Spatial-Frequency Spiking Neural Network for Underwater Object Detection

Long Chen[1,2], Wei Miao[1,3], Xin Gao[1], Yunzhi Zhuge[4], Hongming Xu[5], Yaxin Li[1], Qi Xu[1]*

[1]School of Computer Science and Technology, Dalian University of Technology
[2]Department of Medical Physics and Biomedical Engineering, University College London
[3]Faculty of Information Technology, University of Jyväskylä
[4]School of Information and Communication Engineering, Dalian University of Technology
[5]Faculty of Medicine, Dalian University of Technology
chenlongcv@gmail.com, weimiao@jyu.fi, {gaoxin, zgyz, yaxin.li}@mail.dlut.edu.cn, {mxu, xuqi}@dlut.edu.cn

Abstract

Underwater object detection presents significant challenges due to the unique visual degradations in underwater environments, such as low contrast, poor visibility, and blurry object boundaries. While ANNs have achieved impressive detection accuracy, their high computational cost and power consumption limit their deployment in resource-constrained underwater platforms. In this work, we propose a Spatial-Frequency Spiking Neural Network (SFSNN) that combines the energy-efficient and event-driven nature of Spiking Neural Networks (SNNs) with the discriminative power of spatial-frequency analysis. SFSNN introduces a novel spatial-frequency spiking module that integrates spatial and frequency-domain representations, enhancing edge and texture features crucial for object detection in murky waters. Furthermore, we adapt the YOLOX architecture into a spike-based detector via ANN-to-SNN conversion using signed spiking neurons. Extensive experiments on the RUOD dataset demonstrate that SFSNN achieves superior performance over both SNN- and ANN-based detection models, offering a compelling solution for low-power underwater object detection.

Introduction

Real-time underwater object detection (UOD) (Fu et al. 2023b; Jian et al. 2024) requires low-power, efficient models, especially for applications needing fast response and limited computational resources. While Artificial Neural Networks (ANNs) dominate this field due to their strong feature extraction and support from powerful hardware, they consume high energy and depend on high-performance GPUs/TPUs. In contrast, Spiking Neural Networks (SNNs) (Tavanaei et al. 2019; Yamazaki et al. 2022) offer distinct advantages in energy efficiency and event-driven processing, making them well-suited for certain object detection tasks. Since SNNs operate based on event-driven computation, they are ideal for low-power edge devices such as underwater robots, drones, and neuromorphic chips.

Few studies have focused on developing spiking neural networks for underwater object detection. While some works (Zhang et al. 2022; Sudevan et al. 2024) have introduced basic SNNs for underwater object detection, the research area remains largely unexplored. Nevertheless, several studies (Su et al. 2023; Kim et al. 2020; Miao et al. 2025) have investigated the use of SNNs for generic object detection, demonstrating their feasibility for such tasks. SNN-based object detection methods can be categorized into two main types. The first type involves direct SNN-based object detection, where SNNs are applied directly to the task, using spiking neurons to process image data and generate detection outputs (Su et al. 2023). These approaches are energy-efficient due to event-driven computation and are well-suited for neuromorphic hardware. However, they suffer from limited accuracy compared to ANNs and face training challenges due to the non-differentiability of spiking neurons. The second type is ANN-to-SNN conversion, where pre-trained ANNs are converted into SNNs by approximating activation functions with spiking neurons (Bu et al. 2022; Qu et al. 2024). These methods leverage the performance of pre-trained ANNs and are easier to implement than training SNNs from scratch, but conversion often leads to performance degradation.

Furthermore, directly applying SNN-based generic object detection frameworks to underwater object detection often leads to suboptimal performance due to the challenges unique to the underwater environment (Chen et al. 2024; Liu et al. 2021). Underwater images suffer from attenuation and scattering, resulting in poor contrast and low visibility. The loss of detail, such as blur, diminishes fine features, particularly for objects with indistinct boundaries or low contrast against the background. Blurred images, with the loss of sharp edges, make it difficult for detection algorithms to define object boundaries, ultimately reducing accuracy.

In this work, we explore effective SNN architectures for underwater object detection, aiming to achieve high detection performance with low computational and energy consumptions. We present a Spatial-Frequency Spiking Neural Network (SFSNN) that integrates both spatial and frequency information. The spatial information helps capture the general position and structure of objects, while the frequency information, particularly from high-frequency subbands, enhances features like object edges and textures. This complementary integration improves the overall visibility of objects in challenging underwater environments. Specifically, we incorporate the spike-driven paradigm into the ANN-based de-

*Qi Xu is the corresponding author.

tector YOLOX (Ge et al. 2021) to enhance energy efficiency in underwater object detection. Additionally, we design a spatial-frequency spiking module to extract features across both spatial and frequency domains. Wavelet Transform is employed to transform RGB images into the frequency domain, where the high-frequency subbands (LH, HL, HH) emphasize edge and texture features. By preserving fine-grained details in these subbands, Wavelet Transform significantly improves the model's ability to detect objects in murky or turbid waters where object boundaries are unclear. The main contributions are summarized as follow:

- The first spiking neural network specifically designed for underwater object detection, integrating the spike-driven paradigm into ANNs for energy-efficient detection while tackling the challenges of low visibility and poor contrast in underwater environments.
- A spatial-frequency spiking module is designed to fuse information from both spatial and frequency domains. In the frequency domain, high-frequency components enhance edges and fine details, effectively addressing blurry object boundaries in underwater images.
- The proposed SNN outperforms both SNN- and ANN-based frameworks on the diverse RUOD dataset, demonstrating strong generalization across various underwater environments.

Related Work

Underwater Object Detection

Deep learning-based approaches have become the leading paradigm in underwater object detection due to their strong feature representation capabilities. Early approaches primarily relied on convolutional neural networks (CNNs) for feature extraction, classification, and localization (Chen et al. 2020; Lin et al. 2020). Both two-stage detectors, such as Fast R-CNN (Girshick 2015) and Faster R-CNN (Ren et al. 2016), and one-stage detectors, such as YOLO (Redmon et al. 2016) and SSD (Liu et al. 2016), have been widely employed in this domain (Li et al. 2015; Li, Tang, and Gao 2017; Yang et al. 2021). Two-stage detectors first generate region proposals and then refine them in a second stage for classification and localization. This approach enhances accuracy but comes at the cost of increased computational complexity. In contrast, one-stage detectors integrate classification and localization in a single step, offering a more efficient architecture with faster inference. However, they may exhibit lower accuracy, particularly for detecting small or occluded objects in complex underwater environments.

Recent advancements have introduced transformer-based architectures (Yu et al. 2021), such as Vision Transformers (ViTs) (Chen et al. 2023) and Swin Transformers (Sun et al. 2022), which have demonstrated exceptional performance in handling complex underwater scenes. While transformer-based detection frameworks offer superior accuracy, they are computationally intensive and demand significant power. To improve energy efficiency, spiking neural network have begun to emerge in generic object detection. Various Spiking YOLO variants (Miao et al. 2025; Qu et al. 2024; Su et al. 2023; Bu et al. 2022) have been developed to enhance power efficiency; however, these SNN-based methods still lag behind ANN-based approaches in performance on large-scale generic object detection datasets such as MS COCO (Lin et al. 2014) and PASCAL VOC (Everingham et al. 2010). To date, only the work of (Sudevan et al. 2024) has explored the feasibility of applying Spiking-YOLO (Kim et al. 2020), a generic object detection framework, to the UOD task. However, no practical implementations of SNNs specifically designed for UOD exist, highlighting a significant gap in this research area.

Frequency Information in Image Processing

In image processing, frequency information describes the distribution of intensity variations across an image, reflecting how pixel values change over space (Xu et al. 2020). This information helps differentiate between smooth regions, edges, textures, and noise. Frequency analysis is commonly performed using mathematical transforms such as the Fourier Transform (FT) (Chi, Jiang, and Mu 2020; Yu et al. 2023) and Wavelet Transform (WT) (Huang et al. 2017; Li et al. 2022). While the FT provides a global frequency representation, the WT offers multi-resolution analysis, which is particularly effective for images. Specifically, the wavelet transform delivers a multi-resolution representation that captures both low-frequency components—representing gradual intensity changes (e.g., uniform backgrounds and soft gradients) and conveying overall shape and structure—and high-frequency components, which reflect rapid changes (e.g., sharp edges, fine textures, and noise) and highlight object boundaries and intricate details.

The combination of CNNs and wavelet transforms has gained widespread use in low-level image processing tasks, such as image restoration (Liu et al. 2018), super-resolution (Huang et al. 2017), and compression (Ma et al. 2019). WT's robustness to noise enhances denoising and edge detection, while its ability to provide simultaneous detailed and coarse representations proves invaluable. These studies effectively integrate CNN's feature learning capabilities with the multi-resolution frequency analysis offered by wavelet techniques. In object detection (Strickland and Hahn 1997; Su et al. 2024), WT supports robust feature extraction by capturing both spatial and frequency characteristics of objects, aiding in distinguishing objects from background noise and clutter. More importantly, wavelet-based methods enhance edges and contours, playing a crucial role in detecting object boundaries within images with complex textures.

Method

In this section, we first present the overall SFSNN framework, followed by a detailed description of the spatial-frequency spiking module, which combines spatial and frequency information to enhance detection accuracy. Finally, we introduce the SNN detection module, derived through ANN-to-SNN conversion to improve energy efficiency.

The Overview of SFSNN

Figure 1 illustrates the SFSNN architecture, which comprises a spatial-frequency spiking module for extracting spa-

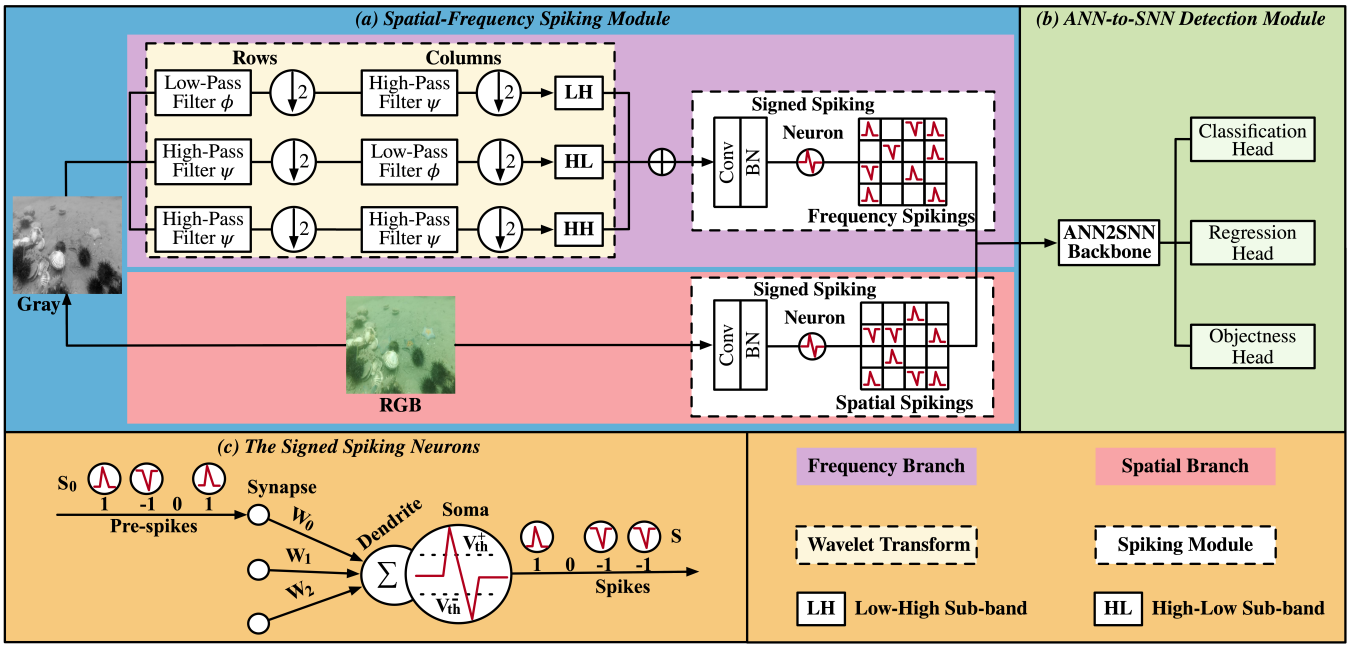

Figure 1: The overview of the proposed SFSNN, which includes: (a) a Spatial-Frequency Spiking Module that extracts both spatial and frequency features, and (b) an ANN-to-SNN detection module designed for energy efficiency. The Spiking Module comprises two branches: the spatial branch captures global structure and color information, while the frequency branch uses Wavelet Transform to extract high-frequency components (LH, HL, HH), enhancing object boundaries and textures. The ANN-to-SNN detection module is adapted from the ANN-based YOLOX framework by replacing standard activations with (c) signed spiking neurons, which provide better energy efficiency and support real-time processing.

tial and frequency features, and an ANN-to-SNN detection module for energy-efficient object classification and localization. In underwater environments, object boundaries often become indistinct due to the adverse physical effects such as light absorption and scattering, which obscure edges and textures. To enhance edge and contour information, we design the spatial-frequency spiking module with two parallel branches: a frequency branch and a spatial branch. The frequency branch employs the Discrete Wavelet Transform (DWT) to decompose the input image $I_{RGB} \in \mathbb{R}^{H\times W\times 3}$ into frequency subbands: LL, LH, HL, and HH. Among these, the LL, LH, and HL subbands emphasize edges and textures, contributing to more accurate object localization. In parallel, the spatial branch utilizes raw RGB image to retain global structural and color information, facilitating object discrimination based on spectral characteristics.

The ANN-to-SNN detection module is adapted from the ANN-based YOLOX framework (Ge et al. 2021), in which traditional activation functions are replaced with signed spiking neurons, as shown in Figure 1 (c). This ANN-to-SNN conversion offers several advantages, including improved energy efficiency, real-time processing capabilities, and greater biological plausibility.

The Spatial-Frequency Spiking Module

The Spatial-Frequency Spiking Module consist of two parallel branches: a frequency branch that captures fine-grained edge and texture details in the frequency domain, and a spatial branch that captures global structure and color information in spatial domain.

In the frequency branch, DWT decomposes an image into low-frequency and high-frequency components using a wavelet function $\psi(x)$ and a scaling function $\phi(x)$. Since DWT operates on single-channel data, the RGB image $I_{RGB}(x, y)$ is first converted to a grayscale image $I_G(x, y)$. The transform produce four subbands:

- **LL (Approximation)**: Low-pass filtering in both directions (captures smooth images).
- **LH (Horizontal details)**: Low-pass filtering along rows and high-pass along columns (detects horizontal edges).
- **HL (Vertical details)**: High-pass filtering along rows and low-pass along columns (detects vertical edges).
- **HH (Diagonal details)**: High-pass filtering in both directions (captures diagonal edges and fine textures).

Each subband is computed through filtering (convolution with $\phi(x)$ or $\psi(x)$) followed by downsampling (by a factor of 2). The decomposition can be expressed as:

$$LL(x,y) = \sum_m \sum_n I_G(m,n) \cdot \phi(x-2m) \cdot \phi(y-2n) \quad (1)$$

$$LH(x,y) = \sum_m \sum_n I_G(m,n) \cdot \phi(x-2m) \cdot \psi(y-2n) \quad (2)$$

$$HL(x,y) = \sum_m \sum_n I_G(m,n) \cdot \psi(x-2m) \cdot \phi(y-2n) \quad (3)$$

$$HH(x,y) = \sum_m \sum_n I_G(m,n) \cdot \psi(x-2m) \cdot \psi(y-2n) \quad (4)$$

Here, $\phi(x)$ is the scaling function (low-pass filter), $\psi(x)$ is the wavelet function (high-pass filter). (m, n) are the spatial coordinates in the image, and the factor 2 in $2m, 2n$ represents downsampling by 2 after filtering.

The LL subband captures the coarse, global structure of the image $I_G(x, y)$, while the LH, HL, and HH subbands preserve rich edge and texture information—crucial for enhancing object localization, particularly in visually degraded underwater environments. Therefore, we use only the high frequency components ($LH(x, y)$, $HL(x, y)$, and $HH(x, y)$) as the input $I_f(x, y)$, in order to emphasize edges and textures. The three subbands are concatenated and subsequently processed by a spiking module, which consists of a convolutional layer $Conv(\cdot)$, a batch normalization layer $BN(\cdot)$, and a spiking neuron layer $S(\cdot)$, to extract deep frequency features. The overall output can be expressed as:

$$F_f(x,y) = S(\text{BN}(\text{Conv}(I_f(x,y)))) \quad (5)$$

Here, $S(\cdot)$ denotes the signed spiking neuron, which acts as a non-linear activation function. Its behavior is detailed in Section 3.3. Unlike conventional activation functions, spiking neurons emit discrete spikes only when the membrane potential exceeds a certain predefined threshold, making them inherently event-driven. This behavior drastically reduces the number of operations and energy consumption.

In the spatial branch, the original RGB image $I_{RGB}(x, y)$ is used as input, as it preserves rich color and structural information that is beneficial for object detection. This color information complements the edge and texture cues captured by the frequency branch. To extract meaningful deep spatial features, we employ a spiking module similar to that used in the frequency branch, consisting of a convolutional layer, batch normalization, and a spiking neuron layer. The output of the spatial branch is given by:

$$F_s(x,y) = S(\text{BN}(\text{Conv}(I_{RGB}(x,y)))) \quad (6)$$

Here, $F_s(x, y)$ denotes the extracted spatial features, and $S(\cdot)$ represents the spiking neuron function, which introduces non-linearity while reducing computational cost through event-driven processing.

The ANN-to-SNN Detection Module

The ANN-to-SNN detection module consists of an SNN-based backbone for feature extraction and three detection heads for object detection. It is adapted from the ANN-based YOLOX framework (Ge et al. 2021) through an ANN-to-SNN conversion process. During this conversion, conventional activation functions in YOLOX are replaced with signed spiking neurons (Kim et al. 2020), which provide notable advantages such as improved energy efficiency and real-time processing capability.

The signed spiking neuron employed in this work produces outputs from the set $\{-1, 0, +1\}$, enabling both excitatory $(+1)$ and inhibitory (-1) spikes. Unlike conventional spiking models such as Integrate-and-Fire (IF) or Leaky Integrate-and-Fire (LIF), which emit only positive spikes, this design allows for a more accurate approximation of both positive and negative activations commonly found in ANNs, such as those produced by Leaky ReLU and tanh functions.

In signed spiking neurons, each neuron i maintains a membrane potential $V_i(t)$, which evolves over time t. The membrane potential is updated at each time step:

$$V_i(t) = V_i(t-1) + z_i(t) \quad (7)$$

where $z_i(t)$ represents the membrane input (or synaptic input current) received by neuron i at time t. For neuron i in layer l, the membrane input is computed as:

$$z_i^l(t) = \sum_{j \in \mathcal{M}^{l-1}} W_{ij}^l \cdot S_j^{l-1}(t) + b_i^l \quad (8)$$

Where $\mathcal{M}^{l-1}$ denotes the set of neurons in the previous layer $l-1$, W_{ij}^l is the synaptic weight from neuron j to neuron i, $S_j^{l-1}(t) \in \{-1, 0, +1\}$ is the ternary spike emitted by presynaptic neuron j at time t, b_i^l is the bias term associated with neuron i in layer l.

After updating $V_i(t)$, the spiking function $S(t)$ determines whether the neuron emits a spike. This function produces a discrete output based on the current membrane potential and the spiking threshold:

$$S(t) = \begin{cases} +1, & \text{if } V_i(t) \geq V_{\text{th}}^+ \quad \text{(positive spike)} \\ -1, & \text{if } V_i(t) \leq V_{\text{th}}^- \quad \text{(negative spike)} \\ 0, & \text{otherwise (no spike)} \end{cases} \quad (9)$$

Here, V_{th}^+ and V_{th}^- denote the positive and negative thresholds, respectively. A spike $S(t) \neq 0$ is generated only when the membrane potential exceeds the thresholds. If $V(t) \in (V_{\text{th}}^-, V_{\text{th}}^+)$, no spike is emitted, and the neuron remains inactive, consuming no energy at that time step. Onece a spike is fired, the membrane potential undergoes a soft reset to preserve residual information for future computation.

$$V(t) = \begin{cases} V(t) - V_{\text{th}}^+ & \text{if } S(t) = +1 \\ V(t) + V_{\text{th}}^- & \text{if } S(t) = -1 \\ V(t) & \text{if } S(t) = 0 \end{cases} \quad (10)$$

Experimental Results

In this section, we first introduce the experimental setup, then compare the proposed SFSNN framework with SOAT methods, and finally perform ablation studies to evaluate the impact of each component on the overall framework.

Experimental Settings

Datasets. To evaluate the proposed SFSNN framework, we utilize the RUOD dataset (Fu et al. 2023b), a comprehensive underwater object detection benchmark featuring a wide range of object types and challenging visual conditions. The dataset comprises 9,800 images for training and 4,200 images for testing, encompassing ten distinct underwater object categories.

Evaluation Metrics. We evaluate model performance using the COCO metrics, which include mean Average Precision (mAP) across a range of Intersection over Union (IoU)

	Methods	Models	Backbones	Params	FLOPs	mAP	$AP_{0.50}$	$AP_{0.75}$	AP_s	AP_m	AP_l
Non-Spiking	Generic	RepPoints	ResNet101	55.82M	256.00G	53.2	82.2	60.1	28.2	44.9	57.8
		FoveaBox	ResNet101	56.68M	268.29G	44.8	80.2	45.2	18.0	37.5	49.1
		ATSS	ResNet101	51.13M	267.26G	54.0	80.3	60.2	18.0	40.0	59.5
		DetectoRS	DResNet50	123.23M	90.05G	53.3	84.1	58.7	**30.8**	46.6	57.8
		YOLOv10	CSPNet	24.40M	120.30G	55.5	84.7	62.5	21.9	**47.0**	60.5
	Underewater	BoostRCNN	ResNet50	45.95M	54.71G	53.9	80.6	59.5	11.6	39.0	59.3
		RFTM	ResNet50	75.58M	91.06G	53.3	80.2	57.7	11.8	39.2	59.3
		ERLNet	SiEdgeR50	45.95M	54.71G	54.8	83.1	60.9	14.7	41.4	59.8
		GCCNet	SwinFT	38.31M	78.93G	56.1	83.2	60.5	11.7	41.9	62.1
		DJLNet	ResNet50	58.48M	69.51G	57.5	83.7	62.5	15.5	41.8	63.1
Spiking	SNN	Spiking-YOLO	TinyYOLO	23.1M	136.9G	49.8	80.7	55.1	17.6	42.5	54.4
		EMS-YOLO	EMSResNet34	33.86M	37.00G	52.0	82.9	58.4	19.3	44.2	57.4
		SpikingYOLOX	SNNCSPNet-L	49.53M	151.69G	57.0	84.2	61.2	11.0	41.2	63.0
	Ours	SF-SNN-Tiny	WaveletSNN-T	**4.42M**	**20.65G**	55.5	83.8	60.0	11.9	40.7	61.1
		SF-SNN-Small	WaveletSNN-S	7.83M	33.74G	56.4	84.2	60.7	11.3	40.8	62.2
		SF-SNN-Large	WaveletSNN-L	49.66M	167.56G	**58.3**	**85.2**	**63.1**	13.4	42.7	**64.3**

Table 1: The quantitative performance of representative detection frameworks on the RUOD dataset. The bold text represents the best performance, while the red text indicates the second-best.

thresholds ($mAP@[0.5:0.05:0.95]$), as well as precision at fixed thresholds ($AP_{0.50}$ and $AP_{0.75}$). To analyze performance across different object scales, we also report Average Precision for small (AP_S), medium (AP_M), and large (AP_L) objects. In addition to accuracy, we assess model efficiency by reporting computational cost in FLOPs (floating point operations) and model complexity in terms of parameter count (Params).

Implementation Details. The SFSNN framework is implemented in PyTorch, utilizing SpikingJelly (Fang et al. 2023) to construct spiking neuron modules. We design three scaled model variants following the YOLOX architecture (Ge et al. 2021): Tiny (**SFSNN-T**), Small (**SFSNN-S**), and Large (**SFSNN-L**). In the frequency branch, the Discrete Wavelet Transform (DWT) is applied to decompose input images into subbands, which are then resized to the original RGB image dimensions before being concatenated. Training is conducted using the Adam optimizer, with an initial learning rate set to 0.01. The learning rate and number of training epochs are adjusted according to the size of each model.

Comparison with State-of-the-Art Methods

We conduct a comprehensive evaluation of our proposed SFSNN against a range of state-of-the-art detection models. These include general-purpose detectors such as YOLOv10 (Wang et al. 2024a), RepPoints (Yang et al. 2019), FoveaBox (Kong et al. 2020), ATSS (Zhang et al. 2020), and DetectoRS (Qiao, Chen, and Yuille 2021), as well as leading underwater detection models like DJLNet (Wang et al. 2024b), GCCNet (Dai et al. 2024), ERLNet (Dai et al. 2023), RFTM (Fu et al. 2023a), and BoostRCNN (Song et al. 2023)). We also compare with top-performing SNN-based models, including SpikingYOLOX (Su et al. 2023), EMS-YOLO (Su et al. 2023), and Spiking-YOLO (Kim et al. 2020)).

As shown in Table 1, underwater-specific detectors such as DJNet and GCCNet outperforms generic detectors, owing to their tailored adaptations for the challenge of underwater environment. In contrast, SNN-based models like Spiking-YOLO and EMS-YOLO lag behind ANN-based detectors in accuracy. This may be attributed to the limited representational capacity of SNNs, which stems from their binary, event-driven signaling. Nevertheless, SNNs provide notable advantages in computational efficiency due to their sparse, asynchronous processing and low-power operation. To improve detection accuracy while retaining efficiency, our proposed SFSNN framework integrates frequency-domain features extracted using the Wavelet Transform. These features help highlight object boundaries and capture localized variations, enhancing object localization performance. Our SFSNN-L achieves SOAT performance, while lightweight SFSNN-T and SFSNN-S models offer a strong balance between accuracy and efficiency.

Models	Frequency	Spatial	mAP	$AP_{0.50}$	$AP_{0.75}$
SFSNN-T			52.6	82.0	56.2
		✓	53.2	82.5	57.5
	✓		51.0	80.9	54.6
	✓	✓	**55.5**	**83.8**	**60.0**
SFSNN-S			54.0	83.0	57.7
		✓	54.9	83.6	58.3
	✓		52.4	81.1	55.8
	✓	✓	**56.4**	**84.2**	**60.7**
SFSNN-L			57.0	84.2	61.2
		✓	57.6	84.7	62.0
	✓		56.2	83.8	60.3
	✓	✓	**58.3**	**85.2**	**63.1**

Table 2: Ablation Study Results of the SF Spiking Module.

Ablation Study of the SF Spiking Module

Ablation experiments on the Spatial-Frequency (SF) Spiking Module are conducted across all three model scales. For each scale, we evalute three configurations: SFSNN with

Category	(a) Baseline-L vs. SFSNN-L										(b) Baseline-S vs. SFSNN-S										(c) Baseline-T vs. SFSNN-T									
	XS	S	M	L	XL	XS	S	M	L	XL	XS	S	M	L	XL	XS	S	M	L	XL	XS	S	M	L	XL	XS	S	M	L	XL
holothurian	30.6	57.9	76.9	81.3	72.8	25.7	58.3	77.2	83.4	75.2	18.7	50.0	75.7	76.2	61.7	24.6	57.3	76.5	81.2	72.1	21.5	51.5	73.8	75.4	62.2	25.6	58.2	76.3	80.0	70.7
echinus	56.9	74.6	82.0	80.6	75.7	50.3	70.5	81.1	81.3	80.9	53.8	75.0	82.0	80.2	69.0	49.2	70.6	80.1	80.3	76.6	55.0	74.5	81.9	79.5	67.9	51.2	72.9	79.3	80.4	76.4
scallop	22.6	69.9	72.3	85.1	89.1	21.4	65.8	71.1	83.5	88.5	16.4	64.2	64.2	78.4	84.7	20.8	66.9	69.6	82.6	87.3	18.9	64.0	60.5	79.6	83.5	19.0	64.2	66.9	80.5	83.4
starfish	60.9	79.6	84.0	86.4	92.5	59.4	76.3	82.1	86.1	91.9	60.4	79.8	82.7	82.8	91.1	60.0	76.1	81.5	84.4	90.1	60.4	78.0	81.8	80.2	88.7	56.9	77.3	82.0	83.7	89.9
fish	4.4	31.7	60.9	77.2	86.9	4.2	30.7	61.2	78.7	85.6	3.8	28.8	55.1	73.6	83.8	3.6	30.0	60.7	80.0	84.8	4.6	29.8	56.5	73.3	84.0	4.2	29.4	60.0	77.9	84.2
corals	14.6	29.9	62.6	78.8	69.1	14.7	30.0	61.3	78.3	69.0	10.9	24.1	61.7	78.5	65.8	13.1	26.8	60.3	76.4	68.5	11.8	23.6	59.9	78.1	65.5	12.7	26.6	60.3	74.8	62.8
diver	75.4	83.3	91.0	95.2	92.9	78.3	86.2	91.2	96.0	94.5	72.0	80.2	87.3	91.3	91.9	70.6	83.6	89.3	93.7	92.4	71.2	80.8	88.2	90.9	91.7	69.9	82.6	89.0	91.4	91.4
cuttlefish	88.7	96.4	95.0	95.9	96.2	87.8	95.0	94.8	96.0	96.1	83.6	93.6	92.7	94.5	94.5	88.0	95.2	94.7	95.2	95.4	86.2	94.2	93.3	95.4	95.0	88.2	95.4	94.7	95.3	94.3
turtle	88.8	94.7	94.9	93.6	93.0	84.1	96.6	95.0	93.6	92.6	85.6	94.4	95.8	94.0	92.8	84.1	95.5	94.7	94.0	93.0	80.7	91.8	94.3	93.9	89.9	83.3	96.0	94.8	94.9	90.7
jellyfish	19.3	52.4	80.9	93.0	92.6	28.8	55.0	83.6	92.0	93.5	13.8	44.6	81.5	90.2	90.1	24.8	50.2	80.4	91.1	91.3	17.2	38.8	79.0	90.5	91.2	24.4	50.7	80.3	89.4	90.3

Table 3: The mean Average Precision comparison between the baseline model (without SF Spiking Module) and our SF-SNNs across different object sizes on the RUOD dataset. The objects are classified into five groups based on size: XS (bottom 10%)=extra-small; S (next 20%)=small; M (next 40%)=medium; L (next 20%)=large; XL (next 10%)=extra-large.

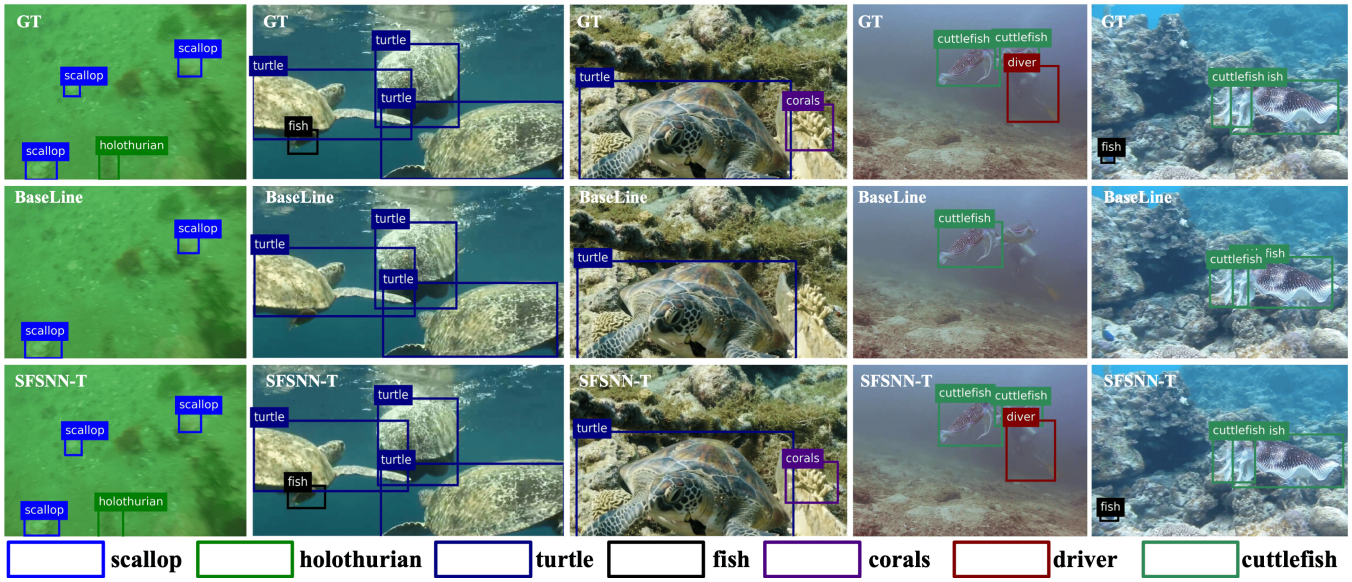

Figure 2: Qualitative comparisons between the baseline and SFSNN-T. The baseline model often struggles to detect small and blurry objects, whereas SFSNN-T shows enhanced capability in identifying them.

only the frequency branch, SFSNN with only the spatial branch, and SFSNN with both branches integrated. Additionally, a model without the SF Spiking Module is included as the **BaseLine** for comparison.

Table 2 presents the performance of each configuration. Two key findings emerge: (1) The spatial branch has a stronger impact because it keeps both color and structural cues from the RGB input—key information for object detection. The frequency branch, which removes color and keeps only intensity changes, offers more limited scene understanding. (2) The frequency branch serves as a strong complementary signal. It helps detect fine details like edges, textures, and small changes that might be missed by the spatial branch alone. Table 3 presents the mAP of the baseline and our SFSNN models across different object sizes. The full SFSNN outperforms the baseline model in almost all size categories, demonstrating the effectiveness of the SF Spiking Module. Qualitative results in Figure 2 further reveal that the baseline often misses small objects, whereas SFSNN-T detects them more accurately. This improvement comes from the SF Spiking Module, which captures high-frequency components such as edges and local intensity variations that are important for detecting small objects.

Analysis of Wavelet Subbands

In the frequency branch, the Discrete Wavelet Transform (DWT) decomposes the input image into four subbands: LL, LH, HL, and HH. To better understand the frequency branch's role, we analyze the impact of different subband combinations while keeping the spatial branch unchanged.

Table 4 shows the performance of SFSNN using different wavelet subbands. The results clearly show that the high-frequency subbands contribute much more to detection accuracy than the low-frequency subband. These high-frequency subbands capture horizontal, vertical, and diagonal edge details, which are crucial for detecting object boundaries and fine textures. In contrast, the low-frequency subband mostly retains smoothed, low-frequency information and exhibits relatively weak performance. In some cases, it even negatively impacts detection. These findings

Models	LL	LH	HL	HH	mAP	$AP_{0.50}$	$AP_{0.75}$
SFSNN-T		✓	✓		54.9	83.1	58.8
	✓	✓	✓		54.6	82.9	58.5
		✓	✓	✓	**55.5**	**83.8**	**60.0**
	✓	✓	✓	✓	55.3	83.7	59.9
SFSNN-S		✓	✓		55.7	84.0	59.9
	✓	✓	✓		55.6	83.8	59.0
		✓	✓	✓	**56.4**	**84.2**	**60.7**
	✓	✓	✓	✓	56.2	83.9	60.2
SFSNN-L		✓	✓		58.0	84.9	62.6
	✓	✓	✓		57.8	84.8	62.4
		✓	✓	✓	**58.3**	**85.2**	**63.1**
	✓	✓	✓	✓	58.2	85.0	62.9

Table 4: The Impact of Frequency Subbands.

Models	Neurons	mAP(0.5:0.95)	$AP_{0.50}$	$AP_{0.75}$
SFSNN-T	IF	54.3	83.0	58.2
	Signed	**55.0**	**83.2**	**58.9**
SFSNN-S	IF	55.3	83.6	59.0
	Signed	**56.4**	**84.2**	**60.7**
SFSNN-L	IF	57.2	84.2	61.5
	Signed	**58.3**	**85.2**	**63.1**

Table 5: The performance comparison between SFSNN models using IF neurons and signed spiking neurons.

suggest that the low-frequency subband can be excluded, allowing the model to concentrate on the more informative high-frequency components without harming accuracy.

Ablation Study of the ANN-to-SNN Module

The proposed ANN-to-SNN detection module is adapted from the ANN-based YOLOX framework by replacing standard activation functions with spiking neurons, thereby converting the ANN architecture into a SNN. We evaluate two types of spiking neurons: the Integrate-and-Fire (IF) neuron (Jin et al. 2023) and the signed spiking neuron (Kim et al. 2020). The key distinction between them lies in how they encode spikes: IF neurons produce binary spikes, indicating only the presence (1) or absence (0) of a spike, while signed spiking neurons encode polarity information, allowing spikes to be positive (+1), negative (-1), or absent (0). This capability enables the explicit representation of bidirectional signals, such as excitatory or inhibitory interactions, at the spike level. Table 5 presents the performance comparisons of the SFSNN framework uisng IF and signed spiking neurons. The results consistently show that signed spiking neurons outperform IF neurons. This advantages stems from their ability to generate both positive and negative spikes, providing a richer and more expressive feature representation than the unidirectional signaling of IF neurons.

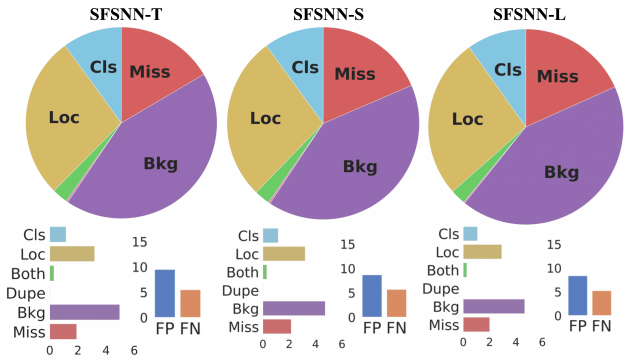

Figure 3: The error types of different SFSNN models. The pie chart represents the relative contribution of each error type, while the bar plots display their absolute contribution.

Conclusion

This work introduces SFSNN, a novel framework designed for underwater object detection. By integrating spatial and frequency-domain features through a spiking spatial-frequency module, our model effectively addresses challenges such as low contrast and blurry boundaries in underwater environments. The adoption of signed spiking neurons further enhances efficiency and performance.

However, SFSNN still has two main limitations. First, it performs noticeably worse when detecting extra-small and small objects compared to large ones. Second, all three SFSNN models show significant background errors (misclassifying background regions as objects) and localization errors (inaccurate object positioning), as illustrated in Figure 3. These issues suggest that the models still struggle with noisy underwater environments. Future work should focus on reducing the impact of such noise, possibly by incorporating image enhancement techniques.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant (No. 62476035, 62206037, and U24B20140), and the Young Elite Scientists Sponsorship Program by CAST under Grant 2024QNRC001.

References

Bu, T.; Ding, J.; Yu, Z.; and Huang, T. 2022. Optimized potential initialization for low-latency spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11–20.

Chen, G.; Mao, Z.; Wang, K.; and Shen, J. 2023. HTDet: A hybrid transformer-based approach for underwater small object detection. *Remote Sensing*, 15(4): 1076.

Chen, L.; Huang, Y.; Dong, J.; Xu, Q.; Kwong, S.; Lu, H.; Lu, H.; and Li, C. 2024. Underwater Object Detection in the Era of Artificial Intelligence: Current, Challenge, and Future. *arXiv preprint arXiv:2410.05577*.

Chen, L.; Liu, Z.; Tong, L.; Jiang, Z.; Wang, S.; Dong, J.; and Zhou, H. 2020. Underwater object detection using Invert Multi-Class Adaboost with deep learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Chi, L.; Jiang, B.; and Mu, Y. 2020. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33: 4479–4488.

Dai, L.; Liu, H.; Song, P.; and Liu, M. 2024. A gated cross-domain collaborative network for underwater object detection. *Pattern Recognition*, 149: 110222.

Dai, L.; Liu, H.; Song, P.; Tang, H.; Ding, R.; and Li, S. 2023. Edge-guided representation learning for underwater object detection. *CAAI Transactions on Intelligence Technology*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88: 303–338.

Fang, W.; Chen, Y.; Ding, J.; Yu, Z.; Masquelier, T.; Chen, D.; Huang, L.; Zhou, H.; Li, G.; and Tian, Y. 2023. Spikingjelly: An open-source machine learning infrastructure platform for spike-based intelligence. *Science Advances*, 9(40): eadi1480.

Fu, C.; Fan, X.; Xiao, J.; Yuan, W.; Liu, R.; and Luo, Z. 2023a. Learning heavily-degraded prior for underwater object detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Fu, C.; Liu, R.; Fan, X.; Chen, P.; Fu, H.; Yuan, W.; Zhu, M.; and Luo, Z. 2023b. Rethinking general underwater object detection: Datasets, challenges, and solutions. *Neurocomputing*, 517: 243–256.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.

Huang, H.; He, R.; Sun, Z.; and Tan, T. 2017. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, 1689–1697.

Jian, M.; Yang, N.; Tao, C.; Zhi, H.; and Luo, H. 2024. Underwater object detection and datasets: a survey. *Intelligent Marine Technology and Systems*, 2(1): 9.

Jin, X.; Zhang, M.; Yan, R.; Pan, G.; and Ma, D. 2023. R-SNN: Region-based spiking neural network for object detection. *IEEE Transactions on Cognitive and Developmental Systems*, 16(3): 810–817.

Kim, S.; Park, S.; Na, B.; and Yoon, S. 2020. Spiking-yolo: spiking neural network for energy-efficient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11270–11277.

Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Li, L.; and Shi, J. 2020. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29: 7389–7398.

Li, X.; Shang, M.; Qin, H.; and Chen, L. 2015. Fast accurate fish detection and recognition of underwater images with fast r-cnn. In *OCEANS 2015-MTS/IEEE Washington*, 1–5. IEEE.

Li, X.; Tang, Y.; and Gao, T. 2017. Deep but lightweight neural networks for fish detection. In *OCEANS 2017-Aberdeen*, 1–5. IEEE.

Li, Z.; Kuang, Z.-S.; Zhu, Z.-L.; Wang, H.-P.; and Shao, X.-L. 2022. Wavelet-based texture reformation network for image super-resolution. *IEEE Transactions on Image Processing*, 31: 2647–2660.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Lin, W.-H.; Zhong, J.-X.; Liu, S.; Li, T.; and Li, G. 2020. Roimix: proposal-fusion among multiple images for underwater object detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2588–2592. IEEE.

Liu, C.; Li, H.; Wang, S.; Zhu, M.; Wang, D.; Fan, X.; and Wang, Z. 2021. A dataset and benchmark of underwater object detection for robot picking. In *2021 IEEE International Conference on Multimedia & Wxpo Workshops (ICMEW)*, 1–6. IEEE.

Liu, P.; Zhang, H.; Zhang, K.; Lin, L.; and Zuo, W. 2018. Multi-level wavelet-CNN for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 773–782.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, 21–37. Springer.

Ma, H.; Liu, D.; Xiong, R.; and Wu, F. 2019. iWave: CNN-based wavelet-like transform for image compression. *IEEE Transactions on Multimedia*, 22(7): 1667–1679.

Miao, W.; Shen, J.; Xu, Q.; Hamalainen, T.; Xu, Y.; and Cong, F. 2025. SpikingYOLOX: Improved YOLOX Object Detection with Fast Fourier Convolution and Spiking Neural Networks. In *The 39th Annual AAAI Conference on Artificial Intelligence*.

Qiao, S.; Chen, L.-C.; and Yuille, A. 2021. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10213–10224.

Qu, J.; Gao, Z.; Zhang, T.; Lu, Y.; Tang, H.; and Qiao, H. 2024. Spiking Neural Network for Ultralow-Latency and High-Accurate Object Detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137–1149.

Song, P.; Li, P.; Dai, L.; Wang, T.; and Chen, Z. 2023. Boosting R-CNN: Reweighting R-CNN Samples by RPN's Error for Underwater Object Detection. *Neurocomputing*.

Strickland, R. N.; and Hahn, H. I. 1997. Wavelet transform methods for object detection and recovery. *IEEE Transactions on Image Processing*, 6(5): 724–735.

Su, Q.; Chou, Y.; Hu, Y.; Li, J.; Mei, S.; Zhang, Z.; and Li, G. 2023. Deep directly-trained spiking neural networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6555–6565.

Su, Y.; Tan, W.; Dong, Y.; Xu, W.; Huang, P.; Zhang, J.; and Zhang, D. 2024. Enhancing concealed object detection in Active Millimeter Wave Images using wavelet transform. *Signal Processing*, 216: 109303.

Sudevan, V.; Zayer, F.; Javed, S.; Karki, H.; De Masi, G.; and Dias, J. 2024. Hybrid-Neuromorphic Approach for Underwater Robotics Applications: A Conceptual Framework. *arXiv preprint arXiv:2411.13962*.

Sun, Y.; Wang, X.; Zheng, Y.; Yao, L.; Qi, S.; Tang, L.; Yi, H.; and Dong, K. 2022. Underwater object detection with swin transformer. In *2022 4th International Conference on Data Intelligence and Security (ICDIS)*, 422–427. IEEE.

Tavanaei, A.; Ghodrati, M.; Kheradpisheh, S. R.; Masquelier, T.; and Maida, A. 2019. Deep learning in spiking neural networks. *Neural Networks*, 111: 47–63.

Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; and Ding, G. 2024a. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*.

Wang, B.; Wang, Z.; Guo, W.; and Wang, Y. 2024b. A dual-branch joint learning network for underwater object detection. *Knowledge-Based Systems*, 293: 111672.

Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y.-K.; and Ren, F. 2020. Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1740–1749.

Yamazaki, K.; Vo-Ho, V.-K.; Bulsara, D.; and Le, N. 2022. Spiking neural networks and their applications: A review. *Brain Sciences*, 12(7): 863.

Yang, H.; Liu, P.; Hu, Y.; and Fu, J. 2021. Research on underwater object recognition based on YOLOv3. *Microsystem Technologies*, 27: 1837–1844.

Yang, Z.; Liu, S.; Hu, H.; Wang, L.; and Lin, S. 2019. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9657–9666.

Yu, H.; Huang, J.; Li, L.; Zhao, F.; et al. 2023. Deep fractional Fourier transform. *Advances in Neural Information Processing Systems*, 36: 72761–72773.

Yu, Y.; Zhao, J.; Gong, Q.; Huang, C.; Zheng, G.; and Ma, J. 2021. Real-time underwater maritime object detection in side-scan sonar images based on transformer-YOLOv5. *Remote Sensing*, 13(18): 3555.

Zhang, F.; Zhong, Y.; Chen, L.; and Wang, Z. 2022. Event-based circular detection for AUV docking based on spiking neural network. *Frontiers in Neurorobotics*, 15: 815144.

Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; and Li, S. Z. 2020. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9759–9768.