95/(%)
1	×	×	×	×	78.7	64.9	49.8
2	✓	×	×	×	82.0	65.5	49.0
3	×	✓	×	×	82.4	67.0	51.9
4	×	×	✓	×	82.2	66.9	51.5
5	×	×	×	✓	83.1	66.1	51.0
6	✓	✓	×	×	80.4	67.2	50.4
7	✓	×	✓	×	84.6	66.5	50.5
8	✓	×	×	✓	84.0	67.4	50.8
9	×	✓	✓	×	81.0	65.4	50.2
10	×	✓	×	✓	84.6	67.3	51.0
11	×	×	✓	✓	84.7	66.1	50.4
12	✓	✓	✓	×	82.0	63.4	42.0
13	✓	✓	×	✓	86.2	67.5	51.4
14	✓	×	✓	✓	84.7	66.1	50.4
15	×	✓	✓	✓	84.9	66.5	51.3
16	✓	✓	✓	✓	87.6	67.7	51.6

mAP50 is the average accuracy of the IoU threshold of 0.5,and mAP50–95 is the average of the *mAP* when the IoU threshold is 0.50 to 0.95 and the step size is 0.05.The same below.

TABLE 3 Comparison experiments with other models.

Model	Precision(%)	Average precision mean	
		mAP50/(%)	mAP50-95/(%)
SSD	54.4	62.2	42.6
YOLOv3	45.7	58.5	34.4
YOLOv5s	82.4	66.5	48.2
YOLOv7	53.6	55.1	33.5
YOLOv8n	78.7	64.9	49.8
YOLOv11n	85.2	63.8	49.3
SDA-HTransYOLOv8n	87.6	67.7	51.6

mAP50 is the average accuracy of the IoU threshold of 0.5,and mAP50–95 is the average of the *mAP* when the IoU threshold is 0.50 to 0.95 and the step size is 0.05. The same below.

4 Conclusion

In response to the challenges faced by image recognition in deep-sea environments, such as dim light, severe water scattering, and blurred target features, this paper proposes an improved YOLOv8n model: SDA-HTransYOLOv8n. By introducing the HyperTransformer module, SDI multi-level feature fusion module, DySample dynamic sampling mechanism, and ADown downsampling module, the model achieves collaborative optimization in the links from feature capture, fusion, and extraction to dimensionality reduction. Compared with the YOLOv8n model, the SDA-HTransYOLOv8n model has achieved significant improvements in Precision, mean average precision (mAP50, mAP50-95), reaching 87.6%, 67.7%, and 51.6% respectively. Meanwhile, when comparing SDA-HTransYOLOv8n with SSD, YOLOv3, YOLOv5, YOLOv7, and YOLOv11, the SDA-HTransYOLOv8n model shows more prominent advantages in detection accuracy. The research method in this paper plays an important technical supporting role in deep-sea biodiversity investigation and assessment, as well as marine ecological environment protection. By improving the accuracy and efficiency of deep-sea biological detection, this study can help scientists gain a more comprehensive understanding of deep-sea ecosystems and provide reliable data support for the sustainable utilization of marine resources and ecological environment protection.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

DC: Writing – review & editing, Conceptualization, Software, Methodology, Writing – original draft, Data curation. XS: Conceptualization, Writing – review & editing, Supervision,

Writing – original draft. ML: Supervision, Writing – original draft, Writing – review & editing. SQ: Writing – original draft, Supervision. ZZ: Supervision, Writing – original draft.

Funding

The author(s) declare financial support was received for the research and/or publication of this article. This research work was supported and sponsored by the National Key Project of Research and Development Program (2024YFC2814400), the National Natural Science Foundation of China (U22A2044).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Bharati, P., and Pramanik, A. (2019). Deep learning techniques—R-CNN to Mask R-CNN: A survey. *Comput. Intell. Pattern Recognition: Proc. CIPR* 2019, 657–668.

Chen, D., Shi, X., Yang, J., and Huang, Y. (2025). Research on method for intelligent recognition of deep-sea biological images based on PSVG-YOLOv8n. *J. Mar. Sci. Eng.* 13, 810. doi: 10.3390/jmse13040810

Cheng, S., Han, Y., Wang, Z., and Sun, X. (2025). An underwater object recognition system based on improved YOLOv11. *Electronics* 14, 201. doi: 10.3390/electronics14010201

Cheung, T. H., and Yeung, D. Y. (2023). A survey of automated data augmentation for image classification: Learning to compose, mix, and generate. *IEEE Trans. Neural Networks Learn. Syst.* 35, 13185–13205. doi: 10.1109/TNNLS.2023.3282258

Costa, C., Fanelli, E., Marini, S., Danovaro, R., and Aguzzi, J. (2020). Global deep-sea biodiversity research trends highlighted by science mapping approach. *Front. Mar. Sci.* 7, 384. doi: 10.3389/fmars.2020.00384

Deng, X., Liao, L., Jiang, P., and Liu, Y. (2023). "Towards scale adaptive underwater detection through refined pyramid grid," in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. (Piscataway: Institute of Electrical and Electronics Engineers (IEEE)) 1–5.

Dong, S., Xie, W., Yang, D., and Li, Y. (2024). "SeaDATE: Remedy dual-attention transformer with semantic alignment *via* contrast learning for multimodal object detection," in *IEEE Transactions on Circuits and Systems for Video Technology*. (Piscataway: Institute of Electrical and Electronics Engineers (IEEE)).

Duan, K., Bai, S., Xie, L., and Li, Y. (2019). "CenterNet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2019*. (Piscataway: Institute of Electrical and Electronics Engineers (IEEE)) 6569–6578.

Fu, L., Wang, Y., Wu, S., and Liu, Z. (2025). TCCFNet: A semantic segmentation method for mangrove remote sensing images based on two-channel cross-fusion networks. *Front. Mar. Sci.* 12, 1535917. doi: 10.3389/fmars.2025.1535917

Han, F., Yao, J., Zhu, H., and Liu, Y. (2020). Marine organism detection and classification from underwater vision based on the deep CNN method. *Math. Problems Eng.* 2020, 3937580. doi: 10.1155/2020/3937580

He, Z., Cao, L., Xu, X., and Wang, Y. (2025). Underwater instance segmentation: A method based on channel spatial cross-cooperative attention mechanism and feature prior fusion. *Front. Mar. Sci.* 12, 1557965. doi: 10.3389/fmars.2025.1557965

Hestness, J., Ardalani, N., and Diamos, G. (2019). "Beyond human-level accuracy: Computational challenges in deep learning," in *Proceedings of the 24th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP 2019)*. (New York: Association for Computing Machinery (ACM)) 1–14.

Hu, D., Yu, M., Wu, X., Hu, J., Sheng, Y., Jiang, Y., et al. (2024). DGW-YOLOv8: A small insulator target detection algorithm based on deformable attention backbone and WIoU loss function. *IET Image Process.* 18, 1096–1108. doi: 10.1049/ipr2.13009

Hu, Z., Cheng, L., Yu, S., Xu, P., Zhang, P., and Han, J. (2025). Underwater target detection with high accuracy and speed based on YOLOv10. *J. Mar. Sci. Eng.* 13, 135. doi: 10.3390/jmse13010135

Kim, J. H., Kim, N., Park, Y. W., and Won, C. S. (2022). Object detection and classification based on YOLO-V5 with improved maritime dataset. *J. Mar. Sci. Eng.* 10, 377. doi: 10.3390/jmse10030377

Lee, M. (2023). Mathematical analysis and performance evaluation of the gelu activation function in deep learning. *J. Mathematics* 2023, 4229924. doi: 10.1155/2023/4229924

Li, B., Li, L., Zhang, Z., and Duan, Y. (2025). LUIEO: A lightweight model for integrating underwater image enhancement and object detection. *IEEE Trans. Circuits Syst. Video Technol.* 35, 1–14. doi: 10.1109/TIM.2025.3563051

Li, X., Sun, W., Ji, Y., and Huang, W. (2024). "A joint detection and tracking paradigm based on reinforcement learning for compact HFSWR," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. (Piscataway: Institute of Electrical and Electronics Engineers (IEEE)).

Li, X., Sun, W., Ji, Y., and Huang, W. (2025). "S2G-GCN: A plot classification network integrating spectrum-to-graph modeling and graph convolutional network for compact HFSWR," in *IEEE Geoscience and Remote Sensing Letters*. (Piscataway: Institute of Electrical and Electronics Engineers (IEEE)).

Liu, W., Lu, H., Fu, H., and Wang, Z. (2023). "Learning to upsample by learning to sample," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2023*. 6027–6037. (Piscataway: Institute of Electrical and Electronics Engineers (IEEE)).

Shi, X., Ren, Y., Tang, J., Fu, W., and Liu, B. (2019). Working tools study for JiaoLong manned submersible. *Mar. Technol. Soc. J.* 53, 56–64. doi: 10.4031/MTSJ.53.2.6

Shi, X., and Wang, H. (2023). Lightweight underwater object detection network based on improved YOLOv4. *J. Harbin Eng. Univ.* 44, 154–160.

Sisodiya, P., and Bhoite, S. (2025). *Enhanced Faster R-CNN for real-time underwater object detection* (Berlin, Germany: ResearchGate Preprint).

Wang, D., Chen, Y., Naz, B., and Li, Y. (2023). Spatial-aware transformer (SAT): Enhancing global modeling in transformer segmentation for remote sensing images. *Remote Sens.* 15, 3607. doi: 10.3390/rs15143607

Wang, H., Frery, A. C., Li, M., and Ren, P. (2023). Underwater image enhancement *via* histogram similarity-oriented color compensation complemented by multiple attribute adjustment. *Intelligent Mar. Technol. Syst.* 1, 12. doi: 10.1007/s44295-023-00015-y

Wang, X., Jiang, X., Xia, Z., and Li, Y. (2022). "Underwater object detection based on enhanced YOLO," in *Proceedings of the 2022 International Conference on Image Processing and Media Computing*. 17–21. (New York: Association for Computing Machinery (ACM)).

Wang, H., Zhang, W., Xu, Y., Li, H., and Ren, P. (2025). WaterCycleDiffusion: Visual-textual fusion empowered underwater image enhancement. *Inf. Fusion*, 103693.

Wazirali, R., and Chaczko, Z. (2016). Hyper edge detection with clustering for data hiding. *J. Inf. Hiding Multimedia Signal Process.* 7, 1–10.

Wei, Y., Wang, Y., Zhu, B., Lin, C., Wu, D., Xue, X., et al. (2024). Underwater detection: A brief survey and a new multitask dataset. *Int. J. Network Dynamics Intell.* 3, 679–692. doi: 10.53941/ijndi.2024.100025

Wen, G., Li, M., Tan, Y., and Zhang, H. (2025). Enhanced YOLOv8 algorithm for leaf disease detection with lightweight GOCR-ELAN module and loss function: WSIoU. *Comput. Biol. Med.* 186, 109630. doi: 10.1016/j.compbiomed.2024.109630

Wu, T., and Dong, Y. (2023). YOLO-SE: Improved YOLOv8 for remote sensing object detection and recognition. *Appl. Sci.* 13, 12977. doi: 10.3390/app132412977

Xie, Y., Yu, Z., Yu, X., and Zheng, B. (2022). Lighting the darkness in the sea: A deep learning model for underwater image enhancement. *Front. Mar. Sci.* 9, 921492. doi: 10.3389/fmars.2022.921492

Yang, H. H., Huang, K. C., and Chen, W. T. (2021). "LaffNet: A lightweight adaptive feature fusion network for underwater image enhancement," in *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA 2021)*. (Piscataway: Institute of Electrical and Electronics Engineers (IEEE)) 685–692.