Sex identification in *Procambarus clarkii* using multi-dimensional feature fusion and enhancement

Lu Liu[1], Hanyu Zhang[2], Jun Yue[1*], Guangjie Kou[1], Zhenbo Li[2], Longchuan Zhuang[3]

(1. *School of Computer and Artificial Intelligence, Ludong University, Yantai 264025, Shandong, China*;
2. *College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China*;
3. *Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai 264003, China*)

Abstract: Accurate sex identification in *Procambarus clarkii* is essential for genetic breeding and aquaculture management, as it helps optimize population structure, improve reproductive efficiency, and support sustainable aquaculture development. However, manual identification is time-consuming, labor-intensive, and prone to errors, especially when subtle visual differences need to be distinguished. To address this problem, this study proposed SCM-DETR, a sex identification method for *Procambarus clarkii* based on multi-dimensional feature fusion and enhancement. A high-resolution imaging system was used to acquire two-dimensional images of *Procambarus clarkii*, and a labeled dataset, the *Procambarus clarkii* gonad dataset (PGD), was constructed. To improve identification performance, a multi-dimensional semantics and details fusion method (MSDM) was designed to integrate high-level semantic information with fine-grained detail features, thereby enhancing feature representation and localization accuracy. In addition, a channel-spatial focus network (CSFN) was introduced to capture discriminative multidimensional features, including texture and color, for more accurate identification of subtle sex-related differences. Experimental results showed that SCM-DETR-R18 achieved 95.8% mAP@0.50 and 64.6% mAP@0.50-0.95 on the PGD, improving by 1.9 and 1.1 percentage points over the baseline model, respectively. The AP values of female and male gonads reached 93.3% and 96.5%, with gains of 3.3 and 1.9 percentage points, respectively. Moreover, the proposed model had the lowest parameter count (21.11 M) among all compared methods. The results of this study demonstrate that SCM-DETR can effectively improve automated sex identification in *Procambarus clarkii* and has good potential for intelligent aquaculture applications.

Keywords: *Procambarus clarkii*, sex identification, detection transformer, multi-dimensional feature fusion, small target detection

DOI: 10.25165/j.ijabe.20261903.10182

Citation: Liu L, Zhang H Y, Yue J, Kou G J, Li Z B, Zhuang L C. Sex identification in *Procambarus clarkii* using multi-dimensional feature fusion and enhancement. Int J Agric & Biol Eng, 2026; 19(3): 139–148.

1 Introduction

Procambarus clarkii is highly favored by consumers because of its desirable taste and rich nutritional value and has considerable economic value due to strong market demand[1,2]. At present, aquaculture is the primary production mode for this species, and accurate sex identification plays an important role in optimizing breeding programs and improving farming efficiency. As shown in Figure 1, the gonads of *P. clarkii* are located in the abdominal region and exhibit distinct morphological characteristics in males and females: female gonads are pinnate, semi-transparent, and plump, whereas male gonads are protruding, milky white, and relatively flat. However, gonad identification is difficult because of the high inter-class similarity to pleopods, as well as interference from motion blur, specular reflection from the carapace, and background noise. In addition, the gonadal region accounts for less than 10% of the image area, which further increases the difficulty of automated detection and segmentation. At present, sex identification relies mainly on manual inspection, which is inefficient and prone to inaccuracies, making it unsuitable for modern breeding and large-scale aquaculture management. More importantly, automated techniques for sex-specific gonad recognition in *P. clarkii* are still limited[3-5], posing a major obstacle to precise and automated sex identification.

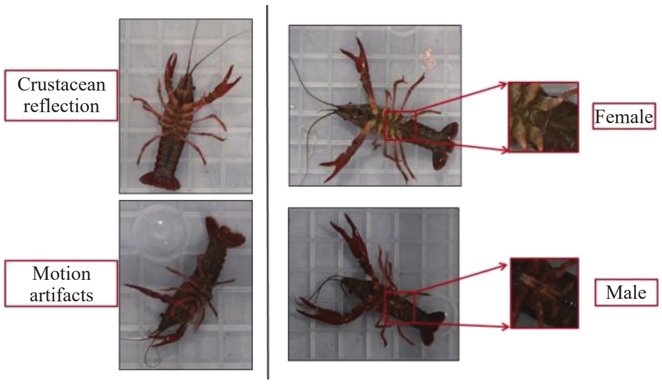


Figure 1 Challenges in sex identification of *Procambarus clarkii*

The gonads of *Procambarus clarkii* exhibit distinct morphological characteristics, making computer vision-based

Received date: 2025-09-09 **Accepted date:** 2026-05-21

Biographies: Lu Liu, MS, research interest: artificial intelligence, Email: 1329820476@qq.com; **Hanyu Zhang**, PhD candidate, research interest: computer science, Email: hanyuu@cau.edu.cn; **Guangjie Kou**, PhD, Professor, research interest: computer vision, Email: kouguangjie@126.com; **Zhenbo Li**, PhD, Professor, research interest: computer graphics, Email: lizb@cau.edu.cn; **Longchuan Zhuang**, PhD, Assistant Researcher, research interest: marine science, Email: lczhuang@yic.ac.cn.

***Corresponding author: Jun Yue**, PhD, Professor, research interest: artificial intelligence technology. School of Computer and Artificial Intelligence, Ludong University, Yantai 264025, Shandong, China. Tel: +86-13562559603, Email: yuejuncn@sohu.com.

identification feasible. Object detection methods[6,7] have proven effective in complex environments and have been increasingly applied in aquaculture. Xing et al.[8] proposed an improved YOLOv8 model integrated with BoT-SORT for sonar-based fish detection and counting, enabling stable real-time monitoring of fish populations in designated areas through joint detection and tracking. Jun et al.[9] developed an enhanced YOLOv7 framework incorporating a dual-layer routing attention mechanism and a Normalized Wasserstein Distance (NWD) loss function, which reduced the model's sensitivity to positional deviations of small targets in blurred images and improved recognition performance for counting dense fish schools under turbid water conditions. Geng et al.[10] introduced a YOLOv8n-based algorithm for detecting the cephalothorax, abdomen, and chelipeds of *Procambarus clarkii*, providing technical support for intelligent and precise grading. Similarly, Chen et al.[11] proposed the R-SINet framework, which integrates a multi-dimensional feature fusion to achieve non-destructive sex identification in oysters. Despite these advances in related aquaculture applications, existing techniques still face significant challenges in fine-grained recognition tasks. Specifically, the identification of *P. clarkii* gonads requires precise fusion of color and texture features as well as contextual correlation analysis, which current methods often fail to achieve effectively. Recent studies have shown that Transformer-based architectures, particularly the Detection Transformer (DETR) model[12-15], provide a promising solution. By leveraging global self-attention mechanisms, these models can capture long-range dependencies and enhance contextual reasoning, thereby improving performance in complex fine-grained biological image analysis tasks.

The global reasoning capability of the Detection Transformer (DETR) model has shown great potential in object detection, and many studies have achieved promising results in different domains based on the DETR framework[16-21]. For instance, Peng et al.[22] developed UAV-DETR for aerial image applications by integrating multi-scale feature fusion, frequency decomposition, and spatial attention calibration modules, which significantly improved the detection of small and occluded targets. Zhang et al.[23] proposed PUFFER-DETR, which employed a Triplet Attention backbone to enhance feature extraction. Combined with similar behavior feature weighting and SHS-FPN cross-scale fusion, this method enabled accurate detection of abnormal fish behavior and supported aquaculturists in evaluating growth conditions. Yang et al.[24] introduced ISTD-DETR, which combined preprocessing with an improved real-time detection transformer to improve the detection of small and low-resolution objects. Huang et al.[25] proposed DQ-DETR, a DETR variant with dynamic queries for tiny object detection, and demonstrated superior performance over existing CNN and DETR-based models on the AI-TOD-v2 dataset.

To address the challenges of sex identification in *Procambarus clarkii*, this study proposed SCM-DETR, an efficient identification method based on an improved detection transformer. The method integrates multi-dimensional semantic information and fine-grained detail features to enhance the representation of gonadal regions in the abdomen near the pleopods, thereby enabling more accurate localization of key targets. In addition, a channel-spatial focus network is designed to extract and strengthen contour, texture, and color cues, which improves gonad recognition accuracy. The main contributions of this paper are summarized as follows:

1) A self-constructed *Procambarus clarkii* gonad dataset (PGD) was established to provide a reliable benchmark for automated gonad identification.

2) A multi-dimensional semantics and details fusion module (MSDM) was proposed to capture multi-level fine-grained information and improve the localization accuracy of gonadal regions.

3) A gonad channel-spatial focus network (CSFN) was designed to integrate multi-dimensional feature maps and enhance gonadal feature representation.

2 Materials and methods

A *Procambarus clarkii* gonad detection model, termed SCM-DETR, is proposed based on the RT-DETR architecture. The model introduces two core modules: the multi-dimensional semantics and details fusion module (MSDM), which integrates multi-level features to improve gonad localization and bounding box regression accuracy, and the channel-spatial focus network (CSFN), which enhances discriminative features such as texture and color through channel and spatial attention mechanisms. The features enhanced by these two modules are then fed into the DETR decoder. Finally, the detection head performs object classification and bounding box regression to generate the final predictions. The overall network architecture is shown in Figure 2.

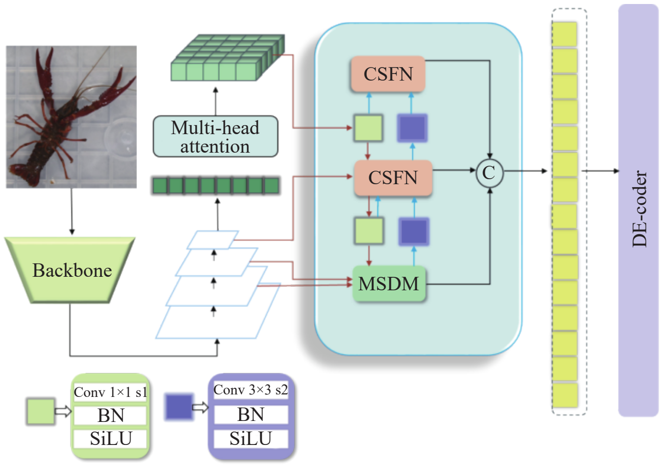


Figure 2 Network architecture of SCM-DETR

2.1 Multi-dimensional semantics and details fusion method

To address the challenge that the gonadal regions of *Procambarus clarkii* are highly concealed and extremely small, conventional CNN-based methods typically rely on stacked convolution and pooling layers for hierarchical feature extraction[26]. Although pooling operations enlarge the receptive field and reduce computational cost, they inevitably decrease the spatial resolution of feature maps, causing the loss of critical local structural details. This issue is particularly severe for gonadal targets, which occupy only a very limited pixel region in images. After multiple downsampling stages, these weak targets are often poorly represented in high-level semantic features, leading to localization deviation and inaccurate boundary prediction.

To overcome these limitations, this study proposed a multi-dimensional semantics and details fusion method (MSDM). Rather than simply combining lightweight convolutional operators, the proposed method is specifically designed for the *Procambarus clarkii* gonad detection task by constructing a multi-dimensional collaborative feature enhancement network that jointly optimizes the extraction of spatial position, weak edge contour, and subtle morphological cues of tiny gonadal targets. This design is particularly effective for the female gonad category, whose boundaries are often blurred and whose local texture contrast is weaker than that of male gonads. The network structure of MSDM is shown in Figure 3.

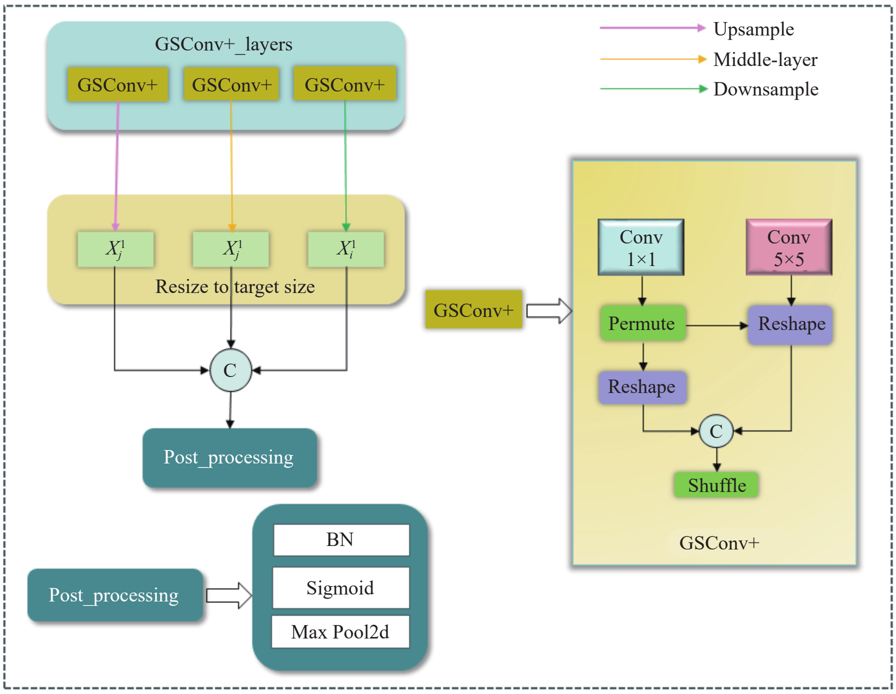


Figure 3 Network architecture of MSDM

In *Procambarus clarkii* gonad detection, the MSDM constructs a collaborative information network to strengthen the extraction of morphological and positional features of tiny gonadal targets. First, the input images are processed by the backbone to generate multi-level hierarchical feature maps. Because different feature levels contain different positional and morphological cues, a three-level feature fusion strategy is adopted to integrate multi-scale information effectively. To further improve the detection of minute gonadal targets, the three input feature maps are first processed by GSConv+ layers[27] for preliminary feature extraction, thereby producing richer feature representations for subsequent fusion.

The input feature map $\boldsymbol{X}$ is then processed by two parallel convolution branches. The first branch is a standard convolution layer (cv1), which reduces the channel dimension by half. This operation preserves essential gonadal features while suppressing background interference from the carapace, thereby improving feature discrimination. The second branch is a 5×5 depthwise separable convolution (cv2), which expands the receptive field while maintaining parameter efficiency. This allows the model to capture contextual information around the gonads and improve localization accuracy.

The channel-reduced features from cv1 are concatenated with the features generated by cv2. By combining feature maps with different receptive fields, the module can more comprehensively capture the subtle differences between gonadal and pleopod morphologies in *Procambarus clarkii*. Subsequently, channel shuffling is applied to rearrange the channel order, thereby promoting interactions among features from different spatial locations and enhancing the representational capacity of gonadal features. Finally, the reduced features are fused again with the depthwise convolutional features to preserve essential original information while incorporating enriched contextual cues. This design improves feature extraction efficiency while reducing the overall parameter count, making it suitable for gonad detection under complex background conditions. The corresponding formulation is as follows:

$$X_{\text{final}} = \text{Shuffle}(\text{Concat}(\sigma(\text{Conv}(X_1, W_1)), \sigma(\text{DepthwiseConv}(X_2, W_2)))) \tag{1}$$

where, $X \in R^{B\times C\times H\times W}$ is the input feature map; X_1 and X_2 are the outputs of the two parallel branches; W_1 and W_2 are the learnable weights of the standard convolution and depthwise convolution branches, respectively; σ is the SiLU activation function; Conv and DWConv denote standard convolution and depthwise separable convolution, respectively; Concat denotes channel-wise concatenation; and Shuffle denotes channel shuffling for inter-branch feature fusion.

Due to the complex backgrounds, variable target scales, and blurred tissue edges commonly observed in acquired images of *Procambarus clarkii* gonadal regions, relying solely on single-scale features makes it difficult to capture both fine-grained details and global semantic information simultaneously. Therefore, feature alignment and fusion across different hierarchical levels are necessary to improve the recognition and localization accuracy of gonadal regions.

In this study, the intermediate-layer feature map X_i^1 was selected as the reference for spatial size alignment, where, H_i, W_i, and c denote its height, width, and number of channels, respectively. During feature fusion, multi-level feature maps from different layers are processed sequentially. According to their relative positions with respect to the reference layer, downsampling, identity mapping, or upsampling is applied to resize all feature maps to the same spatial dimensions as X_i^1.

For shallow-layer features before the reference layer, downsampling is used to achieve dimensional alignment while preserving detailed structural information. The reference-layer feature is retained directly, whereas deep-layer features after the reference layer are upsampled for spatial alignment. After dimensional normalization, multi-level features can be fused in a unified feature space, enabling the integration of shallow-layer details, intermediate-layer structures, and deep-layer semantic information. This improves the representation of gonadal regions and surrounding tissues and enhances the model's robustness against complex background interference. The fused feature is formulated as follows:

$$Y=\begin{cases} F_{\text{avgpool}}\left(X_j^1,\text{target\_size}\right), & \text{if } j<i \\ F_{\text{identity}}\left(X_j^1,\text{target\_size}\right), & \text{if } j=i \\ F_{\text{interpolate}}\left(X_j^1,\text{target\_size}\right), & \text{if } j>i \end{cases} \quad (2)$$

where, $X_i^1 \in R^{H_i \times W_i \times c}$ denotes the input feature map at the i-th level, where i=1, 2, 3 and target_size X_i^1 denotes its spatial size.

In the post-processing stage, max-pooling is used to reduce the spatial resolution of feature maps and preserve salient gonadal features. The activation function improves nonlinear feature representation, while Batch Normalization accelerates model convergence. With the incorporation of MSDM, the model is able to effectively combine semantic information with local detail features, thereby overcoming the limitations of traditional single-path feature extraction and enhancing the representation of gonadal targets.

2.2 Channel-spatial focus network

The primary challenge in gonad detection is to simultaneously capture contour information and fine-grained texture features while enhancing the model's ability to focus on key regions for identifying subtle morphological differences. To address this issue, a Channel-Spatial Focus Network (CSFN) module was designed by integrating channel attention and spatial attention mechanisms. The structure of the CSFN module is shown in Figure 4.

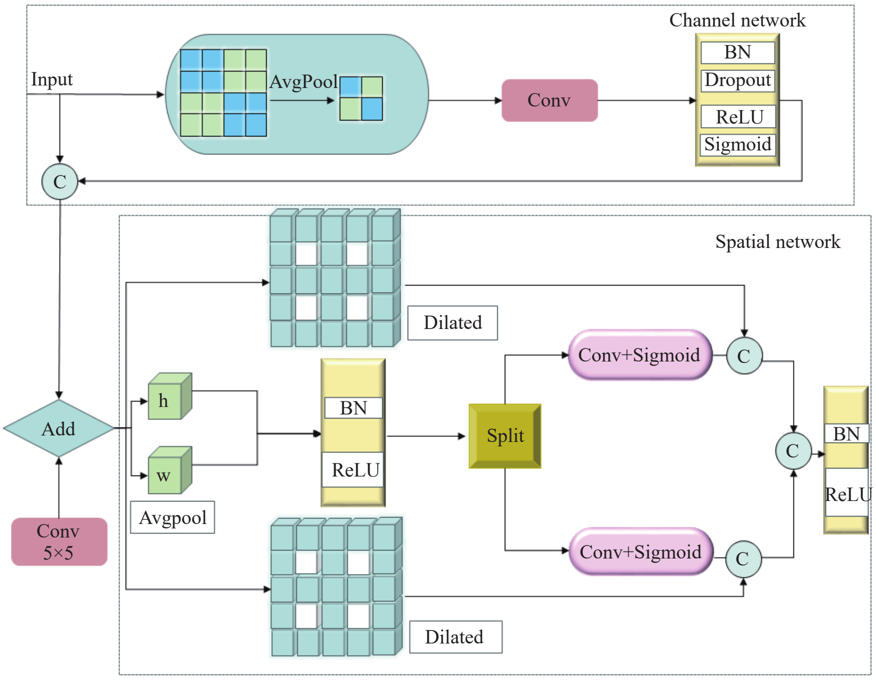


Figure 4 Network architecture of CSFN

2.2.1 Channel network (CN)

In *Procambarus clarkii* gonad detection, the gonads exhibit distinct color and texture characteristics across specific channels. The channel network enhances informative features, suppresses irrelevant responses, and improves the representation of key gonadal regions.

To enhance the focus of the channel network on salient gonadal features in *Procambarus clarkii*, the input feature map is first downsampled by average pooling to reduce spatial dimensions and suppress interference from irrelevant channels. This operation helps preserve representative gonadal information and strengthen the overall morphological representation of the gonadal region. Subsequently, a convolution operation with kernel size k is applied to model inter-channel interactions and generate channel attention weights. These weights are further refined through normalization and nonlinear activation to obtain the final attention coefficients. The resulting attention map is then broadcast to the original feature map for channel-wise reweighting, thereby enhancing the response to distinctive gonadal features. Since the kernel size k is related to the channel dimension C, which typically follows a power-of-two pattern, k is defined proportionally to enable adaptive scaling across different feature levels. The kernel size k is defined as follows:

$$C=\varphi(k)=2^{(\gamma\times k-b)} \quad (3)$$

where, γ is bare scaling parameters that control the relationship between the convolution kernel size k and the channel dimension C[28-30].

$$k=\varphi(C)=\left|\frac{\log_2(C)+b}{\gamma}\right|_{\text{odd}} \quad (4)$$

where, $|\ |_{\text{odd}}$ denotes the nearest odd integer. In this study, γ and b are set to 2 and 1, respectively, according to the nonlinear mapping relationship. As a result, the kernel size can be adaptively determined based on the channel dimension, allowing the channel attention mechanism to capture local cross-channel interactions more effectively and improve feature representation[28-30].

2.2.2 Spatial network (SN)

To effectively distinguish the subtle morphological differences between male and female *Procambarus clarkii* gonads, a Spatial network was designed to enhance spatial feature representation. The network fuses the output of the channel network with that of a single convolutional layer, enabling effective integration of complementary information and improving feature extraction capability. Furthermore, a local attention mechanism is employed to process the input tensor along both horizontal and vertical directions, thereby strengthening the model's sensitivity to subtle gonadal differences.

The statistical characteristics of the gonadal region are obtained by average pooling along the vertical and horizontal axes, which are used to emphasize the pleopod boundary and capture edge texture information, respectively. Nonlinear activation is then applied to strengthen the gradient difference between the gonad and the surrounding tissue. After that, the pooled features are reconstructed through separate F_h and F_w convolutions to generate direction-adaptive spatial attention masks S_h and S_W. By assigning weights in both directions, this design strengthens the response at edge inflection points and improves the representation of jagged gonadal boundaries. The attention weights are generated as follows:

$$S_h = \sigma(F_h(x_{\text{split\_h}})) \tag{5}$$

$$S_W = \sigma(F_w(x_{\text{split\_w}})) \tag{6}$$

where, σ represents the Sigmoid activation function; F_h and F_w denote two 1×1 convolution operations, respectively.

To alleviate the loss of gonadal microstructural information caused by pooling, contextual information from the gonad and surrounding tissues is introduced to reduce false detections. The weighted features are fused by dilated convolution, which enlarges the receptive field to 5×5 while preserving feature map resolution. Then, nonlinear activation is applied to enhance texture variation in the gonadal region and suppress interference from the smooth background. The corresponding formulation is as follows:

$$E_{\text{dilation}} = \left\lfloor \frac{i+2p-k-(k-1)*(d-1)}{s} \right\rfloor + 1 \tag{7}$$

$$F_{\text{CSFN}} = E_{\text{dilation}} \times s_w \times s_h \tag{8}$$

where, E denotes the channel-spatial weight matrix; K denotes the size of the original convolution kernel; d is the expansion factor (d=2); and F_{CSFN} is the feature map after expansion.

By strengthening multi-dimensional information processing, the CSFN can effectively capture both macroscopic contour features and fine-grained texture details, thereby enhancing the representation of key gonadal regions. Through the joint optimization of morphological structure and color-texture information, the module improves the discrimination of subtle differences between male and female gonads in *Procambarus clarkii*, thereby supporting more accurate sex identification.

3 Experiment and result analysis

3.1 Dataset collection and preprocessing

3.1.1 Data collection

To ensure geographic and seasonal representativeness, a total of 333 *Procambarus clarkii* specimens were systematically collected in 12 batches from Yutai County, Jining City, Shandong Province, China (34°53 ′N, 116°39 ′E), between March 2023 and March 2024. The sampling period covered all four seasons to capture image variations under different climatic conditions. The field sex ratio in each batch ranged from 1:1 to 19:11, the ambient temperature during sampling ranged from 2.0°C to 25.2°C, and the storage time ranged from 1.0 to 3.5 h, ensuring the representativeness and diversity of the dataset, as listed in Table 1. After capture, healthy individuals were immediately selected and placed in an oxygenated transport system (dissolved O_2>6 mg/L) under temperature-controlled conditions and then transported to the laboratory through a cold-chain process to minimize transport stress.

Standardized imaging was performed using a Nikon D750 camera positioned 50 cm from the specimen, and all imaging parameters, including focal length, aperture, shutter speed, and ISO, were recorded to ensure reproducibility, as listed in Table 2. Each specimen was then cataloged together with its collection metadata and sex label to establish a complete traceability system. This protocol helped maintain sample stability during imaging, as shown in Figure 5.

Table 1 Sample collection information of *Procambarus clarkii*

Batch No.	Harvest date	Collected N	Male:female ratio	Temperature at harvest/°C	Storage duration/h
1	2023-03-30	108	10:17	22.1	2.0
2	2023-05-11	116	14:15	23.5	2.5
3	2023-06-03	100	2:3	24.0	2.3
4	2023-07-06	120	1:1	25.0	1.0
5	2023-08-07	104	6:7	25.2	2.7
6	2023-09-11	100	13:12	24.8	3.0
7	2023-10-09	120	17:13	17.0	1.8
8	2023-11-06	120	1:1	13.0	2.2
9	2023-12-06	120	13:17	6.0	3.5
10	2024-01-07	104	1:1	2.0	2.1
11	2024-02-17	100	13:12	15.0	1.9
12	2024-03-17	120	19:11	18.0	2.0
Total/Mean	-	1332	-	-	2.3±0.5

Table 2 Camera parameters

Parameters	Value
Camera model	Nikon D750
Focal length	50 mm
Aperture	f/1.8-f/16
Shutter speed	2-30 s
ISO range	800-3200

a. On-site salvage c. Oxygenation for survival e. Annotation data

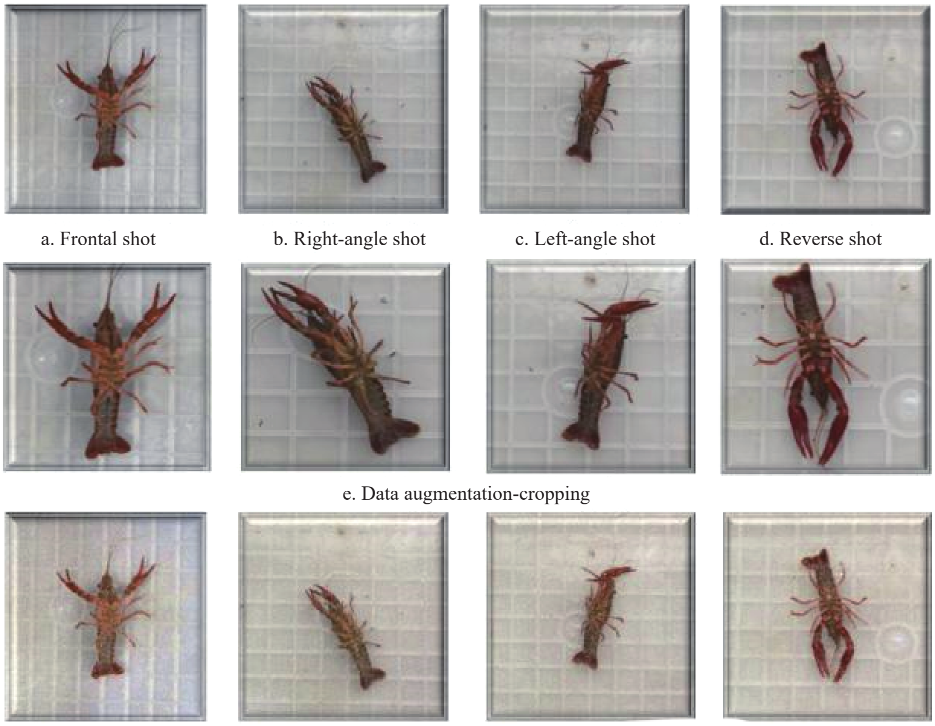
b. Selection d. Scene shooting f. Encapsulation storage

Figure 5 Data collection and imaging process

To comprehensively preserve morphological information for subsequent annotation and model training, a multi-angle imaging protocol was adopted. Each *Procambarus clarkii* specimen was imaged from four standardized perspectives, including frontal, rear, left lateral, and right lateral views. The frontal and rear views were arranged perpendicular to the head-abdomen axis to clearly present the cephalothorax morphology, gonadal structures, and ventral limb positions under static conditions. The left and right lateral views were used to capture the curvature of the abdominal segments, the spatial distribution of ventral limbs, and lateral gonadal characteristics. This multi-view imaging strategy provided more complete spatial and morphological information for gonad identification.

3.1.2 Data preprocessing

To improve dataset diversity and model robustness, the acquired images were further processed by cropping and data augmentation. Cropping was used to focus on key anatomical regions, while augmentation operations, including Gaussian noise, color adjustment, contrast variation, and hue transformation, were applied to simulate illumination changes and imaging artifacts commonly encountered in aquaculture environments. After preprocessing, the dataset was expanded to 5000 images while preserving the core morphological characteristics of the gonadal regions. Representative examples of multi-view original images and augmented samples are shown in Figure 6.


a. Frontal shot b. Right-angle shot c. Left-angle shot d. Reverse shot

e. Data augmentation-cropping

f. Data augmentation-Gaussian noise

Figure 6 Data augmentation processing

All gonad images were manually annotated using LabelMe software to construct the annotated dataset. During annotation, special attention was paid to the morphological differences between male and female gonads, including the branching structure of the testis and the characteristic morphology of the ovary. After annotation, the dataset was randomly divided into training, validation, and test sets at a ratio of 8:1:1. Stratified sampling was adopted to ensure that the sex ratio and maturity distribution in each subset were consistent with the original dataset.

3.2 Experimental settings and evaluation metrics

The experimental platform was configured with an Intel Xeon E3-1230 V2 processor and an NVIDIA GP100GL graphics card. GPU parallel computing was enabled by CUDA 10.2 and cuDNN. The software environment consisted of Python 3.9.10, PyCharm, and the required third-party libraries. The input image size was set to 640×640, the number of training epochs was 50, the batch size was 4, and the learning rate was 1×10^{-4}. Each experiment was repeated three times with different random seeds, and the average results were used for evaluation.

To evaluate the performance of SCM-DETR in *Procambarus clarkii* gonad recognition, Precision, Recall, mAP@0.50, mAP@0.50-0.95, and Parameters (Params) were used as evaluation metrics. Among them, mAP@0.50 denotes the mean Average Precision at an IoU threshold of 0.5, while mAP@0.50-0.95 represents the average mAP over IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05. Precision and Recall are used to measure the accuracy and completeness of detection results, respectively, and Params denotes the total number of trainable parameters in the model. The formulas of these metrics are given as follows:

$$\text{mAP}=\frac{1}{N}\sum_{i=1}^{N}\text{AP}_i \tag{9}$$

where, N denotes the total number of categories; and AP_i denotes the average precision of the i-th category.

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{10}$$

where, False Negative (FN) denotes the number of missed target instances.

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \tag{11}$$

where, True Positive (TP) denotes correctly detected positive instances; and False Positive (FP) denotes incorrectly detected positive instances.

3.3 Comparative experiments and result analysis

To verify the effectiveness of the proposed SCM-DETR, comparative experiments were conducted on the PGD dataset using representative one-stage and Transformer-based detection models, including YOLOv5, YOLOv8, YOLOv9s, YOLOv10s, YOLOv11, DETR, Deformable DETR, DINO, RT-DETR-R18, RT-DETR-R50, and RT-DETR-R101. The comparison results are listed in Table 3.

As listed in Table 3, SCM-DETR-R18 achieved the best overall performance among all compared methods, with a precision of 96.2%, a recall of 96.9%, an mAP@0.50 of 95.8%, and an mAP@0.50:0.95 of 64.6%. Compared with the baseline RT-DETR-R18, SCM-DETR-R18 improved precision, recall, mAP@0.50, and mAP@0.50:0.95 by 1.2, 2.6, 1.9, and 1.1 percentage points, respectively, while slightly reducing the parameter count from 21.2 M to 21.1 M. These results indicate that the proposed multi-dimensional feature fusion and enhancement strategy can effectively improve gonad detection accuracy without increasing model complexity.

Table 3 Comparative experimental results of different detection models

Model	*P*/%	*R*/%	Parameters/ M	mAP@0.50/ %	mAP@ 0.50:0.95/%
YOLOv5[31]	91.0	90.6	25.0	93.9	61.2
YOLOv8[32]	89.2	91.9	30.1	92.1	51.1
YOLOv9s[33]	93.2	92.3	71.7	94.0	62.0
YOLOv10s[34]	80.9	80.2	72.2	84.5	53.9
YOLOv11[35]	81.0	85.9	25.8	90.6	51.0
DETR[36]	88.5	84.7	41.3	88.9	56.8
Deformable DETR[37]	90.6	88.4	50.0	92.3	59.8
DINO[38]	92.1	90.2	48.0	94.0	60.7
RT-DETR-R18[22]	95.0	94.3	21.2	93.9	63.5
RT-DETR-R50[22]	95.1	96.0	42.0	94.5	62.2
RT-DETR-R101[22]	**96.4**	94.9	74.7	95.1	62.0
SCM-DETR-R18	96.2	**96.9**	**21.1**	**95.8**	**64.6**

Compared with the YOLO-series and Transformer-based detection methods, SCM-DETR-R18 achieved a better balance between detection accuracy and model efficiency. Although RT-DETR-R50 and RT-DETR-R101 obtained competitive detection results, their parameter sizes increased considerably. In particular, RT-DETR-R101 achieved a slightly higher precision than SCM-DETR-R18, with a difference of only 0.2 percentage points. However, SCM-DETR-R18 improved recall, mAP@0.50, and mAP@0.50:0.95 by 2.0, 0.7, and 2.6 percentage points, respectively, while using only 21.1 M parameters, which was significantly lower than the 74.7 M parameters of RT-DETR-R101. These results indicate that SCM-DETR-R18 achieved the best overall detection performance with lower model complexity, demonstrating a more favorable trade-off between detection accuracy and parameter scale. Therefore, the proposed model is more suitable for automatic gonadal detection and sex identification of *Procambarus clarkii* under complex imaging conditions.

To further evaluate the performance advantages of SCM-DETR-R18 in male and female gonad detection, RT-DETR-R18, the baseline model, and RT-DETR-R101, a strong comparison model, were selected for focused comparison. In Table 4, the average mAP@0.50 and mAP@0.50:0.95 represent the mean detection performance of the three categories, namely Crayfish, female, and male. As shown in Table 4, for the female category, SCM-DETR-R18 achieved an AP_female@0.50 of 93.3%, representing improvements of 3.3 and 2.7 percentage points over RT-DETR-R18 and RT-DETR-R101, respectively. Its mAP_female@0.50:0.95 reached 46.3%, with corresponding improvements of 1.8 and 5.2 percentage points, respectively. Considering that female gonads are generally characterized by small target size, blurred boundaries, and weak texture contrast, these results demonstrate the effectiveness of SCM-DETR-R18 in enhancing fine-grained feature representation and localization capability. For the male category, SCM-DETR-R18 achieved AP_male@0.50 and mAP_male@0.50:0.95 values of 96.5% and 60.4%, respectively, which were 1.9 and 1.4 percentage points higher than those of RT-DETR-R18. Compared with RT-DETR-R101, although the AP_male@0.50 of SCM-DETR-R18 was 1.0 percentage point lower, its mAP_male@0.50:0.95 was 1.4 percentage points higher, indicating better boundary localization accuracy under the stricter multi-IoU threshold evaluation.

Table 4 Class-wise detection performance comparison on the PGD dataset

Index	Model	YOLOv5	YOLOv8	YOLOv9s	YOLOv10s	YOLOv11	DETR
mAP@ 0.50/%	Crayfish	99.0	96.0	98.1	87.2	95.2	92.5
	Female	85.0	84.2	86.5	82.1	86.0	84.3
	Male	97.6	96.2	97.5	84.3	90.7	90.0
	Average	93.9	92.1	94.0	84.5	90.6	88.9
mAP@ 0.50:0.95/ %	Crayfish	87.0	74.2	89.0	76.5	74.0	81.2
	Female	40.5	34.1	40.0	37.2	35.0	36.2
	Male	56.1	45.1	57.0	48.1	44.0	53.1
	Average	61.2	51.1	62.0	53.9	51.0	56.8
mAP@ 0.50/%	Crayfish	95.8	97.2	97.1	97.5	97.3	**97.5**
	Female	86.4	88.5	90.0	91.0	90.6	**93.3**
	Male	94.7	96.3	94.6	95.0	**97.5**	96.5
	Average	92.3	94.0	93.9	94.5	95.1	**95.8**
mAP@ 0.50:0.95/ %	Crayfish	84.1	86.1	87.0	86.8	85.9	**87.2**
	Female	39.7	41.8	44.5	41.9	41.1	**46.3**
	Male	55.7	54.3	59.0	57.9	59.0	**60.4**
	Average	59.8	60.7	63.5	62.2	62.0	**64.6**

Overall, SCM-DETR-R18 improved the average detection performance of the Crayfish, female, and male categories under the same R18 backbone and enhanced the recognition and localization accuracy of male and female gonads, with a particularly notable improvement in the female category. Moreover, compared with RT-DETR-R101 using a deeper backbone, SCM-DETR-R18 still showed advantages in female gonad detection and in the stricter localization metric for the male category. These results indicate that the proposed model effectively alleviates the challenges caused by small targets, weak textures, and blurred boundaries in *Procambarus clarkii* gonad detection, thereby improving the accuracy and robustness of male and female gonad detection.

To further verify the ability of SCM-DETR to focus on key gonad-related features, YOLOv10s, YOLOv11, RT-DETR-R18, and RT-DETR-R50 were selected for heatmap visualization. YOLOv10s and YOLOv11 represent recent YOLO-series detectors; RT-DETR-R18 is the baseline model in this study, and RT-DETR-R50 is a comparison model with a deeper backbone. Figure 7 shows the attention distributions of gonad regions from the PGD dataset. The gonad location is marked with a yellow bounding box, and the heatmap is visualized using a color gradient from blue, cyan, and yellow to red, where warmer colors indicate higher attention responses.

As shown in Figure 7, the YOLO-series models mainly focused on locally salient regions, such as claws and tails, while their coverage of the gonad region within the yellow bounding box was insufficient. The RT-DETR-series models expanded the attention range over the gonad region, but uneven attention distribution and insufficient boundary responses remained. In contrast, SCM-DETR produced more continuous and concentrated high responses in the gonad region. Its heatmap more completely covered the gonad contour within the yellow bounding box and significantly reduced low-attention regions. These results indicate that SCM-DETR can more effectively capture fine-grained discriminative features of gonad regions and improve the perception of small targets.

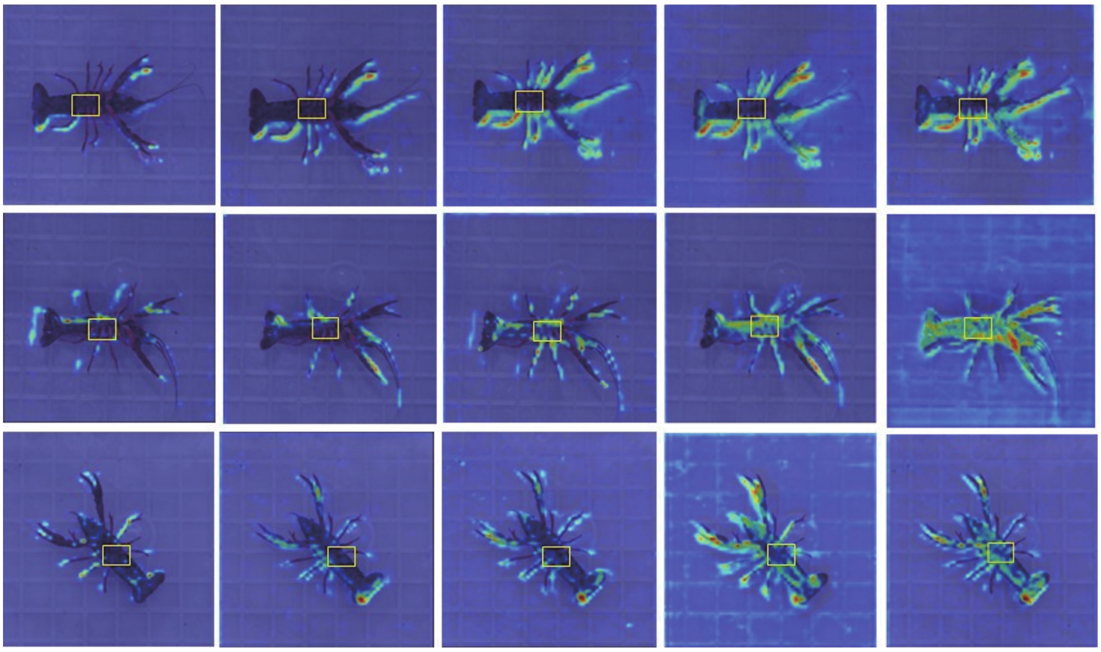


Figure 7 Comparison of visual heat maps

3.4 Generalization evaluation on the VisDrone dataset

To further evaluate the generalization capability of SCM-DETR-R18 in public small-object detection scenarios, additional experiments were conducted on the VisDrone dataset. The VisDrone dataset contains a large number of small[39], densely distributed, and partially occluded objects in aerial images, which share certain challenges with *Procambarus clarkii* gonad detection, such as small target size, complex backgrounds, and weak visual features. Therefore, this dataset was used to further assess the adaptability of the proposed model to complex small-object detection tasks.

As shown in Table 5, SCM-DETR-R18 achieved a precision of 56.3%, a recall of 38.8%, an mAP@0.50 of 40.6%, and an mAP@0.50:0.95 of 23.9% on the VisDrone dataset, obtaining the best results among all compared models. Compared with the baseline RT-DETR-R18, SCM-DETR-R18 improved precision, recall, mAP@0.50, and mAP@0.50:0.95 by 3.4, 0.9, 1.5, and 0.3 percentage points, respectively, while reducing the number of parameters from 26.84 M to 23.2 M. Compared with RT-DETR-R101, which uses a deeper backbone, SCM-DETR-R18 still improved mAP@0.50 and mAP@0.50:0.95 by 0.2 and 0.1 percentage points, respectively, while reducing the parameter count by 63.6 M. These results indicate that SCM-DETR-R18 achieves better small-object detection performance with lower model complexity.

Overall, SCM-DETR-R18 not only performs well in the *Procambarus clarkii* gonad detection task but also achieves stable performance on the public VisDrone dataset. This further demonstrates that the proposed multi-dimensional feature fusion and enhancement strategy can improve feature representation for small-scale, densely distributed, and complex-background targets, indicating good generalization capability and application potential.

3.5 Ablation experiment

As listed in Table 6, Model 1 used only the RT-DETR-R18 baseline model and achieved a precision of 95.0%, a recall of 94.3%, an mAP@0.50 of 93.9%, and an mAP@0.50:0.95 of 63.5%, providing the baseline results for the subsequent ablation analysis. After introducing MSDM, Model 2 shows improvements in all metrics. The precision and recall increased to 95.9% and 96.0%, respectively, while mAP@0.50 and mAP@0.50:0.95 increased to 94.9% and 63.9%, respectively. This indicates that MSDM enhances feature representation for gonad regions by integrating high-level semantic information and fine-grained detail features. After further introducing the channel branch CN of CSFN in Model 3, the model performance continued to improve, with precision, recall, mAP@0.50, and mAP@0.50:0.95 reaching 96.1%, 96.4%, 95.3%, and 64.2%, respectively. This suggests that CN helps strengthen important channel responses and improves the discriminative ability for gonad targets. In Model 4, the spatial branch SN was further added to construct the complete SCM-DETR-R18, which achieved the best performance, with precision, recall, mAP@0.50, and mAP@0.50:0.95 values of 96.2%, 96.9%, 95.8%, and 64.6%, respectively. Compared with the baseline model, the complete model improved precision, recall, mAP@0.50, and mAP@0.50:0.95 by 1.2, 2.6, 1.9, and 1.1 percentage points, respectively.

Table 5 Detection performance comparison of different models on the VisDrone dataset

Model	*P*/%	*R*/%	Parameters/M	mAP@0.50/%	mAP@0.50:0.95/%
YOLOv5[31]	34.2	24.6	28	23.5	12.6
YOLOv8[32]	36.5	25.3	35.7	24.3	13.2
YOLOv9s[33]	45.6	33.9	79.3	34.8	20.2
YOLOv10s[34]	49.7	38.7	82.2	39.7	23.7
YOLOv11[35]	47.8	36.4	31.6	38.2	22.5
DETR[36]	44.8	32.5	41.3	35.4	20.8
Deformable DETR[37]	50.5	36.9	52.1	38.6	22.9
DINO[38]	51.2	37.4	53.5	39	23.3
RT-DETR-R18[22]	52.9	37.9	26.84	39.1	23.6
RT-DETR-R50[22]	55.8	38.5	47.1	39.5	23.8
RT-DETR-R101[22]	56	38.6	86.8	40.4	23.8
SCM-DETR-R18	**56.3**	**38.8**	**23.2**	**40.6**	**23.9**

These results demonstrate that MSDM, CN, and SN all contribute to improving model performance. Their collaborative effect further enhances feature representation, target localization, and overall detection capability for *Procambarus clarkii* gonad regions.

Table 6 Ablation study results of SCM-DETR-R18

Serial number	RT-DETR-R18	MSDM	CSFN (CN)	CSFN (SN)	*P*/%	*R*/%	mAP@0.50/%	mAP@0.50-0.95/%
1	√				95.0	94.3	93.9	63.5
2	√	√			95.9	96	94.9	63.9
3	√	√	√		96.1	96.4	95.3	64.2
4	√	√	√	√	**96.2**	**96.9**	**95.8**	**64.6**

Figure 8 presents the heatmap visualization of the ablation models for *Procambarus clarkii* gonad detection. The gonad region is marked with a yellow bounding box in the original image, providing a visual reference for evaluating the attention distributions of different models. For the RT-DETR-R18 baseline model, the attention response around the gonad region is relatively scattered, and the coverage of key gonad areas is insufficient. After introducing MSDM, the model shows enhanced attention to both the boundary and central regions of the gonad, indicating that the fusion of semantic and detailed features improves the representation of weak-texture targets.

With the further introduction of the channel branch CN in CSFN, the attention distribution becomes more continuous and uniform within the gonad region, suggesting that CN strengthens discriminative channel responses related to gonad morphology and texture. After adding the spatial branch SN, the complete SCM-DETR-R18 produces a more concentrated and complete attention response over the gonad region, with improved coverage of the overall gonad contour within the yellow bounding box. These progressive heatmap changes demonstrate that the proposed modules gradually guide the model from scattered local attention to more accurate and complete gonad feature perception, thereby verifying the effectiveness of MSDM, CN, and SN in improving detection performance.

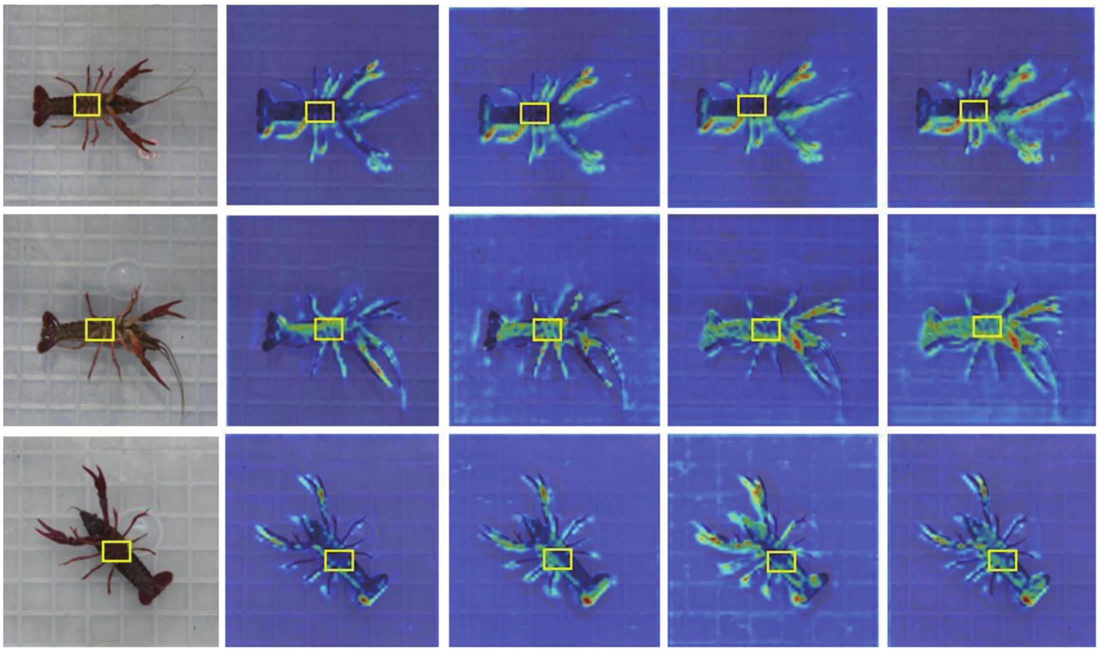


Figure 8 Visual comparison chart of the ablation experiment

4 Conclusions

This study proposed SCM-DETR, an end-to-end detection framework for sex identification of *Procambarus clarkii* based on gonad detection. The proposed model was designed to address the challenges of small target size, weak texture contrast, blurred boundaries, and partial occlusion in gonad images. In SCM-DETR, the multi-dimensional semantics and details fusion method (MSDM) was introduced to integrate high-level semantic information with fine-grained detail features, thereby improving feature representation and localization accuracy. In addition, the Channel-Spatial Focus Network (CSFN) was developed to enhance channel- and spatial-level discriminative features, enabling the model to better capture subtle differences in gonad morphology, color, and texture between female and male individuals.

Experiments on the self-built PGD dataset showed that SCM-DETR-R18 achieved an mAP@0.50 of 95.8% and an mAP@0.50:0.95 of 64.6%. Specifically, the AP@0.50 values for female and male gonads reached 93.3% and 96.5%, respectively. Compared with the baseline RT-DETR-R18, SCM-DETR-R18 improved mAP@0.50 and mAP@0.50:0.95 by 1.9 and 1.1 percentage points, respectively, while maintaining a relatively low parameter count. These results indicate that the proposed model achieves a favorable balance between detection accuracy and model complexity. Therefore, SCM-DETR can provide an effective technical solution for automated sex identification of *Procambarus clarkii* and has potential application value in intelligent aquaculture management.

Future work will focus on improving the generalization capability and functional applicability of SCM-DETR. Larger datasets covering different growth stages, imaging conditions, and culture environments will be collected to enhance model robustness. Multi-label learning and fine-grained classification strategies will also be incorporated to support more comprehensive biological trait analysis. In addition, cross-species validation will be conducted to further evaluate the scalability of the proposed method for other aquatic species.

Acknowledgements

The authors acknowledge that this work was financially supported by the Shandong Province Key R&D Program Project (Grant No. 2022TZXD005), the Shandong Province Science and Technology SMEs Innovation Capacity Enhancement Project

(Grant No. 2021TSGC1003), the General Program of the National Natural Science Foundation of China (Grant No. 42476107), and the General Program of the Shandong Provincial Natural Science Foundation (Grant No. ZR2024MD065).

[References]

[1] Campana S E. Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods. Journal of Fish Biology, 2001; 59(2): 197–242.

[2] Hutchinson C E, TenBrink T T. Age determination of the Yellow Irish Lord: Management implications as a result of new estimates of maximum age. North American Journal of Fisheries Management, 2011; 31(6): 1116–1122.

[3] Han G F, Liu Y J, Cai J W, Duan E Z, Shi Z F, Zhao S D, et al. Deep learning-based gender recognition in Cherry Valley ducks through sound analysis. Animals, 2024; 14. doi: 10.3390/ani14203017.

[4] Nontunha N, Tinikul R, Chaiyamoon A, Vetkama W, Thongbuakaew T, Chaichotranunt S, et al. The effects of prostaglandin E2 on gonadal development and germ cell proliferation, and its presence during the gonadal cycle in the sea cucumber, *Holothuria scabra*. Aquaculture, 2022; 555: 738201.

[5] Lang N, Snæbjarnarson V, Cole E, Aodha O M, Igel C, Belogine S, et al. From coarse to fine-grained open-set recognition. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024; pp.17804–17814. doi: 10.1109/CVPR52733.2024.01686.

[6] Zhang S, Chi C, Yao Y, Lei Z, Li S Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle: IEEE, 2020; pp.9756–9765. doi: 10.1109/CVPR42600.2020.00978.

[7] Wang Y, Li X, Weng S, Zhang G, Yue H, Feng H, et al. KD-DETR: Knowledge distillation for detection transformer with consistent distillation points sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024; pp.16016–16025.

[8] Xing B W, Sun M, Liu Z C, Guan L W, Han J T, Yan C X, et al. Sonar fish school detection and counting method based on improved YOLOv8 and BoT-SORT. Journal of Marine Science and Engineering, 2024; 12(6): 964.

[9] Li Y J, Hu Z Y, Tu W, Zhang P, Wei S X, Yu H, et al. Dense fish population counting detection based on improved YOLOv7. Journal of Guangdong Ocean University, 2024; 44(2): 115–123. (in Chinese)

[10] Geng C X, Wang A M, Yan T Y, Yu A N, Zhang H X, Zhang W X, et al. A *Procambarus clarkii* morphological feature recognition algorithm based on improved YOLOv8n loss function. Fishery Modernization, 2024; 51(5): 72–80. (in Chinese)

[11] Chen Y F, Yue J, Li Z B, Yang J M, Wang W J. Pacific oyster gonad identification and grayscale calculation based on unapparent object detection. In: Liu Q, et al. (Ed.). Pattern Recognition and Computer Vision (PRCV 2023), 2023; 14436: 94–106.

[12] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Computer Vision – ECCV 2020, 2020; 12346: 213–229.

[13] Hou X Q, Liu M Q, Zhang S L, Wei P, Chen B D. Salience DETR: Enhancing detection transformer with hierarchical salience filtering refinement. 2024 IEEE/CVF Conference Computer Vision Pattern Recognition (CVPR), Seattle: IEEE, 2024; pp.17574–17583. doi: 10.1109/CVPR52733.2024.01664.

[14] Zhao Y A, Lv W Y, Xu S L, Wei J M, Wang G Z, Dang Q Q. DETRs beat YOLOs on real-time object detection. In: 2024 IEEE/CVF Conference Computer Vision and Pattern Recognition (CVPR), Seattle: IEEE, 2024; pp.16965–16974. doi: 10.1109/CVPR52733.2024.01605.

[15] Li Y X, Hou Q B, Zheng Z H, Cheng M M, Yang J, Li X. Large selective kernel network for remote sensing object detection. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris: IEEE, 2023; pp.16748–16759. doi: 10.1109/ICCV51070.2023.01540.

[16] Zhu X Z, Su W J, Lu L W, Li B, Wang X G, Dai J F. Deformable DETR: Deformable transformers for end-to-end object detection. arXiv preprint, 2020; arXiv: 2010.04159.

[17] Meng D P, Chen X K, Fan Z J, Zeng G, Li H Q, Yuan Y H. Conditional DETR for fast training convergence. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal: IEEE, 2021; pp.3651–3660. doi: 10.1109/ICCV48922.2021.00363.

[18] Liu S, Li F, Zhang H, Yang X, Qi X B, Su H, et al. DAB-DETR: Dynamic anchor boxes are better queries for DETR. arXiv preprint, 2022; arXiv: 2201.12329.

[19] Wang Y, Zhang Y, Jiang Y, Yuan L. DQ-DETR: Dual query detection transformer for phrase extraction and grounding. Proc Eur Conf Comput Vis (ECCV), 2024.

[20] Jia D, Yuan Y H, He H D, Wu X P, Yu H J, Lin W H. DETRs with hybrid matching. In: 2023 IEEE/CVF Conference Computer Vision and Pattern Recognition (CVPR), Vancouver: IEEE, 2023; pp.19702–19712. doi: 10.1109/CVPR52729.2023.01887.

[21] Chen Q, Chen X K, Wang J, Zhang S, Yao K, Feng H C. Group DETR: Fast DETR training with group-wise one-to-many assignment. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris: IEEE, 2023; pp.6610–6619. doi: 10.1109/ICCV51070.2023.00610.

[22] Zhang H X, Zhang B, Liu K, Gan Z X, Zhu G N. UAV-DETR: Efficient end-to-end object detection for unmanned aerial vehicle imagery. 2025 IEEE/RSJ International Conference on Intelligence Robots and Systems (IROS), 2025; pp.15143-15149. doi: 10.1109/IROS60139.2025.11246176.

[23] Zhang Y X, Hu Z Y, Liu J H, Li Y J, Lin J J, Wang Y, et al. PUFFER-DETR: Tiger puffer similar abnormal behavior recognition based on transformer. Aquacultural Engineering, 2025; 111: 102552.

[24] Yang H Y, Wang J, Bo Y M, Wang J C. ISTD-DETR: A deep learning algorithm based on DETR and super-resolution for infrared small target detection. Neurocomputing, 2025; 621: 129289.

[25] Huang Y X, Liu H I, Shuai H H, Cheng W H. DQ-DETR: DETR with dynamic query for tiny object detection. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G. (Ed.) In: Computer Vision – ECCV 2024. Lecture Notes in Computer Science, 2024; 15134: 290–305.

[26] Girshick R. Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV), Santiago: IEEE, 2015; pp.1440–1448. doi: 10.1109/ICCV.2015.169.

[27] Li H L, Li J, Wei H B, Liu Z, Zhan Z F, Ren Q L. Slim-neck by GSConv: A lightweight design for real-time detector architectures. Journal of Real-Time Image Process, 2024; 21(3): 62.

[28] Kang M, Ting C M, Ting F F, Phan R C W, et al. ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell instance segmentation. Image and Vision Computing, 2024; 141: 105057.

[29] Konopczynski T, Heiman R, Woźnicki P, Gniewek P, Duvernoy M-C, Hallatscheck O, et al. Instance segmentation of densely packed cells using a hybrid model of U-Net and Mask R-CNN. In: Rutkowski L, Scherer R, Korytkowski M, Pedrycz W, Tadeusiewicz R, Zurada J M, et al. (Ed.) Artificial Intelligence and Soft Computing (ICAISC 2020), Part I, 2020; pp.626–635. doi: 10.1007/978-3-030-61401-0_58.

[30] Wang J, Zhang Z T, Wu M H, Ye Y G, Wang S, Cao Y, et al. Improved BlendMask: Nuclei instance segmentation for medical microscopy images. IET Image Process, 2023; 110(10): 2284–2296.

[31] Zhao L J, Jia Y F, Yin W K, Li Z H, Liu C D, Luo H, et al. Recognition and localization method of maize weeding robot based on improved YOLOv5. Int J Agric Biol Eng, 2025; 18(2): 248–258.

[32] Varghese R, Sambath M. YOLOv8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai: IEEE, 2024; pp.1–6. doi: 10.1109/ADICS58448.2024.10533619.

[33] Wang C Y, Yeh I H, Liao H Y M. YOLOv9: Learning what you want to learn using programmable gradient information. In: Computer Vision - ECCV 2024, 2024; pp.1–21. doi: 10.1007/978-3-031-72751-1_1.

[34] Liu R. YOLOv10 tomato ripening detection enhanced by convolutional neural network attention mechanism. In: 2024 4th International Conference on Computer Science and Blockchain (CCSB), Shenzhen: IEEE, 2024; pp.90–93. doi: 10.1109/CCSB63463.2024.10735542.

[35] Song Q, Liu S W, Dai K H, Bai K. YOLOv11-DEC: An improved YOLOv11 model for UAV detection in complex contexts. In: 2025 28th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Compiegne: IEEE, 2025; pp.2404–2409. doi: 10.1109/CSCWD64889.2025.11033339.

[36] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Computer Vision – ECCV 2020, 2020; pp.213–229. doi: 10.1007/978-3-030-58452-8_13.

[37] Zhu X Z, Su W J, Lu L W, Li B, Wang X G, Dai J F. Deformable DETR: Deformable transformers for end-to-end object detection. arXiv preprint, 2020; arXiv: 2010.04159, 2020.

[38] Zhang H, Li F, Liu S L, Zhang L, Su H, Zhu J, et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. arXiv preprint, arXiv: 2203.03605, 2022.

[39] Cao Y R, He Z J, Wang L J, Wang W G, Yuan Y X, Zhang D W. VisDrone-DET2021: The vision meets drone object detection challenge results. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal: IEEE, 2021; pp.2847–2854. doi: 10.1109/ICCVW54120.2021.00319.