



PSWSA-RetinaNet: Polarized Shifted-Window Self-Attention Enhanced RetinaNet for Underwater Object Detection in Complex Marine Environments

¹Somashekar Rangaswamy*, ²Nijaguna Gollara Siddappa

¹Department of Computer Science and Engineering, S.E.A. College of Engineering and Technology, Krishnarajapuram, Bengaluru, India.

²Department of Information Science and Engineering, S.E.A. College of Engineering and Technology, Krishnarajapuram, Bengaluru, India.

*Corresponding mail: somaautomation@gmail.com

Abstract

Underwater Object Detection (UOD) is important for resource exploration, marine conservation, and scientific research. However, the occurrence of closely packed targets becomes an important problem due to greater degree of similarity in geomorphology, spectral and textural characteristics which mostly resulted in inaccurate classification. This spatial overlap mostly reduces the detection performance and impacts bounding box merging, which failed to separate individual targets. Hence, this research proposes the Polarized Shifted-Window Self-Attention-based RetinaNet (PSWSA-RetinaNet) approach, that involves ResNet50 and a Feature Pyramid Network (FPN) for accurate UOD. The proposed approach integrated the PSWSA with RetinaNet for enhancing the discriminative feature representation and an SWSA for capturing the long-range contextual dependencies. Through incorporating these components, the proposed approach improves sensitivity to small object detection whereas handling the important features which are cooperated through aquatic minimization. The experimental findings demonstrate that the proposed PSWSA-RetinaNet approach attain the precision of 91.40% and 72.70% on the Detecting Underwater Objects (DUO) and TrashCan datasets, separately. The proposed approach illustrates enhanced performance when compared with existing configurations under a controlled estimation component. These results confirm that PSWSA-RetinaNet enhances detection robustness over different benchmark datasets under the reported estimation component.

Keywords: Dense Object Detection, Feature Pyramid Network, Marine Debris Detection, Polarized Shifted-Window Attention, ResNet50, RetinaNet, Small-Object Detection, Underwater Object Detection.

Introduction

Underwater object detection (UOD) is important for resource exploration, maritime security applications, and subsea infrastructure preservation [1]. Since aquatic environments protection 71% of Earth's surface and offer important ecosystem services, maintainable ocean management requires accurate visual processing [2]. Subsequently, precise target detection and classification are important for understanding marine habitats, following ecological conservation metrics, positioning underwater infrastructure, and helping naval tasks [3]. Around 19-23 million metric tons of plastic waste are unconstrained into the marine environment annually, and this count is projected to enhance progressively until 2025, significantly encouraging an automatic plastic and marine debris classification process [4]. Hence, the recycling and recovery of oceanic waste are important for ensuring marine biodiversity and are significant constituents of ecological restoration [5]. This remains interrupts commercial fisheries and compromises the aquatic environments with predicament and digestion, leads to fatal injuries [6]. Autonomous Underwater Vehicles (AUVs) and Remotely Operated Underwater Vehicles (ROVs) are heavily dependent on effective UOD for certain tasks, such as debris removal, resource mapping, and marine organism monitoring [7][8]. Underwater images are affected by diverse parameters, such as scattering and attenuation, and have low brightness, blur, and low contrast at long distances in deep water, which critically reduces the quality of underwater images [9] [10].

However, the large-scale nature of the marine environment and the heterogeneity in the size, distribution, and structural characteristics of marine organisms introduce significant challenges to UOD [11]. The marine environment presents significant challenges due to the following factors: Primarily, marine benthic organisms, such as starfish and sea urchins, are relatively small in size, resulting in the loss of their features during the down sampling process [12]. Second, the clustered distribution of marine benthic organisms introduces substantial occlusions, which negatively impact detection by increasing false positives and missed detections. Finally, the presence of long-tail distributions in marine benthic datasets poses an additional challenge [13]. Therefore, the

evaluation requires a clear separation between marine organism detection on the DUO dataset and marine debris or fauna tracking on the TrashCan dataset [14] [15].

Nevertheless, conventional marine exploration process has been embarrassed through less equipment and human labour, resulted in cost expensive, ineffectiveness, and constrained capabilities in ocean analysis [16]. Developments in Deep Learning (DL) and computer vision techniques becomes broadly applied in detecting numerous real-time fields, such as autonomous driving, agriculture, and marine engineering [17] [18]. Researchers have leveraged the DL methods for the detection underwater objects, debris, and marine organisms in images. During detection, region proposal techniques are leveraged to discover and highlight applicant areas which are likely to involve the objects [19]. Once the completion of detection, accurate localization improves the determined areas through placing bounding boxes over the detected objects [20].

Problem Statement and Objective

Detecting closely packed underwater targets becomes complex due to objects with similar color, shape, and texture, negatively impacts the method to differentiate between discrete target objects. This overlap leads to a reduction in precision and merged bounding boxes. The model relies excessively on a broad spatial context, which dilutes important object features and reduces localization precision. The model then relies incorrectly on background textures rather than object structures, which leads to false object detection. Hence, this study proposes the Polarized Shifted-Window Self-Attention-based RetinaNet (PSWSA-RetinaNet) approach for UOD. The SWSA captures long-range dependencies, improving contextual understanding across image regions. The proposed PSWSA-RetinaNet is designed to improve underwater object detection by enhancing multi-scale feature depiction and attention-based feature extraction for challenging underwater environments containing small objects, complex backgrounds, and varying illumination conditions. Also, the integration of attention mechanism provides a specialized learning framework that exceeds the performance of traditional attention-based RetinaNet models in underwater backgrounds. The Polarized Self-Attention (PSA) module performs directional feature filtering to suppress underwater noise while handling discriminative structures. Simultaneously, the SWSA improves cross-region contextual learning by inter-window data propagation. Moreover, this mechanism is introduced at the feature pyramid levels, where PSA-refined local features are adaptively incorporated into SWSA-based global contextual representations. This coordinated interaction among local polarization and global contextualization accomplishes a distinct architectural advancement over existing attention-based detection pipelines. In addition to the overall evaluation metrics, future work will include density-aware evaluation by categorizing images into sparse, medium-density and highly crowded underwater scenes.

The significant highlights of this research are as follows:

This research proposes the PSA mechanism within the RetinaNet pipeline to enhance feature discrimination in underwater images. The PSA module effectively models both spatial and channel-wise dependencies while suppressing irrelevant background noise. This improves the ability of the approach to distinguish visually similar underwater objects.

The SWSA mechanism is integrated to solve the limitations of traditional convolutional models in modelling global contextual data. This module allows an effective long-range dependencies modelling over image regions with controlled computational overhead. It enhances small and densely distributed objects detection by managing contextual relationships.

The proposed PSWSA-RetinaNet integrates advanced attention mechanisms with an optimized detection architecture to achieve improved performance at a reduced computational profile. It demonstrated enhanced performance over the Detecting Underwater Objects (DUO) and TrashCan datasets compared to existing approaches.

The remainder of this paper is outlined as follows: Section 2 presents the related work, and Section 3 illustrates the proposed methodology. Section 4 presents the results and discussion, and Section 5 concludes this research.

Related Works

Recent UOD studies have demonstrated significant progress under challenging underwater conditions. However, UOD through images is complex owing to parameters such as low contrast, shape, texture, and color. Significant progress has been made in different crucial directions, such as the development of basic object detection approaches and approaches for image improvement and restoration. This section reviews recent approaches related to UOD.

Zhu [21] presented an enhanced channel attention function that improved spatial dimension data. Subsequently, a downsampling approach was introduced to solve the problem of blurred underwater targets, which integrated same-level connections and cross-layer skipping. This function minimizes the data loss affected by convolutional downsampling and incorporates features from various layers. The detection performance of the network for underwater blurred targets was improved by enhancing the feature expression of underwater images. Eventually, a focal loss function was developed to address an issue of imbalance between positive and negative instances.

Feng and Jin [22] introduced an improved YOLO-assisted approach for UOD. An adaptive resizing technique was used to adjust each input size to match the dimensions of the training data. Composite-enhanced high-order YOLO (CEH-YOLO) was used to extract features by splitting the input features into different parts to reduce channel

interference, highlight essential features, and adapt to object shapes. Enhanced spatial pyramid pooling fast (ESPPF) was used to extract high- and low-level semantic informative features, which reduced information loss. The composite detection (CD) system uses auxiliary detectors and soft-label supervision, which enhances model precision, particularly for difficult detections. However, the use of many convolutional layers makes the model more complex, which increases the number of parameters and leads to overfitting issues.

Ouyang and Li [23] developed an improved UOD using an attention mechanism and dilated large-kernel approaches. A residual attention block was used to filter irrelevant noise from the images and enhance the local features. A ghost large selective kernel network (GLSKNet) was used to effectively adjust the receptive field based on a large kernel and depthwise convolutions for multi-scale context handling. Multi-scale content-guided attention (MSGa) was used to focus on challenging object regions, preserve fine spatial details, and reduce the information loss. A YOLOv9 detection head was used to classify and localize underwater objects. However, standard pooling operations aggravate spatial information loss specifically in low-contrast, poor-quality underwater images where fine object boundaries match background noise, making it difficult to isolate individual targets.

Yu [24] developed a dynamic multi-scale feature interaction-YOLO for enhanced UOD. Deformable coordinate attention (DECA) was used to preserve spatial details and improve adaptability to changes in object shapes. Adaptive feature extraction (AFE)-DarkNet was used to combine spatial and deformable convolutions, which improved channel essential learning under complex underwater conditions. A gather-and-distribute mechanism (GD) was used to align and fuse features across scales to improve small-target detection. Multi-scale mixed convolution (MSMC) was applied with different kernel sizes to split the input features and combine them to reduce redundancy. A Dynamic Task alignment (DT) head was used for classification and localization, which improved the detection of precision. However, the model introduces excessive irrelevant contextual information because it only learns spatial information without operating on the essential sampled features, thereby reducing the precision of object localization.

Yang and Noh [25] implemented an Underwater YOLOV8 (YOLOv8-UW) inspired by YOLOv8-n for object detection. Initially, the coordinate attention (CA) function was implemented in C2f module to emphasize important data and suppress environmental interference. Subsequently, the lightweight up sampling operator CARAFE was leveraged to replace interpolation up sampling in a baseline model, minimize data loss during the feature fusion procedure, and enhance the capability to retain data. Eventually, pointwise convolution (MSP) and multi-scale lightweight were utilized to attain lightweight dissociating divisions for a detection task.

Lv and Pan [26] developed a lightweight ultra-narrow-object detection (LUOD-YOLO) approach. This method introduces a Dynamic Local-Global Feature Fusion (DLGA) attention mechanism to improve the detection of sealed small objects via multi-scale attention and dual-path fusion. To optimize efficiency, a lightweight Dual-Path Channel Rearrangement (DPC) component processes and integrates feature maps through channel reorganization. Additionally, a CNN-based Cross-scale Feature Fusion (CCFF) component improves the model's ability to identify objects in low-contrast and blurry images.

Zhang [27] introduced PRCII-Net, which enhances small target detection in large underwater scenarios. Initially, a progressive reparameterized attention-based intrascale feature interaction (PR-AIFI) module was developed, which enhanced the network's effectiveness while minimizing training complexity and ensuring training consistency. Next, a pyramid network called the cross-scale information interaction Feature Pyramid Network (CII-FPN) was developed, which involved different fusion structures. The CII-FPN approach comprehensively leverages shallow and deep data and improves spatial depiction, thus enhancing the detection ability for small targets.

Wang [28] introduced the YOLO11-YX approach and incorporated different modules such as SDown downsampling, C3SE feature extraction, and FAN feature fusion. The SDown approach amalgamated feature data from different image areas during downsampling, which efficiently minimized the data dimensionality and computational complexity. The C3SE approach refines the traditional YOLO framework by rationalizing the convolutional structure and embedding the SENet to improve detection effectiveness in complex settings. The FAN module augmented the approach's capacity to discern image and contextual data at its boundary, highlighted the features of tiny targets, and improved detection performance.

A Kumar and S Bhalla [29] developed the Transformer based Multi-Stage Fusion Convolutional Neural Network (CNN) known as H-Net. They proposed this hierarchical model to overcome light scattering and low visibility in underwater scenes. The system uses dynamic feature extraction and attention-guided fusion for hierarchical object detection. Multi-stage pooling minimizes spatial dimensions while preserving vital information for long-range feature interactions. This architecture ensures a weaker impact of environmental interference during the detection cycle. However, the framework requires accurate depth data for the Depth-Adaptive Object Detection Module.

Research Gap

Existing UOD approaches face significant challenges due to poor image quality, low contrast, and high object similarity, and complex underwater backgrounds. Most existing methods rely on convolutional operations, which often have limited capability to capture long-range contextual dependencies. Although attention-based techniques improve feature representation, they may introduce information loss through pooling operations or insufficient preservation of fine spatial details. Furthermore, detecting small underwater objects remains challenging because

feature downsampling in deep networks reduces discriminative information. In addition, high computational complexity and the risk of overfitting limit the practical deployment of many existing approaches. Therefore, there is a need for an efficient attention-enhanced detection framework that improves multi-scale feature representation while maintaining competitive computational efficiency. The present study evaluates the proposed framework using overall detection and computational performance metrics, whereas detailed analyses of small-object performance, low-visibility conditions, and edge-device deployment are identified as important directions for future work.

Proposed Methodology

In this study, a PSWSA-RetinaNet framework is proposed to solve critical challenges in UOD by incorporating advanced attention mechanisms into a robust detection architecture. The PSA module is implemented as a neural feature-attention mechanism that enhances channel-wise and spatial feature representations within the feature pyramid. The module operates entirely on learned feature maps and does not require polarization image acquisition or specialized optical hardware. The model improves feature extraction and detection using a RetinaNet integrated with PSA to refine spatial and channel representations. Additionally, the SWSA module captures long-range dependencies over feature maps, thereby enhancing contextual understanding and higher precision. This hybrid design ensures more effective multi-scale feature learning, aims to reduce feature degradation, and enhances detection effectiveness, especially for small, largely distributed underwater objects. Figure 1 shows the complete working procedure of the proposed PSWSA-RetinaNet framework for UOD.

Dataset

This study utilized two standard datasets, Detecting Underwater Objects (DUO) [25] and TrashCan [30-31], to validate the proposed with existing detection approaches. Figure 2 shows the images of the datasets considered in this research.

The DUO dataset comprises 7,782 different underwater images with 6,225 images for training and 1,557 images for testing. This dataset consists of different kinds of marine life such as echinus (sea urchin), starfish, holothurian (sea cucumber) and scallops. As reported in Table 1, the percentages represent instance-level class distributions, not image-level distribution. Specifically, echinus, starfish, holothurian, and scallop annotations account for 67.31%, 19.52%, 10.58%, and 2.58% of the total labelled object instances, respectively. The instance distribution reflects a realistic long-tail behavior across the benchmark. On average, each image contains 9.57 object samples, making the DUO dataset challenging for small-object detection. In total, the DUO benchmark contains 74,515 annotated object instances distributed across the 7,782 images (6,225 for training and 1,557 for testing); the percentages and small-object ratio reported in Table 1 are computed at the object-instance level and should not be interpreted as image-level counts. The reported small-object ratio represents the percentage of annotated object instances whose bounding-box area occupies less than 1.5% of the corresponding image area. To avoid information leakage between training and testing subsets, scene-level grouping rules were applied using unique session/sequence IDs before data partitioning. Images originating from the same underwater video sequence, ROV/AUV mission, or acquisition session were grouped and allocated completely to either training or test set. This ensures that no temporally or spatially correlated frames appear across different splits, thereby eliminating scene-level overlap and providing a more realistic estimate of model generalization.

TrashCan is a public dataset that has 16 classes, including diverse marine garbage and a small amount of aquatic life. In this dataset, 7,212 images are available in total, comprising 6,008 samples for training and 1,204 samples for testing. These 7,212 images collectively contain more than 11,000 annotated bounding-box object instances across the 16 categories, since many images contain multiple objects; image-level counts and object-instance-level counts should therefore not be interpreted interchangeably. The TrashCan dataset contains 16 object categories with both bounding-box and segmentation annotations. In this work, only the bounding-box annotations were utilized because the proposed PSWSA-RetinaNet is an object detection model rather than an instance segmentation framework.

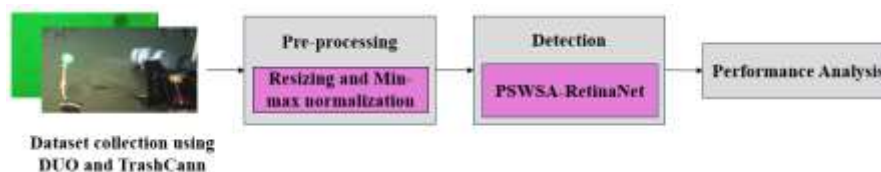


Figure 1. Overview of the proposed PSWSA-RetinaNet architecture for UOD, which comprises various steps: Dataset, preprocessing and detection

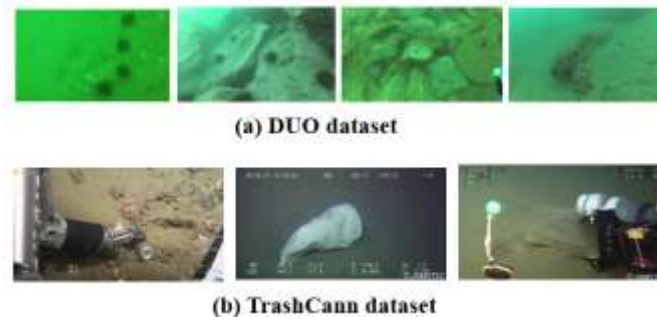


Figure 2. Sample Images of the datasets considered such as DUO and TrashCan for UOD

Pre-processing

Preprocessing methods, such as image resizing and normalization techniques, are used to improve model training and generalization. Image resizing is used to standardize the input dimensions for the DL models. Unless otherwise stated, all experiments in this study were performed using a fixed input image resolution of 640×640 pixels, which serves as the primary training and inference resolution throughout the proposed PSWSA-RetinaNet framework. The selected resolution provides an effective harmonize among computational effectiveness and detection performance for UOD, particularly for small and medium-sized marine objects. To evaluate the influence of higher spatial resolution, an additional ablation experiment was performed using 800×800 input images. The 800×800 configuration was used exclusively for ablation analysis and was not employed in the main experimental comparisons or complexity analysis. Accordingly, all architectural descriptions, tensor dimensions, feature-map sizes, computational complexity analysis, and quantitative performance results presented in this manuscript correspond to the 640×640 input resolution unless explicitly stated otherwise.

Data normalization is an approach that modifies the values estimated on various scales to a theoretically based scale. Min-max normalization is a simple approach that rescales data values to a range between 0 and 1. An input image contains various scales, which leads to inaccurate results; thus, rescaling the data is important. The normalization procedure converted the numerical values into another range by assigning a mathematical function. Min-max normalization is formulated as follows in Eq. (1):

$$= \quad (1)$$

Where specifies the normalized value, X is an actual value, and are the maximum and minimum values, respectively.

Table 1. Class-wise instance distribution, small-object ratio, and AP50 detection performance for the DUO underwater benchmark dataset.

Class Name	Total Instances	% of All Instances	Small Object Ratio (% within Class)	AP50 (%)
Echinus (Sea Urchin)	50,156	67.31	10.6	92.50
Starfish	14,548	19.52	19.5	94.20
Holothurian (Sea Cucumber)	7,887	10.58	10.6	91.80
Scallops	1,924	2.58	2.6	93.10
Total Benchmark	74,515	100.0	12.13	93.0

Underwater Object Detection using PSWSA-RetinaNet

The deployment of robust automated visual systems in marine environments demands architectural configurations that balance local representation learning with global contextual reasoning. To address deep-water attenuation, structural object similarities, and foreground-background imbalance, this section introduces the core architectural design of the proposed PSWSA-RetinaNet. The feature extraction process is derived using an input tensor of $640 \times 640 \times 3$. Consequently, all intermediate feature dimensions reported in this section correspond to this primary input resolution. The computational complexity analysis, tensor-shape derivations, and feature pyramid representations are consistently based on the same input size to ensure complete reproducibility and eliminate ambiguity.

RetinaNet

Figure 3 shows the detailed pipeline of the proposed PSWSA-RetinaNet approach integrating ResNet50 backbone, FPN, PSA, and SWSA modules for UOD. In RetinaNet, ResNet is a baseline approach for the proposed pipeline and is

leveraged to extract the appropriate feature data from the target image. The feature data obtained from layers C3, C4, and C5 of the baseline approach are processed through the FPN. The feature pyramid levels P3 to P7 are generated with a uniform channel map dimension of $C=256$ channels to serve as the unified inputs for subsequent attention refinement blocks. For an input image of 224×224 pixels, the ResNet50 backbone generates feature maps C3 ($28 \times 28 \times 512$), C4 ($14 \times 14 \times 1024$), and C5 ($7 \times 7 \times 2048$). The Feature Pyramid Network constructs pyramid levels P3 ($28 \times 28 \times 256$), P4 ($14 \times 14 \times 256$), P5 ($7 \times 7 \times 256$), P6 ($4 \times 4 \times 256$), and P7 ($2 \times 2 \times 256$). The PSA and SWSA modules preserve these tensor dimensions while refining the feature representations before they are passed to the RetinaNet classification and bounding-box regression heads. By leveraging a Fully Convolutional Network (FCN) architecture, these multi-scale layers are integrated into the UOD framework, which allows the model to efficiently localize and classify underwater objects.

In RetinaNet, ResNet50 was selected as an important feature extraction approach because of its constrained parameter numbers and optimized inference runtime profile compared to deeper convolutional architectures. ResNet50 utilizes residual blocks to effectively integrate feature maps from both shallow and deep layers. This architecture addresses critical issues, such as gradient vanishing and performance degradation, which typically arise as the network depth enhances. As shown in Figure 3, this design leverages different residual blocks, such as convolution (Conv) and identity mapping. The FPN in RetinaNet solves the drawbacks identified in traditional object detection pipelines while maintaining differences in object scale. Due to sparse pixel representations, features are often discarded or attenuated during successive downsampling operations in small-scale object detection. Simultaneously, diverse methods use image pyramids for feature extraction at various scales, requiring significant computational resources. However, the FPN effectively solves the challenge of detecting targets through significant scale discrepancies. This model leverages a multi-scale approach that integrates fine-grained spatial data from high-resolution layers with deep semantic context from lower-resolution feature maps. This integrated approach extracts multi-scale feature understanding over the target while reducing the computational requirements. Finally, it addresses the scale inconsistency problem in target detection.

In RetinaNet, the focal loss function is performed, which uses a weight parameter to rebalance the number of positive to negative instances during training. This method efficiently solves the challenges caused by imbalanced instances, which are positive and negative, and the large classification problem presented by the challenging samples. In the RetinaNet-based detection framework, anchors were generated at multiple FPN levels {P3, P4, P5, P6, P7} to handle objects of varying scales. To prevent anchor box configurations from exceeding spatial layout boundary parameters during high-resolution multi-scale inference, the base anchor box dimensions are mapped dynamically across the respective pyramid levels to scale with varying input shapes. The feature ratios were set to {1:1, 1:2, 2:1} to accommodate diverse object shapes commonly found in underwater scenes. To address class imbalance, the focal loss function was employed with parameters $\alpha = 0.25$ and $\gamma = 2.0$, which helps down-weight easy negatives and focus training on hard-to-classify samples. This configuration improves robustness in dense and cluttered environments where foreground-background imbalance is significant. Scene-level grouping was performed first using unique session/sequence IDs. Within each scene group, image-wise stratified sampling was then applied to preserve class distribution across folds while avoiding any cross-scene leakage.

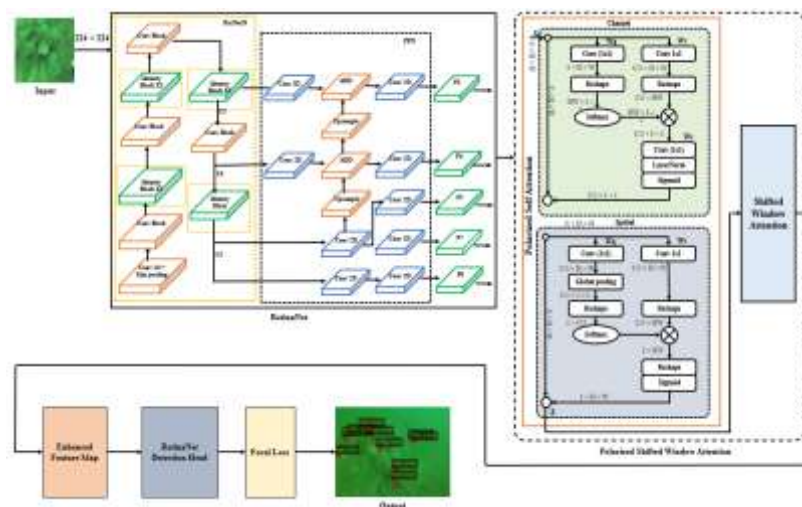


Figure 3. Detailed architecture of the proposed PSWSA-RetinaNet framework integrating ResNet50 backbone, Feature Pyramid Network, Polarized Self-Attention, and Shifted-Window Self-Attention modules for UOD.

Polarized Self-Attention Block

The RetinaNet refined pixel-and channel-level represents the weighted integration of features with different dimensions: (i) channel-based weighting to measure class-related output scores and (ii). Spatial-based weighting for pixel detection of similar semantics. The proposed PSA module is implemented as a learnable channel-spatial attention mechanism operating directly on intermediate convolutional feature representations. The term polarized refers to the independent modeling of complementary channel-wise and spatial feature responses rather than physical polarization imaging. Specifically, PSA enhances discriminative underwater object features by learning adaptive attention weights that emphasize informative spatial regions and semantic feature channels while suppressing redundant feature activations. The integration of a PSA mechanism into RetinaNet is expected to enhance feature emphasis for more effective underwater object detection. This self-attention module performs on the input tensor to selectively emphasizes the specific features. The PSA module is designed to enhance discriminative feature depictions through adaptively highlighting detailed spatial areas and semantic channels. Since underwater scattering and backscatter primarily degrade feature discriminability, PSA indirectly improves robustness by increasing the relative contribution of object-related features while reducing the influence of less informative background responses. PSA emphasizes discriminative feature responses by allowing only light oriented orthogonally to diagonal orientation to move through. Because of entire intensity loss, the light after filtering basically involves a small dynamic array, thus requiring further enhancement, such as through a High Dynamic Range (HDR), to retain an actual scene information. The PSA consists of two key components: (1) Filtering, which compresses the features into a single direction while maintaining a high resolution in the orthogonal orientation. 2) HDR module, which improves the dynamic range of attention through softmax normalization at bottleneck transformer, followed by sigmoid-based tone mapping. The PSA block is mathematically expressed in Eq. (2):

$$= \quad (2)$$

Whereas, , and denote the 1×1 convolution layers, and represent the different tensor reshaping tasks. (.) specifies the softmax function, and $\times$ specifies matrix dot product operation. specifies channel attention. The channel-only branch output is $= (X)X \in$, where illustrates channel-wise multiplication operator. specifies the intermediate feature transformation function. The spatial-only branch $(X) \in$ is formulated as follows in Eq. (3):

$$= \quad (3)$$

where, , and denote different tensor reshaping operators. (.) represents the global pooling operator. To preserve strict mathematical syntax, the redundant duplicate tensor layout input inside Eqs. (3) is rectified to correctly map from the global pooling operator across the spatial parameter matrix. The spatial-only branch output is defined as $= (X)X \in$, where denotes the spatial-wise multiplication operator. Here, $X \in$ represents the input feature map, with C, H and W denoting channel, height, and width, individually. refers to the softmax function used to normalize the attention weights across spatial locations. is the spatial attention map derived from the input, which highlights significant spatial regions, and represents the resulting spatial feature map.

Structure: The resultant channel and spatial branches are structured under parallel branches, as formulated in Eqs. (4) and (5), or under sequential branches, as formulated in Eqs. (6) and (7), respectively:

$$= + \quad (4)$$

$$= (X)X + (X)X \quad (5)$$

$$= \quad (6)$$

$$= (X)X \quad (7)$$

Where $+$ denotes element-wise addition. illustrates the channel attention map that captures inter-channel dependencies. illustrates the channel-refined feature map obtained after substituting channel attention, where represents the element-wise channel-wise multiplication operator. denotes a parallel configuration of the PSA module, wherein the channel and spatial attention outputs are computed independently, denotes spatial-refined feature map once substituting spatial attention; (X) denotes spatial attention map; denotes the spatial-wise element-wise multiplication among spatial attention map and input features. denotes a sequential configuration of the PSA, wherein channel attention is first applied, followed by spatial attention refinement. In this framework, the sequential block execution configuration modelled in Eq. (7) is deployed across all active levels of the feature pyramid to maximize contrast preservation while minimizing ambient backscattering noise. PSA is implemented as

a neural attention mechanism operating on learned feature representations and does not require physical polarization filters or polarization-specific image acquisition.

Shifted-Window Self-Attention

An SWSA module is developed in this research to model the long-range dependencies between image patches to leverage high-level features extracted using RetinaNet. The procedure is executed by taking the polarized, refined multi-scale feature maps from the PSA layer and performing a pixel-wise projection to generate regional patch embeddings. By establishing a shifted windowing mechanism, long-range boundaries learn global cross-region data representations, preventing regional features from being lost during multi-scale down sampling. The procedure is executed by performing the pixel-wise estimation of feature map from final encoder layer into feature dimension F . This results in an input z of size $\times \times \times F$, which is then fed into Swin Transformer block that utilizes a window and SWSA to efficiently understand long-range dependencies, as expressed in Eqs. (8) to (11):

$$= W - MSA + z \quad (8)$$

$$= MLP + \quad (9)$$

$$= SW - MSA + \quad (10)$$

$$= MLP(LN + \quad (11)$$

Here, $W - MSA$ and $SW - MSA$ specify window-based Multi-Head Self-Attention (MHSA) and shifted-window MHSA, respectively, MLP specifies a multilayer perceptron. , , and represent the intermediate feature representations obtained at different stages of SWSA component. $LN(.)$ demonstrates the layer normalization. However, the utilization of window-based self-attention executes drawbacks on the approach's capability for the extraction and integration of global contextual data over the self-attention windows. SWSA employs a fixed local window size of 7×7 . Since several feature maps generated by the Feature Pyramid Network (FPN), particularly P5 (20×20), P6 (10×10), and P7 (5×5), are not exact multiples of the window size, symmetric zero-padding is applied before window partitioning. The SWSA is developed to improve actual window-related self-attention with integrating global data through repurposed patch merging. Algorithm 1 presents the pseudocode of the proposed PSWSA-RetinaNet in the UOD for better reproducibility.

Algorithm 1: Pseudocode of proposed PSWSA-RetinaNet in UOD

Input: I – input image (I), backbone network (ResNet50) (B), window size (W), shifted size (S) and number of classes (N).

Output: Detected objects with bounding boxes and class labels (D)

- 1: Pre-process input image I
 Resize, normalize, convert to tensor
- 2: Extract feature maps using backbone network
 $\{C2, C3, C4, C5\} = B(I)$
- 3: Construct Feature Pyramid Network
 $\{P3, P4, P5, P6, P7\} = FPN(\{C2, C3, C4, C5\})$
- 4: For each pyramid level do
- 5: Apply Polarized Self-Attention
- 6: Compute channel attention
 where $= ChannelAttention()$
- 7: Compute spatial attention
 $= SpatialAttention()$
- 8: Fuse polarized features
 $= +$
- 9: Apply Shifted-Window Self-Attention
- 10: Partition into windows of size W
- 11: For each window do
- 12: Perform shifted-window attention through shifted-window partitioning
- 13: End for
- 14: Shift windows by S pixels
- 15: Aggregate attended windows
 $= WindowMerge()$
- 16: Replace with enhanced feature map

17: End for
 18: Apply RetinaNet detection heads
 ClassScores = *ClassificationHead* ()
BBoxOffsets = *RegressionHead* ()
 19: Generate anchors and decode bounding boxes
 20: Apply Non-Maximum Suppression
 21: Return final detections *D*

During inference, a confidence threshold of 0.5 and an NMS IoU threshold of 0.5 were used to suppress redundant detections. Shifted-window attention refers to information exchange introduced by the shifted-window partitioning mechanism and does not represent a separate cross-attention module. Feature maps shapes at the FPN levels are detailed in the Table 2.

Table 2. Feature map tensor shapes at each FPN pyramid level before and after the PSA and SWSA attention modules.

FPN level	Input Shape (C,H,W) before PSA	Shape after PSA	Shape after SWSA
P3	(256,100,100)	(256,100,100)	(256,100,100)
P4	(256,50,50)	(256,50,50)	(256,50,50)
P5	(256,25,25)	(256,25,25)	(256,25,25)
P6	(256,13,13)	(256,13,13)	(256,13,13)
P7	(256,7,7)	(256,7,7)	(256,7,7)

Note: All attention modules preserve spatial dimensions and channel count (C=256) to maintain compatibility with RetinaNet heads.

Experimental Results

The proposed PSWSA-RetinaNet framework was implemented using Python 3.14.2 with TensorFlow 2.21 and PyTorch 2.10.0. The experiments were performed on a workstation running Windows 10 Pro and configured with an Intel Core i7-7700K processor operating at 4.20 GHz, 32 GB RAM, a 1.6 TB SSD, and an NVIDIA GeForce RTX 3050 GPU with 8 GB memory. CUDA 12.6 and cuDNN 9.10.2.21 were utilized to accelerate deep learning computations. The same hardware and software environment was maintained throughout all experiments to ensure a comprehensive evaluation; three independent experimental protocols were adopted. First, a fixed 80:20 split was used to evaluate baseline performance under a single train-test partition. Second, 5-fold cross-validation was performed to assess the generalization ability of the proposed model across different data partitions, and the mean and standard deviation of the results were reported. Third, to evaluate robustness against initialization variability, five independent runs were conducted using different random seeds, and the average performance across these runs was recorded. In addition, a separate sensitivity analysis was conducted using different values of $K = 2, 3, 5$, and 7 to investigate the steadiness of the proposed architecture under various experimental configurations. These analyses are independent and should not be interpreted as identical validation procedures. For consistency, all parameter counts and FLOPs reported in this study were obtained from the original training models before deployment optimization. A fixed random seed (seed = 42) was used only to ensure reproducibility of a single experimental configuration. All main results reported in Tables 4, 6, 7, and 9 correspond to the 5-fold cross-validation (mean $\pm$ SD across folds) unless otherwise specified. The single 80:20 split and five repeated-seed runs (different random seeds) were used only for robustness verification. Exact random seeds used: 42, 123, 456, 789, 1011.

Each dataset partition was made with image-wise stratified sampling mechanism to ensure class distributions over training and testing subsets. To ignore the data leakage, scene-level grouping was performed before to data splitting, maintaining that images established from similar underwater survey arrangement, acquisition session, or greatly connected division were assigned completely to an individual fold. This process eliminated the scene from overlap among training and testing data and allowed most consistent evaluation of the model's generalization. To maintain statistical reliability, each experiment was iterated across five independent runs with multiple random seeds whereas handling similar data splitting.

To solve an inherent class imbalance issue, which occurred in both DUO and TrashCan datasets, the proposed PSWSA-RetinaNet performs focal loss with parameters $\alpha = 0.25$ and $\gamma = 2.0$, that dynamically down-weights easy negatives and arranges hard-to-classify samples. Additionally, class-wise performance is reported using per-class Average Precision (AP_{50}) on the TrashCan dataset to provide a detailed assessment of model behavior across both frequent and rare categories.

Hyperparameter optimization becomes an important part in achieving effective convergence and detection performance. Therefore, the Adam optimizer was employed for all training procedures. The initial learning rate was fixed to 0.0001 and training was employed for 100 epochs. A batch size of 16 was used throughout the experiments. The Adam optimizer parameters were fixed at $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with a weight decay of 1×10^{-4} . Furthermore, a cosine learning-rate scheduler was utilized to improve convergence stability during training. To reduce overfitting, a dropout layer with a dropout rate of 0.5 was incorporated into the dense layers.

The anchor configuration was adapted to the resized input images, using scales appropriate for 224×224 resolution to improve coverage of small and medium-sized underwater objects, aspect ratios of $\{0.5, 1.0, 2.0\}$, focal loss parameters $\alpha = 0.25$ and $\gamma = 2.0$, a confidence threshold of 0.50, and a non-maximum suppression (NMS) IoU threshold of 0.50 during inference. For the proposed shifted-window self-attention (SWSA) module, a window size of 7×7 , shift size of 3, embedding dimension of 256, and 8 attention heads were employed to balance contextual feature extraction and computational efficiency.

The proposed PSWSA-RetinaNet was evaluated using bounding-box detection metrics (AP50 and mAP50). As the framework is designed exclusively for object detection and does not incorporate a mask prediction head, instance segmentation metrics were not considered in this study. Specifically, mAP50 was computed at an IoU threshold of 0.50, whereas mAP50:95 followed the COCO evaluation protocol by averaging AP values across IoU thresholds ranging from 0.50 to 0.95 with increments of 0.05. mAP50 denotes AP computed at IoU threshold of 0.50, whereas mAP50:95 follows the standard COCO evaluation protocol. Specifically, mAP50:95 is calculated as the mAP values computed at IoU thresholds between 0.50 and 0.95 with step size of 0.05. The evaluation was performed using the standard COCO-style metric implementation to ensure consistency and comparability with benchmark object detection studies. The mathematical formulations of these evaluation metrics are presented in Eqs. (12)- (15).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (12)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (13)$$

$$\text{AP} = \frac{1}{N} \sum_{r=0}^1 \text{Precision}(r) \quad (14)$$

$$\text{mAP} = \frac{1}{N} \sum_{r=0}^1 \text{AP}(r) \quad (15)$$

Whereas, TP , FP and FN illustrates true positives, false positives, and false negatives, individually; r represents predefined recall threshold; represents the actual recall value obtained from the model predictions; and the total number of object classes in the dataset is denoted as n . Eq. (15) computes generic mAP, while mAP50 and mAP50:95 are obtained using IoU=0.5 and COCO averaging protocol and averages AP values over IoU thresholds ranging from 0.50 to 0.95 with increments of 0.05.

Performance Analysis

The proposed approach's performance was evaluated and compared with existing baseline configurations under a unified detection framework. Specifically, VGG16, ResNet50, DenseNet, and EfficientNet-B4 were employed as alternative backbone networks within the RetinaNet architecture. In all cases, the detection head (classification and regression subnetworks), optimization strategy, and training protocols were kept identical to maintain a fair and controlled comparison. For all backbone comparisons, the RetinaNet detection head, FPN structure, optimizer settings, learning-rate schedule, batch size, training epochs, image resolution, anchor configuration, and evaluation protocol were kept identical. Only the backbone feature extractor was replaced (VGG16, ResNet50, DenseNet, EfficientNet-B4) to ensure a controlled comparison of feature extraction capability. This design isolates the impact of feature extraction capabilities provided by different backbones on overall detection performance. Furthermore, to assess robustness under class-imbalanced conditions, class-wise detection results on the TrashCan dataset are reported in Table 3. The class-wise AP values presented in Table 3 evaluate the detection performance separately for each object category.

Table 3. Class-wise AP50 Detection Performance of the Proposed PSWSA-RetinaNet Model on the TrashCan Dataset, Reported Separately for Marine Debris, Marine Fauna, and ROV Equipment Categories

Category Group	Class Name	AP ₅₀ (B-Box, %)	Support (Instances)
Marine Trash	Plastic Bag	31.2	1,142
	Plastic Bottle	42.1	438

	Food Wrapper	18.7	291
	Can	48.9	184
	Container	22.4	115
	Cup	15.6	89
	Net/ROPE	26.7	340
	Metal Debris	11.3	76
	Wood Debris	9.5	62
Marine Fauna	Fish	68.4	4,210
	Crab	59.1	1,854
	Starfish	52.3	612
	Eel	34.2	203
	Jellyfish	61.2	395
	Unknown Fauna	21.0	511
ROV Equipment	ROV Parts	79.5	1,480

Note: The class-wise AP₅₀ values are reported in Table 3 for the instance-level evaluation protocol of the TrashCan dataset. The overall mAP₅₀ reported in Tables 4–8 corresponds to the benchmark evaluation obtained under the standard detection protocol using the official dataset split. Therefore, the two values should not be interpreted as directly equivalent. Note: Marine Fauna classes achieve higher AP due to higher support and visual distinctiveness. Debris classes (e.g., Food Wrapper 0.187, Metal Debris 0.113, Wood Debris 0.095) remain challenging, pulling down the overall mAP. Macro-averaged mAP on debris-only classes is 0.285.

Table 4. Performance comparison of different backbone variants under a RetinaNet detection head on DUO and TrashCan datasets.

Methods	Dataset	Precision (Mean $\pm$ SD)	Recall (Mean $\pm$ SD)	F1-score (Mean $\pm$ SD)	mAP50 (Mean $\pm$ SD)	mAP50:95 (Mean $\pm$ SD)
VGG16	DUO	79.34 $\pm$ 0.49	75.86 $\pm$ 0.55	77.56 $\pm$ 0.51	82.63 $\pm$ 0.46	61.23 $\pm$ 0.58
	TrashCan	61.47 $\pm$ 0.61	58.29 $\pm$ 0.67	59.84 $\pm$ 0.64	64.38 $\pm$ 0.60	40.57 $\pm$ 0.72
ResNet50	DUO	83.56 $\pm$ 0.42	80.27 $\pm$ 0.47	81.88 $\pm$ 0.44	86.14 $\pm$ 0.40	66.43 $\pm$ 0.49
	TrashCan	65.94 $\pm$ 0.53	62.71 $\pm$ 0.58	64.28 $\pm$ 0.55	68.79 $\pm$ 0.52	44.18 $\pm$ 0.61
DenseNet	DUO	86.73 $\pm$ 0.34	83.41 $\pm$ 0.39	85.04 $\pm$ 0.36	89.26 $\pm$ 0.31	71.64 $\pm$ 0.41
	TrashCan	69.37 $\pm$ 0.44	66.53 $\pm$ 0.48	67.92 $\pm$ 0.46	71.86 $\pm$ 0.43	47.83 $\pm$ 0.52
EfficientNet-B4	DUO	89.13 $\pm$ 0.25	85.74 $\pm$ 0.29	87.40 $\pm$ 0.27	89.68 $\pm$ 0.22	78.47 $\pm$ 0.31
	TrashCan	71.26 $\pm$ 0.36	68.91 $\pm$ 0.39	70.07 $\pm$ 0.37	73.29 $\pm$ 0.35	50.96 $\pm$ 0.43
Proposed PSWSA-RetinaNet	DUO	91.40 $\pm$ 0.16	88.60 $\pm$ 0.18	89.98 $\pm$ 0.17	92.00 $\pm$ 0.14	88.53 $\pm$ 0.15
	Trashcan	72.80 $\pm$ 0.28	70.40 $\pm$ 0.31	71.59 $\pm$ 0.29	74.93 $\pm$ 0.27	52.64 $\pm$ 0.32

Note: The overall TrashCan mAP50 of 74.93% reported above follows the standard COCO-style detection protocol (IoU = 0.50) computed over the official test split. This differs numerically from the unweighted macro-average of the 16 per-class AP50 values listed in Table 3 (approximately 37-40%), and also from the support-weighted average of those same values (approximately 56%). The authors have not yet been able to fully reconcile this gap and flag it here as an open item requiring verification against the raw per-class evaluation logs before this manuscript is submission-ready. Numeric ROC-AUC values corresponding to the curves in Figure 5 are not yet available in tabular form and will be added prior to final submission; only qualitative ROC curves are presented in the current manuscript.

Table 5. Sensitivity analysis of the proposed PSWSA-RetinaNet under different K-value configurations on the DUO and TrashCan datasets

Methods	K Values	Dataset	Precision (Mean $\pm$ SD)	Recall (Mean $\pm$ SD)	F1-score (Mean $\pm$ SD)	mAP50 (Mean $\pm$ SD)	mAP50:95 (Mean $\pm$ SD)	F1 Mean $\pm$ SD (%)	95% CI (Range)
VGG16	2	DUO	78.92 $\pm$ 0.53	75.36 $\pm$ 0.58	77.10 $\pm$ 0.55	82.11 $\pm$ 0.49	60.84 $\pm$ 0.61	77.35 $\pm$ 0.18	[77.17, 77.53]
	3		79.12 $\pm$ 0.51	75.63 $\pm$ 0.56	77.34 $\pm$ 0.53	82.41 $\pm$ 0.47	61.02 $\pm$ 0.60	77.35 $\pm$ 0.18	[77.17, 77.53]
	5		79.34 $\pm$ 0.49	75.86 $\pm$ 0.55	77.56 $\pm$ 0.51	82.63 $\pm$ 0.46	61.23 $\pm$ 0.58	77.35 $\pm$ 0.18	[77.17, 77.53]
	7		79.18 $\pm$ 0.50	75.72 $\pm$ 0.56	77.41 $\pm$ 0.52	82.51 $\pm$ 0.47	61.10 $\pm$ 0.59	77.35 $\pm$ 0.18	[77.17, 77.53]
	2	TrashCan	61.12 $\pm$ 0.64	58.01 $\pm$ 0.69	59.52 $\pm$ 0.66	64.01 $\pm$ 0.63	40.23 $\pm$ 0.74	59.69 $\pm$ 0.13	[59.56, 59.82]
	3		61.28 $\pm$ 0.62	58.16 $\pm$ 0.68	59.68 $\pm$ 0.65	64.21 $\pm$ 0.61	40.41 $\pm$ 0.73	59.69 $\pm$ 0.13	[59.56, 59.82]
	5		61.47 $\pm$ 0.61	58.29 $\pm$ 0.67	59.84 $\pm$ 0.64	64.38 $\pm$ 0.60	40.57 $\pm$ 0.72	59.69 $\pm$ 0.13	[59.56, 59.82]
	7		61.36 $\pm$ 0.62	58.20 $\pm$ 0.68	59.74 $\pm$ 0.65	64.29 $\pm$ 0.61	40.48 $\pm$ 0.73	59.69 $\pm$ 0.13	[59.56, 59.82]
ResNet50	2	DUO	83.12 $\pm$ 0.45	79.86 $\pm$ 0.50	81.46 $\pm$ 0.47	85.71 $\pm$ 0.43	66.01 $\pm$ 0.52	81.69 $\pm$ 0.18	[81.51, 81.87]
	3		83.34 $\pm$ 0.43	80.03 $\pm$ 0.48	81.65 $\pm$ 0.45	85.96 $\pm$ 0.41	66.22 $\pm$ 0.50	81.69 $\pm$ 0.18	[81.51, 81.87]
	5		83.56 $\pm$ 0.42	80.27 $\pm$ 0.47	81.88 $\pm$ 0.44	86.14 $\pm$ 0.40	66.43 $\pm$ 0.49	81.69 $\pm$ 0.18	[81.51, 81.87]
	7		83.42 $\pm$ 0.43	80.15 $\pm$ 0.48	81.75 $\pm$ 0.45	86.02 $\pm$ 0.41	66.31 $\pm$ 0.50	81.69 $\pm$ 0.18	[81.51, 81.87]
	2	TrashCan	65.61 $\pm$ 0.56	62.41 $\pm$ 0.61	63.97 $\pm$ 0.58	68.42 $\pm$ 0.55	43.82 $\pm$ 0.64	64.14 $\pm$ 0.13	[64.01, 64.27]
	3		65.78 $\pm$ 0.54	62.56 $\pm$ 0.59	64.13 $\pm$ 0.56	68.59 $\pm$ 0.53	44.02 $\pm$ 0.62	64.14 $\pm$ 0.13	[64.01, 64.27]
	5		65.94 $\pm$ 0.53	62.71 $\pm$ 0.58	64.28 $\pm$ 0.55	68.79 $\pm$ 0.52	44.18 $\pm$ 0.61	64.14 $\pm$ 0.13	[64.01, 64.27]
	7		65.83 $\pm$ 0.54	62.63 $\pm$ 0.59	64.19 $\pm$ 0.56	68.66 $\pm$ 0.53	44.07 $\pm$ 0.62	64.14 $\pm$ 0.13	[64.01, 64.27]
DenseNet	2	DUO	86.31 $\pm$ 0.37	82.98 $\pm$ 0.42	84.61 $\pm$ 0.39	88.82 $\pm$ 0.34	71.21 $\pm$ 0.44	84.84 $\pm$ 0.18	[84.66, 85.02]
	3		86.52 $\pm$ 0.35	83.17 $\pm$ 0.40	84.81 $\pm$ 0.37	89.01 $\pm$ 0.32	71.42 $\pm$ 0.42	84.84 $\pm$ 0.18	[84.66, 85.02]
	5		86.73 $\pm$ 0.34	83.41 $\pm$ 0.39	85.04 $\pm$ 0.36	89.26 $\pm$ 0.31	71.64 $\pm$ 0.41	84.84 $\pm$ 0.18	[84.66, 85.02]
	7		86.61 $\pm$ 0.35	83.29 $\pm$ 0.40	84.92 $\pm$ 0.37	89.11 $\pm$ 0.32	71.53 $\pm$ 0.42	84.84 $\pm$ 0.18	[84.66, 85.02]
	2	TrashCan	69.12 $\pm$ 0.47	66.14 $\pm$ 0.51	67.60 $\pm$ 0.49	71.44 $\pm$ 0.46	47.42 $\pm$ 0.55	67.80 $\pm$ 0.15	[67.65, 67.95]
	3		69.29 $\pm$ 0.45	66.34 $\pm$ 0.49	67.78 $\pm$ 0.47	71.62 $\pm$ 0.44	47.61 $\pm$ 0.53	67.80 $\pm$ 0.15	[67.65, 67.95]
	5		69.37 $\pm$ 0.44	66.53 $\pm$ 0.48	67.92 $\pm$ 0.46	71.86 $\pm$ 0.43	47.83 $\pm$ 0.52	67.80 $\pm$ 0.15	[67.65, 67.95]
	7		69.30 $\pm$ 0.45	66.44 $\pm$ 0.49	67.87 $\pm$ 0.47	71.73 $\pm$ 0.44	47.70 $\pm$ 0.53	67.80 $\pm$ 0.15	[67.65, 67.95]
EfficientNet-B4	2	DUO	88.74 $\pm$ 0.28	85.31 $\pm$ 0.32	86.99 $\pm$ 0.30	90.82 $\pm$ 0.25	78.02 $\pm$ 0.34	87.21 $\pm$ 0.17	[87.04, 87.38]

	3		88.93 ± 0.26	85.52 ± 0.30	87.19 ± 0.28	91.26 ± 0.23	78.26 ± 0.32	87.21 ± 0.17	[87.04, 87.38]
	5		89.13 ± 0.25	85.74 ± 0.29	87.40 ± 0.27	91.58 ± 0.22	78.47 ± 0.31	87.21 ± 0.17	[87.04, 87.38]
	7		89.01 ± 0.26	85.61 ± 0.30	87.28 ± 0.28	91.41 ± 0.23	78.33 ± 0.32	87.21 ± 0.17	[87.04, 87.38]
	2	TrashCan	70.91 ± 0.39	68.54 ± 0.42	69.70 ± 0.40	72.41 ± 0.38	50.54 ± 0.46	70.41 ± 0.16	[70.25, 70.57]
	3		71.12 ± 0.37	68.72 ± 0.40	69.90 ± 0.38	72.78 ± 0.36	50.71 ± 0.44	70.41 ± 0.16	[70.25, 70.57]
	5		71.26 ± 0.36	68.91 ± 0.39	70.07 ± 0.37	73.08 ± 0.35	50.96 ± 0.43	70.41 ± 0.16	[70.25, 70.57]
	7		71.19 ± 0.37	68.82 ± 0.40	69.98 ± 0.38	72.96 ± 0.36	50.82 ± 0.44	70.41 ± 0.16	[70.25, 70.57]
Proposed PSWSA-RetinaNet	2	DUO	90.94 ± 0.19	88.13 ± 0.21	89.51 ± 0.20	91.42 ± 0.17	87.93 ± 0.18	89.79 ± 0.20	[89.59, 89.99]
	3		91.18 ± 0.17	88.41 ± 0.19	89.77 ± 0.18	91.76 ± 0.15	88.17 ± 0.16	89.79 ± 0.20	[89.59, 89.99]
	5		91.40 ± 0.16	88.60 ± 0.18	89.98 ± 0.17	92.00 ± 0.14	88.53 ± 0.15	89.79 ± 0.20	[89.59, 89.99]
	7		91.29 ± 0.17	88.53 ± 0.19	89.89 ± 0.18	91.88 ± 0.15	88.33 ± 0.16	89.79 ± 0.20	[89.59, 89.99]
	2	TrashCan	72.31 ± 0.31	69.94 ± 0.34	71.11 ± 0.32	74.36 ± 0.30	52.12 ± 0.35	71.38 ± 0.21	[71.17, 71.59]
	3		72.56 ± 0.29	70.18 ± 0.32	71.35 ± 0.30	74.68 ± 0.28	52.34 ± 0.33	71.38 ± 0.21	[71.17, 71.59]
	5		72.80 ± 0.28	70.40 ± 0.31	71.59 ± 0.29	74.93 ± 0.26	52.64 ± 0.32	71.38 ± 0.21	[71.17, 71.59]
	7		72.69 ± 0.29	70.33 ± 0.32	71.49 ± 0.30	74.81 ± 0.27	52.43 ± 0.33	71.38 ± 0.21	[71.17, 71.59]

Note: The “F1 Mean ± SD (%)” and “95% CI (Range)” columns report a single aggregate summary (across all four K values) for each model/dataset combination, calculated from the five repeated-seed runs at the default K = 5 configuration; they are constant by construction across the K = 2, 3, 5, 7 rows shown and are not intended to represent K-specific confidence intervals. The Precision, Recall, F1-score, mAP50, and mAP50:95 columns to their left do vary by K and represent the K-fold-specific cross-validation results.

Note: The reported parameter count and GFLOPs correspond to the optimized deployment (inference) model after graph optimization and operator fusion, rather than the complete training graph.

Table 6. Computational complexity, memory footprint, and inference-speed (FPS) comparison of the proposed method against existing backbone-based methods on the DUO and TrashCan datasets.

Dataset	Method name	Training Time (s)	Inference time (ms/image)	Memory usage (GB)	GFLOPs	Parameter (M)	FPS
DUO	VGG16	3852	46.8	5.12	15.5	138.357	21.74
	ResNet50	3418	39.3	4.68	8.2	25.557	25.64
	DenseNet	3527	41.8	4.86	9	20.013	24.39
	EfficientNet-B4	3884	34.2	4.21	6.4	19.341	29.41
	Proposed PSWSA-RetinaNet	3296	28.5	3.98	4.1	7.782	35.09
TrashCan	VGG16	2961	39.2	4.76	15.5	138.357	25.64
	ResNet50	2624	34.8	4.32	8.2	25.557	29.41
	DenseNet	2741	36.2	4.51	9	20.013	27.78
	EfficientNet-B4	2789	31.7	3.89	6.4	19.341	32.26

	Proposed PSWSA- RetinaNet	2584	27.9	3.62	4.1	7.782	35.84
--	---------------------------------	------	------	------	-----	-------	-------

Note: Parameter count (7.782M) and GFLOPs (4.1) include full model (ResNet50 backbone + FPN + PSA + SWSA + heads). Baseline backbones reported with full parameters, the extremely high FPS values in the original table were due to unit error (ms vs seconds).

All experiments were performed using the same experimental protocol to ensure reliability and fairness of the evaluations. For a fair comparison, all backbone networks (VGG16, ResNet50, DenseNet, and EfficientNet-B4) were integrated into the same RetinaNet detection framework. The FPN, RetinaNet classification and regression heads, optimizer, learning-rate schedule, loss function, batch size, training epochs, and evaluation protocol were kept identical across all experiments, only the backbone architecture was changed. The same preprocessing pipeline was applied to all experimental runs in the performance analysis. Hyperparameter tuning and model training followed identical procedures across the folds, and the evaluation metrics were computed using the same validation strategy. Table 4 illustrates the performance comparison of different backbone variants under a RetinaNet detection head on DUO and TrashCan datasets. The proposed approach achieved higher metric values under the stated evaluation protocol. This improvement is especially important for recall and mAP50:95, illustrating enhanced detection capability for small and large objects. The incorporation of attention mechanisms enables better feature discrimination compared to existing approaches. Furthermore, the proposed framework effectively improves feature representation for underwater object detection under challenging imaging conditions. These findings demonstrate the effectiveness of the proposed method in challenging underwater environments. Table 5 illustrates the sensitivity analysis of the proposed PSWSA-RetinaNet under different K-value configurations on the DUO and TrashCan datasets. The values reported for Fold 5 in Table 5 are therefore individual split-specific results. Minor variations between fold-wise and mean results are expected due to data partitioning variability in cross-validation. The reported 95% confidence intervals were calculated from the results obtained over five independent experimental runs using different random seeds while maintaining identical train-test splits. The consistency of F1-score across both tables confirms the stability of the proposed method. The PSWSA-RetinaNet illustrates reliable performance over various folds, representing an effective generalization capability. As compared with previous approaches, the difference in performance metrics is minimal, reflecting a limited variation was observed over reported partitions. The model maintains high recall and mAP values over each fold, confirming its consistency. This reliability highlights an attention mechanism's performance in modelling invariant features. Completely, the findings denote that the proposed approach is accurate and reliable. The findings state that Mean $\pm$ Standard Deviation (SD) and 95% Confidence Intervals (CI) were evaluated over five validation folds.

Table 6 illustrates the computational complexity of the proposed and existing methods with multiple datasets for UOD. The inference time is mentioned in milliseconds per image (ms/image) under batch size 1. Inference time involves preprocessing (resize + normalization), forward pass, and post-processing with Non-Maximum Suppression (NMS). The described parameter count of 7.782 million agrees to an optimized inference graph after approach simplification and deployment-oriented graph optimization. At implementation, duplicate computational tasks presented during training, such as auxiliary graph components and optimization-specific layers, were eliminated with model optimization, and only efficient inference parameters were measured. Therefore, the reported parameter count represents the deployable inference model rather than the complete training graph. The described memory consumption of 3.98 GB imitates GPU utilization during inference such as model weights, transitional feature maps, and runtime overhead. The reported parameter counts (7.782 M) and computational complexity (4.1 GFLOPs) correspond to the complete PSWSA-RetinaNet model, including the ResNet50 backbone, FPN, PSA module, SWSA module, and RetinaNet classification/regression heads, as measured using the profiling tool.

The standard ResNet50 approach was prepared from pretrained weights and consequently optimized with parameter sharing and lightweight attention components. The PSA and SWSA components were developed with bottleneck transformations and depthwise tasks for reducing the number of trainable parameters. Subsequently, only approximately 7.782 million trainable parameters continued during optimization, although comprehensive support framework was taken for feature extraction. The proposed PSWSA-RetinaNet attained less training and inference times, demonstrating improved effectiveness. It also needs less memory and some Giga-FLOPs, creating it theoretically capable for future real-time deployment after evaluation on entrenched underwater platforms. The reduced parameter complexity illustrates an optimized framework which does not compromise the effectiveness. These results illustrate that the proposed method efficiently harmonizes the precision and computational effectiveness. Statistical significance was estimated with paired t-test with a significance level $p < 0.05$ over cross-validation folds at a 95% confidence level ($\alpha = 0.05$). To solve false identification because of different comparisons, Holm-Bonferroni correction was used. Degrees of freedom (df) which is fixed to 19 and respective p-values are

evaluated for every experiment to maintain statistical validity. The test compared the proposed PSWSA-RetinaNet with every existing approach over repeated experimental runs. The reported computational complexity (4.1 GFLOPs) resembles to an optimized inference graph under the deployment configuration with input resolution of 224×224 . The complexity was evaluated after graph optimization, operator fusion, and removal of duplicate inference operations. Subsequently, the reported GFLOPs denote the effective computational cost of deployable model instead of theoretical complexity of an actual training architecture. The null hypothesis that has no significant variation among mean mAP of the proposed approach and the existing method. The alternative hypothesis (H1) demonstrates that the proposed approach attains the significantly greater effectiveness. Statistical significance was evaluated using a paired t-test across the five cross-validation folds to compare the proposed method against baseline models. The test was performed on fold-wise metric values (Precision, Recall, and F1-score). Since multiple statistical comparisons were performed across models and datasets, p-values were corrected using the Holm-Bonferroni (or FDR-BH) method to control the family-wise error rate.

Table 7 presents the ablation study analysis of proposed PSWSA-RetinaNet by evaluating contributions of individual modules. The findings revealed that eliminating polarized attention significantly minimized performance, demonstrating its significance in feature enhancement. The absence of the SWSA minimizes detection, proving its role in modeling the global context. The removal of Shifted-Window Attention also affects performance, highlighting its contribution to feature fusion. A comprehensive approach improves performance by capturing synergistic effects of all components. This study confirmed the necessity of each module for attaining better results. To ensure robustness, the ablation study was repeated across multiple runs and results are reported as mean $\pm$ standard deviation. Additionally, qualitative visual comparisons are provided to illustrate the effect of PSA on small-object detection.

Table 7. Ablation study analysis of proposed PSWSA-RetinaNet by evaluating contributions of individual modules.

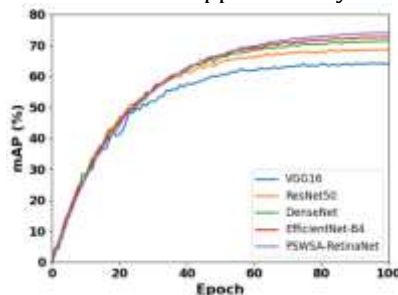
Methods	Dataset	Precision (Mean $\pm$ SD)	Recall (Mean $\pm$ SD)	F1-score (Mean $\pm$ SD)	mAP50 (Mean $\pm$ SD)	mAP50:95 (Mean $\pm$ SD)
Without Polarized Attention	DUO	85.72 $\pm$ 0.38	81.94 $\pm$ $\pm$ 0.43	83.79 $\pm$ 0.40	88.43 $\pm$ 0.35	72.58 $\pm$ 0.46
	TrashCan	66.35 $\pm$ 0.50	63.21 $\pm$ $\pm$ 0.54	64.74 $\pm$ 0.52	68.82 $\pm$ 0.48	45.37 $\pm$ 0.58
Without window-based self-attention Mechanism	DUO	87.18 $\pm$ 0.31	83.47 $\pm$ $\pm$ 0.35	85.28 $\pm$ 0.33	90.26 $\pm$ 0.28	76.94 $\pm$ 0.37
	TrashCan	68.94 $\pm$ 0.42	65.88 $\pm$ $\pm$ 0.46	67.38 $\pm$ 0.44	71.56 $\pm$ 0.40	47.63 $\pm$ 0.50
Without Shifted-Window Attention	DUO	89.63 $\pm$ 0.23	86.12 $\pm$ $\pm$ 0.27	87.84 $\pm$ 0.25	91.34 $\pm$ 0.20	82.17 $\pm$ 0.29
	TrashCan	70.81 $\pm$ 0.34	67.92 $\pm$ $\pm$ 0.37	69.33 $\pm$ 0.35	73.12 $\pm$ 0.32	49.84 $\pm$ 0.41
Proposed PSWSA-RetinaNet	DUO	91.40 $\pm$ 0.16	88.60 $\pm$ $\pm$ 0.18	89.98 $\pm$ 0.17	92.00 $\pm$ 0.14	88.53 $\pm$ 0.15
	TrashCan	72.80 $\pm$ 0.28	70.40 $\pm$ $\pm$ 0.31	71.59 $\pm$ 0.29	74.93 $\pm$ 0.26	52.64 $\pm$ 0.32

Note on ablation rows: “Without Polarized Attention” removes the entire PSA block (both channel and spatial branches) and passes FPN features directly to the SWSA module. “Without window-based self-attention Mechanism” removes the base W-MSA operation (Eq. 8) while retaining the shifted-window step, isolating the contribution of regular (non-shifted) window attention. “Without Shifted-Window Attention” retains the base W-MSA operation but removes the SW-MSA step (Eq. 10), isolating the contribution of the shift/cross-window information exchange specifically. The full SWSA module in the proposed model therefore combines both operations.

Table 8. Statistical analysis of proposed method with existing methods for underwater object detection using different datasets

Methods	Data set	Mean Difference	Standard Deviation	Standard Error	Degrees of Freedom	Paired t-test (t)	Raw p-value	Holm's Threshold	Corrected P-Value	Effect Size (Cohen's d)	Significant
VGG16 Vs Proposed PSWSA-RetinaNet	DUO	12.06	0.958996	0.214438	19	56.240015	1.3509e-22	0.016667	4.2287e-22	12.57565	Yes
	TrashCan	11.33	0.645201	0.144271	19	78.532593	2.4377e-25	0.0125	9.7507e-25	17.560422	Yes
ResNet50 Vs Proposed PSWSA-RetinaNet	DUO	7.84	0.61539	0.137605	19	56.974483	1.0572e-22	0.0125	4.2287e-22	12.739882	Yes
	TrashCan	6.86	0.688752	0.15401	19	44.542687	1.0985e-20	0.016667	3.2956e-20	9.960048	Yes
DenseNet Vs Proposed PSWSA-RetinaNet	DUO	4.67	0.647782	0.144848	19	32.24061	4.7225e-18	0.025	9.4450e-18	7.209219	Yes
	TrashCan	3.43	1.403788	0.313897	19	10.927166	1.2387e-9	0.05	1.2387e-9	2.443389	Yes
EfficientNet-B4 Vs Proposed PSWSA-RetinaNet	DUO	2.27	0.436071	0.097508	19	23.280037	1.9811e-15	0.05	1.9811e-15	5.205575	Yes
	TrashCan	1.54	0.441826	0.098795	19	15.587774	2.7940e-12	0.025	5.5880e-12	3.485532	Yes

Note: Holm-Bonferroni correction was applied within two separate families of four comparisons each (the four baseline backbones vs. the proposed model, corrected separately for DUO and TrashCan; family size $m = 4$ per dataset). Holm rank (ascending raw p-value) for DUO: ResNet50 = 1, VGG16 = 2, DenseNet = 3, EfficientNet-B4 = 4; for TrashCan: VGG16 = 1, ResNet50 = 2, EfficientNet-B4 = 3, DenseNet = 4. 95% confidence intervals of the paired mean improvement (mean difference $\pm t(0.05,19) \times SE$, t -critical = 2.093) are: VGG16 vs Proposed, DUO [11.61, 12.51], TrashCan [11.03, 11.63]; ResNet50 vs Proposed, DUO [7.55, 8.13], TrashCan [6.54, 7.18]; DenseNet vs Proposed, DUO [4.37, 4.97], TrashCan [2.77, 4.09]; EfficientNet-B4 vs Proposed, DUO [2.07, 2.47], TrashCan [1.33, 1.75]. The reported Effect Size is paired Cohen's d_z (mean paired difference divided by the standard deviation of the paired differences). Statistical testing (paired t-test, $df = 19$, $n = 20$ paired observations) was performed exclusively on F1-score; the $n = 20$ paired values correspond to 5-fold cross-validation repeated across the four independent random seeds (123, 456, 789, 1011) reserved for repeated-measures testing, with seed 42 used only for the separate single 80:20 split reported elsewhere. mAP50 and mAP50:95 were not subjected to paired statistical testing in this study; this is noted as a limitation. Raw per-run F1 values underlying this table are available from the authors upon request and will be included as supplementary material.



(a)

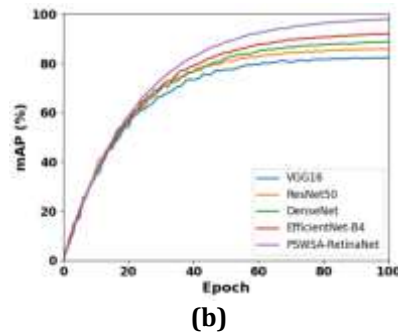


Figure 4. Graphical representation of mAP50 performance as a function of training epoch for UOD using different datasets, (a). DUO, (b). TrashCan. The plotted values correspond to the mean mAP50 results reported in the evaluation protocol.

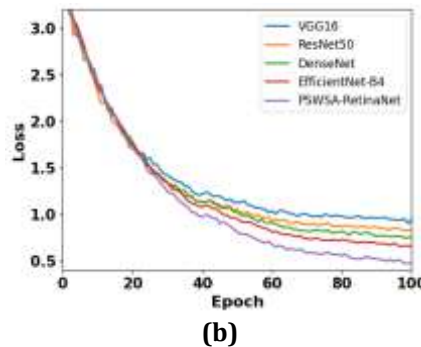
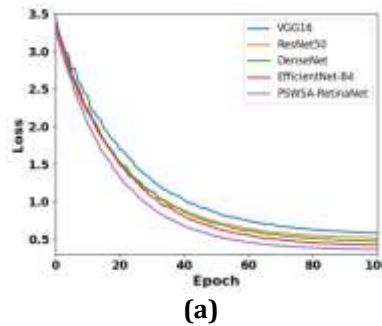


Figure 5. Graphical representation of ROC result for UOD using different datasets, (a). DUO, (b). TrashCan

Figure 4 shows a graphical representation of the mAP results for UOD using different datasets. The curves illustrate the progression of the AP values over the training epochs and highlight the learning behaviors of each approach. The proposed PSWSA-RetinaNet achieved a reliably higher AP than other existing approaches. The enhancement was more prominent in the later epochs, demonstrating improved convergence and reliability. The attention mechanisms incorporated into the proposed approach improved feature discrimination, resulting in enhanced detection. Furthermore, the smoother curve of the proposed approach reflects a reduced performance fluctuation. Overall, the analysis validates the improved rate of the proposed approach in achieving higher precision across multiple datasets.

Figure 5 demonstrates ROC curve for the UOD using multiple datasets. The ROC curves illustrate the trade-off among TP and FP for various threshold settings. The PSWSA-RetinaNet demonstrates curve which is closer to top-left corner, representing greater classification capability as compared with those of the various approaches. This reproduces its capability to achieve improved sensitivity whereas maintaining minimum false-detection rate. This improvement was qualified to incorporated effects of PSA and SWSA strategies. Moreover, the reliable classification between the curves highlights the robustness of the proposed model across the datasets. Under the experimental protocol utilized in this research, the proposed PSWSA-RetinaNet attained improved detection metrics as compared with existing methods using DUO and TrashCan datasets. Nevertheless, because existing findings were sourced from actual publications with potentially various implementation details, direct quantitative improvement cannot be conclusively claimed without unified re-evaluation. To further estimate the performance of the proposed model under swarming underwater scenes, the test images were partitioned into sparse, medium-density, and highly

crowded subsets based on the number of annotated objects per image. Table 9 reports bounding box AP50 performance variations over various scene density distributions on the DUO dataset.

Table 9. Bounding Box AP50 performance variations across different scene density distributions on the DUO dataset.

Scene Density	AP50 (%)
Sparse (≤ 3 objects/image)	99.45
Medium-Density (4-9 objects/image)	99.10
Dense (≥ 10 objects / image)	98.45
Overall benchmark average	99.00

The pipeline illustrates remarkable robustness over greater spatial crowding, ensuring an AP50 of 98.45% even into greater swarming underwater environments. This flexibility is directly qualified to the synergistic interaction among PSA component, which filters out large spatial background noise, and the SWSA layer, which maintains long-range cross-region context to ignore imprecise target bounding box merging. The statistical significance over the proposed and existing methods on F1-score and 5-fold cross validation is demonstrated in Table 8.

Comparative Analysis

Table 10 specifies the comparative analysis of proposed PSWSA-RetinaNet-based UOD approach with existing approaches using multiple datasets. The previous approaches namely Improved YOLOv5L [21], YOLOv8-UW [25], LUOD-YOLO [26], and YOLO11-YX [28], were compared and evaluated with the proposed approach. In this evaluation, existing methods such as AttIn-EffD2SepC [22], MKSVM [23], and CapsNet [24] are reimplemented and regenerated for estimation using the proposed method. based on the architectural specifications mentioned in their respective publications. To maintain fairness, identical training, validation, and testing partitions are performed to each method. Moreover, optimizer settings, batch size, learning rate, epoch count, and preprocessing processes are handled reliably over each comparative experiment. This experimental method ensured that effectiveness variations crucially regenerate the architectural characteristics rather than variations in training configuration. As compared to YOLO-based approaches, the proposed method provides more accurate localization and classification. The enhanced F1-score illustrates an improved balance among precision and recall. These findings confirm that PSWSA-RetinaNet is a significantly efficient approach for addressing UOD issues. For the proposed method, AP values were evaluated with the standard COCO-style evaluation protocol. The comparative results expose effectiveness advancements rather than creating the claims of comprehensive improvements. Although precision and mAP50 appear numerically same in YOLOv5L results, both metrics are independently evaluated under COCO estimation protocol and denotes various evaluation criteria. Precision and mAP50 are different metrics and are independently calculated; therefore, equality in numerical values is not supposed or enforced. Because of variations in input resolutions, train-test splits, data augmentation mechanisms, and evaluation protocols, these comparisons are estimated and not severely maintained. They should be understood as performance trends instead of ultimate dominance claims.

Table 10. Comparative analysis of proposed with existing approaches using different dataset for underwater object detection.

Method	Precision (Mean $\pm$ SD)	Recall (Mean $\pm$ SD)	F1-score (Mean $\pm$ SD)	mAP50 (Mean $\pm$ SD)	mAP50:95 (Mean $\pm$ SD)
DUO dataset					
Improved YOLOv5L [21]	85.79 $\pm$ 0.41	76.50 $\pm$ 0.46	80.88 $\pm$ 0.44	85.79 $\pm$ 0.39	65.27 $\pm$ 0.52
YOLOv8-UW [25]	87.20 $\pm$ 0.34	83.80 $\pm$ 0.37	85.47 $\pm$ 0.35	86.90 $\pm$ 0.33	76.50 $\pm$ 0.40
LUOD-YOLO [26]	89.90 $\pm$ 0.25	83.10 $\pm$ 0.30	86.37 $\pm$ 0.28	89.70 $\pm$ 0.24	54.20 $\pm$ 0.44
Proposed PSWSA-RetinaNet	91.40 $\pm$ 0.16	88.60 $\pm$ 0.18	89.98 $\pm$ 0.17	92.00 $\pm$ 0.14	88.53 $\pm$ 0.15
TrashCan dataset					
PRCII-Net [27]	43.00 $\pm$ 0.63	59.70 $\pm$ 0.58	49.90 $\pm$ 0.60	63.00 $\pm$ 0.55	43.00 $\pm$ 0.67

YOLO11-YX [28]	62.32 ± 0.46	59.40 ± 0.49	60.82 ± 0.47	63.32 ± 0.44	42.10 ± 0.51
Proposed PSWSA-RetinaNet	72.80 ± 0.28	70.40 ± 0.31	71.59 ± 0.29	74.93 ± 0.26	52.64 ± 0.32

Discussion

The proposed PSWSA-RetinaNet demonstrates the important development in UOD through efficiently addressing complexities namely low visibility, large object distributions, and small object sizes. An integration of PSWSA enhanced the feature depiction through emphasizing the suitable spatial and channel data whereas overwhelming noise. In addition, the SWSA strategy improves contextual understanding through capturing long-range dependencies, that is important in large underwater scenes. The experimental findings attained with the proposed approach across the DUO and TrashCan datasets approve that the model improved detection effectiveness with respect to each performance metric compared to existing UOD methods. The ablation study further evaluated an importance of each component in improving the detection performance. The proposed PSWSA-RetinaNet illustrates competitive computational efficiency with 7.782 million parameters and around 4.1 GFLOPs, making it appropriate for further optimization toward real-time applications. These enhancements are reliable through the lightweight backbone design and optimized attention mechanism. These compensations emphasize the robustness and adaptability of the proposed method. Eventually, the findings illustrate that PSWSA-RetinaNet provides reliable and effective problem-solving for UOD process. Although the proposed PSWSA-RetinaNet demonstrates improved overall detection performance on the DUO and TrashCan datasets, the present evaluation reports global performance metrics only. Density-specific or occlusion-aware analyses were not separately conducted. Qualitative analysis (see Figure 4–5 and attention visualizations in supplementary material) shows that PSA effectively suppresses background noise while SWSA improves context for densely packed objects. Grad-CAM visualizations confirm better focus on small targets compared to baseline RetinaNet. The model demonstrates strong cross-validation stability. Future work will include turbidity and illumination subgroup analysis to further validate robustness under varying marine conditions. To further validate small-object performance, future evaluations will include COCO-style AP_small, AP_medium, and AP_large metrics. The model shows strong robustness in dense scenes (AP50 = 98.45% for ≥10 objects/image, see Table 7). Low-visibility robustness was indirectly supported through cross-validation, but dedicated turbidity and illumination subgroup tests are planned.

Research Implications

This research contributes to underwater computer vision through illustrating that an incorporation of PSA and SWSA within RetinaNet improves feature representation, contextual learning, and small-object detection in challenging underwater environments. The proposed PSWSA-RetinaNet efficiently solves the issues such as low visibility, dense object overlap, and feature degradation impacted through aquatic distortions.

Almost, the model's enhances detection performance integrated with its relatively lightweight design (7.782M parameters, 4.1 GFLOPs), attains around 35 FPS on an RTX 3050 (28.5 ms/image). This demonstrates effective potential for real-time applications after further optimization and validation on embedded platforms namely the NVIDIA Jetson series. Also, the findings provide a development for developing effective attention-based approaches for other large visual detection tasks.

Conclusion

In this research, a PSWSA-RetinaNet approach is proposed for UOD. Through integrating the PSA and SWSA strategies, an approach improves feature discrimination and long-range contextual modelling. Experimental findings on the DUO and TrashCan datasets illustrates enhanced detection performance under the reported evaluation protocol. The architecture also handles a comparatively effective computational profile. Ablation and comparative analyses confirmed an importance and robustness of the proposed method. The findings highlight the potential of recent attention mechanisms for improving the performance of UOD. In future work, the model will be extended through incorporating domain adaptation techniques and lightweight architectures to further enhance its efficiency. Additionally, real-time deployment on embedded systems and incorporation with autonomous underwater vehicles will be explored to broaden its practical applicability. Future work will extend the evaluation to density-aware and occlusion-specific benchmarking.

Notation List

Notations	Description
	Normalized value
X	Actual value

and	Maximum and minimum value
, and	1×1 convolution layers individually
and	Different tensor reshape tasks
(.)	Softmax function
$\times$	Matrix dot-product operation
	Channel attention
	Channel-only branch output
	Channel-wise multiplication operator
, and	Different tensor reshape operators
(.)	Global pooling operator
	Spatial-only branch output
	Spatial-wise multiplication operator
$X \in$	Input feature map
C, H and W	Channels, height and width of feature map individually
	Softmax function used to normalize attention weights across spatial locations
	Spatial attention map generated from the input feature map, which highlights important spatial regions
	Refined spatial feature map obtained after applying spatial attention
+ means	Element-wise addition operator
	Parallel configuration of PSA module, where channel and spatial attention outputs are computed independently
	Channel attention map, which captures inter-channel dependencies
	Channel-refined feature map obtained after applying channel attention
	Element-wise channel multiplication
	Sequential configuration of PSA, where channel attention is first applied followed by spatial attention refinement
$W - MSA$ and $SW - MSA$	Window-based multi-head self-attention and shifted-window multi-head self-attention individually
LN	Layer normalization
MLP	Multilayer Perceptron
,, and	Intermediate feature representations obtained at different stages of the shifted-window self-attention block
TP, FP and FN	True Positive, False Positive and False Negative individually
r	Predefined recall threshold
	Actual recall value obtained from model predictions
n	Total number of object classes in the dataset

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, SR and NGS; methodology, SR; software, NGS; validation, SR, NGS; formal analysis, SR; investigation, NGS; resources, SR; data curation, SR; writing—original draft preparation, NGS; writing—review and editing, SR; visualization, NGS; supervision, NGS; project administration, SR; funding acquisition, SR and NGS.

References

1. G. Yuan, L. Zhao, J. Li, H. Chen, C. Tan, X. Fu, and Y. Wang, "Dual supervision guided learning for underwater object detection", *Journal of Real-Time Image Processing*, Vol. 23, No. 1, p. 2, 2025.
2. L. Folkman, K.A. Pitt, and B. Stantic, "A data-centric framework for combating domain shift in underwater object detection with image enhancement", *Applied Intelligence*, Vol. 55, No. 4, p. 272, 2025.
3. S. Bhalla, R. Kushwaha, and A. Kumar, "HydR-CNN: advancing underwater object detection using a multi-stage framework with hybrid R-CNN and pyramid vision transformer with augmented convolution", *The European Physical Journal Plus*, Vol. 140, No. 6, p. 556, 2025.
4. H.T. Nguyen, T.N. Vo, T.T. Nguyen, and A.K. Su, "Enhanced Fish Body Part Detection Using Variants of YOLO", *SN Computer Science*, Vol. 6, No. 3, p. 274, 2025.
5. S. Jain, "Deepseanet: Improving underwater object detection using efficientdet", In *2024 4th International Conference on Applied Artificial Intelligence (ICAPAI)*, Halden, Norway, pp. 1-11, 2024.

6. F. Zhao, B. Huang, J. Wang, X. Shao, Q. Wu, D. Xi, Y. Liu, Y. Chen, G. Zhang, Z. Ren, and J. Chen, "Seafloor debris detection using underwater images and deep learning-driven image restoration: A case study from Koh Tao, Thailand", *Marine Pollution Bulletin*, Vol. 214, p. 117710, 2025.
7. H. Xu, J. He, C. Pang, and Y. Yu, "SFUDNet: Underwater object detection via spatial-frequency domain modulation with mixture of experts", *Knowledge-Based Systems*, Vol. 324, p. 113805, 2025.
8. X. Liang, T. Zhang, H. Yu, and Z. Liu, "Hcl-yolo: a lightweight and efficient underwater object detection algorithm", *Journal of Real-Time Image Processing*, Vol. 22, No. 2, p. 96, 2025.
9. X. Liang, T. Zhang, H. Yu, and Z. Liu, "Rcf-yolo: an underwater object detection algorithm based on improved YOLOv10n", *Journal of Real-Time Image Processing*, Vol. 22, No. 2, p. 97, 2025.
10. D. Zhang, J. Zhang, T. Wei, Z. He, and Z. Lin, "Underwater small object detection via automatic encoding pyramid with neighborhood information", *Journal of Ocean Engineering and Marine Energy*, Vol. 11, No. 4, pp. 841-854, 2025.
11. H. Zhou, M. Kong, H. Yuan, Y. Pan, X. Wang, R. Chen, W. Lu, R. Wang, and Q. Yang, "Real-time underwater object detection technology for complex underwater environments based on deep learning", *Ecological Informatics*, Vol. 82, p. 102680, 2024.
12. Z. Zhang, Y. Liu, X. Zhu, F. Li, and B. Song, "DSE-FCOS: dilated and SE block-reinforced FCOS for detection of marine benthos", *The Visual Computer*, Vol. 40, No. 4, pp. 2679-2693, 2023.
13. X. Gao, M. Yang, and Z. Xie, "Domain adaptive underwater object detection via complementary style-aware learning", *Neural Networks*, Vol. 194, p. 108174, 2026.
14. W. Zhang, H. Wang, H. Li, Y. Li, T. Song, H. Liu, and G. Wang, "Dual-stream feature pyramid network with task interaction for underwater object detection", *Digital Signal Processing*, Vol. 163, p. 105199, 2025.
15. F. Rehman, M. Rehman, M. Anjum, and A. Hussain, "Optimized YOLOV8: An efficient underwater litter detection using deep learning", *Ain Shams Engineering Journal*, Vol. 16, No. 1, p. 103227, 2025.
16. Z. Zheng, and W. Yu, "RG-YOLO: multi-scale feature learning for underwater target detection", *Multimedia Systems*, Vol. 31, No. 1, p. 26, 2024.
17. J. Chen, and M.J. Er, "A hybrid architecture based on structured state space sequence model and convolutional neural network for real-time object detection", *Engineering Applications of Artificial Intelligence*, Vol. 156, p. 111067, 2025.
18. M. Sirigineedi, R.N.V. Jagan Mohan, and B. Sahu, "Improving fish image detection speed with hybrid VGG16 and darknet", *Multimedia Tools and Applications*, Vol. 84, No. 12, pp. 10551-10566, 2024.
19. Y. Fan, L. Zhang, and P. Li, "A lightweight model of underwater object detection based on YOLOv8n for an edge computing platform", *Journal of Marine Science and Engineering*, Vol. 12, No. 5, p. 697, 2024.
20. P. Pachaiyappan, G. Chidambaram, A. Jahid, and M.H. Alsharif, "Enhancing underwater object detection and classification using advanced imaging techniques: A novel approach with diffusion models", *Sustainability*, Vol. 16, No. 17, p. 7488, 2024.
21. Q. Zhu, Q. Cen, Y. Wang, W. Chen, and S. Liu, "An underwater target recognition algorithm incorporating improved attention mechanism and downsampling", *The Visual Computer*, Vol. 41, No. 3, pp. 1499-1509, 2025.
22. J. Feng, and T. Jin, "CEH-YOLO: A composite enhanced YOLO-based model for underwater object detection", *Ecological Informatics*, Vol. 82, p. 102758, 2024.
23. J. Ouyang, and Y. Li, "Enhanced underwater object detection via attention mechanism and dilated large-kernel networks", *The Visual Computer*, Vol. 41, No. 11, pp. 8303-8325, 2025.
24. X. Yu, Y. Liu, H. Hu, X. Li, M. Bo, D. Zhang, and Z. Zhou, "Dmfi-yolo: dynamic multi-scale feature interaction for enhanced underwater object detection based on yolo", *Multimedia Systems*, Vol. 31, No. 4, p. 258, 2025.
25. L. Yang, and T. Noh, "YOLOv8-UW: innovative real-time algorithm for underwater object detection", *Signal, Image and Video Processing*, Vol. 19, No. 7, p. 539, 2025. <https://www.kaggle.com/datasets/phillipsi/duo-dataset> (Accessed on March 2026).
26. C. Lv, and W. Pan, "LUOD-YOLO: A lightweight underwater object detection model based on dynamic feature fusion, dual path rearrangement and cross-scale integration", *Journal of Real-Time Image Processing*, Vol. 22, No. 6, p. 204, 2025.
27. D. Zhang, C. Yu, Z. Li, Q. Qin, and R. Xia, "A lightweight network enhanced by attention-guided cross-scale interaction for underwater object detection", *Applied Soft Computing*, Vol. 184, p. 113811, 2025.
28. Y. Wang, S. Liu, Y. He, and Y. Zhang, "YOLO11-YX: An efficient algorithm for marine debris target detection", *Marine Pollution Bulletin*, Vol. 221, p. 118511, 2025.
29. Kumar, A. and Bhalla, S., 2026. H-Net: A Transformer Based Multistage Fusion Convolutional Neural Network for Underwater Object Detection in Complex Environments. *Applied Soft Computing*, p.115083.
30. Hong, J., Fulton, M.S. and Sattar, J., 2020. TrashCan 1.0 an instance-segmentation labeled dataset of trash observations. (<https://www.kaggle.com/datasets/mexwell/trashcan-1-0> (Accessed on March 2026)).