frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 16 April 2025
DOI 10.3389/fmars.2025.1539371

OPEN ACCESS

EDITED BY
Muhammad Yasir,
China University of Petroleum (East China), China

REVIEWED BY
Tao Xu,
Henan Institute of Science and Technology, China
Chen Congcong,
Southeast University, China

*CORRESPONDENCE
Shanwen Zhang
wjdw716@163.com

RECEIVED 04 December 2024
ACCEPTED 18 March 2025
PUBLISHED 16 April 2025

CITATION
Wang Z, Guo J, Zhang S and Zhang Y (2025) Sonar-based object detection for autonomous underwater vehicles in marine environments.
Front. Mar. Sci. 12:1539371.
doi: 10.3389/fmars.2025.1539371

Sonar-based object detection for autonomous underwater vehicles in marine environments

Zhen Wang[1,2], Jianxin Guo[1], Shanwen Zhang[1]* and Yucheng Zhang[3]

[1]College of Electronic Information, Xijing University, Xi'an, China, [2]College of Computer Science, Northwestern Polytechnical University, Xi'an, China, [3]College of Computer Science, Xijing University, Xi'an, China

Sonar image object detection plays a crucial role in obstacle detection, target recognition, and environmental perception in autonomous underwater vehicles (AUVs). However, the complex underwater acoustic environment introduces various interferences, such as noise, scattering, and echo, which hinder the effectiveness of existing object detection methods in achieving satisfactory accuracy and robustness. To address these challenges in forward-looking sonar (FLS) images, we propose a novel multi-level feature aggregation network (MLFANet). Specifically, to mitigate the impact of seabed reverberation noise, we designed a low-level feature aggregation module (LFAM), which enhances key low-level image features, such as texture, edges, and contours in the object regions. Given the common presence of shadow interference in sonar images, we introduce the discriminative feature extraction module (DFEM) to suppress redundant features in the shadow regions and emphasize the object region features. To tackle the issue of object scale variation, we designed a multi-scale feature refinement module (MFRM) to improve both classification accuracy and positional precision by refining the feature representations of objects at different scales. Additionally, the CIoU-DFL loss optimization function was constructed to address the class imbalance in sonar data and reduce model computational complexity. Extensive experimental results demonstrate that our method outperforms state-of-the-art detectors on the Underwater Acoustic Target Detection (UATD) dataset. Specifically, our approach achieves a mean average precision (mAP) of 81.86%, an improvement of 7.85% compared to the best-performing existing model. These results highlight the superior performance of our method in marine environments.

KEYWORDS

autonomous underwater vehicles, forward-looking sonar, marine object detection, feature extraction, feature fusion

1 Introduction

As an important underwater exploration means, sonar technology is widely used in the field of marine resource development (Zhang et al., 2022b), marine scientific studies (Grzadziel, 2020), and national defense security (Hansen et al., 2011). A forward-looking sonar (FLS) system can realize the positioning, imaging, and recognition of underwater targets by transmitting sound waves and receiving echo information (Liu et al., 2015), so it has significant advantages in underwater object detection and monitoring tasks. FLS image object detection (Karimanzira et al., 2020) refers to using computer vision and signal processing technology to perform object detection and recognition on the image data obtained by sonar devices to achieve the classification, positioning, and tracking of underwater objects. Different from natural scene images, sonar images are affected by the underwater environment and terrain. As shown in Figure 1, there are serious interferences, such as seabed reverberation noise, sediment shadow region, and background clutter information, in the sonar image. Moreover, FLS images commonly contain underwater objects with different scales and weak feature information, which presents great challenges for sonar object detection.

Compared to object detection in natural scene images, sonar image object detection faces unique challenges due to severe noise interference, complex environments, substantial variations in object scales, and weak saliency of object features. These factors often lead to low detection accuracy, missed detections, and false positives. To address these issues, many methods based on hand-crafted feature extraction combined with classifiers have been proposed. These approaches rely on algorithms for extracting features such as edges, contours, and textures from sonar image regions of interest, followed by classifiers such as support vector machine (SVM) (Chandra and Bedi, 2021), AdaBoost (Collins et al., 2002), and K-nearest neighbors (KNN) (Zhang and Zhou, 2007) for object recognition. For example, Abu and Diamant (2019) developed an object detection framework for synthetic aperture sonar (SAS) images based on unsupervised statistical learning. In the context of FLS images, Zhou et al. (2022b) combined fuzzy C-means and K-means clustering to extract target features through global clustering. Kim and Yu (2017) employed multi-scale feature extraction to obtain Haar-like features from sonar target regions, leveraging AdaBoost to cascade weak classifiers for detection. In efforts to address noise interference, Xinyu et al. (2017) applied fast curve transforms to filter noise and K-means clustering for object region pixel extraction. Zhang et al. (2023) used non-local mean filtering to remove speckle noise and applied super-pixel segmentation to delineate object contours. Although these hand-crafted feature-based methods combined with classifiers have been widely used in sonar object detection, they are limited by their applicability to simple underwater scenes or single-object detection. In more complex underwater acoustic environments and multi-class object detection scenarios, these methods exhibit shortcomings such as insufficient robustness, poor real-time performance, and limited ability to meet high-precision detection requirements.

Benefiting from the robust feature extraction and representation capabilities of convolutional neural networks (CNNs) (Gu et al., 2018), CNN-based methods have gained widespread use in object detection tasks, achieving significantly improved detection performance (Li et al., 2021). These methods leverage frameworks similar to those used in natural scene object detection, such as Faster R-CNN (Ren et al., 2016), You Only Look Once, Version 3 (YOLOv3) (Redmon and Farhadi, 2018), and FPN (Lin et al., 2017a), to detect various types of sonar images, including forward-looking sonar, side-scan sonar, and synthetic aperture sonar. For example, based on the FPN framework, Li et al. (2024) proposed a dual spatial attention network that utilizes a multi-layer convolutional structure to extract features at different scales, with the attention mechanism enhancing feature representation to improve sonar object detection accuracy. To address sonar object detection in complex underwater acoustic environments, Zhao et al. (2023) introduced a composite backbone network that extracts multi-level feature information. Their method uses the shuffle convolution

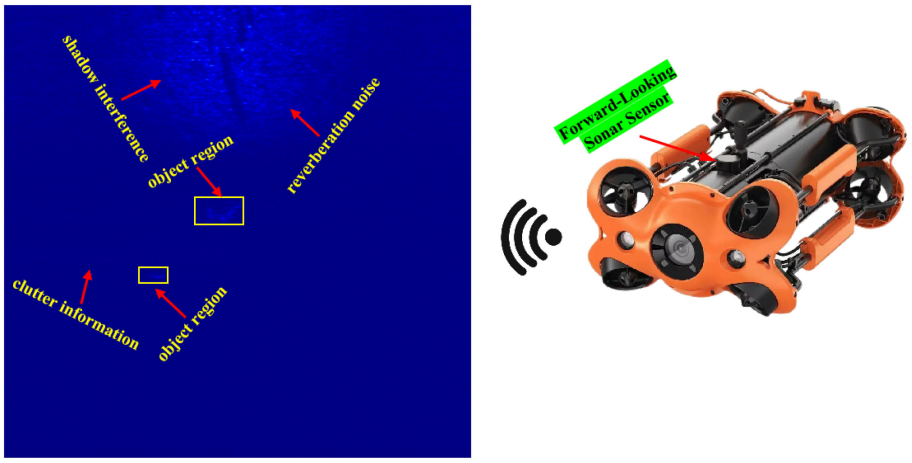

FIGURE 1
Example of a forward-looking sonar image containing object region, seabed reverberation noise, clutter information, and shadow interference.

block attention mechanism and multi-scale feature fusion module to suppress redundant feature interference. Inspired by the two-stage object detection network architecture, Wang et al. (2022d) developed the sonar object detection model, which includes multi-level feature extraction and fusion modules to handle both forward-looking and side-scan sonar detection challenges. Building on the YOLO series of detectors, Zhang et al. (2022a) incorporated the coordinate attention mechanism to extract spatial position features from sonar image regions. They also employed model pruning and compression techniques to enhance the real-time performance of their detector. Yasir et al. (2024) proposed the YOLOShipTracker for ship detection, which has achieved better results in tiny object detection in complex scenes. For tiny object detection, Wang et al. (2022c) introduced the multi-branch shuffle module to reconstruct features at different scales, along with a mixture attention mechanism to strengthen feature representation of small object regions and mitigate clutter interference. Combining CNNs with transformer models, Yuanzi et al. (2022) proposed the TransYOLO detector, which integrates a cascade structure to capture texture and contour features from sonar images, utilizing the attention mechanism for multi-scale feature fusion. Kong et al. (2019) developed the YOLOv3-DPFIN, which achieves effective sonar object detection in complex underwater environments. Their approach employs dense connections for multi-scale feature transmission and the cross-attention mechanism to enhance object region features while reducing reverberation noise interference.

Although CNN-based sonar object detection methods have shown significant improvements over traditional hand-crafted feature extraction techniques, they still face challenges in certain difficult scenarios, such as seabed reverberation noise, shadow interference, object scale variation, and tiny object detection. It is well established that CNN-based object detection methods achieve excellent performance primarily due to their powerful feature extraction capabilities. However, the inherent characteristics of sonar images, such as noise and interference, significantly hinder the feature extraction process of CNN models, making it difficult to fully capture the valuable information necessary for effective sonar image object detection. As illustrated in Figure 2, we provide visualization results of convolution feature heatmaps in challenging scenarios involving seabed reverberation noise interference, shadow interference, clutter, and multi-scale object transformations. These visualizations clearly demonstrate how these interference factors disrupt the feature extraction process of CNN models, leading to a notable decline in detection accuracy across different categories of sonar objects. To address the challenge of sonar image object detection in complex marine acoustic environments, we propose a multi-level feature aggregation network (MLFANet) for FLS image detection. Different from traditional CNN-based methods, MLFANet is specifically designed for challenging sonar detection tasks. The main contributions of this study are as follows:

- **Low-Level Feature Aggregation Module (LFAM):** We introduce the LFAM, a novel module that enhances low-level features and suppresses the impact of seabed reverberation noise, improving feature extraction and object detection in noisy underwater environments. The LFAM significantly enhances the robustness of sonar object detection in the presence of acoustic interference.
- **Discriminative Feature Extraction Module (DFEM):** To handle large-scale shadow regions, we designed the DFEM, which filters redundant features and refines object region representations. The DFEM improves the accuracy of object localization and classification, making MLFANet more efficient in detecting objects even in highly cluttered or shadowed regions.
- **Multi-Scale Feature Refinement Module (MFRM):** We developed the MFRM to address the challenge of object

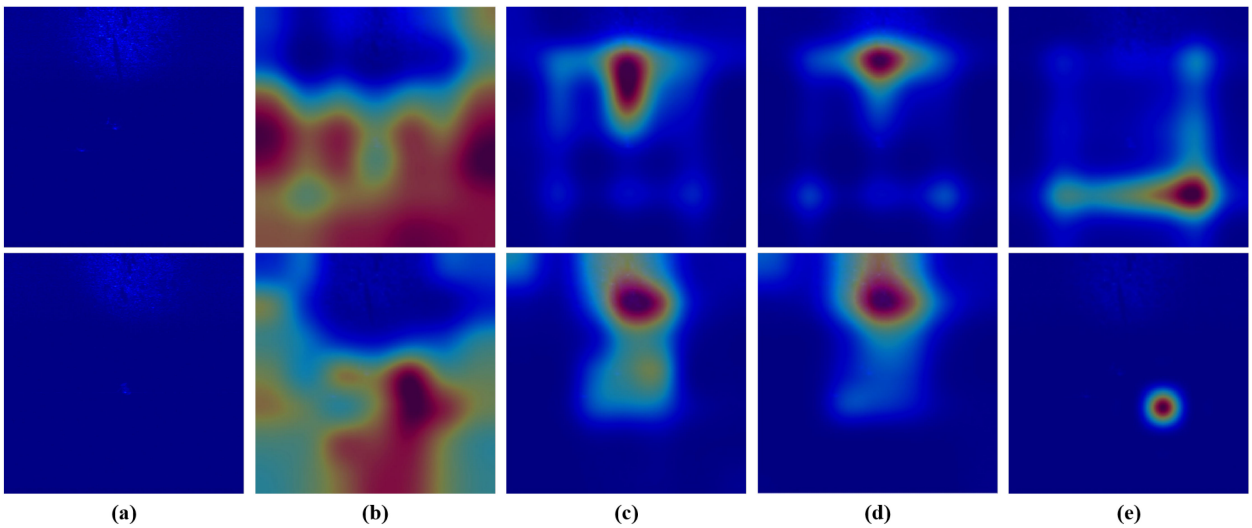

FIGURE 2
Visualization of convolution feature heat map under different interference scenes. **(a)** Two original FLS images. **(b)** Seabed reverberation noise interference. **(c)** Shadow interference. **(d)** Clutter information interference. **(e)** Multi-scale object transformation.

scale variation. The MFRM extracts and fuses fine-grained multi-scale features, enabling the network to handle objects of various sizes more effectively, ensuring that small, medium, and large objects are all accurately detected.

- **CIoU-DFL Loss Function:** To address the issue of object category imbalance in sonar datasets, we propose the CIoU-DFL loss function. This loss function optimizes the model by improving the accuracy of bounding box predictions and reducing computational complexity, particularly for challenging sonar image datasets with skewed category distributions.
- **Extensive Experimental Validation:** We perform extensive experiments on the Underwater Acoustic Target Detection (UATD) dataset, demonstrating that MLFANet outperforms existing state-of-the-art methods in terms of both efficiency and accuracy. Our results highlight the effectiveness of MLFANet in real-world sonar object detection tasks, particularly in complex underwater environments.

The article is organized as follows. Section 2 presents an overview of related works. Section 3 introduces the proposed MLFANet framework and related components. Section 4 presents the experimental results and analysis. Finally, the conclusion is drawn in Section 5.

2 Related works

2.1 Multi-scale feature extraction

For CNN-based object detection methods, multi-scale features play an important function in improving model detection accuracy, fusing global context information, and enhancing model robustness and generalization. Currently, widely used multi-scale feature extraction methods include constructing multi-scale convolution structures (Mustafa et al., 2019), using feature pyramid networks (Lin et al., 2017a), and designing adaptive extraction strategies (Zhou et al., 2022a). Guo et al. (2020) constructed AugFPN to obtain semantic multi-scale features and used residual feature augmentation to highlight the object region feature information. Ma et al. (2020) used the cascade structure to extract multi-scale context information and used feature parameter sharing to establish the correlation of different scale features. To reduce the detail information loss in the multi-scale feature extraction process, Kim et al. (2018b) achieved feature restoration by constructing the global relationship between channel and spatial features. Jiang et al. (2024) used the dense feature pyramid network for small object detection, which uses the multi-scale parallel structure to obtain different scale feature information of the multi-scale object region. MFEFNet (Zhou et al., 2024) uses the efficient spatial feature extraction module to obtain context semantic information and uses a progressive feature extraction strategy to obtain multi-scale features of context information. Tang et al. (2022) constructed a scale-aware feature pyramid structure to obtain multi-scale feature information of the object deformation region and used the feature alignment module to solve the feature offset problem. However, these multi-scale feature extraction methods focus on the extraction of spatial and semantic features, ignoring the important contribution of low-level feature information. Especially for FLS image object detection, low-level features can effectively improve the positioning precision of the object detection model. In this article, we construct the LFAM to obtain low-level multi-scale feature information of the FLS image to improve positioning and recognition accuracy for the sonar object region.

2.2 Contextual feature fusion

Since the contextual information can provide more object region and background information, it can effectively improve the detection accuracy of the object detection model for small object categories. FLS image object detection is a typical small object detection scene, so it is essential to fully mine and fuse the global context feature information. Currently, the commonly used context feature fusion methods include the context feature pyramid (Kim et al., 2018a), global context model (Du et al., 2023), and multi-scale context structure (Wang et al., 2022a). Liang et al. (2019) used the feature pyramid structure to obtain multi-scale context feature information and performed context feature fusion using a spatial-channel reconstruction strategy. Cheng et al. (2020) constructed a cross-scale feature fusion framework to extract local context features and used the region feature aggregation module to achieve context feature fusion. Lu et al. (2021) used the multi-layer feature fusion module to obtain context feature information and introduced a dual-path attention mechanism and multi-scale receptive field module for context feature fine-grained fusion. CANet (Chen et al., 2021) uses a patch attention mechanism to obtain context patch spatial feature information and uses feature mapping and semantic enhancement modules to filter the valuable information of context features. Dong et al. (2022) used deformable convolution and feature pyramids to obtain multi-scale global information and the multi-level feature fusion module is used to fuse local-global context features. These aforementioned context feature fusion methods can effectively fuse feature information of different scales to improve the feature representation for the object region. However, for FLS image object detection, due to the interference of shadow region and clutter information, the existing context feature fusion method cannot solve the feature redundancy problem. In this article, we design the DFEM to suppress redundant feature representation and achieve context feature fusion.

2.3 Visual attention mechanism

An important component of an object detection model, the visual attention mechanism enhances feature representation, solving object deformation and feature correlation modeling. Currently, the attention modules widely used in object detection models include the spatial attention mechanism (Zhu et al., 2019),

channel attention mechanism (Wang et al., 2020), and self-attention mechanism (Shaw et al., 2018). Gong et al. (2022) used the self-attention mechanism to obtain the robust invariant feature information of the object region to enhance the small object region feature representation. Wang et al. (Wang and Wang, 2023) constructed a pooling and global feature fusion self-attention mechanism to obtain the feature correlation and used the feature adaption module for fine-grained feature fusion. Zhu et al. (2018) constructed a cascade attention mechanism to obtain global receptive field information and used dual encoder-decoder attention to reduce feature information loss. Miao et al. (2022) used cross-context attention to obtain local-global feature information and used a spatial-channel attention module to enhance different scale features. To accurately detect multi-scale objects with complex backgrounds, Xiao et al. (2022) designed a pixel attention mechanism to model the pixel correlation information of different object regions and used the self-attention mechanism to enhance the pixel region feature representation. Although the existing visual attention mechanism can effectively enhance the model feature representation and solve the object scale variable problem, for FLS image object detection, due to the serious interference of clutter information and underwater terrain in the object region, the existing attention mechanism struggles to fine-grain enhance the object region feature information, so it cannot obtain satisfactory detection results for small object categories. To solve this problem, inspired by the deformable convolution and attention mechanism, we construct the MFRM to improve the detection accuracy for multi-scale sonar objects by extracting the robust invariant feature information of the object region.

3 Methodology

To solve the problem of object detection in FLS sonar images, based on the YOLOX (Ge et al., 2021) detector, we constructed MLFANet to detect different object categories in sonar images. As shown in Figure 3, the proposed MLFANet introduces the LFAM, DFEM, and MFRM on the basis of the YOLOX detector. Specifically, to improve the object detection performance in complex seabed reverberation noise interference scenes, the LFAM is used to enhance the shallow feature information (*C*1, *C*2, and *C*3) of the backbone network, so that the model can obtain more feature information that is conducive to improving the object positioning precision. Then, to suppress redundant feature representation in deep feature information (*C*4 and *C*5), the DFEM is used to obtain valuable information on deep features to optimize the sonar object detection effect under shadow interference conditions. Moreover, to improve the recognition accuracy of the detector for different categories of sonar objects, we introduce the DFEM into the neck structure, which performs fine-grained fusion of multi-scale feature maps by generating attention weights to further enhance the representation ability of the feature maps and alleviate clutter noise information interference. For the model parameter optimization, we combine CIoU (Zheng et al., 2020) and the DLF (Li et al., 2020) loss function to solve the problem of sample category imbalance and model computational complexity.

3.1 Low-level feature aggregation module

Since the interference of signal intensity difference and reverberation noise, there are many dark areas in sonar images, which makes it difficult for the existing feature extraction network (Elharrouss et al., 2022) to obtain low-level feature information such as texture, edge, and contour of sonar object regions, and the obtained low-level features lead to serious loss in the process of convolutional feature transmission. To solve this problem, we designed the LFAM and embedded it into the backbone network to compensate for the feature information loss of deep convolution by mining the low-level features obtained in the shallow

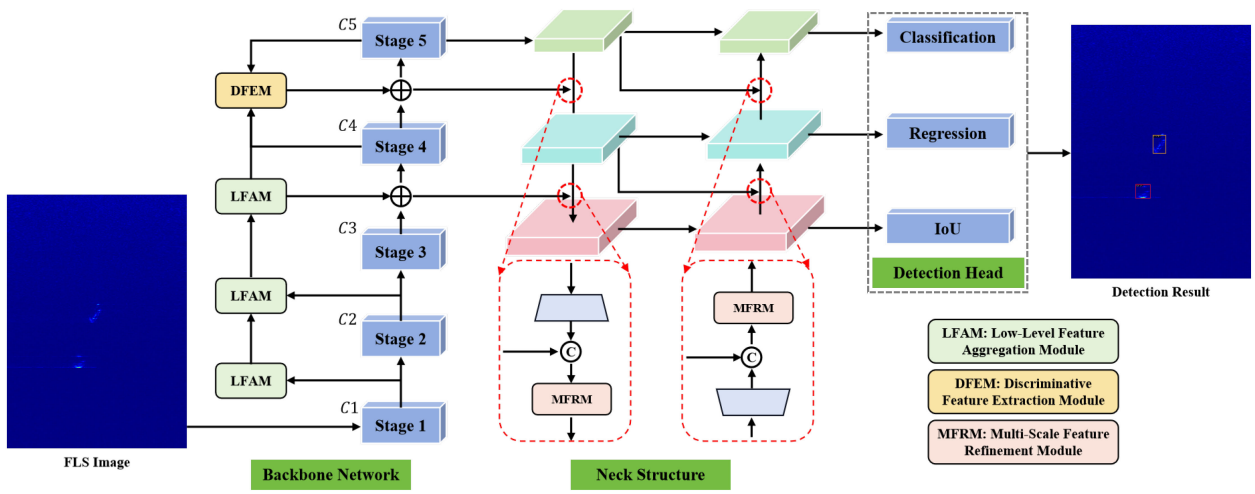

FIGURE 3
The overall architecture of the proposed multi-level feature aggregation network (MLFANet), including the low-level feature aggregation module (LFAM), discriminative feature extraction module (DFEM), and multi-scale feature refinement module (MFRM).

convolution stages. The LFAM is designed to enhance low-level feature information, such as texture, edges, and contours, while suppressing seabed reverberation noise that commonly disrupts the feature extraction process.

The specific structure of the LFAM is shown in Figure 4, where the backbone network consists of five convolution stages, and denotes the feature map obtained in the *lth* convolution stage and $l \in [1, 5]$. The proposed LFAM takes the feature maps C_1, C_2, and C_3 obtained in the shallow convolution stage as input features, and performs feature fusion in turn to generate the aggregation feature map $\boldsymbol{G} \in \mathbb{R}^{C\times H\times W}$, so that it can retain more low-level feature information. The specific fusion process is as follows:

$$\boldsymbol{G} = K_{3\times3}(K_{3\times3}(C_1) \oplus C_2) \oplus C_3 \tag{1}$$

where $K_{3\times3}(\cdot)$ represents the 3 × 3 convolution function for feature map resolution and feature channel adjustment, and $\oplus$ denotes the element-by-element summation operation. The aggregate feature map is used as the output of parallel pooling, which uses different pooling layers to obtain the context information of the aggregate feature map to extract more discriminative low-level features. The parallel pooling consists of different pooling functions, namely $1 \times W$ strip pooling, $H \times 1$ strip pooling, and $S \times S$ spatial pooling and residual connection. For the aggregate feature map $\boldsymbol{G}$ with a size of $H \times W$, the feature map is averaged using strip pooling with a pooling range of $(1, W)$ and $(H, 1)$, which compresses the feature map and encodes feature information along the vertical and horizontal directions. Furthermore, the use of strip pooling establishes long-distance dependencies between discretely distributed feature regions for spatial dimension information in the vertical and horizontal directions and obtains low-level feature information such as edges and contours of the object region in the global dimension. The calculation of strip pooling is as follows:

$$y_w = \frac{1}{H}\sum_{0\le i<H} \boldsymbol{G}(i, W) \tag{2}$$

$$y_h = \frac{1}{H}\sum_{0\le j<W} \boldsymbol{G}(H, j) \tag{3}$$

where $y_w \in \mathbb{R}^{C\times1\times W}$ and $y_h \in \mathbb{R}^{C\times H\times1}$ represent the feature tensors obtained by strip pooling with sizes of 1 × 1 and 3 × 3, respectively. The one-dimensional convolution is used to integrate the adjacent feature information inside the feature tensor, and the bilinear interpolation operation is used to recover the spatial information of feature tensor y_w and y_h. To generate low-level features with rich edges and contours, the feature tensor is fused by using the element-by-element multiplication operation. The calculation process is as follows:

$$z_1 = \mathcal{F}_{ex}(f_{3\times1}(y_w)) \oplus \mathcal{F}_{ex}(f_{1\times3}(y_h)) \tag{4}$$

where $\mathcal{F}_{ex}(\cdot)$ represents the bilinear interpolation operation, and $f_{3\times1}(\cdot)$ and $f_{1\times3}(\cdot)$ represent the one-dimensional convolution operation with the size of 3 × 1 and 1 × 3, respectively. Moreover, the parallel pooling introduces spatial pooling with a range of $S \times S$, which can use rectangular pooling windows to detect densely distributed object region feature information and obtain texture feature information of sonar objects in the local receptive field range. The residual connection is used to preserve the original spatial information of the aggregate feature map $\boldsymbol{G}$, and it is fused with the spatial pooling feature to generate low-level texture feature tensor z_2. The specific calculation process is as follows:

$$z_2 = P_{s\times s}(\boldsymbol{G}) \oplus \boldsymbol{G} \tag{5}$$

where $P_{s\times s}(\cdot)$ denotes the spatial pooling with a size of 1 × 1. For feature tensors z_1 and z_2, the 3 × 3 convolution is used to further extract detailed information, and the feature stitching operation is used to generate feature map $z_3 \in \mathbb{R}^{C\times H\times W}$ with

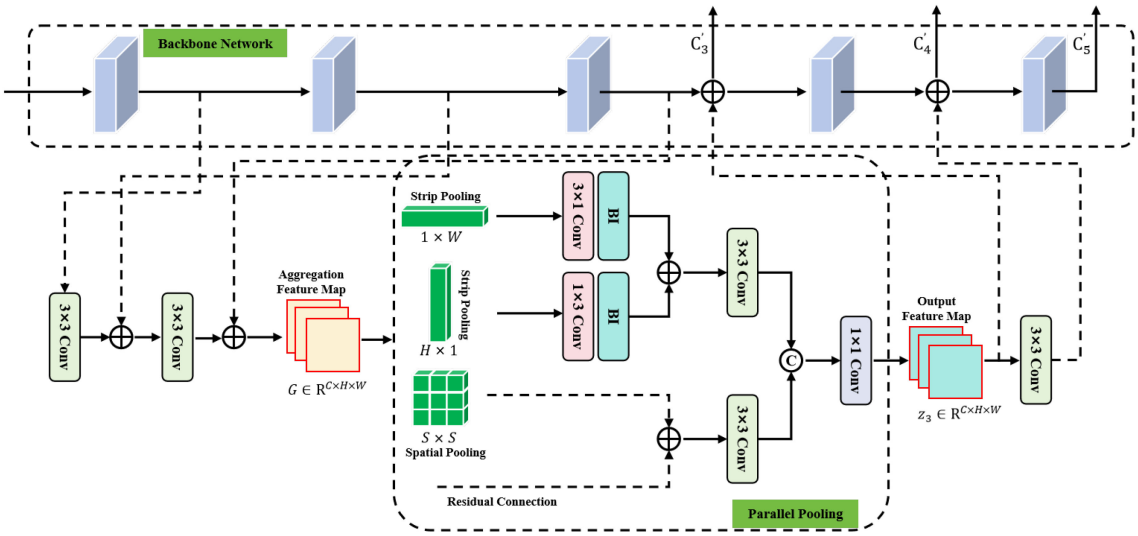

FIGURE 4
The specific structure of the low-level feature aggregation module (LFAM) includes 3×3 convolution, 1 × 1 convolution, 3 × 1, 1 × 3 one-dimensional convolution, element-by-element summation, channel stitching, and bilinear interpolation operation.

more discriminative information. The calculation is as follows:

$$z_3 = K_{1\times1}([K_{3\times3}(z_1); K_{3\times3}(z_2)]) \tag{6}$$

where $K_{1\times1}(\cdot)$ and $K_{3\times3}(\cdot)$ represent convolution operations with sizes of 1×1 and 3×3, respectively, and $[\cdot;\cdot]$ denotes the feature stitching operation on the channel dimension. The feature map z_3 is fused with the features C_3 and C_4 in the deep convolution stage of the backbone network, and input to the subsequent convolution stage to compensate for low-level feature information loss. The feature maps $C_3^{'}$, $C_4^{'}$, and $C_5^{'}$ generated by the fuse operation can retain more effective edge, contour, and texture feature information, which is beneficial for improving the positioning precision for different object categories. The generation process of feature maps $C_3^{'}$, $C_4^{'}$, and $C_5^{'}$ is calculated as follows:

$$C_3^{'} = C_3 \oplus z_3 \tag{7}$$

$$C_4^{'} = K_{3\times3}(z_3) \oplus \mathcal{F}_{\text{conv}}^{4}(C_3^{'}) \tag{8}$$

$$C_5^{'} = \mathcal{F}_{\text{conv}}^{5}(C_4^{'}) \tag{9}$$

where $K_{3\times3}(\cdot)$ represents the convolution operation with a size of 3 × 3, and $\mathcal{F}_{\text{conv}}^{l}(\cdot)$ denotes the *lth* convolution stage. The LFAM leverages feature aggregation and parallel pooling operations to extract discriminative low-level feature information. By preserving key spatial details and reducing noise interference, LFAM enhances the model's ability to detect object boundaries and localization precision.

3.2 Discriminative feature extraction module

Due to the redundant feature interference in the feature extraction process of the convolution operation (Qin et al., 2020), it is difficult to retain valuable tiny object region information. To solve this problem, we propose the DFEM, as shown in Figure 5. The DFEM improves the robustness of feature extraction in shadowed and cluttered regions by suppressing redundant features and enhancing salient object features. For the deep feature information (C_4 and C_5) obtained by the backbone network, given the specific feature mapping $\boldsymbol{X} \in \mathbb{R}^{C\times W\times H}$, where C, H, and W represent the number of channels, width, and height of the feature map, respectively. To mine the local regions with discriminative attributes in convolution features, the obtained deep features are divided into k regions along the W dimension, where each region feature is defined as $\boldsymbol{X}_i \in \mathbb{R}^{C\times W/k\times H}$. The feature description importance factor corresponding to each region is calculated as

$$\boldsymbol{a}_i = \text{SoftMax}(\mathcal{F}_{\text{GAP}}(K_{1\times1}(\boldsymbol{X}_i))) \tag{10}$$

where $K_{1\times1}(\cdot)$ represents the convolution operation with a size of 1 × 1, $\mathcal{F}_{\text{GAP}}(\cdot)$ denotes the global average pooling function, and the softmax function is used for feature normalization. The high importance factor indicates that the region feature significance is strong. By comparing the importance factor of different regions, the region with strong discrimination feature description in W dimension can be located. We use the descriptor $\boldsymbol{Y}$ to denote the positioning region and separate it from the initial feature $\boldsymbol{X}$. The region $\boldsymbol{Y}$ is uniformly split into n sub-regions along the H dimension, and $\boldsymbol{Y}_j \in \mathbb{R}^{C\times W/k\times H/n}$ is used to denote the feature information of each sub-region, where $j \in [1, 2, \ldots, n]$. The calculation of the importance factor for sub-region feature description is as follows:

$$\boldsymbol{b}_j = \text{SoftMax}\left(\mathcal{F}_{\text{GPA}}(K_{1\times1}(\boldsymbol{Y}_j))\right) \tag{11}$$

The normalized importance factor of each sub-region can be used to discriminate the sub-region $\boldsymbol{Y}_J^{'} \in \mathbb{R}^{C\times W/k\times H/n}$ with important feature information in the feature mapping $\boldsymbol{X}$. By using the above feature discrimination process, it can effectively solve the deviation problem of feature extraction and enhance the localization ability for the discriminant feature region. To further mine the valuable information in the feature map, we use the discriminative feature enhancement-suppression strategy to preprocess the sub-region feature $\boldsymbol{Y}_J^{'}$, and obtain the feature maps $\boldsymbol{Y}_e \in \mathbb{R}^{C\times W/k\times H}$ and $\boldsymbol{Y}_s \in \mathbb{R}^{C\times W/k\times H}$. The calculation is as follows:

$$\boldsymbol{Y}_e = \boldsymbol{Y} + \alpha \times (\boldsymbol{E} \otimes \boldsymbol{Y}) \tag{12}$$

$$\boldsymbol{Y}_s = \boldsymbol{S} \otimes \boldsymbol{Y} \tag{13}$$

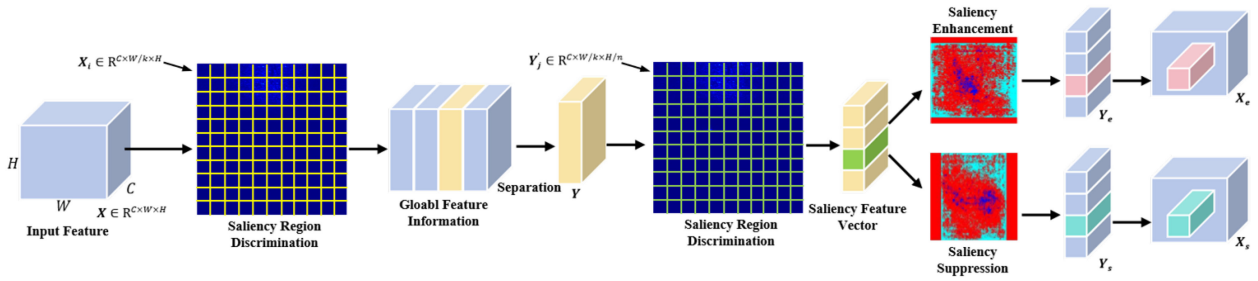

FIGURE 5
The specific structure of the discriminative feature extraction module (DFEM) includes saliency region discrimination, global feature information extraction, saliency region selection, and saliency feature enhancement and suppression.

where $\otimes$ represents the element-by-element multiplication, and the specific calculation for features $\boldsymbol{E}$ and $\boldsymbol{S}$ are as follows:

$$\begin{cases} \boldsymbol{E} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)^{\mathrm{T}} \\ \boldsymbol{e}_j = \begin{cases} \boldsymbol{b}_j, \text{if} \boldsymbol{b}_j = \max(\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n) \\ 0, \text{otherwise} \end{cases} \end{cases} \tag{14}$$

$$\begin{cases} \boldsymbol{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n)^{\mathrm{T}} \\ \boldsymbol{s}_j = \begin{cases} 1-\beta, \text{if} \boldsymbol{b}_j = \max(\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n) \\ 1, \text{otherwise} \end{cases} \end{cases} \tag{15}$$

where α and β denote the coefficients used to control feature enhancement and suppression, respectively. The original feature $\boldsymbol{Y}$ is replaced by feature maps $\boldsymbol{Y}_e$ and $\boldsymbol{Y}_s$, and fused with feature $\boldsymbol{X}_i$ along the $\boldsymbol{W}$ dimension to generate the discriminative enhancement feature $\boldsymbol{X}_e \in \mathbb{R}^{C\times W\times H}$ and the discriminative suppression feature $\boldsymbol{X}_s \in \mathbb{R}^{C\times W\times H}$, respectively. By using a discriminative enhancement operation, it can effectively suppress redundant feature representation to improve the detection accuracy for tiny object categories in sonar images. The DFEM improves the robustness of feature extraction in shadowed and cluttered regions by suppressing redundant features and enhancing salient object features.

3.3 Multi-scale feature refinement module

Due to interference in underwater environments, FLS images contain serious object deformation problems, which makes it difficult for the object detection network to extract fine-grained feature information from the object region, and it is prone to lose the valuable feature information in the shadow region. To solve this problem, we constructed the MFRM and embedded it into the neck structure of the detector to enhance the feature extraction capacity for the deformation object regions. The MFRM consists of region location branch and feature refinement branch, and the specific structure is shown in Figure 6. The MFRM addresses the challenge of detecting objects at varying scales by extracting robust, scale-invariant features and refining multi-scale feature representations. The region location branch is used to position the range of object region, which uses 7 × 7 convolution to obtain local feature information and extract the valuable feature region information for the input feature map $\boldsymbol{X} \in \mathbb{R}^{C\times W\times H}$. The 7 × 7 convolution kernel provides a larger receptive field compared to smaller kernels (e.g., 3 × 3 or 5 × 5), enabling the extraction of richer local feature information. Parallel dilated convolution with different dilation coefficients is used to expand the range of receptive fields and stitch the dilated convolution features to aggregate fine-grained context information. To generate the region attention map, the 3 × 3 convolution is used to encode the context information to obtain the object region features. The calculation is as follows:

$$\boldsymbol{U}_1 = \mathcal{K}_{3\times3}([\mathcal{F}^{6}_{\mathrm{atr}}(K_{7\times7}(\boldsymbol{X})); \mathcal{F}^{12}_{\mathrm{atr}}(K_{7\times7}(\boldsymbol{X}))]) \tag{16}$$

where $K_{3\times3}(\cdot)$ and $K_{7\times7}(\cdot)$ represent convolution operations with sizes of 3 × 3 and 7 × 7, respectively; $\mathcal{F}^{6}_{\mathrm{atr}}(\cdot)$ and $\mathcal{F}^{12}_{\mathrm{atr}}(\cdot)$ denote the dilation coefficients of 6 and 12; $[\cdot;\cdot]$ represents the feature splicing operation on the spatial dimension. The feature refinement branch obtains the fine-grained feature information of the object region through the feature cross-dimensional interaction. This branch performs different global adaptive pooling operations on the input feature map $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$ to obtain global spatial feature information and perform feature space compression. Specifically, 1 × 1 global adaptive average pooling is used to compress the global feature spatial information, 3×3 global adaptive average pooling is used to enhance the global feature representation, and 2 × 2 global adaptive maximum pooling is used to enhance the feature structure information and refine the global feature information obtained by the global adaptive average pooling. The feature tensor obtained by the different pooling operations is converted into vector representation using feature reconstruction to achieve a cross-dimensional interaction of feature

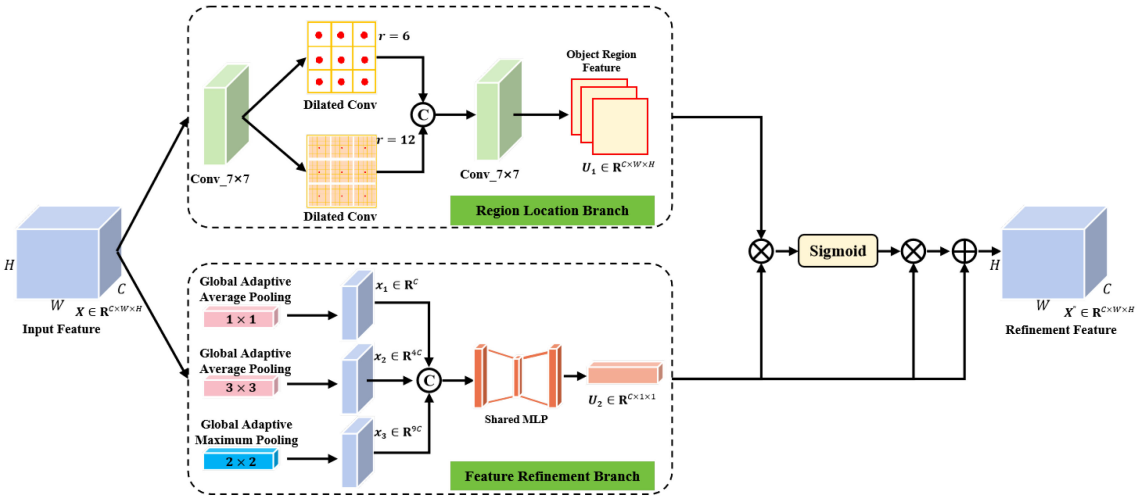

FIGURE 6
The specific structure of the multi-scale feature refinement module (LFAM) includes a region location branch and feature refinement branch.

information on the spatial dimension and fuse it with the object region features retained on the channel dimension to generate one-dimensional feature vectors $x_1 \in \mathbb{R}^C$, $x_2 \in \mathbb{R}^{4C}$ and $x_3 \in \mathbb{R}^{9C}$. The one-dimensional feature vector is spliced to obtain the feature vector $x_4 \in \mathbb{R}^{14C}$ that aggregates rich cross-dimensional interaction feature information. The specific calculation of this process is as follows:

$$\boldsymbol{x}_1 = \mathcal{F}_{\text{resize}}\left(\boldsymbol{P}^1_{\text{avg}}(\boldsymbol{X})\right) \tag{17}$$

$$\boldsymbol{x}_2 = \mathcal{F}_{\text{resize}}\left(\boldsymbol{P}^2_{\text{avg}}(\boldsymbol{X})\right) \tag{18}$$

$$\boldsymbol{x}_3 = \mathcal{F}_{\text{resize}}\left(\boldsymbol{P}^3_{\text{max}}(\boldsymbol{X})\right) \tag{19}$$

$$\boldsymbol{X}_c = [\boldsymbol{x}_1; \boldsymbol{x}_2; \boldsymbol{x}_3] \tag{20}$$

where $P^n_{\text{avg}}(\cdot)$ represents the global adaptive average pooling function with a size of $n \times n$, $P^n_{\text{max}}(\cdot)$ represents the global adaptive maximum pooling function with a size of $n \times n$, and $\mathcal{F}_{\text{resize}}(\cdot)$ feature reconstruction operation. The multi-layer perceptron composed of the fully connected layer and non-linear activation function is used to encode the feature vector $\boldsymbol{X}_c$ to generate the feature descriptor $\boldsymbol{U}_2 \in \mathbb{R}^{C\times1\times1}$. The specific calculation process is as follows:

$$\boldsymbol{U}_2 = \text{MLP}(\boldsymbol{X}_c) = \mathcal{F}_1(\delta(\mathcal{F}_2(\boldsymbol{X}_c))) \tag{21}$$

where $\mathcal{F}_1 \in \mathbb{R}^{C/r\times C}$ and $\mathcal{F}_2 \in \mathbb{R}^{C\times C/r}$ represent different fully connected functions, and set $r = 32$; δ denotes the ReLU activation function. Element-by-element multiplication is used to fuse the region attention mapping $\boldsymbol{U}_1$ and the feature descriptor $\boldsymbol{U}_2$, and the Sigmoid function is used to normalize the feature values to the range of $(0, 1)$ to generate the attention weight $\boldsymbol{M}$. The original feature map $\boldsymbol{X}$ is weighted to achieve object feature adaptive optimization to highlight the object region feature information and reduce the seabed reverberation noise interference. The specific calculation is as follows:

$$\boldsymbol{M} = \sigma(\boldsymbol{U}_1 \otimes \boldsymbol{U}_2) \tag{22}$$

$$\boldsymbol{Y} = \boldsymbol{X} \oplus (\boldsymbol{X} \otimes \boldsymbol{M}) \tag{23}$$

where $\otimes$ represents element-by-element multiplication, σ denotes the Sigmoid activation function, $\oplus$ denotes element-by-element summation, and $\boldsymbol{Y}$ represents the multi-scale refinement feature map. The MFRM uses a dual-branch architecture to effectively model object regions at different scales. The region location branch focuses on coarse object localization, while the feature refinement branch enhances fine-grained feature details through cross-dimensional feature interactions. This ensures that objects of different sizes, from small to large, are accurately detected and classified.

3.4 Loss function optimization

To optimize the proposed MLFANet detector, we combined CIoU (Zheng et al., 2020) and DLF (Li et al., 2020) to calculate the regression loss of the bounding box. The constructed loss function uses DLF loss to obtain the loss probability of the bounding box and object label by calculating the cross-entropy function. The distribution probability of the bounding box is restored as the prediction box, and CIoU is used to calculate the loss value of the prediction box and truth box to achieve the optimization of the prediction box generation process. The calculation of CIoU is as follows:

$$\mathcal{L}_{\text{CIoU}} = 1 - \text{IoU} + \frac{\rho^2(b, b^{gt})}{(c_w)^2 + (c_h)^2} + \frac{4}{\pi}\left(\arctan\frac{w_{gt}}{h_{gt}} - \arctan\frac{w}{h}\right)^2 \tag{24}$$

where IoU represents the intersection in the union of the prediction bounding box and truth bounding box; $\rho^2(b, b^{gt})$ denotes the Euclidean distance between the prediction box and the truth box; h and w represent the height and width of the prediction box; h_{gt} and w_{gt} represent the height and width of the truth box; c_h and c_w denote the height and width of the minimum bounding box consisting of the prediction box and truth box. Since CIoU ignores the problem of sample imbalance, smaller positional offsets lead to significant decreases in IoU values for small object regions in sonar images, while large size object regions will produce an IoU difference. Moreover, since the calculation process involves the solution of inverse trigonometric function, it increases the model computational complexity. To solve this problem, we introduce the normalized Wasserstein distance (NWD) position regression loss function, which uses the two-dimensional Gaussian distribution to calculate the similarity between the prediction box and truth box. The loss calculation process can reflect the true distance between the prediction box and object region distribution, and it has strong robustness to the object scale scaling, so it is more suitable for solving the tiny object detection problem. The specific calculation of the NWD position loss function is as follows:

$$N_a = [cx_a, cy_a, w_a/2, h_a/2]^T \tag{25}$$

$$N_b = [cx_b, cy_b, w_a/2, h_a/2] \tag{26}$$

$$W_2^2(N_a, N_b) = \| (N_a, N_b) \|_2^2 \tag{27}$$

$$\mathcal{L}_{\text{NWD}}(N_a, N_b) = \exp\left(-\sqrt{W_2^2(N_a, N_b)}/C\right) \tag{28}$$

where C denotes the number of object categories; $W_2^2(N_a, N_b)$ denotes the distance measure; N_a and N_b denote the Gaussian distributions modeled by $A = (cx_a, cy_a, w_a, h_a)$ and $B = (cx_b, cy_b, w_b, h_b)$, respectively. Since CIoU is suitable for large size object categories, we combine CIoU and NWD to construct the loss optimization function. The specific calculation is as follows:

$$\mathcal{L}_{\text{CIoU\_NWD}} = \alpha \cdot \mathcal{L}_{\text{CIoU}} + (1 - \alpha) \cdot \mathcal{L}_{\text{NWD}} \tag{29}$$

where α represents the adaptive weight adjustment coefficient, $\mathcal{L}_{\text{CIoU}}$ and $\mathcal{L}_{\text{NWD}}$ denote the CIoU loss function and the NWD loss function, respectively.

4 Experiments and analysis

In this section, we present a detailed description of the forward-looking sonar image dataset, model training strategy, experimental parameter setting, evaluation metrics, ablation studies, and robustness analysis.

4.1 FLS image dataset

To verify the effectiveness and feasibility of the proposed method, we conducted experimental verification on the UATD dataset (Qin et al., 2020) in a real-scene underwater acoustic environment. The dataset was released in 2022 and was provided by Peng Cheng Laboratory, Shenzhen, China. It used Tritech Gemini 1200ik multi-beam forward-looking sonar for image collection. The sonar operates at two acoustic frequencies, 720kHZ for lone-range object detection, and 1,200kHz for enhanced high-resolution imaging at shorter ranges. The data collection sites were located in Golden Pebble Beach in Dalian (39.0904292°N, 122.0071952°E) and Haoxin Lake in Maoming (21.7011602°N, 110.8641811°E).

The dataset contains 9,200 high-resolution original forward-looking sonar images and corresponding manual annotation information. To improve the readability of the sonar images, we performed Gaussian filtering and pseudo-color enhancement on the original images, as shown in Figure 7. The annotation object categories provided by the dataset contain a cube, ball, cylinder, human body model, tire, circle cage, square cage, metal bucket, plane model, and ROV, and the corresponding physical entities and sizes are shown in Figure 8. We present the statistical information of the number of different object categories in Figure 9a, from which it can be seen that the dataset has a serious category imbalance problem. To further analyze the dataset, we calculated the area and aspect ratio of the rectangular label boxes of different object categories, and drew the corresponding histogram, as shown in Figures 9b, c. It can be seen that the different object category sizes were diverse, as the minimum area covered 12 pixels, and the maximum area included 38,272 pixels; the rectangle minimum ratio of length/width was 0.22, and the maximum ratio was 7.95. From the above statistical information, it can be shown that the dataset poses a great challenge to the sonar image object detection task.

4.2 Training strategies and implementation details

The specific details of the dataset and hyperparameters in the experiment are described as follows.

4.2.1 Dataset setting

For the 9,200 forward-looking sonar images contained in the UATD dataset, we randomly split them into the training, verification, and testing sets based on the ratio of 7:2:1. Specifically, the training set contained 6,440 images, the verification set contained 1,840 images, and the testing set contained 920 images. To further improve the model robustness and generalization performance, data augmentation methods including random rotation, image deformation, brightness transformation, image sharpening, and adding noise were used to supplement the number of training set samples. The use of data augmentation can also alleviate the overfitting problem in the

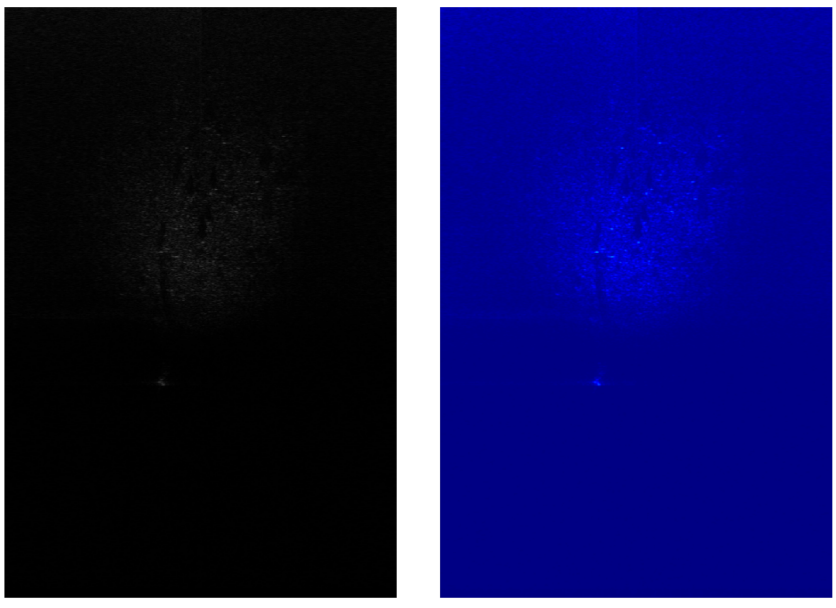

FIGURE 7
The original forward-looking sonar image and the preprocessed image from the UATD dataset. **(a)** original image. **(b)** preprocessed image.

FIGURE 8
The physical sonar target entities and their corresponding size in the UATD dataset. The size is measured in meters.

model training process. Moreover, limited by the device memory, we uniformly scaled the original sonar image to 512×512 pixels in the training process and maintained the original image size for the verification and testing sets.

4.2.2 Training strategies

The experiments were conducted on a workstation equipped with an Intel i9-12900T CPU, 64GB RAM, an NVIDIA GeForce RTX 4090 GPU, and the Ubuntu 18.04 operating system. The code was implemented using the PyTorch 2.1.0 and MMDetection 3.2.0 frameworks. All models were trained and evaluated on the UATD dataset using the same training, validation, and testing splits to ensure fairness. During training, input images were resized to 512×512 pixels, and data augmentation techniques, including random horizontal flipping, random rotation, and color jittering, were applied equally to all models to improve robustness and prevent overfitting. Mixed precision training was employed to enhance training speed and memory efficiency.

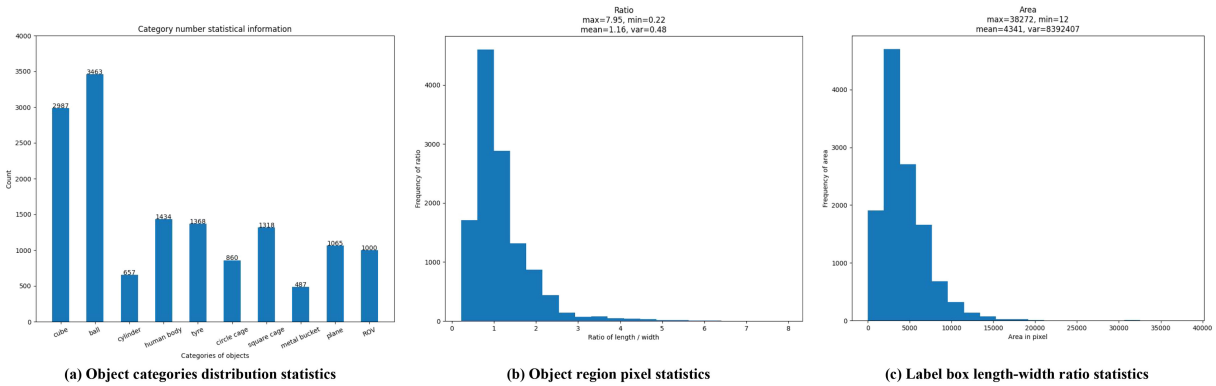

FIGURE 9
An overview of the detailed statistical information of the UATD dataset. **(a)** Object categories distribution statistics. **(b)** Object region pixel statistics. **(c)** Label box length-width ratio statistics.

For the proposed MLFANet, we used a ResNet-50 or ResNet-101 backbone pre-trained on ImageNet. The batch size was set to 8, and the optimizer was SGD with momentum (0.9) and a weight decay of 0.0001. The initial learning rate was set to 0.02 and reduced by a factor of 10 at epochs 8 and 11, with a total of 12 training epochs (1× schedule). To further optimize performance, we adopted a three-stage training strategy: (1) pre-training the backbone on ImageNet with a batch size of 32 and an initial learning rate of 0.001, decayed every 1,000 iterations; (2) fine-tuning the pre-trained backbone on the sonar image dataset with a batch size of 8, an initial learning rate of 0.001, and decay applied every 500 iterations; and (3) training the entire model with a batch size of 16, an initial learning rate of 0.0001, and decay applied every 2,000 iterations. This staged strategy ensured optimal parameter learning and mitigated overfitting.

For the baseline models, we used their standard configurations as described in their original implementations. For example, Faster R-CNN, RetinaNet, Cascade R-CNN, Dynamic R-CNN, and DH R-CNN were trained with a ResNet-50 backbone, a batch size of 8, an initial learning rate of 0.02 (reduced by a factor of 10 at epochs 8 and 11), and 12 training epochs. CenterNet was trained with a ResNet-101 backbone, a batch size of 16, an initial learning rate of 0.01 (reduced at epochs 30 and 45), and 50 training epochs. The DETR-based models (e.g., DETR, DAB-DETR, Sparse R-CNN, and CO-DETR) used AdamW optimizers, with a batch size of 4 and an initial learning rate of 0.0001 for the transformer and 0.00001 for the backbone. These models were trained for 50 epochs, with learning rate reductions at epoch 40. ViTDet used a ViT-B backbone, a batch size of 8, an initial learning rate of 0.0001, and was trained for 36 epochs, with learning rate reductions at epochs 24 and 30. By using consistent preprocessing, training splits, and hyperparameters tailored to each model, we ensured a fair and comprehensive comparison across all methods.

4.3 Evaluation metrics

To quantitatively evaluate the effectiveness and advantages of the proposed sonar object detection model, we used the precision, recall, average precision (AP), false alarm rate (FAR), F1 score, and frames-per-second (FPS) metrics commonly used in natural scene image object detection tasks as the evaluation metrics. First, we defined TP, FP, TN, and FN as true positive, false positive, true negative, and false negative. Specifically, TP indicates the model correctly detects the sonar object, FP denotes a non-object is falsely detected as the object region, TN indicates the model correctly predicts the non-object category, and FN denotes the object region is mistakenly predicted as a non-object. The calculation of different evaluation metrics is as follows.

1) The precision is defined as the proportion of the model's correct object detection to overall detection results.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{30}$$

2) The recall is defined as the proportion of model correct object detection to the truth annotation object.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{31}$$

3) The AP is defined as the area under the precision-recall (PR) curve used to evaluate the model performance.

$$\text{AP}_{\text{IoU}} = \int_0^1 P(R)\, d(R) \tag{32}$$

where IoU denotes the intersection-over-union threshold used to determine whether the detection result belongs to TP or FP. For the sonar object detection task, we set the IoU to 0.5. Additionally, the evaluation metrics AP^s, AP^m, and AP^l of the Microsoft COCO dataset (Lin et al., 2014) were used to further refine the evaluation and analyze model performance.

4) The FAR evaluates the prediction result credibility by calculating the proportion of FP in all the results.

$$\text{FAR} = \frac{\text{FP}}{\text{TP} + \text{FP}} \tag{33}$$

5) The F1 score is defined as the harmonic mean of precision and recall and can assess the comprehensive performance of the object detection model.

$$\text{F1\_score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{34}$$

6) The FPS represents the speed of the object detection model to process a single frame image per second.

$$\text{FPS} = 1/T_{\text{single}} \tag{35}$$

where T_{single} denotes the time taken to process a single forward-looking sonar image.

4.4 Comparison experiments and analysis

To demonstrate the advantages of the proposed forward-looking sonar object detector MLFANet, we compared it with 11 state-of-the-art object detection models on the UATD dataset. The compared methods can be classified into CNN-based methods and Transformer-based methods. Specifically, the CNN-based methods included Faster R-CNN (Girshick, 2015), RetinaNet (Lin et al., 2017b), Cascade R-CNN (Cai and Vasconcelos, 2019), CenterNet (Duan et al., 2019), Dynamic R-CNN (Zhang et al., 2020), DH R-CNN (Wang et al., 2022b), and Spare R-CNN (Sun et al., 2023); the Transformer-based methods included DETR (Carion et al., 2020), ViTDet (Li et al., 2022), DAB-DETR (Liu et al., 2022) and CO-DETR (Zong et al., 2023). To ensure experiment fairness, the compared methods were retrained on the UATD dataset and used the same training strategy and parameter settings as the proposed methods.

The comparative analysis included quantitative comparison, qualitative comparison, and model complexity analysis. The details are as follows:

4.4.1 Quantitative analysis

The quantitative comparison of different object detection methods was performed on the testing set of the UATD dataset. The performance quantitative analysis results of different methods are shown in Table 1. From the analysis results, compared with other object detection models, the proposed MLFANet obtained the optimal results on multiple evaluation metrics. Additionally, for metrics AP^l, AP^m, and AP^s, the proposed method reached 62.79%, 58.24%, and 45.36%, respectively, which further explains the comprehensive performance advantages of our MLFANet. Specifically, compared with the CNN-based optimal model CenterNet (Duan et al., 2019) and Transformer-based optimal model CO-DETR (Zong et al., 2023), the proposed method was 6.53% and 2.85% higher for the AP metric, respectively. For the CNN-based methods, such as Faster R-CNN (Girshick, 2015), RetinaNet (Lin et al., 2017b), and Cascade R-CNN (Cai and Vasconcelos, 2019), the AP values only reached 32.53%, 29.75%, and 34.97%, respectively, and were accompanied by higher FAR values. The reason for this phenomenon is that the seabed reverberation noise and clutter information contained in the sonar image seriously interfere with the feature extraction process of the CNN model, and the use of a simple convolution operation cannot fully extract the valuable feature information. Moreover, the weak and dark light characteristics of the sonar image object region diminish the positioning and recognition of the CNN-based methods, so they cannot achieve the ideal detection accuracy. Since the Transformer model has better global feature extraction and modeling effect, compared with the CNN-based method, the Transformer-based method has a slight advantage for the sonar image object detection task. For example, compared with Dynamic R-CNN (Girshick, 2015), ViTDet (Li et al., 2022) was 8.40% and 6.86% higher for the AP and F1 score, respectively. Furthermore, for the metrics AP^l and AP^m, the optimal Transformer-based model CO-DETR (Zong et al., 2023) reached 58.93% and 54.68%, indicating that the method can accurately detect large/medium size objects in sonar images. However, the imaging characteristics of sonar images cause redundant information interference in the global information correlation modeling process of the Transformer-based method, which makes it difficult to achieve satisfactory results for small object detection. For instance, the AP^s values of ViDet (Li et al., 2022), DAB-DETR (Liu et al., 2022), and CO-DETR (Zong et al., 2023) were only 41.32%, 39.76%, and 42.18%, and these methods have high false alarm rates. The reason for this problem is that the Transformer model only focuses on global feature information extraction, ignoring the important value of local feature information, resulting in false discrimination of small object region features as background information features. To verify the detection accuracy of different object detection models for different object categories in sonar images, we randomly selected 1,200 images from the UATD dataset as experimental data. As shown in Table 2, the mean AP (mAP) value of the proposed MLFANet was 81.86%, which is better than all the compared methods. The quantitative results further illustrate the superior detection performance of the proposed method compared to other object detection models. For the AP value of each sonar object category, we can conclude that for the tiny object categories, i.e. the ball, circle cage, and tire, the optimal CNN-based model CenterNet (Duan et al., 2019) only reached 61.28%, 39.78%, and 30.12%, and the optimal Transformer-based model CO-DETR (Zong et al., 2023) only reached 62.85%, 45.63%, and 35.92%. For the large-size object categories, i.e., the cube, plane, and metal bucket, the experimental results in Table 2 show that

TABLE 1 Performance comparison of different object detection methods on the testing set of the UATD dataset, where the score in bold is the highest score.

Model	Backbone	Precision	Recall	F1 score	AP	AP50	AP75	AP^l	AP^m	AP^s	FAR
Faster R-CNN	ResNet-50	0.8245	0.8547	0.8393	0.3253	0.8013	0.2179	0.4768	0.4312	0.3147	0.1755
RetinaNet	ResNet-50	0.7852	0.8165	0.8005	0.2975	0.7928	0.1852	0.4573	0.4127	0.3052	0.2148
Cascade R-CNN	ResNet-50	0.8564	0.8872	0.8715	0.3497	0.8417	0.2368	0.4892	0.4562	0.3387	0.1436
CenterNet	ResNet-101	0.8864	0.8953	0.8908	0.3958	0.8736	0.2873	0.5579	0.5124	0.3865	0.1136
Dynamic R-CNN	ResNet-50	0.8426	0.8692	0.8557	0.3375	0.8327	0.2295	0.4936	0.4457	0.3249	0.1574
DH R-CNN	ResNet-50	0.8647	0.8873	0.8758	0.3589	0.8562	0.2674	0.5183	0.4618	0.3621	0.1353
DETR	ResNet-50	0.8958	0.9267	0.9110	0.4122	0.8893	0.3275	0.5724	0.5218	0.4018	0.1042
Sparse R-CNN	ResNet-101	0.8782	0.8879	0.8830	0.3624	0.8624	0.2587	0.5276	0.4835	0.3512	0.1218
ViTDet	ViT-B	0.9128	0.9385	0.9255	0.4215	0.9032	0.3386	0.5597	0.5197	0.4132	0.0872
DAB-DETR	ResNet-50	0.9067	0.9249	0.9157	0.4037	0.8924	0.3194	0.5482	0.4973	0.3976	0.0933
CO-DETR	Swin-L	0.9273	0.9486	0.9378	0.4326	0.9162	0.3417	0.5893	0.5468	0.4218	0.0727
MLFANet (Ours)	ResNet-50	0.9438	0.9652	0.9543	0.4583	0.9548	0.3578	0.6142	0.5679	0.4427	0.0562
	ResNet-101	**0.9521**	**0.9716**	**0.9617**	**0.4611**	**0.9602**	**0.3792**	**0.6279**	**0.5824**	**0.4536**	**0.0479**

TABLE 2 Comparison of category detection accuracy of different object detection methods, where the score in bold is the highest score.

Model	Backbone	Cube	Ball	Cylinder	HB	Plane	CC	SC	MB	Tire	ROV	mAP
Faster R-CNN	ResNet-50	0.8126	0.5247	0.7582	0.6978	0.8668	0.3576	0.6632	0.8345	0.2864	0.7238	0.6516
RetinaNet	ResNet-50	0.7834	0.4873	0.7763	0.6504	0.8174	0.2865	0.5983	0.7853	0.2981	0.7124	0.6195
Cascade R-CNN	ResNet-50	0.8345	0.5562	0.7956	0.7126	0.8972	0.4128	0.6895	0.8672	0.2573	0.7559	0.6779
CenterNet	ResNet-101	0.8872	0.6128	0.8325	0.7382	0.9203	0.3978	0.7326	0.8763	0.3012	0.8057	0.7105
Dynamic R-CNN	ResNet-50	0.8257	0.5369	0.8154	0.6893	0.8438	0.3736	0.6782	0.8325	0.2297	0.7354	0.6561
DH R-CNN	ResNet-50	0.8536	0.5738	0.7862	0.7025	0.9056	0.4265	0.7024	0.8614	0.2895	0.7564	0.6858
DETR	ResNet-50	0.8842	0.6297	0.8537	0.7458	0.9159	0.4758	0.7253	0.8713	0.3158	0.8126	0.7230
Sparse R-CNN	ResNet-101	0.8264	0.5261	0.7436	0.6951	0.8397	0.3695	0.6915	0.8427	0.2697	0.7327	0.6537
ViTDet	ViT-B	0.8976	0.6385	0.8423	0.7626	0.9234	0.4425	0.7456	0.9057	0.3387	0.8051	0.7302
DAB-DETR	ResNet-50	0.8653	0.5642	0.8535	0.7715	0.9386	0.3871	0.7167	0.8976	0.3254	0.8385	0.7158
CO-DETR	Swin-L	0.8946	0.6285	0.8761	0.7869	0.9527	0.4563	0.7315	0.9143	0.3592	0.8274	0.7428
MLFANet (Ours)	ResNet-50	0.9473	0.7942	0.9182	0.8213	0.9738	0.5138	0.7826	0.9485	0.4895	0.8573	0.8046
	ResNet-101	**0.9584**	**0.8157**	**0.9203**	**0.8306**	**0.9814**	**0.5264**	**0.8014**	**0.9526**	**0.5123**	**0.8615**	**0.8161**

HB, CC, SC, and MB denote the human body, circle cage, square cage, and metal bucket.

these compared object detection models still cannot achieve satisfactory detection accuracy. In contrast, the proposed MLFANet obtained AP values of 95.84%, 98.14%, and 95.26% for the large-size object categories, respectively. Additionally, for the other object categories such as cylinder, human body, square cage, and ROV, the proposed method achieved AP values of 92.03%, 83.06%, 80.14%, and 86.15%, which are the optimal results for all compared methods. The quantitative analysis results in Tables 1 and 2 show that the proposed method has significant advantages in solving sonar image object detection tasks. The reason is that MLFANet fully considers the interference of seabed reverberation noise, shadow region, and clutter information in the sonar images, and proposes corresponding solutions, so it can obtain better object detection accuracy. To further intuitively compare the performance of different object detection models, we drew the PR curve of different object detection models for comparison. The PR curve in Figure 10 demonstrates the performance of MLFANet compared to baseline models across various classification thresholds. The PR curve of MLFANet exhibits a higher AUC, indicating its ability to achieve both high precision and high recall. This is particularly

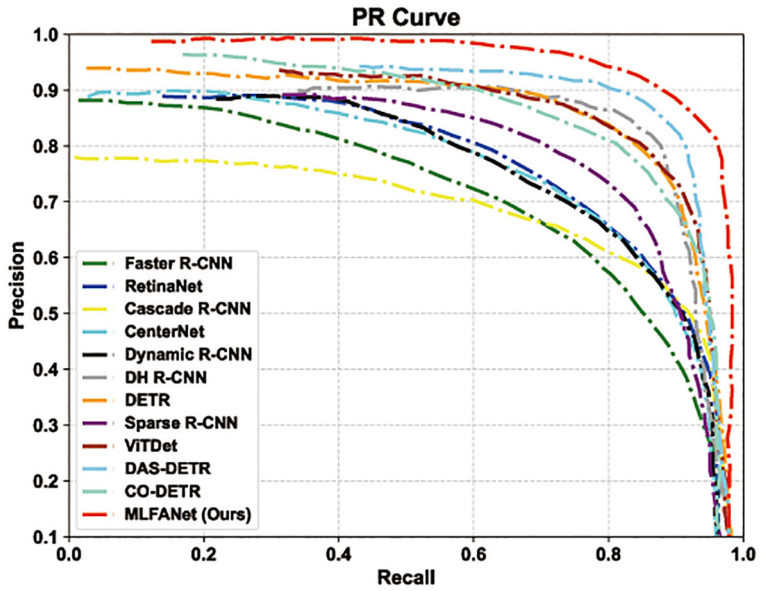

FIGURE 10
Comparison of PR curves for different object detection models.

important for FLS images, where the presence of noise, shadows, and reverberation can lead to false positives or missed detections. Compared to baseline models, MLFANet maintains a more gradual decline in precision as recall increases, reflecting its robustness to challenging underwater conditions. This is attributed to the integration of the LFAM, DFEM, and MFRM, which together enhance feature representation and reduce noise interference. Additionally, the CIoU-DFL loss function contributes to this improved performance by addressing class imbalance and refining object localization and classification. The precision value of MLFANet was the highest among all models, further supporting the superior performance of the proposed framework. This analysis highlights the effectiveness of MLFANet in achieving a favorable precision-recall trade-off, making it well-suited for underwater object detection.

4.4.2 Qualitative analysis

To further demonstrate the effectiveness of the proposed MLFANet, we visualized the prediction results of sonar images under different scene conditions contained in the UATD dataset. As shown in Figures 11–13, these scenes include seabed reverberation noise interference, shadow region interference, and object scale variation. It can be seen from the prediction results that the proposed method can accurately locate and recognize the different categories of sonar objects in the test images with high confidence scores. In contrast, the compared methods suffer from location deviation, high false alarm rate, and recognition failures. Additionally, as shown in Table 3, we present the confidence scores of different object detection models for the object categories in the test images. Following this, we present a detailed analysis of the different object detection model prediction results under three underwater scene conditions and the advantages of the constructed sonar object detector. The qualitative comparison results effectively illustrate the advantages of the proposed method for sonar object detection.

4.4.2.1 Superiority in scenes with seabed reverberation noise interference

The irregularity of underwater terrain seriously affects the propagation and reflection of sound waves on the seabed, so a forward-looking sonar image is disturbed by seabed reverberation noise. As shown in Figure 11, under the interference of seabed reverberation noise, it is difficult for the compared object detection models to obtain satisfactory detection results. For example, for CNN-based object detection models, Faster R-CNN (Girshick, 2015) and RetinaNet (Lin et al., 2017b) could not correctly detect all object categories in sonar images, resulting in false detection and missing detection. The reason is that the non-linear characteristics of seabed reverberation noise interfere with the detection and recognition process of CNN-based methods. For the Transformer-based object detection models, ViTDet (Li et al., 2022) and CO-DETR (Zong et al., 2023) obtained relatively better detection results. However, the results in Figure 11 show that these methods still struggle to accurately detect small-size object categories. In contrast, MLFANet effectively suppress the seabed reverberation noise interference on the feature extraction process,

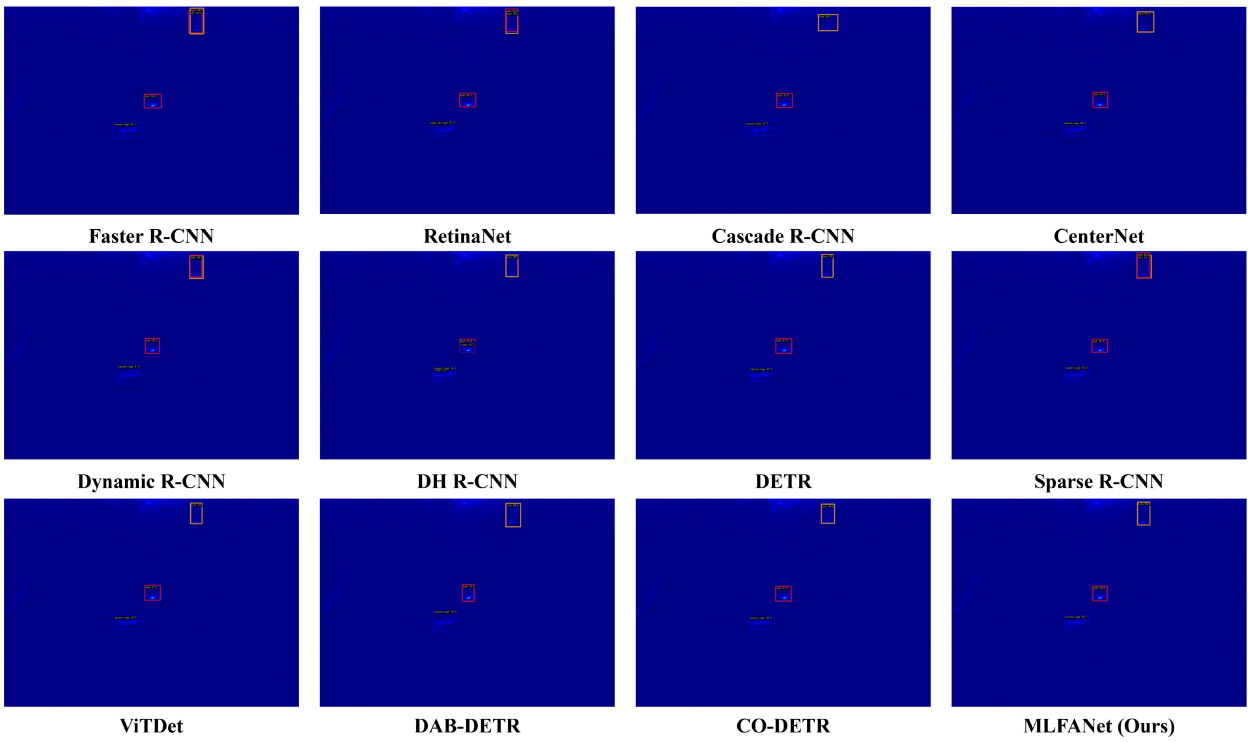

FIGURE 11
Visualization detection results of different object detection models in seabed reverberation noise interference scenes.

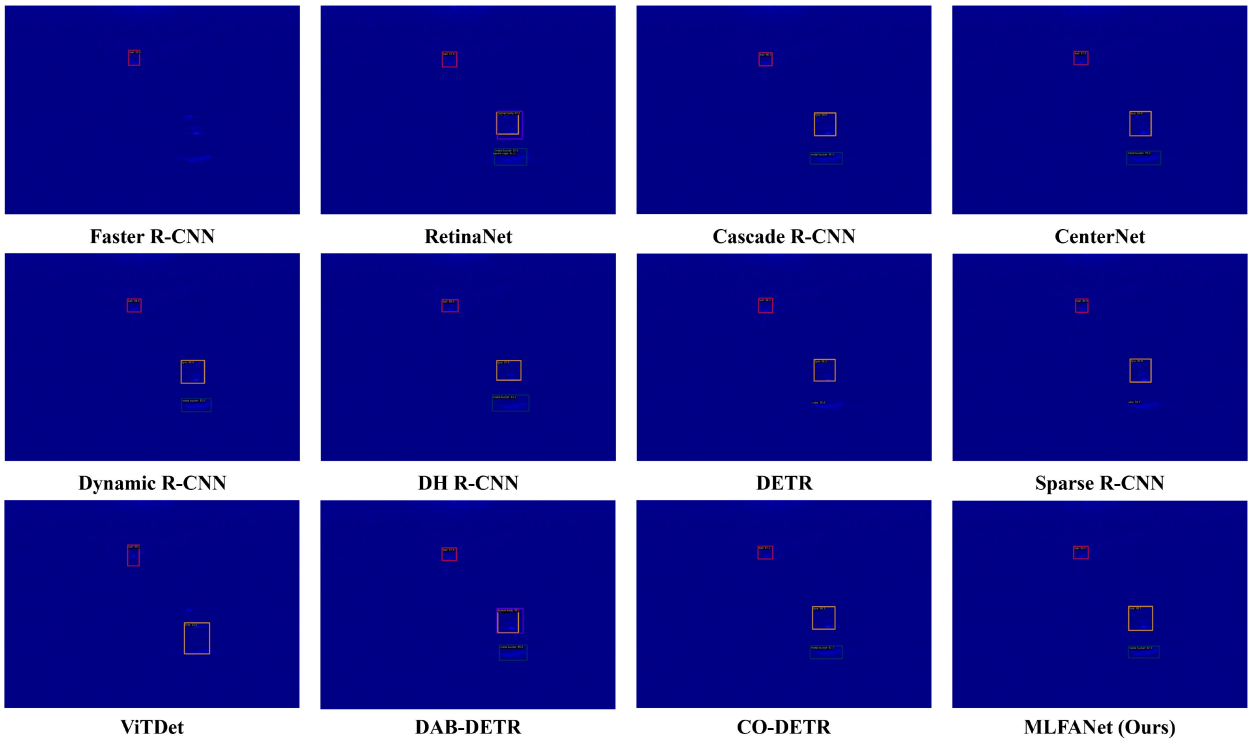

FIGURE 12
Visualization detection results of different object detection models on shadow region interference scenes.

successfully detects different object categories and obtains the higher confidence score. Moreover, in environments with strong seabed reverberation noise, MLFANet occasionally misclassifies noise patterns as objects due to their similar intensity and texture. Future work could focus on integrating advanced noise suppression or training with adversarial noise augmentation to mitigate this issue.

4.4.2.2 Superiority in scenes with shadow region interference

Since the underwater object has the characteristics of absorption, reflection, and scattering of sonar signal, it is difficult for the acoustic wave to directly penetrate the object entity, so the shadow interference region is formed in the reverse of the object region. The existence of the shadow region causes object occlusion, so it is difficult for the object detection model to accurately extract the edge, contour, and detail feature information. As shown in Figure 12, in the shadow interference scene, the compared sonar object detection models struggled to accurately locate and identify the object category and obtained a lower confidence score. Among the competitors, for CNN-based methods, CenterNet (Duan et al., 2019) obtained relatively better detection results. The reason is that the model uses a center point detection strategy to locate the object region, which can effectively alleviate the shadow region interference on the object feature extraction process. For the Transformer-based methods, CO-DETR obtained the optimal detection results. The reason is that it suppresses the representation of redundant feature information in the shadow region through global context modeling, and uses the position encoder mechanism to improve the object positioning accuracy. The proposed method obtains the optimal detection effect, which suppresses and filters the shadow feature interference by focusing on the discriminative feature information of the object region to improve the location and recognition accuracy. In addition, the objects located in regions with strong shadow interference are sometimes missed due to low contrast and insufficient discriminative features. Introducing adaptive contrast enhancement or attention mechanisms could help improve detection in such regions.

4.4.2.3 Superiority in scenes with object multi-scale transformation

Due to the influence of different object entities, object distance transformation, sonar beam angle, and object motion state, there are complex object scale transformation phenomena in the forward-looking sonar image. The variable object scale puts forward higher requirements for the multi-scale feature extraction capability of the object detection model. However, the existing object detection methods can only solve the multi-scale feature extraction problem of natural scene images, while multi-scale feature extraction for sonar images still cannot achieve satisfactory performance. As shown in Figure 13, for sonar images with different scale objects, the compared methods had false alarms and missing detection problems. Among competitors, Cascade R-CNN (Cai and Vasconcelos, 2019) and Dynamic R-CNN (Zhang et al., 2020), which use multi-scale feature extraction strategies, achieved relatively better results. The reason is that these methods

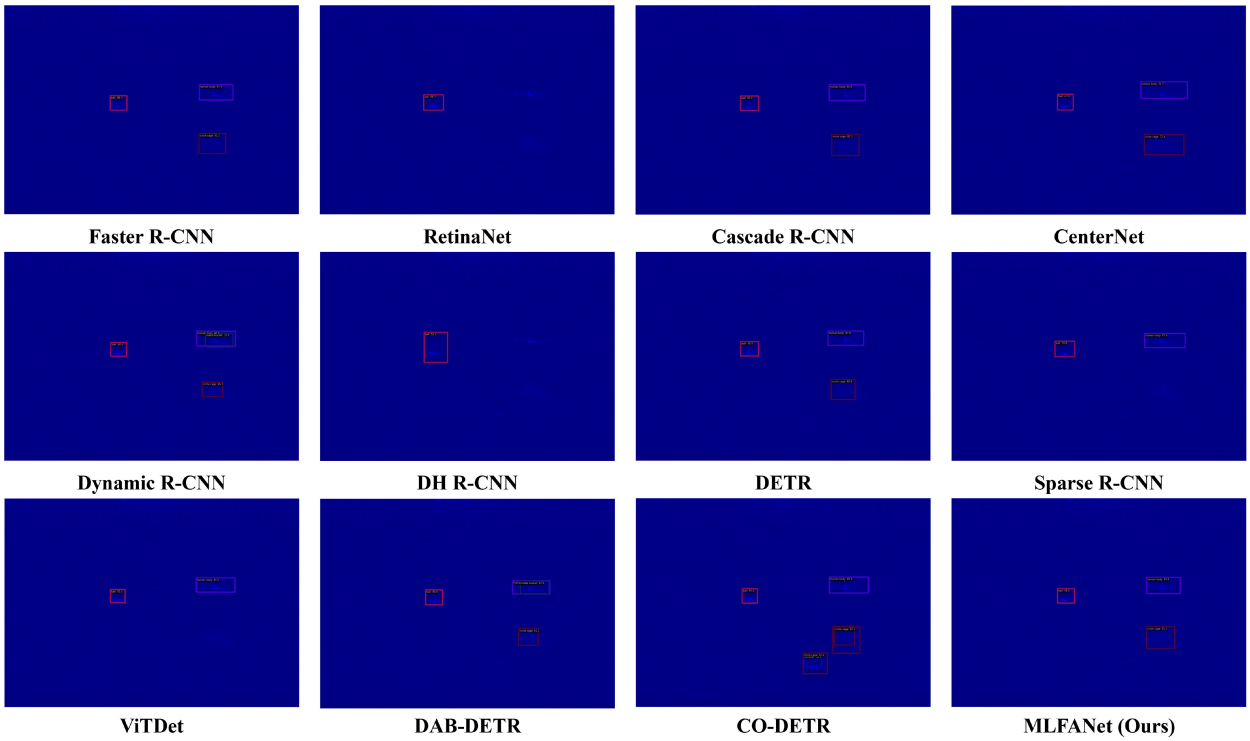

FIGURE 13
Visualization detection results of different object detection models on object multi-scale transformation scenes.

TABLE 3 Comparison of the confidence scores of different object detection methods.

Models	Reverberation noise scenes			Shadow interference scenes			Object scale transformation		
	NA	NO	Confidence	NA	NO	Confidence	NA	NO	Confidence
Faster R-CNN	3	4	80.2%, 99.3%, 64.9%, 49.0%	3	2	68.1%, 34.6%	3	2	95.2%, 45.2%
RetinaNet	3	5	42.1%, 57.3%, 99.3%, 53.6%, 96.2%	3	5	93.8%, 67.0%, 58.3%, 30.6%, 81.2%	3	1	98.7%
Cascade R-CNN	3	3	61.7%, 87.5%, 68.6%	3	3	86.2%, 94.0%, 49.4%	3	3	99.0%, 94.8%, 81.1%
CenterNet	3	3	81.0%, 83.2%, 51.0%	3	3	97.9%, 94.8%, 75.3%	3	3	47.0%, 31.7%, 72.4%
Dynamic R-CNN	3	4	97.6%, 98.2%, 34.3%, 28.6%	3	1	69.0%	3	3	98.3%, 93.9%, 91.2%
DH R-CNN	3	5	36.0%, 61.7%, 44.3%, 93.8%, 86.8%	3	3	98.6%, 97.3%, 61.2%	3	1	52.2%
DETR	3	3	87.4%, 97.0%, 95.5%	3	3	98.2%, 98.3%, 38.4%	3	3	98.0%, 87.6%, 88.8%
Sparse R-CNN	3	4	91.6%, 98.9%, 48.6%, 82.0%	3	3	89.9%, 97.8%, 50.7%	3	2	70.6%, 45.6%
ViDet	3	3	84.9%, 97.9%, 93.8%	3	3	98.3%, 98.8%, 92.0%	3	4	98.9%, 32.9%, 46.6%, 96.7%
DAB-DETR	3	3	93.9%, 93.9%, 88.2%	3	4	97.8%, 45.7%, 78.3%, 85.0%	3	4	98.0%, 43.4%, 54.2%, 61.2%
CO-DETR	3	3	96.6%, 97.4%, 94.8%	3	3	97.2%, 98.3%, 92.2%	3	6	99.8%, 99.3%, 40.2%, 83.9%, 31.2%
MLFANet	3	3	99.7%, 99.8%, 99.6%	3	3	99.0%, 99.5%, 92.3%	3	3	99.0%, 94.8%, 81.1%

HB, CC, SC, and MB denote the human body, circle cage, square cage, and metal bucket.

construct a multi-scale feature extraction structure, which can alleviate the influence of object scale transformation. In contrast, we can observe from Figure 12 that the Transformer-based object detection models were less effective for object scale variable scenarios. Taking the DAB-DETR (Liu et al., 2022) detector as an example, it only focuses on the efficient modeling of global information and ignores the extraction of scale-invariant features, which leads to missing detection and false alarm problems. The proposed MLFANet can effectively detect the different scale object categories in sonar images and obtain a higher confidence score. The reason is that the multi-scale feature refinement module can accurately locate sonar objects with different scales and obtain the robust invariant feature information in sonar image. Moreover, MLFANet struggles with extreme scale variations, leading to missed detections of very small objects or fragmented detections of very large targets. Developing more robust multi-scale feature fusion techniques or scale-invariant detection mechanisms could address this limitation.

4.4.3 Performance in small sample scenarios

To evaluate the potential of MLFANet for small sample learning, we conducted experiments by reducing the training dataset size to simulate limited data conditions. Specifically, 50%, 25%, and 10% of the original training data were used, while the test set remained unchanged. The performance of MLFANet and four representative baseline models (Faster R-CNN, RetinaNet, DETR, and CO-DETR) under these conditions is summarized in Table 4. The results in Table 4 demonstrate that MLFANet consistently outperforms the baseline models across all training data fractions. Notably, in extremely small sample conditions (10% training data), MLFANet achieved an AP of 30.85%, significantly surpassing Faster R-CNN (18.57%), RetinaNet (16.42%), DETR (24.17%), and CO-DETR (26.58%). This highlights the robustness and effectiveness of MLFANet in low-data conditions.

4.4.4 Computational complexity analysis

Since the sonar image object detection task has high requirements for algorithm real-time performance, we compared and analyzed the computational complexity of different object detection models, and the specific results are shown in Figure 14. Table 5 presents the number of parameters, FLOPs (Floating Point Operations), and FPS for each model on a workstation equipped with an NVIDIA RTX 4090 GPU. It can be seen from the comparison results that the CNN-based methods have advantages in computational complexity and real-time performance compared with the Transformer-based object detection models. To take the Transformer-based method ViTDet (Li et al., 2022) as an example, its calculation parameter reached 5,792 MB, and the inference speed

TABLE 4 Model performance verification under small sample conditions.

Data fraction	Model	AP (%)	AP50 (%)	AP75 (%)	AP^l (%)	AP^m (%)	AP^s (%)	FAR (%)
100%	Faster R-CNN	32.53	80.13	21.79	47.68	43.12	31.47	17.55
	RetinaNet	29.75	79.52	18.52	45.73	41.27	30.52	21.48
	DETR	41.22	88.92	32.75	57.24	52.18	40.18	10.42
	CO-DETR	43.26	91.62	34.17	58.92	54.68	42.18	7.27
	MLFANet (Ours)	46.11	96.02	37.92	62.79	58.24	45.36	4.79
50%	Faster R-CNN	29.20	76.80	18.50	43.12	39.84	28.74	19.80
	RetinaNet	26.85	74.30	15.67	41.45	37.20	26.32	23.67
	DETR	36.80	85.20	28.90	50.72	46.10	36.40	12.80
	CO-DETR	39.40	88.10	30.70	53.34	49.36	39.20	9.72
	MLFANet (Ours)	42.50	93.80	34.80	57.30	53.60	42.10	6.30
25%	Faster R-CNN	24.72	63.18	14.52	37.11	32.84	23.45	23.42
	RetinaNet	22.17	64.82	12.33	33.84	30.11	21.52	26.64
	DETR	31.25	76.41	23.74	43.55	39.62	30.18	15.63
	CO-DETR	33.84	80.12	26.24	46.85	42.62	33.51	12.92
	MLFANet (Ours)	37.58	86.74	30.66	49.25	46.13	36.35	8.87
10%	Faster R-CNN	18.57	55.42	9.83	25.17	22.52	15.68	27.92
	RetinaNet	16.42	51.27	8.28	21.84	19.43	14.36	30.65
	DETR	24.17	61.74	17.98	33.65	30.24	24.06	18.73
	CO-DETR	26.58	65.97	20.42	36.84	33.41	26.17	15.83
	MLFANet (Ours)	30.85	74.26	24.74	41.58	38.92	31.74	10.48

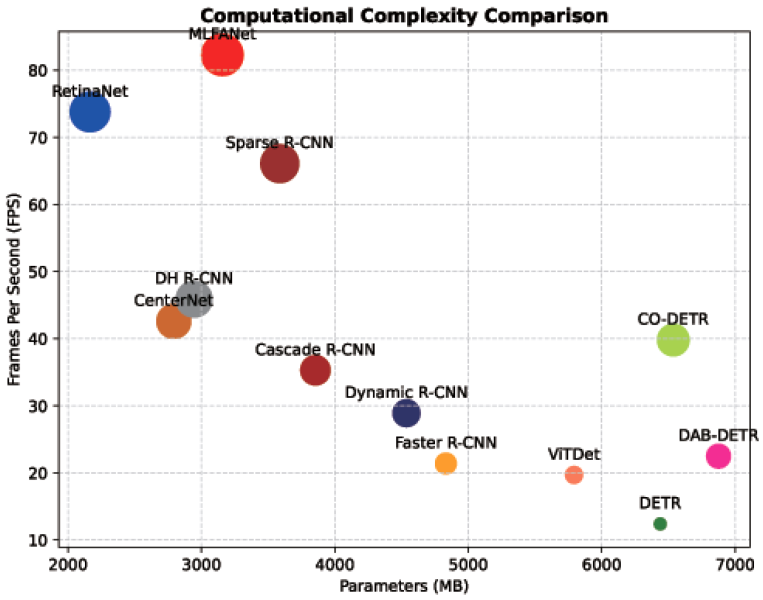

FIGURE 14
Comparison of our method with other object detection models in calculation parameters and inference speed.

was only 19.7 FPS. The reason is that the self-attention mechanism used in the Transformer model requires the calculation of the correlation of each pixel spatial position information, which increases the model inference time and calculation parameters. For the CNN-based methods, to take the Cascade R-CNN (Cai and Vasconcelos, 2019) as an example, the number of calculation parameters was 3,854 MB, and the inference speed reached 35.3 FPS. Although this method outperforms several Transformer-based object detection models, it still fails to address the real-time requirements of the sonar object detection task. In contrast, the computational parameter of the proposed MLFANet reached 3,157 MB, and the inference speed was 82.3 FPS, which was significantly better than the other object detection models. The proposed method can achieve an advantage because it constructs the corresponding feature extraction and fusion module for the sonar image, and effectively alleviates the influence of redundant feature and noise information on the inference process of the object detection model. To further validate the feasibility of MLFANet for deployment on embedded devices, experiments were conducted on an NVIDIA Jetson Xavier NX. The model was optimized using quantization techniques to reduce memory consumption and computational overhead. After optimization, MLFANet achieved an inference speed of 27.4 FPS with a memory footprint of 2.60 GB on the Jetson Xavier NX. These results demonstrate that MLFANet meets the real-time requirements of embedded systems, making it practical for AUV applications such as obstacle avoidance and object tracking.

TABLE 5 Comparison of the computational complexity of different object detection models.

Model	Parameter (MB)	FLOPs (G)	FPS (512×512 image size)
Faster R-CNN	4,138	207.1	32.6
RetinaNet	3,815	198.7	35.9
Cascade R-CNN	3,854	223.2	35.3
CenterNet	3,384	189.2	41.2
Dynamic R-CNN	4,052	204.7	30.1
DH R-CNN	4,327	212.3	29.6
DETR	5,748	350.2	19.4
Sparse R-CNN	4,877	298.1	22.3
ViTDet	5,792	420.5	17.8
DAB-DETR	6,015	368.4	18.5
CO-DETR	5,674	324.5	19.7
MLFANet (Ours)	3,157	183.4	82.3

4.5 Ablation study and analysis

To demonstrate the effectiveness of the important components LFAM, DFEM, and MFRM in the constructed MLFANet, we performed an ablation study on the UATD testing set, and the specific quantitative analysis results are shown in Table 6. In the experiment, we used the YOLOX detector (Ge et al., 2021) as the baseline model and verified the detector performance improvement by adding different components. Additionally, since the different constructed components are mainly for feature extraction and fusion of sonar images, we present the feature map visualization results of the different component modules in Figure 15. The specific analysis of the ablation study is as follows.

4.5.1 Effect of the LFAM

The constructed LFAM aims to fully exploit the low-level feature information such as texture, edge, and contour in the sonar image to improve the discriminating ability of the model for object region and background information. As shown in Table 6, when the LFAM was embedded into the baseline model, it achieved 68.74% (18.5% ↑) mAP on the testing set. Additionally, each object category experienced a corresponding increase in AP value, for example, the ball category had an increase of 25.58%, and the circle cage category had an increase of 22.13%. The feature visualization results corresponding to Figure 15 further show that the LFAM can make the model focus on feature extraction in the sonar object region and significantly enhance the model's feature representation ability for low-level feature information.

4.5.2 Effect of the DFEM

To filter the redundant feature information interference in the feature extraction process, the DFEM was constructed, which obtains the discriminative attributes of the object region by enhancing the local feature information representation in deep convolution. As shown in Table 6, when the DFEM was introduced into the baseline model, its mAP on the testing set reached 70.60%. Moreover, the DFEM enhanced the small object region feature representation, so that the AP values of the ball, circle cage and tire small object categories increased by 30.74%, 25.59%, and 9.60% respectively, and the AP values of the cube, plane and metal bucket large-size object categories increased by 15.69%, 11.47%, and 12.87% respectively. Combined with the LFAM and DFEM, the baseline model achieved significant performance improvement. Compared with the initial results, the

TABLE 6 Quantitative evaluation of the ablation study with different components, where the score in bold is the highest score.

Methods	Cube	Ball	Cylinder	HB	Plane	CC	SC	MB	Tire	ROV	mAP
Baseline	0.7153	0.3274	0.5865	0.4317	0.7528	0.2154	0.5171	0.7349	0.2054	0.5366	0.5024
+LFAM	0.8543	0.5832	0.7941	0.7352	0.8759	0.4367	0.7185	0.8437	0.2681	0.7639	0.6874
+DFEM	0.8722	0.6348	0.8364	0.7281	0.8675	0.4713	0.6985	0.8636	0.3014	0.7863	0.7061
+MFRM	0.8657	0.5962	0.8046	0.7524	0.8854	0.3762	0.6735	0.8893	0.3267	0.8127	0.6983
+LFAM+DFEM	0.8875	0.6584	0.8536	0.7782	0.9107	0.4685	0.7639	0.9164	0.4172	0.8311	0.7485
+LFAM+DFEM+MFRM	**0.9318**	**0.7653**	**0.8735**	**0.8094**	**0.9512**	**0.4956**	**0.7855**	**0.9381**	**0.4597**	**0.8535**	**0.7864**

NA denotes the number of actual objects and NO denotes the number of objects detected.

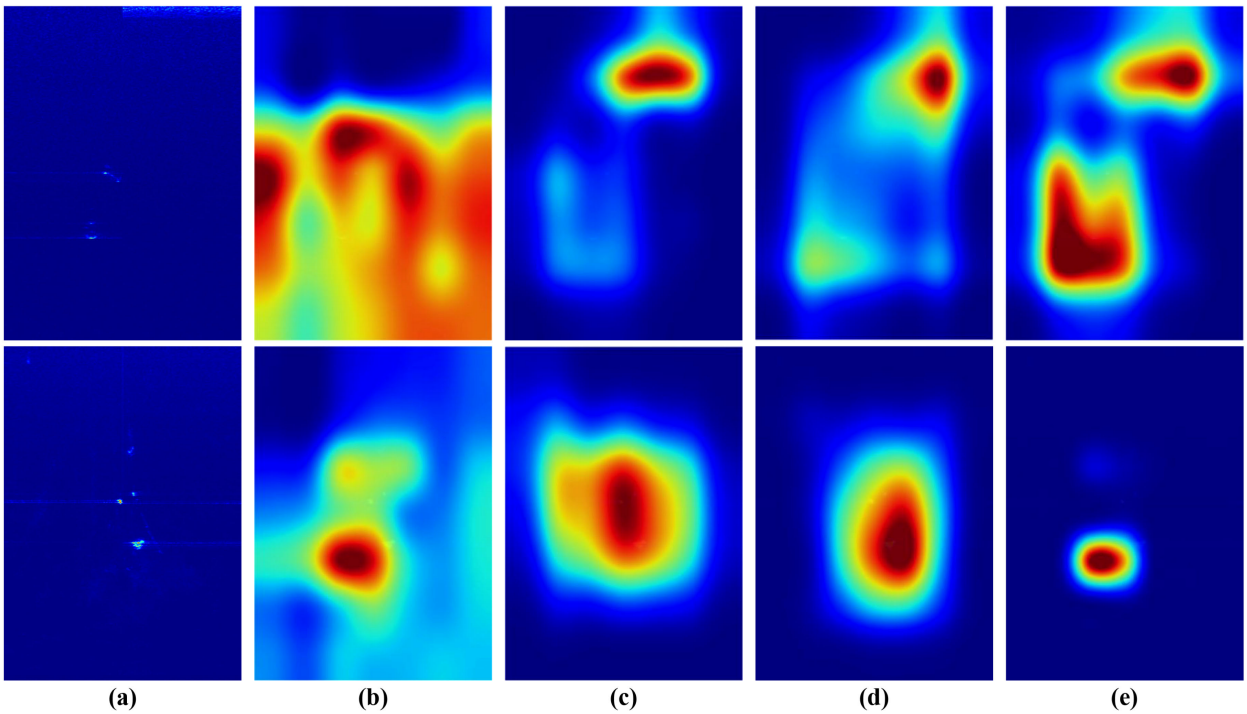

FIGURE 15
Visualizing the feature extraction effects of the LFAM, DFEM, and MFRM. **(a)** original forward-looking sonar images. **(b)** Feature extraction results from baseline. **(c)** Feature extraction results with the LFAM. **(d)** Feature extraction results with the DFEM. **(e)** Feature extraction results with the MFRM.

mAP increased by 24.61%, and the AP values for the cylinder, human body, square cage, and ROV increased by 26.71%, 34.65%, 24.64%, and 29.45%, respectively. The feature visualization results in Figure 15 show that DFEM can effectively filter the redundant feature information interference to improve the sonar object detection accuracy in clutter and shadow information interference scene.

4.5.3 Effect of the MFRM

To solve the problem of multi-scale feature extraction in seabed reverberation noise and shadow region interference scene, the MFRM was constructed, which obtains the scale-invariant features of sonar images by region location branch and feature refinement branch. Different from placing the LFAM and the DFEM in the feature extraction stage, we embedded the MFRM into the neck structure of the detector. As shown in Table 6, when placing the MFRM in the baseline model, it increased the mAP by 19.59%. Additionally, the model obtained a significant boost in AP values for object categories with different scales, for example, it increased by 26.88%, 16.08%, and 12.13% for the ball, circle cage, and tire, respectively. From the results shown in Figure 15, it can be observed that the use of the MFRM effectively improved the model's receptive field deformation ability, so that it could obtain the discriminative feature information of object regions with different scales. Notably, when combining the LFAM, DFEM, and MFRM, the baseline model performance was optimized, and the mAP value on the UATD testing set reached 78.64%, which further demonstrates the effectiveness of the different components in improving the detector performance.

5 Conclusion

To solve the problem of forward-looking sonar image object detection in complex underwater acoustic environment, in this article, we propose a novel multi-level feature aggregation network (MLFANet) to achieve an underwater sonar image object detection task. The proposed MLFANet contains three innovative modules, the LFAM, DFEM, and MFRM. Specifically, the LFAM is used to enhance the low-level feature information representation of sonar images to alleviate the influence of seabed reverberation noise on the feature extraction process. The DFEM enhances the saliency of object region features in deep convolution by constructing the correlation of local-global features to filter shadow and clutter information interference. The MFRM uses the region location and feature refinement branches to extract robust invariant feature information of different scale objects to solve the problem of underwater object multi-scale variation. To demonstrate the effectiveness and advantages of the proposed method, we conducted a series of experiments on a real-scene sonar image dataset, and MLFANet achieved better performance than the existing state-of-the-art methods. The ablation studies further validate the effectiveness and feasibility of the proposed different innovation modules. Although the proposed method can obtain better detection performance, it requires more training samples. Therefore, in future work, we intend to explore the forward-looking sonar image object detection method in small sample conditions.

Data availability statement

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

Author contributions

ZW: Writing – original draft, Writing – review & editing. JG: Writing – review & editing. SZ: Writing – review & editing. YZ: Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work is supported by the National Natural Science Foundation of China (61671465), the National Natural Science Foundation of China (61624931), the Neural Science Foundation of Shaanxi Province (2021JM-537), the Youth Talent Support Program of Shaanxi Science and Technology Association (23JK0701), the Xi'an Science and Technology Planning Projects (20240103), and the China Postdoctoral Science Foundation under Grant (2024M754225).

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.

Publisher's note

References

Abu, A., and Diamant, R. (2019). A statistically-based method for the detection of underwater objects in sonar imagery. *IEEE Sensors J.* 19, 6858–6871. doi: 10.1109/JSEN.7361

Cai, Z., and Vasconcelos, N. (2019). Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 1483–1498. doi: 10.1109/TPAMI.2019.2956516

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Glasgow, UK. 213–229.

Chandra, M. A., and Bedi, S. (2021). Survey on svm and their application in image classification. *Int. J. Inf. Technol.* 13, 1–11. doi: 10.1007/s41870-017-0080-1

Chen, L., Liu, C., Chang, F., Li, S., and Nie, Z. (2021). Adaptive multi-level feature fusion and attentionbased network for arbitrary-oriented object detection in remote sensing imagery. *Neurocomputing* 451, 67–80. doi: 10.1016/j.neucom.2021.04.011

Cheng, G., Si, Y., Hong, H., Yao, X., and Guo, L. (2020). Cross-scale feature fusion for object detection in optical remote sensing images. *IEEE Geosci. Remote Sens. Lett.* 18, 431–435. doi: 10.1109/LGRS.8859

Collins, M., Schapire, R. E., and Singer, Y. (2002). Logistic regression, adaboost and bregman distances. *Mach. Learn.* 48, 253–285. doi: 10.1023/A:1013912006537

Dong, X., Qin, Y., Gao, Y., Fu, R., Liu, S., and Ye, Y. (2022). Attention-based multi-level feature fusion for object detection in remote sensing images. *Remote Sens.* 14, 3735. doi: 10.3390/rs14153735

Du, B., Huang, Y., Chen, J., and Huang, D. (2023). "Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images." in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, Vancouver, Canada. 13435–13444. doi: 10.1109/CVPR52729.2023.01291

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., and Tian, Q. (2019). "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, Seoul, South Korea. 6569–6578.

Elharrouss, O., Akbari, Y., Almaadeed, N., and Al-Maadeed, S. (2022). Backbones-review: Feature extraction networks for deep learning and deep reinforcement learning approaches. *arXiv preprint arXiv:2206.08016*. doi: 10.48550/arXiv.2206.08016

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*. doi: 10.48550/arXiv.2107.08430

Girshick, R. (2015). "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision* (ICCV), Santiago, Chile. 1440–1448.

Gong, H., Mu, T., Li, Q., Dai, H., Li, C., He, Z., et al. (2022). Swin-transformer-enabled yolov5 with attention mechanism for small object detection on satellite images. *Remote Sens.* 14, 2861. doi: 10.3390/rs14122861

Grzadziel, A. (2020). Results from developments in the use of a scanning sonar to support diving operations from a rescue ship. *Remote Sens.* 12, 693. doi: 10.3390/rs12040693

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., et al. (2018). Recent advances in convolutional neural networks. *Pattern recognition* 77, 354–377. doi: 10.1016/j.patcog.2017.10.013

Guo, C., Fan, B., Zhang, Q., Xiang, S., and Pan, C. (2020). "Augfpn: Improving multi-scale feature learning for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Seattle, WA, USA. 12595–12604.

Hansen, R. E., Callow, H. J., Sabo, T. O., and Synnes, S. A. V. (2011). Challenges in seafloor imaging and mapping with synthetic aperture sonar. *IEEE Trans. Geosci. Remote Sens.* 49, 3677–3687. doi: 10.1109/TGRS.2011.2155071

Jiang, L., Yuan, B., Du, J., Chen, B., Xie, H., Tian, J., et al. (2024). Mffsodnet: Multi-scale feature fusion small object detection network for uav aerial images. *IEEE Trans. Instrumentation Measurement* 73, 1–14. doi: 10.1109/TIM.2024.3381272

Karimanzira, D., Renkewitz, H., Shea, D., and Albiez, J. (2020). Object detection in sonar images. *Electronics* 9, 1180. doi: 10.3390/electronics9071180

Kim, B., and Yu, S.-C. (2017). "Imaging sonar based real-time underwater object detection utilizing adaboost method," in *2017 IEEE Underwater Technology (UT) (IEEE)*, Busan, South Korea. 1–5.

Kim, S.-W., Kook, H.-K., Sun, J.-Y., Kang, M.-C., and Ko, S.-J. (2018a). "Parallel feature pyramid network for object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany. 234–250.

Kim, Y., Kang, B.-N., and Kim, D. (2018b). "San: Learning relationship between convolutional features for multi-scale object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany. 316–331.

Kong, W., Hong, J., Jia, M., Yao, J., Cong, W., Hu, H., et al. (2019). Yolov3-dpfin: A dual-path feature fusion neural network for robust real-time sonar target detection. *IEEE Sensors J.* 20, 3745–3756. doi: 10.1109/JSEN.7361

Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans. Neural Networks Learn. Syst.* 33, 6999–7019. doi: 10.1109/TNNLS.2021.3084827

Li, Y., Mao, H., Girshick, R., and He, K. (2022). "Exploring plain vision transformer backbones for object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Tel Aviv, Israel. 280–296.

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., et al. (2020). Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* 33, 21002–21012. doi: 10.5555/3495724.3497487

Li, Z., Xie, Z., Duan, P., Kang, X., and Li, S. (2024). Dual spatial attention network for underwater object detection with sonar imagery. *IEEE Sensors J.* 24, 6998–7008. doi: 10.1109/JSEN.2023.3336899

Liang, X., Zhang, J., Zhuo, L., Li, Y., and Tian, Q. (2019). Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis. *IEEE Trans. Circuits Syst. Video Technol.* 30, 1758–1770. doi: 10.1109/TCSVT.76

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017a). "Feature pyramid networks for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Honolulu, HI, USA, 2117–2125.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). "Focal loss for dense object detection," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, Venice, Italy, 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, 740–755.

Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., et al. (2022). Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*. doi: 10.48550/arXiv.2201.12329

Liu, X., Zhou, F., Zhou, H., Tian, X., Jiang, R., and Chen, Y. (2015). A low-complexity real-time 3-d sonar imaging system with a cross array. *IEEE J. Oceanic Eng.* 41, 262–273. doi: 10.1109/JOE.2015.2439851

Lu, X., Ji, J., Xing, Z., and Miao, Q. (2021). Attention and feature fusion ssd for remote sensing object detection. *IEEE Trans. Instrumentation Measurement* 70, 1–9. doi: 10.1109/TIM.2021.3118092

Ma, W., Wu, Y., Cen, F., and Wang, G. (2020). Mdfn: Multi-scale deep feature learning network for object detection. *Pattern Recognition* 100, 107149. doi: 10.1016/j.patcog.2019.107149

Miao, S., Du, S., Feng, R., Zhang, Y., Li, H., Liu, T., et al. (2022). Balanced single-shot object detection using cross-context attention-guided network. *Pattern recognition* 122, 108258. doi: 10.1016/j.patcog.2021.108258

Mustafa, H. T., Yang, J., and Zareapoor, M. (2019). Multi-scale convolutional neural network for multi-focus image fusion. *Image Vision Computing* 85, 26–35. doi: 10.1016/j.imavis.2019.03.001

Qin, Y., Yan, C., Liu, G., Li, Z., and Jiang, C. (2020). Pairwise gaussian loss for convolutional neural networks. *IEEE Trans. Ind. Inf.* 16, 6324–6333. doi: 10.1109/TII.9424

Redmon, J., and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*. doi: 10.48550/arXiv.1804.02767

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Shaw, P., Uszkoreit, J., and Vaswani, A. (2018). Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*. doi: 10.48550/arXiv.1803.02155

Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., et al. (2023). Sparse r-cnn: An end-to-end framework for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 15650–15664. doi: 10.1109/TPAMI.2023.3292030

Tang, L., Tang, W., Qu, X., Han, Y., Wang, W., and Zhao, B. (2022). A scale-aware pyramid network for multi-scale object detection in sar images. *Remote Sens.* 14, 973. doi: 10.3390/rs14040973

Wang, J., Feng, C., Wang, L., Li, G., and He, B. (2022c). Detection of weak and small targets in forward-looking sonar image using multi-branch shuttle neural network. *IEEE Sensors J.* 22, 6772–6783. doi: 10.1109/JSEN.2022.3147234

Wang, Z., Guo, J., Zeng, L., Zhang, C., and Wang, B. (2022d). Mlffnet: Multilevel feature fusion network for object detection in sonar images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–19. doi: 10.1109/TGRS.2022.3214748

Wang, B., Ji, R., Zhang, L., and Wu, Y. (2022a). Bridging multi-scale context-aware representation for object detection. *IEEE Trans. Circuits Syst. Video Technol.* 33, 2317–2329. doi: 10.1109/TCSVT.2022.3221755

Wang, D., Shang, K., Wu, H., and Wang, C. (2022b). Decoupled r-cnn: Sensitivity-specific detector for higher accurate localization. *IEEE Trans. Circuits Syst. Video Technol.* 32, 6324–6336. doi: 10.1109/TCSVT.2022.3167114

Wang, C., and Wang, H. (2023). Cascaded feature fusion with multi-level self-attention mechanism for object detection. *Pattern Recognition* 138, 109377. doi: 10.1016/j.patcog.2023.109377

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q. (2020). "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Seattle, WA, USA, 11534–11542.

Xiao, J., Guo, H., Yao, Y., Zhang, S., Zhou, J., and Jiang, Z. (2022). Multi-scale object detection with the pixel attention mechanism in a complex background. *Remote Sens.* 14, 3969. doi: 10.3390/rs14163969

Xinyu, T., Xuewu, Z., Xiaolong, X., Jinbao, S., and Yan, X. (2017). "Methods for underwater sonar image processing in objection detection," in *2017 International conference on computer systems, electronics and control (ICCSEC)*, Dalian, China, 941–944.

Yasir, M., Liu, S., Pirasteh, S., Xu, M., Sheng, H., Wan, J., et al. (2024). Yoloshiptracker: Tracking ships in sar images using lightweight yolov8. *Int. J. Appl. Earth Observation Geoinformation* 134, 104137. doi: 10.1016/j.jag.2024.104137

Yuanzi, L., Xiufen, Y., and Weizheng, Z. (2022). Transyolo: high-performance object detector for forward looking sonar images. *IEEE Signal Process. Lett.* 29, 2098–2102. doi: 10.1109/LSP.2022.3210839

Zhang, M., Cai, W., Wang, Y., and Zhu, J. (2023). A level set method with heterogeneity filter for side-scan sonar image segmentation. *IEEE Sensors J.* 24, 584–595. doi: 10.1109/JSEN.2023.3334765

Zhang, H., Chang, H., Ma, B., Wang, N., and Chen, X. (2020). "Dynamic r-cnn: Towards high quality object detection via dynamic training," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Glasgow, UK. 260–275.

Zhang, H., Tian, M., Shao, G., Cheng, J., and Liu, J. (2022a). Target detection of forward-looking sonar image based on improved yolov5. *IEEE Access* 10, 18023–18034. doi: 10.1109/ACCESS.2022.3150339

Zhang, Y., Zhang, H., Liu, J., Zhang, S., Liu, Z., Lyu, E., et al. (2022b). Submarine pipeline tracking technology based on auvs with forward looking sonar. *Appl. Ocean Res.* 122, 103128. doi: 10.1016/j.apor.2022.103128

Zhang, M.-L., and Zhou, Z.-H. (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition* 40, 2038–2048. doi: 10.1016/j.patcog.2006.12.019

Zhao, Z., Wang, Z., Wang, B., and Guo, J. (2023). Rmfenet: Refined multi-scale feature enhancement network for arbitrary oriented sonar object detection. *IEEE Sensors J.* 23, 29211–29226. doi: 10.1109/JSEN.2023.3324476

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., and Ren, D. (2020). "Distance-iou loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI conference on artificial intelligence*, New York, USA. 34. 12993–13000.

Zhou, T., Si, J., Wang, L., Xu, C., and Yu, X. (2022b). Automatic detection of underwater small targets using forward-looking sonar images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–12. doi: 10.1109/TGRS.2022.3181417

Zhou, K., Zhang, M., Wang, H., and Tan, J. (2022a). Ship detection in sar images based on multi-scale feature extraction and adaptive feature fusion. *Remote Sens.* 14, 755. doi: 10.3390/rs14030755

Zhou, L., Zhao, S., Wan, Z., Liu, Y., Wang, Y., and Zuo, X. (2024). Mfefnet: A multi-scale feature information extraction and fusion network for multi-scale object detection in uav aerial images. *Drones* 8, 186. doi: 10.3390/drones8050186

Zhu, X., Cheng, D., Zhang, Z., Lin, S., and Dai, J. (2019). "An empirical study of spatial attention mechanisms in deep networks," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, Seoul, South Korea. 6688–6697.

Zhu, Y., Zhao, C., Guo, H., Wang, J., Zhao, X., and Lu, H. (2018). Attention couplenet: Fully convolutional attention coupling network for object detection. *IEEE Trans. Image Process.* 28, 113–126. doi: 10.1109/TIP.2018.2865280

Zong, Z., Song, G., and Liu, Y. (2023). "Detrs with collaborative hybrid assignments training," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, Paris, France. 6748–6758.