2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2025)
October 19-25, 2025. Hangzhou, China

2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 979-8-3315-4393-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/IROS60139.2025.11247598

WFDA: Wavelet-based Frequency Decomposition and Aggregation for Underwater Object Detection

Xueting Liu, *Student Member, IEEE* Chunying Li*, *Member, IEEE* Shuxiang Guo*, *Fellow, IEEE*

***Abstract*—Underwater Object Detection (UOD) techniques are critical for Autonomous Underwater Vehicle (AUV), which must operate in harsh underwater environments characterized by low visibility while satisfying the lightweight and real-time constraints required for vehicle-mounted systems. Current methods typically rely on underwater image enhancement combined with object detection to adapt to underwater conditions. However, these approaches mainly focus on the spatial domain, often overlooking the frequency-domain characteristics of the underwater environment. This oversight limits the removal of noise factors, such as scattering, blurring, distortion, and uneven illumination, and diminishes the focus of object edges and textures. Additionally, the increased parameter size and higher computational cost render them less suitable for real-time detection. To address these, the Wavelet-based Frequency Decomposition and Aggregation Network (WFDA) was proposed, which leverages the Wavelet Transform (WT) to decompose features into high- and low-frequency components for effective feature modeling and fusion-based downsampling. Specifically, the Wavelet-Based Feature Decomposition Modeling (WDM) module utilized multi-level wavelet decomposition to hierarchically model features across different frequency bands, while the Wavelet-Based Feature Aggregation Downsampling (WAD) module refined and extracted core features through single-level wavelet decomposition combined with channel aggregation. Evaluations on four public datasets demonstrate that WFDA achieved state-of-the-art (SOTA) performance and efficiency, making it well-suited for real-time, high-accuracy detection on robotic platforms. Code is available at https://github.com/Mariiiioooo/WFDA.**

***Index Terms*—Underwater Object Detection (UOD), Autonomous Underwater Vehicle (AUV), Wavelet Transform (WT), Frequency-domain Characteristics, lightweightness**

I. INTRODUCTION

UOD is crucial for marine exploration [1], [2], resource development [3], and for applications involving AUV and bionic Amphibious Underwater Robots (ASR) [4] to perform tasks autonomously. These systems rely on efficient and accurate detection models that can meet the rigorous real-time processing requirements of underwater environments [5]. Meanwhile, the inherent challenges of underwater imaging, such as scattering, blurring, distortion, uneven illumination, and floating particles, make feature extraction and detection in UOD more difficult than in traditional object detection methods. Existing object detection methods are classified into two main categories: two-stage detectors and single-stage detectors. Two-stage detectors (Faster-RCNN etc.) [6] use RoIPooling or RoIAlign to crop features from the backbone network. In contrast, single-stage detectors, such as YOLO [7], directly regress both the bounding box coordinates and the object classes in a single step, which is the focus of this paper.

Even though current object detection methods have achieved considerable success, their reliance on extracting and optimizing local features in the spatial domain often results in in key feature attention. This limitation diminishes their ability to effectively perceive critical textures, boundaries, and noise within the spatial distribution of objects, particularly in underwater environments.As a result, feature extraction may suffer from biases and deficiencies. Moreover, traditional object detection methods exhibit limited robustness to common challenges in harsh underwater environments [8], such as scattering, blurring, distortion, and noise. This further hampers the effective

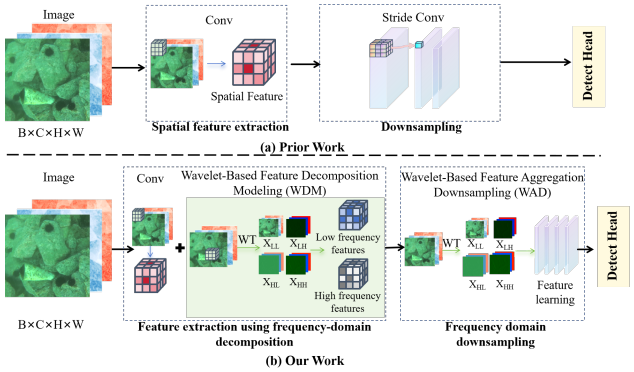


Fig. 1. Different methods of UOD (a) Prior Work: Feature extraction with convolution and stride-based downsampling, limited by poor handling of features and noise. (b) Our Work: Feature extraction using frequency-domain high-low decomposition modeling and aggregation downsampling, effectively capturing features and improving robustness to noise and blur.

* Corresponding authors: Chunying Li and Shuxiang Guo

Research supported in part by the Shenzhen Science and Technology Program under Grant RCBS20231211090725048, Shenzhen, China, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515011007, Guangdong, China, in part by the High level of special funds under Grant G03034K003 from Southern University of Science and Technology, Shenzhen, China.

Xueting Liu and Chunying Li are with the Dept. of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China (liuxt2023@mail.sustech.edu.cn; licy@sustech.edu.cn).

Shuxiang Guo is with the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China, also with the Aerospace Center Hospital, School of Life Science and the Key Laboratory of Convergence Medical Engineering System and Healthcare Technology, Ministry of Industry and Information Technology, Beijing Institute of Technology, Beijing, China (guo.shuxiang@sustech.edu.cn).

extraction and attention to underwater object features, leading to challenges in performance.

Inspired by previous works [9], [10], we argue that mining frequency information in underwater images is essential. Different frequency components can reflect the distinct characteristics of objects in the underwater environment and help differentiate key information from redundant noise. Given the subtle and complex nature of object properties in underwater settings, leveraging both low-frequency and high-frequency information is crucial for enhancing detection performance. Specifically, in underwater images, low-frequency information helps preserve the overall structure of objects and background context, while high-frequency information highlights fine textures and key edge features, aiding in the recovery of details that may be blurred or obscured by noise. Decomposing and modeling both low- and high-frequency information allows for fine-grained, specialized feature extraction from each frequency band, while the aggregation of these features helps retain important frequency information and effectively mitigate the impact of noise on high-frequency features. Therefore, there is an urgent need for new methods that can effectively exploit the full spectrum of frequency information to overcome these limitations, reducing the effects of noise and blurring in harsh underwater environments and enhancing the ability to detect underwater objects.

In this paper, we propose a Wavelet-based Frequency Decomposition and Aggregation Network (WFDA) that leverages the Haar Wavelet Transform (WT) to decompose features into high- and low-frequency components and model them hierarchically, thereby enhancing the attention to key features of underwater objects. Specifically, WFDA introduces the Wavelet-Based Feature Decomposition Modeling (WDM) to perform decomposition of high- and low-frequency features in different directions. The two-level decomposition extracts multi-frequency features, enabling specialized extraction, selection and fusion of features across different frequency bands. Consequently, WDM helps improve the extraction of key textures and edges while reducing the impact of high-frequency noise. Additionally, WFDA introduces the Wavelet-Based Feature Aggregation Downsampling (WAD) to achieve adaptive dynamic feature downsampling. This module uses decoupled spatial convolutions and wavelet-based downsampling to replace traditional stride convolutions, facilitating better interaction patterns. Wavelet downsampling can decompose high- and low-frequency features and adaptively achieve channel aggregation of different frequency features, thereby improving the retention of critical fine-grained details. Importantly, since the WT does not require learnable parameters and neither the WDM nor the WAD introduces complex training structures, the approach incurs minimal additional computational burden, making it well-suited for lightweight model requirements.

To evaluate the effectiveness of WFDA, we assessed its performance on four public available datasets: URPC2020 [11], RUOD [12], TrashCan [13], and Brackish [14]. The experimental results demonstrate that the proposed WFDA achieves SOTA performance on URPC2020 [11] and TrashCan [13]. The contributions of this paper are summarized as follows:

- WFDA is the first study to apply the wavelet transform to underwater object detection, enhancing key feature extraction and preservation through high-low frequency decomposition.
- The Wavelet-Based Feature Decomposition Modeling (WDM) module is proposed to decompose, selectively extract, and fuse multi-frequency features, enhancing key feature extraction while reducing high-frequency noise.
- The Wavelet-Based Feature Aggregation Downsampling (WAD) module adaptively aggregates and downsamples multi-frequency features, preserving key details while reducing feature distortion.

II. RELATED WORK

A. Underwater Object Detection

UOD is a specialized technology for underwater environments, which are challenging due to factors like light scattering, absorption, and color distortion [15]. Existing underwater object detection methods [16], [17] have attempted to use Underwater Image Enhancement (UIE) as a preprocessing step for object detection, but this doesn't always improve detection performance and can sometimes degrade it [18]. Because image enhancement alter the image distribution, causing domain shifts that harm object detection performance. To address this, Dai *et al*. [19] proposed GCC-Net, which combines features from both original and enhanced images for better domain fusion. Other studies like ROIMIX [20] and contrastive learning [21] have sought domain generalization through data augmentation. Zhou *et al*. [22] introduced AMSP-UOD to reduce underwater noise in feature extraction, but traditional CNNs still struggle with the complex textures and noise in underwater images. Unlike existing methods, we proposed leveraging WT to improve UOD performance, as shown in the Fig. 2. WT retains the time-frequency information of images, enabling it to capture both low-frequency and high-frequency features. This capability allows it to effectively capture both the global structure and fine details of underwater targets, significantly enhancing the accuracy and robustness of UOD.

B. Application of Wavelet Transform in Visual Tasks

WT has seen increasing integration into neural networks due to its multi-scale analysis capabilities. It has been applied across time-frequency analysis of ECG signals [23], image super-resolution, and face reconstruction [24], [25] and generative models [9]. WT has also facilitated the development of multi-level wavelet convolutional neural networks (MWCNN) [26], balancing computational efficiency with multi-scale analysis. For architectural enhancements, Xu *et al*. [27] introduced the Haar wavelet transform downsampling module (HWD) to enhance image segmentation across diverse datasets and CNN architectures.

Additionally, Finde *et al*. [28] proposed Wavelet Convolutions for Large Receptive Fields, a lightweight alternative to depthwise convolutions that expands receptive fields. Despite these advancements, existing methods are designed for specific tasks, prioritizing either computational efficiency [27] or architectural refinement. They fail to fully optimize interactions between high- and low-frequency features, limiting their adaptability in complex, noise-prone environments. To address these limitations, we proposed a wavelet-based approach that decomposes features into high- and low-frequency components and integrates them through hierarchical modeling. This enhances attention to critical features, improving feature representation and robustness.

III. METHOD

A. Overview

The overall pipeline of WFDA is shown in Fig. 2, For an input image $I \in \mathbb{R}^{H\times W\times 3}$, our model first processes the features through a WDM module, which employs wavelet-domain processing to enhance the feature representation. Next, the enhanced feature map is fed into the WAD to reduce spatial resolution while preserving critical boundary information and high-frequency details. Finally, a detection head is used to output the classification and localization results. All components will be detailed in the following sections.

B. Wavelet Transform (WT)

WT is a technique utilized for analyzing signals across multiple resolutions. The Haar wavelet transform is a widely utilized and orthogonal transformation. The 1-stage one-dimensional Haar transform is defined by the following wavelet basis and scaling functions:

$$\begin{cases}\phi_1(x) = \frac{1}{\sqrt{2}}\phi_{1,0}(x) + \frac{1}{\sqrt{2}}\phi_{1,1}(x),\\ \psi_1(x) = \frac{1}{\sqrt{2}}\phi_{1,0}(x) - \frac{1}{\sqrt{2}}\phi_{1,1}(x).\end{cases} \tag{1}$$

The general Haar basis function is formulated as:

$$\phi_{j,k}(x) = \sqrt{2^j}\phi(2^j x - k), \quad k = 0, 1, \ldots, 2^j - 1, \tag{2}$$

where j and k denotes the scale level and the spatial shift of the Haar function. Additionally, The fundamental scaling function $\phi_{0,0}(x)$ is given by:

$$\phi_{0,0}(x) = \phi_0(x) = \begin{cases}0, & x < 0,\\ 1, & 0 \le x < 1,\\ 0, & x \ge 1.\end{cases} \tag{3}$$

Thus, the 1-stage Haar transform can be rewritten using 0-stage Haar basis functions:

$$\begin{cases}\phi_1(x) = \phi_0(2x) + \phi_0(2x-1),\\ \psi_1(x) = \phi_0(2x) - \phi_0(2x-1).\end{cases} \tag{4}$$

When applied to a two-dimensional signal (such as a image $F \in \mathbb{R}^{H\times W}$), the Haar wavelet transform produces four components, each with half the spatial resolution of the original signal. The filtering process is performed both horizontally and vertically, as follows:

$$F = F_{\mathrm{LL}},\, F_{\mathrm{LH}},\, F_{\mathrm{HL}},\, F_{\mathrm{HH}} = \mathrm{DWT}(F), \tag{5}$$

where F_{LL} represents the low-frequency components, while $F_{\mathrm{LH}}, F_{\mathrm{HL}}$, and F_{HH} represent the high-frequency details, such as edges and textures.

In this paper, WT is applied to decompose feature maps into low-frequency and high-frequency details, allowing the model to better capture edges, textures, and subtle variations in the image, thus improving object detection performance.

C. Wavelet-Based Feature Decomposition Modeling (WDM)

To fully leverage the extracted features, we proposed WDM, which specialized extraction and fusion of different frequency features can be carried out, refines the feature representation by capturing subtle texture variations and boundary cues. In WDM, we have set up a Wavelet-Transform Bottleneck (WTBottleneck) block along with a traditional convolutional Bottleneck. WDM is fully configurable within the model architecture: By setting the use-wavelet parameter to True, all convolutional blocks automatically adopt the WTBottleneck. If use-wavelet is False, the standard convolution-based feature extractor is used instead. This design provides flexibility, allowing WDM to be easily adapted to various architectures. In our backbone, the parameter controlling the wavelet-enhanced variant (use-wavelet) is set to True for all blocks.

The WTBottleneck accepts an input feature map $F \in \mathbb{R}^{B\times C\times H\times W}$, where B is the batch size, C the number of channels, and H and W are the spatial dimensions. Its goal is to enhance feature extraction by preserving both global structure and fine details using a wavelet-based approach combined with conventional convolutional processing.

A 3×3 convolution is applied to F to reduce its dimensionality and prepare the features for enriching the feature extraction process. And then, the input feature map F is decomposed into four sub-bands using DWT. These sub-bands are then reshaped into a tensor $\tilde{F} \in \mathbb{R}^{B\times 4C\times \frac{H}{2}\times \frac{W}{2}}$.

A depth-wise convolution with learnable scaling factors was applied to further refine the extracted multi-frequency features:

$$\tilde{F}' = \gamma \odot \mathrm{Conv}_{\mathrm{dw}}(\tilde{F}), \tag{6}$$

where γ represents the learnable scaling factors that dynamically adjust the contribution of different frequency components and $\odot$ denotes element-wise multiplication. By processing each frequency sub-band separately, this step ensures that different frequency components receive adaptive feature enhancement.

Next, the processed tensor is reshaped back into its sub-band format: $F' \in \mathbb{R}^{B\times C\times 4\times \frac{H}{2}\times \frac{W}{2}}$. An Inverse Discrete Wavelet Transform (IDWT) is then applied to reconstruct the feature map to the original resolution, to realize specialized extraction and fusion of different frequency features:

$$\hat{F} = \mathrm{IDWT}(F'_L + \hat{F}_{next}, F'_H), \tag{7}$$

Where $F'_L \in \mathbb{R}^{B\times C\times 1\times \frac{H}{2}\times \frac{W}{2}}$ is the low-frequency component, $F'_H \in \mathbb{R}^{B\times 3\times C\times \frac{H}{2}\times \frac{W}{2}}$ represents the three high-frequency components, and $\hat{F}_{next}$ is the output of

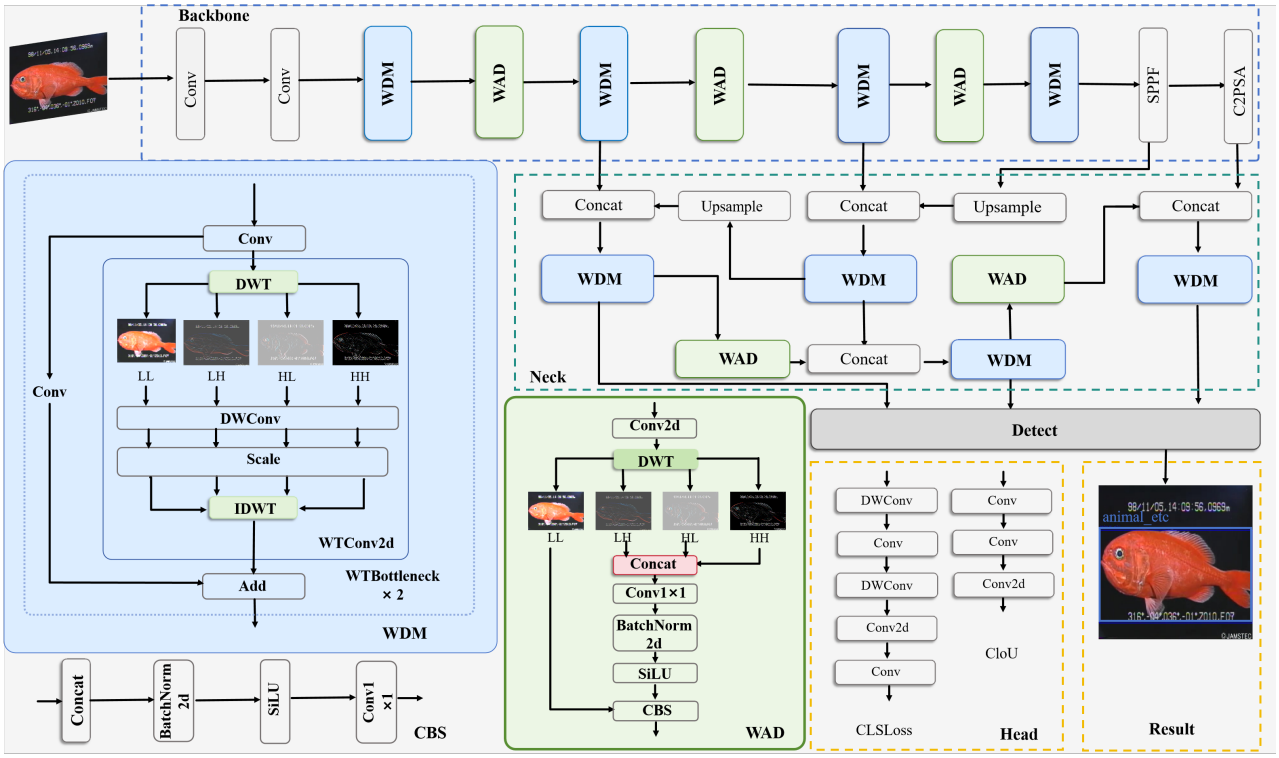


Fig. 2. The overall framework of WFDA. For underwater images, we first extract frequency-domain high-low decomposition, capturing both global structures and detailed information using WDM. Then, WAD is applied to the image, dynamically adaptive aggregation of features from different frequencies, preserving the integrity of both high-frequency and low-frequency features. Finally, fuse these features to enhance the ability to perceive information, optimizing object detection performance.

the next level of wavelet decomposition. The reconstructed feature is $\hat{F} \in \mathbb{R}^{B\times C\times H\times W}$.

By preserving the essential low-frequency content while selectively refining high-frequency details, this reconstruction process ensures that both subtle texture variations and boundary cues are effectively retained.

To further enhance the extracted features, a residual connection was introduced. First, compute a Standard Feature Map (SFM) using a standard 3×3 convolution:

$$\text{SFM} = \text{Conv}_{\text{base}}(F). \tag{8}$$

The wavelet-based output is fused with the standard feature map using a residual connection:

$$Y = \alpha \cdot \text{SFM} + \hat{F}, \tag{9}$$

where α is a learnable weight that balances the contribution of conventional convolutional features and wavelet-based features. This fusion ensures that WDM can fully exploit both spatial domain and frequency domain features, resulting in a more comprehensive representation.

Finally, in order to refine the extracted features, two sequential Wavelet-Transform Bottleneck (WTBottleneck) blocks are applied:

$$Y' = \text{WTBottleneck}^{(2)}(Y). \tag{10}$$

Local feature extraction using standard convolution, Multi-frequency refinement through wavelet transforms, Residual learning to preserve important information while suppressing noise.

D. Wavelet-Based Feature Aggregation Downsampling (WAD)

Traditional downsampling methods, such as strided convolution and pooling, often lose high-frequency details that are crucial for accurately capturing object boundaries and fine textures. In contrast, WAD leverages the DWT to decompose the input feature map into frequency-specific sub-bands, and adaptively aggregate channel features based on different frequency information, which preserves the integrity of both high-frequency and low-frequency components. This ensures that critical structural and texture details are maintained during resolution reduction, thereby enhancing the robustness of features for object detection in noisy and blurred underwater environments.

Given an input feature map F, prior to applying the WT, a 3×3 convolution is applied to reduce input dimensionality and prepare features for decomposition:

$$F' = \text{Conv}_{3\times 3}(F), \tag{11}$$

where F' has half the number of channels compared to the input F, ensuring computational efficiency while preserving essential information.

Then, the decomposition process generates four subbands through DWT. Since high-frequency sub-bands $(F_{\text{LH}}, F_{\text{HL}}, F_{\text{HH}})$ contain vital texture and edge information, we refine these features before merging them. First, the

high-frequency components are concatenated:

$$F_{\text{HF}} = \text{Concat}(F_{\text{LH}}, F_{\text{HL}}, F_{\text{HH}}). \quad (12)$$

A lightweight 1×1 convolution is applied to enhance these features, followed by batch normalization and activation:

$$F'_{\text{HF}} = \sigma(\text{BN}(\text{Conv}_{1\times1}(F_{\text{HF}}))), \quad (13)$$

where $\text{Conv}_{1\times1}$ reduces channel dimensions to minimize redundancy. $\text{BN}(\cdot)$ normalizes the feature maps to stabilize training. $\sigma(\cdot)$ is the SiLU activation function, which introduces non-linearity and improves feature expressiveness.

To fully exploit both global structure (low-frequency) and fine details (high-frequency), feature fusion is performed before passing the downsampled features to the next stage. The final output is computed as:

$$F_{\text{out}} = \sigma(\text{BN}(\text{Conv}_{1\times1}(\text{Concat}(F_{\text{LL}}, F'_{\text{HF}})))), \quad (14)$$

where F_{LL} ensures the preservation of global contextual information, while F'_{HF} retains crucial texture and boundary details. The final 1×1 convolution refines the fused features and projects them into the desired output dimensionality.

E. Loss Function

Our detection loss is computed as a weighted sum of the box loss, classification loss (Cls loss), and distribution focal loss (DFL loss). The overall loss is computed as:

$$\text{loss}_{\text{total}} = \lambda_{\text{box}}\,\text{BoxLoss} + \lambda_{\text{cls}}\,\text{ClsLoss} + \lambda_{\text{dfl}}\,\text{DFLLoss}, \quad (15)$$

where the box loss optimizes the difference between the predicted and ground truth bounding boxes, the classification loss ensures accurate object class predictions, and the DFL loss addresses regression uncertainty.

F. Overall Framework

Our model begins by processing an input image through WDM that enhances feature representation multi-scale wavelet-domain information. A residual fusion mechanism, combined with sequential bottleneck blocks, further refines the features. The enriched features are then passed into WAD, which effectively reduces spatial resolution while preserving critical boundary cues and high-frequency details through discrete wavelet decomposition and reconstruction. Finally, detection head produces the final classification and localization outputs. This design not only improves the ability to capture both global structure and fine details but also enhances robustness in challenging scenarios, such as noisy or blurred environments, by maintaining essential texture and edge information throughout the network.

IV. EXPERINMENT RESULT

A. Datasets and Evaluation Metrics

1) Datasets: We evaluate WFDA on URPC2020 [11], RUOD [12], TrashCan [13], and brackish [14]. URPC [11] is a dataset used in the Underwater Robot Professional Contest (URPC), comprising 5543 underwater images for training, 1,200 images from its B-list answers as the test set, covering four types of underwater organisms: echinus, holothurian, scallop, and starfish. RUOD [12] includes 9,800 training images and 4,200 test images, with three underwater detection challenges: fog effect, color cast, and light interference, encompassing 10 types of underwater targets such as fish, divers, and starfish. TrashCan [13] is an instance segmentation annotation dataset for underwater garbage, comprising 16 categories, including garbage, remotely operated vehicles, and a diverse array of underwater flora and fauna. Brackish [14] is captured in temperate brackish water environments, encompassing six categories such as large fish, crabs, jellyfish.

2) Evaluation Metrics: The results in this paper follow the standard COCO-style Average Precision (AP) metrics, including AP, AP_{50} (IoU=0.5), and AP_{75} (IoU=0.75). AP is computed by averaging across multiple IoU thresholds ranging from 0.5 to 0.95 with an interval of 0.05.

B. Implementation Details

The proposed model was trained on a single NVIDIA GeForce RTX 4090 GPU. We used the SGD optimizer with a weight decay of 0.0005 and momentum of 0.937. Training configurations included an input image size of 640×640, a batch size of 16, and a fixed random seed (0) to ensure reproducibility. The initial learning rate was set to 0.01 with SGD-compatible scheduling, and the model was trained for 300 epochs. No pre-trained weights were used during initialization ("Weights: None").

C. Comparisons with the SOTA

We compared WFDA with several SOTA methods on URPC2020 [11], Brackish [14], TrashCan [13], and RUOD [12] datasets. The results are shown in Table I and Table II.

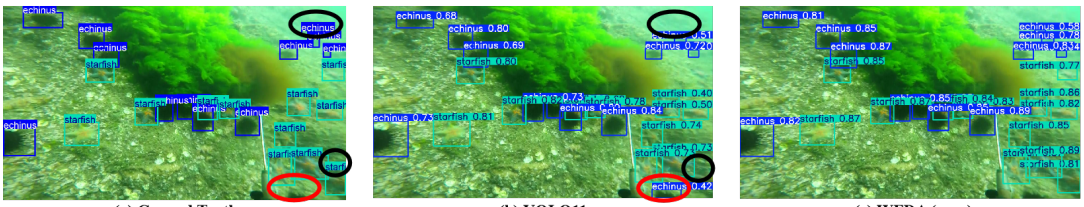
1) Results on URPC [11]: WFDA (Ours) demonstrates outstanding performance on the URPC dataset, outperforming several SOTA methods. Specifically, WFDA has a maximum AP_{50} of 90.4 and AP_{75} of 79.4. As shown in the tableI, WFDA consistently outperforms all other methods in both AP_{50} and AP_{75} metrics. Compared to the SOTA GCC-Net [19] method, our model achieved a performance gain of 2.6% in AP_{50} and 3.1% in AP_{75}. Furthermore, compared to the YOLO11 [7] model, our model demonstrates improvements of 1.2% in AP_{50} and 5.5% in AP_{75}. Moreover, as shown in the Fig. 3, the black circles indicate missed detections, while the red circles indicate false detections. Compared to YOLO11, our method not only aligns more closely with the ground truth but also has lower rates of false detections and missed detections. These results validate the effectiveness of our proposed method.

WFDA establishes a WDM architecture that combines wavelet-based decomposition with convolutional processing. This hybrid design allows our model to effectively capture both global structure (low-frequency) and fine-grained details (high-frequency). Additionally, we introduced a WAD that preserves both high-frequency and low-frequency features during resolution reduction, dynamically and adaptively retains the core parts of high-frequency and low-frequency featuressignificantly improving detection performance in underwater environments. Moreover, as evident from the

TABLE I

PERFORMANCE COMPARISON ON THE URPC DATASET. **BOLD** AND UNDERLINE INDICATE THE BEST AND SECOND-BEST RESULTS IN EACH COLUMN. GOD IS GENERIC OBJECT DETECTION, UOD IS UNDERWATER OBJECT DETECTION

	Methods	AP↑	AP_{50}↑	AP_{75}↑	echinus	holothurian	scallop	starfish	Params (M)↓	GFLOPs↓
GOD	RetinaNet [29]	57.8	78.9	63.0	79.4	68.2	51.1	74.9	55.38	83.2
	Faster R-CNN [6]	61.2	82.7	69.6	70.4	61.4	41.9	71.4	41.3	120
	Cascade R-CNN [30]	61.6	80.5	72.4	69.0	61.9	41.9	72.0	88.15	140
	DetectoRS [31]	60.8	81.4	69.3	69.5	60.9	41.1	70.2	123.23	90.03
	YOLOv7 [32]	49.8	85.4	64.2	73.7	66.3	50.8	74.5	6.2	13.8
	YOLOv8 [33]	59.1	88.1	72.8	75.2	64.8	54.7	70.9	3.2	8.7
	YOLO11 [7]	60.8	89.2	73.9	75.5	64.3	55.2	73.5	**2.6**	**6.3**
UOD	Boosting R-CNN [34]	63.7	79.0	72.3	70.0	64.3	46.6	74.8	45.95	169
	RoIAttn [35]	62.2	82.8	69.5	70.7	62.2	40.5	71.4	55.23	331.7
	ERL-Net [36]	63.7	81.9	72.2	70.8	66.6	45.4	73.7	45.95	54.8
	GCC-Net [19]	**69.1**	87.8	76.3	75.2	68.2	56.3	76.7	36.74	300.79
	AMSP-UOD [22]	40.1	78.5	–	**87.5**	60.6	42.5	77.5	10.4	23.8
	WFDA (Ours)	66.3	**90.4**	**79.4**	82.3	**70.9**	**60.0**	**80.0**	**2.6**	7.2



(a) Ground Truth (b) YOLO11 (c) WFDA (ours)

Fig. 3. Visualization comparison results with YOLO11, red circles indicate false positives, and black circles indicate missed detections.

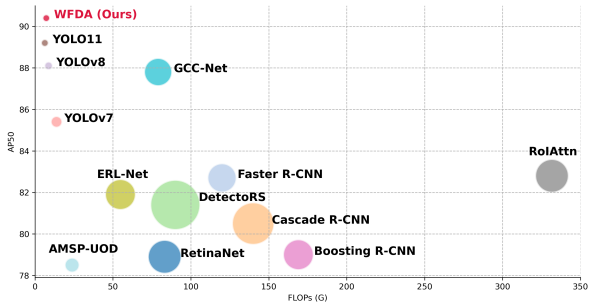


Fig. 4. Algorithm-Performance Comparison with the SOTA. The size of a circle represents its parameter count.

Table II and Fig. 4, WFDA has a relatively low number of GLOPs (Floating Point Operations), which translates to lower computational cost during inference. This makes it more efficient for deployment on AUVs with limited computational power, offering a practical solution in real-world applications compared to other larger, more computationally intensive models. WFDA strikes an efficient balance between accuracy and speed, making it particularly well-suited for real-time decision-making in dynamic underwater environments.

2) Results on TrashCan [13], RUOD [12] and Brackish [14] : For these datasets, we present the performance of WFDA and compare it with several SOTA methods. As shown in TableII, WFDA demonstrates superior performance across all three datasets. Specifically, it achieves the highest AP and AP_{50} scores in the TrashCan [13] datasets, outperforming other methods such as GCC-Net [19] and ERL-Net [36]. In the RUOD [12], WFDA achieves SOTA at AP_{50} of 86.1. In the Brackish [14] dataset, WFDA achieves AP of 82.6 and AP_{50} of 98.5, which are better than the second-best results. These results highlight the effectiveness of the proposed architecture in capturing both local and global features in challenging underwater environments.

TABLE II

BENCHMARKING RESULTS BETWEEN WFDA AND OTHER SOTA METHODS ON TRASHCAN [13], RUOD [12] AND BRACKISH [14].

	Methods	TrashCan		RUOD		Brackish	
		AP↑	AP_{50}↑	AP↑	AP_{50}↑	AP↑	AP_{50}↑
GOD	RetinaNet [29]	29.4	53.8	48.2	74.3	74.0	91.3
	Faster R-CNN [6]	31.2	55.3	54.3	74.9	61.2	62.9
	Cascade R-CNN [30]	33.6	52.7	56.0	79.7	80.7	98.5
	YOLOv7 [32]	26.1	48.8	47.2	76.4	71.3	90.5
	YOLOv8 [33]	44.2	61.5	58.0	82.1	80.1	97.7
	YOLO11 [7]	45.1	61.9	58.1	81.9	80.4	98.1
UOD	Boosting R-CNN [34]	36.8	57.6	52.3	79.8	79.6	97.4
	RoIAttn [35]	32.5	56.8	49.7	75.2	79.3	91.2
	ERL-Net [36]	37.0	58.9	53.2	80.5	**85.4**	**98.8**
	GCC-Net [19]	41.3	61.2	56.1	83.2	80.5	98.3
	AMSP-UOD [22]	–	–	**65.2**	**86.1**	–	–
	WFDA (Ours)	**46.9**	**63.9**	63.4	**86.1**	82.6	98.5

D. Ablation Study

In this section, we present a series of ablation experiments conducted on the TrashCan [13] dataset to evaluate the effectiveness of the components in our proposed method. We use YOLO11 as the baseline due to its strong performance and efficient inference speed.

1) Impact of Model Components: As shown in the TableIII, the original YOLO11 model achieves 45.1 AP and 61.9 AP_{50}. By incorporating the WAD, we observe a performance improvement, with AP increasing to 45.8 and AP_{50} to 62.6. This shows that retaining both high-frequency and low-frequency information during downsampling significantly improves detection accuracy. Adding the WDM further enhances performance, reaching 46.0 AP and 63.3 AP_{50}. This indicates that optimize the extraction and fusion of different frequency features through WDM can extract key detail features from different levels. The full model, incorporating both the WAD and WDM, achieves the highest performance with 46.9 AP and 63.9 AP_{50}. These results validate the importance of each component in the model, and highlight the significant improvement in performance when combining the WAD and WDM.

TABLE III
ABLATION STUDY RESULTS. BOLD INDICATE THE BEST RESULTS IN EACH COLUMN.

Module Combination	AP↑	AP_{50}↑
Baseline (YOLO11)	45.1	61.9
+ WAD	45.8	62.6
+ WDM	46.0	63.3
Complete Model	**46.9**	**63.9**

2) Impact of the WAD : To assess the impact of the WAD, we conducted four ablation experiments, each focusing on different configurations of the module, as shown in the TableIV. The results indicate that using feature aggregation and optimally placing the 3×3 Conv significantly improves performance. Without any 3×3 Conv or aggregation, the baseline achieves 45.1 AP. Adding aggregation after DWT boosts performance (AP 46.1), and applying the 3×3 Conv before DWT with aggregation yields the best results (AP 46.9) with a moderate GFLOPs cost, demonstrating the effectiveness of the WAD.

TABLE IV
IMPACT OF THE WAD. FEATURE AGGREGATION: 1×1 CONV

Configuration		AP↑	AP_{50}↑	GFLOPs↓
3×3 Conv Position	Feature Aggregation			
×	×	45.1	61.9	6.3
×	✓	44.2	60.3	**5.5**
After DWT	×	45.3	62.0	9.2
After DWT	✓	46.1	62.7	7.1
Before DWT	✓	**46.9**	**63.9**	7.3

TABLE V
ABLATION STUDY: IMPACT OF WDM PLACEMENT IN BACKBONE.

Configuration	AP↑	AP_{50}↑	GFLOPs↓
No WDM (baseline)	45.1	62.9	**6.3**
+ WDM at P1, P2	45.4	63.3	6.7
+ WDM at P3	45.7	62.7	6.6
+ WDM at P4, P5	**47.1**	63.3	6.6
Complete Model	46.9	**63.9**	7.2

3) Impact of the WDM: In the ablation study, we examine the impact of placing the WDM at various stages within the backbone. As shown in the TableV, the results indicate that the placement of WDM has significant effects on both performance and computational efficiency.Placing the WDM in the early layers (P1 and P2) enhances performance by capturing local features and textures early in the network, leading to slight improvements in AP and AP_{50}. When the WDM is applied to the middle layer (P3), it refines mid-level features, resulting in a moderate performance gain. In the deeper layers (P4 and P5), the WDM improves the model's ability to extract more abstract features, which leads to higher AP and AP_{50} scores. However, the complete model shows the best overall performance. This is because it effectively captures both local and global features early on, providing more informative input to the subsequent layers and enhancing the overall feature extraction process.

V. CONCLUSION

In this paper, the Wavelet-based Frequency Decomposition and Aggregation Network (WFDA) was proposed to enhance fine-grained detail discrimination and mitigate noise effects in underwater object detection. The approach leveraged the inherent high-low frequency decomposition of wavelet transforms to design two specialized modules: the WDM, which performs frequency-specific feature extraction, selection, and fusion, and the WAD, which adaptively aggregates and preserves core features from different frequency bands. Extensive experiments on four public datasets demonstrated that WFDA achieves state-of-the-art performance in terms of AP and AP_{50}, while maintaining high computational efficiency to meet both lightweight and high-performance requirements. Furthermore, ablation studies have validated the effectiveness of the proposed framework. However, WFDA still faces certain limitations, including its dependence on predefined wavelet bases that may not be optimal for all underwater scenarios. In future work, we plan to explore adaptive wavelet basis learning to further improve detection performance under challenging underwater conditions.

REFERENCES

[1] S. Guo, T. Fukuda, and K. Asaka, "A new type of fish-like underwater microrobot," *IEEE/ASME Transactions on Mechatronics*, vol. 8, no. 1, pp. 136–141, 2003.

[2] H. Yin, S. Guo, A. Li, L. Shi, and M. Liu, "A deep reinforcement learning-based decentralized hierarchical motion control strategy for multiple amphibious spherical robot systems with tilting thrusters," *IEEE Sensors Journal*, vol. 24, no. 1, pp. 769–779, 2024.

[3] A. Li, S. Guo, and C. Li, "An improved motion strategy with uncertainty perception for the underwater robot based on thrust allocation model," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 64–71, 2025.
[4] C. Li and S. Guo, "Characteristic evaluation via multi-sensor information fusion strategy for spherical underwater robots," *Information Fusion*, vol. 95, pp. 199–214, 2023.
[5] F. Zocco, T.-C. Lin, C.-I. Huang, H.-C. Wang, M. O. Khyam, and M. Van, "Towards more efficient efficientdets and real-time marine debris detection," *IEEE Robotics and Automation Letters*, vol. 8, no. 4, pp. 2134–2141, 2023.
[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
[7] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024.
[8] O. A. Aguirre-Castro, E. E. García-Guerrero, O. R. López-Bonilla, E. Tlelo-Cuautle, D. López-Mancilla, J. R. Cárdenas-Valdez, J. E. Olguín-Tiznado, and E. Inzunza-González, "Evaluation of underwater image enhancement algorithms based on retinex and its implementation on embedded systems," *Neurocomputing*, vol. 494, pp. 148–159, 2022.
[9] F. Guth, S. Coste, V. De Bortoli, and S. Mallat, "Wavelet score-based generative modeling," *NeurIPS*, vol. 35, pp. 478–491, 2022.
[10] Z. Li, Z.-S. Kuang, Z.-L. Zhu, H.-P. Wang, and X.-L. Shao, "Wavelet-based texture reformation network for image super-resolution," *IEEE Transactions on Image Processing*, vol. 31, pp. 2647–2660, 2022.
[11] C. Liu, H. Li, S. Wang, M. Zhu, D. Wang, X. Fan, and Z. Wang, "A dataset and benchmark of underwater object detection for robot picking," in *Proceedings of the 2021 IEEE International Conference on Multimedia (ICME)*, 2021, pp. 1–6.
[12] C. Fu, R. Liu, X. Fan, P. Chen, H. Fu, W. Yuan, M. Zhu, and Z. Luo, "Rethinking general underwater object detection: Datasets, challenges, and solutions," *Neurocomputing*, vol. 517, pp. 243–256, 2023.
[13] J. Hong, M. Fulton, and J. Sattar, "Trashcan: A semantically-segmented dataset towards visual detection of marine debris," *CoRR*, vol. abs/2007.08097, 2020.
[14] M. Pedersen, J. Haurum, R. Gade, T. Moeslund, and N. Madsen, "Detection of marine animals in a new underwater dataset with varying visibility," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 18–26.
[15] J. Zhou, J. Sun, C. Li, Q. Jiang, M. Zhou, K.-M. Lam, W. Zhang, and X. Fu, "Hclr-net: hybrid contrastive learning regularization with locally randomized perturbation for underwater image enhancement," *International Journal Of Computer Vision*, vol. 132, no. 10, pp. 4132–4156, 2024.
[16] R. Liu, Z. Jiang, S. Yang, and X. Fan, "Twin adversarial contrastive learning for underwater image enhancement and beyond," *IEEE Transactions on Image Processing*, vol. 31, pp. 4922–4936, 2022.
[17] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3227–3234, 2020.
[18] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo, "Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4861–4875, 2020.
[19] L. Dai, H. Liu, P. Song, and M. Liu, "A gated cross-domain collaborative network for underwater object detection," *Pattern Recognition*, vol. 149, p. 110222, 2024.
[20] W.-H. Lin, J.-X. Zhong, S. Liu, T. Li, and G. Li, "Roimix: proposal-fusion among multiple images for underwater object detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2588–2592.
[21] Y. Chen, P. Song, H. Liu, L. Dai, X. Zhang, R. Ding, and S. Li, "Achieving domain generalization for underwater object detection by domain mixup and contrastive learning," *Neurocomputing*, vol. 528, pp. 20–34, 2023.
[22] J. Zhou, Z. He, K.-M. Lam, Y. Wang, W. Zhang, C. Guo, and C. Li, "Amsp-uod: When vortex convolution and stochastic perturbation meet underwater object detection," in *AAAI*, vol. 38, no. 7, 2024, pp. 7659–7667.
[23] T. Wang, C. Lu, Y. Sun, M. Yang, C. Liu, and C. Ou, "Automatic ecg classification using continuous wavelet transform and convolutional neural network," *Entropy*, vol. 23, no. 1, p. 119, 2021.
[24] T. Wang, Y. Xiao, Y. Cai, G. Gao, X. Jin, L. Wang, and H. Lai, "Ufsrnet: U-shaped face super-resolution reconstruction network based on wavelet transform," *Multimedia Tools and Applications*, vol. 83, no. 25, pp. 67 231–67 249, 2024.
[25] C. Korkmaz and A. M. Tekalp, "Training transformer models by wavelet losses improves quantitative and visual performance in single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 6661–6670.
[26] T. Wu, W. Li, S. Jia, Y. Dong, and T. Zeng, "Deep multi-level wavelet-cnn denoiser prior for restoring blurred image with cauchy noise," *IEEE Signal Processing Letters*, vol. 27, pp. 1635–1639, 2020.
[27] G. Xu, W. Liao, X. Zhang, C. Li, X. He, and X. Wu, "Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation," *Pattern Recognition*, vol. 143, p. 109819, 2023.
[28] S. E. Finder, R. Amoyal, E. Treister, and O. Freifeld, "Wavelet convolutions for large receptive fields," in *Proceedings of the European Conference on Computer Vision*, 2024.
[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.
[30] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6154–6162.
[31] S. Qiao, L.-C. Chen, and A. Yuille, "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution," *arXiv preprint arXiv:2006.02334*, 2020.
[32] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7464–7475.
[33] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023.
[34] P. Song, P. Li, L. Dai, T. Wang, and Z. Chen, "Boosting r-cnn: Reweighting r-cnn samples by rpn's error for underwater object detection," *Neurocomputing*, vol. 530, pp. 150–164, 2023.
[35] X. Liang and P. Song, "Excavating roi attention for underwater object detection," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2022.
[36] L. Dai, H. Liu, P. Song, H. Tang, R. Ding, and S. Li, "Edge-guided representation learning for underwater object detection," *CAAI Transactions on Intelligence Technology*, vol. 9, no. 5, pp. 1078–1091, 2024.