AMSP-UOD: When Vortex Convolution and Stochastic Perturbation Meet Underwater Object Detection

Jingchun Zhou[1*], Zongxin He[2*], Kin-Man Lam[3], Yudong Wang[4], Weishi Zhang[1], Chunle Guo[5], Chongyi Li[5†]

[1] School of Information Science and Technology, Dalian Maritime University
[2] School of Computer Science and Engineering, Huizhou University
[3] Department of Electrical and Electronic Engineering, Hong Kong Polytechnic University
[4] School of Electrical and Information Engineering, Tianjin University, China
[5] VCIP, CS, Nankai University
zhoujingchun03@gmail.com, hikari0608@outlook.com, enkmlam@polyu.edu.hk, yudongwang@tju.edu.cn, teesiv@dlmu.edu.cn, {guochunle, lichongyi}@nankai.edu.cn

Abstract

In this paper, we present a novel Amplitude-Modulated Stochastic Perturbation and Vortex Convolutional Network, AMSP-UOD, designed for underwater object detection. AMSP-UOD specifically addresses the impact of non-ideal imaging factors on detection accuracy in complex underwater environments. To mitigate the influence of noise on object detection performance, we propose AMSP Vortex Convolution (AMSP-VConv) to disrupt the noise distribution, enhance feature extraction capabilities, effectively reduce parameters, and improve network robustness. We design the Feature Association Decoupling Cross Stage Partial (FAD-CSP) module, which strengthens the association of long and short range features, improving the network performance in complex underwater environments. Additionally, our sophisticated post-processing method, based on Non-Maximum Suppression (NMS) with aspect-ratio similarity thresholds, optimizes detection in dense scenes, such as waterweed and schools of fish, improving object detection accuracy. Extensive experiments on the URPC and RUOD datasets demonstrate that our method outperforms existing state-of-the-art methods in terms of accuracy and noise immunity. AMSP-UOD proposes an innovative solution with the potential for real-world applications. Our code is available at: https://github.com/zhoujingchun03/AMSP-UOD.

Introduction

Recently, underwater object detection (UOD) has gained attention in the fields of marine technology, deep-sea exploration, and environmental protection. Precise detection of biological, geological, and man-made structures in deep-sea environments is vital for human society and environmental conservation (Xu et al. 2023) (Zhuang et al. 2022). However, challenges in seawater, such as transparency, color, temperature, and suspended particles, combined with varying marine environments and target object types, reduce the accuracy of object detection.

Due to light absorption and scattering (Zhou et al. 2023b; Zhang et al. 2022), underwater imaging often suffers from quality degradation compared to object detection in high-quality images. This impacts the performance of Convolutional Neural Network (CNN)-based object detectors. The key challenges include 1) the lack of underwater object detection datasets hindering the training of deep learning models, 2) degradation factors, such as light absorption and scattering, leading to low contrast and color distortion (Zhou et al. 2023a) (Guo et al. 2022), 3) the difficulty in extracting rich details from small and clustered underwater objects, and 4) class imbalance, making the challenging for object detectors to learn features for classes with a small-sample size (Fu et al. 2023). To address these challenges, new detectors capable of accurate localization and classification in complex underwater environments are required. This research aims to advance ocean science and deep-sea exploration technology and holds practical value in environmental protection and resource development.

In this paper, we propose the AMSP-UOD network, crafted to tackle non-ideal imaging factors in underwater environments. Utilizing the optical imaging model $I = H(J, B, t) + N$ (I represent observed image, J represent raw scene, B represents backscatter, and t represent transmission map), we discern that underwater images combine a degradation function H with noise N. To remove noise, we propose the AMSP-VConv. This strategy not only reduces parameters but also bolsters the network's robustness. We further implement the FAD-CSP to improve feature extraction in degraded environments. Our post-processing strategy, which relies on NMS, is designed to optimize the detection of dense clusters of underwater objects. Experimental results on the URPC (Liu et al. 2021) and RUOD (Fu et al. 2023) datasets showcase the effectiveness of our method. Overall, AMSP-UOD presents an innovative solution for UOD with potential real-world applications.

The main contributions of this paper are as follows:

(1) We propose a novel single-stage UOD network. In the backbone, we design the AMSP-VConv to address the impact of noise and other degradations in underwater object

*These authors contributed equally.
†Corresponding author

detection. In the neck, the FAD-CSP boosts long and short distance feature connection, enhancing performance in complex underwater environments. Furthermore, an NMS-based post-processing method is introduced to enhance the detection performance of the network in complex underwater scenarios like dense waterweed clusters and fish schools.

(2) Our AMSP strategy refines network through parameter adjustments, enhancing detection by distinguishing between ideal and non-ideal imaging factors.

(3) Experimental results on public datasets and UOD competition datasets reveal that our method outperforms state-of-the-art UOD techniques in terms of both detection accuracy and speed. Ablation studies demonstrate that AMSP-VConv possesses superior noise resistance and interpretability, offering a novel solution for noise processing in detection tasks and computer vision.

Related Work

The UOD task focuses on detecting objects in underwater images. Deep learning has significantly advanced information fusion (Ma et al. 2023), image enhancement (Liu et al. 2023), and object detection (Chen et al. 2022). In many cases, these methods outperform traditional approaches, in terms of speed and accuracy (Liu et al. 2016; Ren et al. 2015; Redmon et al. 2016). However, underwater environments introduce image degradations due to factors like light attenuation. Underwater robots also need efficient algorithms due to limited resources. Existing UOD techniques are either anchor-based or anchor-free, with variations in their approach.

Anchor-Based Methods

Single-Stage Methods: These methods predict the object's location and type directly, ensuring faster performance. Examples include SSD (Liu et al. 2016) that leverages feature pyramids for multi-scale perception, RetinaNet (Lin et al. 2017) using Focal Loss for sample weight adjustment, and NAS-FPN (Ghiasi et al. 2019) that refines feature pyramid network structures. While efficient, they can struggle with precise object boundary localization in UOD tasks, especially in challenging conditions or with limited data. Data augmentation is often used to enhance generalization.

Multi-Stage Methods: These techniques split detection into two stages: region proposal and object classification with bounding box prediction. Examples include Faster R-CNN (Ren et al. 2015), Cascade R-CNN (Cai and Vasconcelos 2018), DetectoRS (Qiao, Chen, and Yuille 2020), and Dynamic R-CNN (Zhang et al. 2020a). They enhance accuracy using cascaded detectors, novel pyramid networks, and balanced learning (Ren et al. 2015; Cai and Vasconcelos 2018; Qiao, Chen, and Yuille 2020; Zhang et al. 2020a). However, their computational demands pose challenges for on-the-go applications.

Anchor-Free Methods

Key-Point Based Methods: These techniques use key-points, either predefined or self-learned, for detection, offering finer object boundary detail. Examples are Reppoints (Yang et al. 2019) for learning object-related features, Grid (Tian et al. 2019) for grid-guided detection, and CenterNet (Zhou, Koltun, and Krähenbühl 2020) and ExtremeNet (Zhou, Zhuo, and Kr"ahenb"uhl 2019) that use multiple key-points. While effective in general OD, their application in UOD is challenging due to limited underwater datasets (Fu et al. 2023), manual annotations, and computational demands conflicting with UOD's typical scenarios.

Center-Point Based Methods: These methods focus on predicting object center points, ideal for dense and fast detections. Notably, YOLO (Redmon et al. 2016) approaches detection as a single regression task, optimizing dense object detection. Enhancements include per-pixel prediction and feature abstraction (Redmon et al. 2016; Tian et al. 2019; Zhu, He, and Savvides 2019; Kong et al. 2020; Liu et al. 2019). However, their scalability for various object sizes is limited, and they may not excel in tasks needing precise boundary localization, like specific underwater robot operations.

Proposed AMSP-UOD Network

The underwater environment is marked by complexity due to various regular and irregular degradation factors, including marine biological activity, human activity and current movement (Chou et al. 2021). These factors create unpredictable noise patterns, posing challenges to models attempting to perceive and model underwater degradation scenes. Underwater noise is complex compared (Li et al. 2019) to typical noise conditions and requires a higher parameter count to denoise, but this increases the risk of overfitting. Instead of focusing on modeling noise, we propose a novel UOD network, namely ASMP-UOD (in Figure 1). Our approach aims to disrupt noise and reduce parameters, focusing on extracting ideal features rather than increasing the burden of noise analysis. Unlike previous methods that struggled with complex scenarios, ASMP-UOD is designed to better adapt to regular underwater scenes.

Anti-Noise Capability of AMSP and VConv

Convolution and its variants (Chollet 2017)(Han et al. 2020) are crucial for feature extraction but often struggle in scenarios with noise interference or complex scenarios. The challenge lies in distinguishing between background features and target object features, limiting detection accuracy. To deal these issues, we design a novel AMSP-VConv to mitigate noise interference, enhancing the network's adaptability in underwater scenarios.

Inspired by the vortex phenomenon in turbulent water flows, which disrupts continuity through rapid rotation, AMSP-VConv introduces 'vortices' in the information flow to break the interference caused by noise. This innovation improves the network's ability to differentiate background and target features, enhancing detection in complex underwater environments.

In Figure 2, we present the complete structure of AMSP-VConv. Starting with an input tensor F_{in} of the size $[b, c, h, w]$ (b: batch size, c: number of channels, h: height,

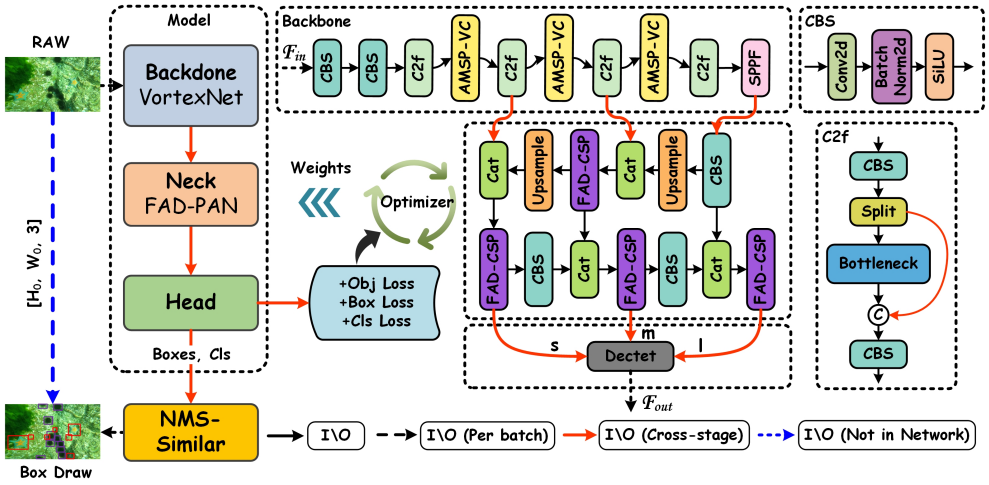

Figure 1: AMSP-UOD network architecture: AMSP-VConv for underwater noise elimination; FAD-PAN for information analysis, FAD-CSP for semantic feature decoupling; NMS-Similar for merging traditional and Soft-NMS for efficient dense scene detection.

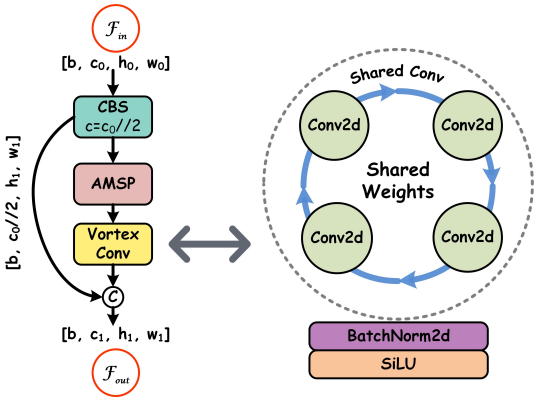

Figure 2: AMSP Vortex Convolution. (a) The AMSP-VConv structure, (b) an expanded diagram of VConv, featuring the uniquely designed Shared Conv with BN, complemented by the SiLU activation function.

w: width), it is processed by the combination of convolution, batch normalization, and the SiLU activation function (CBS) structure. This structure is designed to capture latent associations. By utilizing a kernel size of size 3 and a step size of 1, yielding an output tensor X of size $[b, c//2, h, w]$. The transformation can be expressed as follows:

$$X = CBS(F_{in}) = \delta(BatchNormal(Conv(F_{in}))) \quad (1)$$

where δ represents the SiLU activation function. As illustrated in Equations (2) and (3), we introduce the Amplitude Modulation and Shuffling Perturbation (AMSP) strategy in the subsequent steps. This strategy infuses random perturbations into the original grouped structure of associated features within X, thereby disrupting the association between noise and regular features. It is crucial to highlight that, while the AMSP strategy introduces these perturbations, it does not annihilate the features. Instead, it preserves a majority of the feature associations and induces a random shuffling among channels. This mechanism effectively serves the connection between noise samples and regular features, especially in the higher-level channels.

$$T = AM_t(X) = \begin{bmatrix} c_1 & c_2 & \cdots & c_t \\ c_{t+1} & c_{t+2} & \cdots & c_{2t} \\ \cdots & \cdots & \cdots & \cdots \\ c_{kt+1} & c_{kt+2} & \cdots & c_{kt+t} \end{bmatrix} \quad (2)$$

$$Y = SP_t(T) = \begin{bmatrix} c_{a_0t+1} & c_{a_0t+2} & \cdots & c_{a_0t+t} \\ c_{a_1t+1} & c_{a_1t+2} & \cdots & c_{a_1t+t} \\ \cdots & \cdots & \cdots & \cdots \\ c_{a_kt+1} & c_{a_kt+2} & \cdots & c_{a_kt+t} \end{bmatrix} \quad (3)$$

$$\{a_0, a_1, \ldots, a_k\} = \{0, 1, \ldots, k\} \quad (4)$$

As depicted in Equations (2) and (3), the process involves two primary operations: Amplitude Modulation (AM) and Shuffling Perturbation (SP). AM is responsible for mapping the information to higher dimensions, SP perturbs these features. Here, we divide the channels into $k+1$ groups of t channels each, c_i denotes the i-th channel, and Y is the output of the AMSP, which is aligned with the dimensions of the intermediate variable T.

$$Z = Concat(VConv(Y)) \quad (5)$$

$$Z^{'} = \delta(BatchNormal(Z)) \quad (6)$$

The VConv processes the reconstructed result Z to optimize the extracted features. Drawing a parallel with the ideal state of water vortices, vortex convolution comprises multiple spiral lines (group convolutions) with a fixed spacing (shared convolution parameters). These group convolutions capture and extract features according to global and local imaging rules, removing isolated noise.

$$F_{out} = Concat(X, Z^{'}) \quad (7)$$

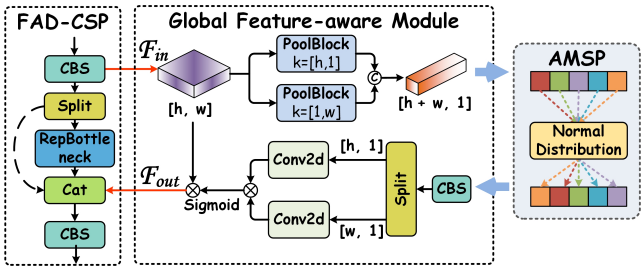

Figure 3: FAD-CSP structure. The FAD-CSP module is built on a cross-stage network, comprising an efficient Global Feature-Aware (GFA) and a local decoupling-focused RepBottleneck. The essence of FAD-CSP lies in creating an efficient decoupling network through the interaction of long and short distance features.

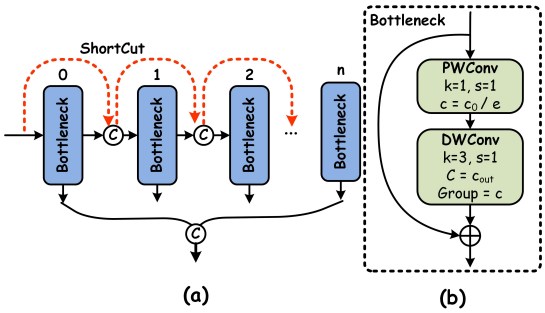

Figure 4: (a) The represents the RepBottleneck structure with n=3, and (b) the detailed design of Bottleneck, a residual structure composed of pointwise convolution and depthwise convolution.

To ensure the integrity of the correct feature semantic information, we employ residual connections to concatenate the original associated features X and $Z^{'}$ to obtain the final output F_{out}, with shape $[b, c, h, w]$. This method can better adapt to feature attenuation and noise effects, thereby acquiring complete and correct features of the ideal degradation scenario under the guidance of the gradient optimizer.

Feature Association Decoupling CSP

In order to extract features at different distances for enhancing adaptability to underwater environments, we introduce a feature association and decoupling module based on a cross-stage network (FAD-CSP). This module is designed to incorporate the novel global feature-aware approach to extract long-range global features, while utilizing the optimized RepBottleneck as a sampling module to capture short-range local features.

Global Feature-Aware Representation: In convolutional operations for global feature processing, a deeper network structure is usually required to extract rich feature information. This often increases the likelihood of the network getting trapped in local optima. To address this issue, we devised an efficient global feature-aware module and seamlessly integrated it into the FAD-CSP network using an attention mechanism. The structure of the proposed global feature-aware module is depicted in Figure 3. For a given input tensor F_{in}, we processing it through a bar-shaped pooling group, which compresses salient features into a one dimensional space. This method not only trades longer distance feature correlations, but also exhibits much lower computational overhead compared to convolution. This process can be expressed as follows:

$$v_c^h = AvgPool_h(F_{in}) + MaxPool_h(F_{in}) \tag{8}$$
$$v_c^w = AvgPool_w(F_{in}) + MaxPool_w(F_{in}) \tag{9}$$

The variables v_h and v_w are introduced into the subsequent stage of global feature-aware processing. Utilizing the AMSP strategy, they undergo a random alternation. This procedure generates a distribution map highlighting global prominent features as follows:

$$y_f = CBS(AMSP(Cat_c([v_c^h, v_c^w]))) \tag{10}$$
$$y_{c/r}^h, y_{c/r}^w = split_c(y_f) \tag{11}$$

where c denotes the channel count of this intermediary value and r is the scaling ratio. The CBS is employed and Cat_c denotes concatenation by channels to rebuild feature relationships y_f, extracting accurate long-distance feature associations.

$$y_c^h = Conv(y_{c/r}^h), y_c^w = Conv(y_{c/r}^w) \tag{12}$$

To capture long-range features from the attention map, we incorporate a weighted gradient flow. This ensures the retention of valuable information in the output and upholds the consistency within the original feasible solution domain. Specifically, the adaptive weighting stems from redistributing two distinct linear feature sets sourced from independent conventional mappings and harnessing the capabilities of the Sigmoid function. This method allows for adaptive weighting of the features, according to the disparities in feature importance within specific regions, ensuring a refined adjustment to feature changes across different areas.

$$A_f = Sigmoid(y_c^h \times y_c^w) \tag{13}$$
$$F_{out} = A_f \bigodot F_{in} \tag{14}$$

Ultimately, the global attention A_f, obtained by weighting the product of the strip attention maps $y_c^h \times y_c^w$ followed by the Sigmoid function, is multiplied with the input F_{in}. This generates an expanded solution domain F_{out}, reinterpreted by the global feature perception decoupling module. It provides the network with a richer and optimized feature representation. Introducing the attention map allows the network to better understand and process features from different regions while retaining key information. This aids the network in achieving global optima, enhancing the performance of the UOD task.

RepBottleneck: This is an efficient residual structure, as shown in Figure 4 (b), which uses a combination of depthwise separable convolutions and residual connections to reduce the number of network parameters, aggregating local

features within the receptive field. Our RepBottleneck is an optimization over the Bottleneck List, addressing the deficiency in global feature-awareness. RepBottleneck focuses on local representations at a short distance. The proposed RepBottleneck is depicted in Figure 4(a), which interconnects multiple Bottlenecks and ShortCuts, to enhance the degree of association between local features. It is expressed as follows:

$$I_n = \begin{cases} Bottleneck((I_0), & if\ n = 1 \\ Bottleneck(Cat(I_{n-1}, I_{n-2})), & if\ n \neq 1 \end{cases} \quad (15)$$

where I_n represent the output of RepBottleneck and I_0 represent the input of RepBottleneck.

Eventually, FAD-CSP obtains rich local features, abstract global features, and separated primitive features. As shown in Figure 3, FAD-CSP uses CBS to associate long and short distance features related to the target, decoupling irrelevant degraded features and realizing the improvement of detection accuracy.

Non-Maximum Suppression-Similar

In dense underwater environments, two primary challenges in detection are overlapping objects with similar features and overlapping bounding boxes for the same target, leading to inaccuracies in traditional NMS methods. While Soft-NMS (Bodla et al. 2017) retains more boxes, it increases computational time. To overcome these issues, we propose an NMS method based on aspect ratio similarity, called NMS-Similar. This method combines traditional NMS's speed with SoftNMS's precision, using a unique aspect ratio threshold and optimized greedy strategy. The suppression mechanism for each object is as follows:

$$S_i = S_i e^{-IoU(M,b_i)^2/\sigma} \quad (16)$$

$$L' = (IoU(b_i, L) <= N_t) \text{ and } (Sim(M, L) > N_s) \quad (17)$$

$$Sim(M, b_i) = \frac{\vec{M} \cdot \vec{b_i}}{\|\vec{M}\|\|\vec{b_i}\|} \quad (18)$$

where S_i is the current detection box's confidence, M is the highest confident box, Intersection over Union (IoU) measures overlap between predicted and ground-truth boxes, N_t is the preset IoU threshold, Sim calculates the aspect ratio similarity, $\vec{M}$ represents the length and width of M. N_s is the preset similarity threshold, and σ is a Gaussian weighting function. L and L' represent the remaining and recalculated detection boxes, respectively. Equation (17) adjusts the suppression counts for non-maximum confidence boxes by introducing an aspect ratio threshold to exclude similar detection boxes. The threshold strategy takes into account that object detection boxes for the same object at different scales share similar aspect ratios. During the computation process, similar detection boxes are precluded in advance, reducing the suppression time in dense scenes while ensuring detection accuracy.

Experimental Results

We elaborate on our experimental setup and comparative analyses. Experiments reveal that our approach significantly enhances the network's accuracy and resistance to noise, especially in challenging underwater conditions.

Implementation Details

Our experiments run on an Intel Xeon E5-2650 v4 @ 2.20G CPU and an Nvidia Tesla V100-PCIE-16GB GPU with the Ubuntu 20.04 LTS operating system and Python 3.10 environment built on Anaconda, with a network architecture based on Pytorch 2.0.1 build. The hyperparameters are shown in Table 1.

In addition, if not specified, the comparison experiments are performed using the traditional NMS method.

Type	Setting	Type	Setting
Image size	640	Weights	None
Batch-size	16	Seeds	0
Optimizer	SGD	LR	0.01
Epochs	300	Early-stop	True

Table 1: Hyperparameter settings

Evaluation Metrics and Datasets

We adopt AP and AP50 as the primary metrics for model accuracy evaluation, with precision (P) and recall (R) as supplementary indicators. To showcase the generalizability of our network, we trained it on the URPC (Zhanjiang) (Liu et al. 2021) dataset, from the 2020 National Underwater Robotics Professional Competition and the extensive RUOD dataset. The URPC dataset contains 5,543 training images across five categories, with 1,200 images from its B-list answers serving as the test set. The RUOD dataset (Fu et al. 2023) contains various underwater scenarios and consists of 10 categories. It includes 9,800 training images and 4,200 test images.

Visual Comparisons

Figure 5 visualizes the object detection results of different detection frameworks on the URPC (Zhanjiang) dataset. Many of these frameworks struggle to accurately detect smaller objects, with some even mistakenly identifying the background as a target. The Faster R-CNN (Ren et al. 2015), RetinaNet (Lin et al. 2017), and PAA methods exhibit false positives by detecting kelp as seagrass. In contrast, the YOLO methods (Redmon et al. 2016) miss some objects, failing to detect certain starfish. Our method excels in detecting smaller objects without any false positives or missed detections.

Quantitative Comparisons

In Table 3, the performance of various versions of AMSP-UOD on the URPC and RUOD datasets is presented. Notably, while our AMSP-VConv version indicates slightly reduced stability and precision compared to the Our-Standard version in balanced scenarios, it showcases enhanced detection capability in more degenerative conditions (URPC).

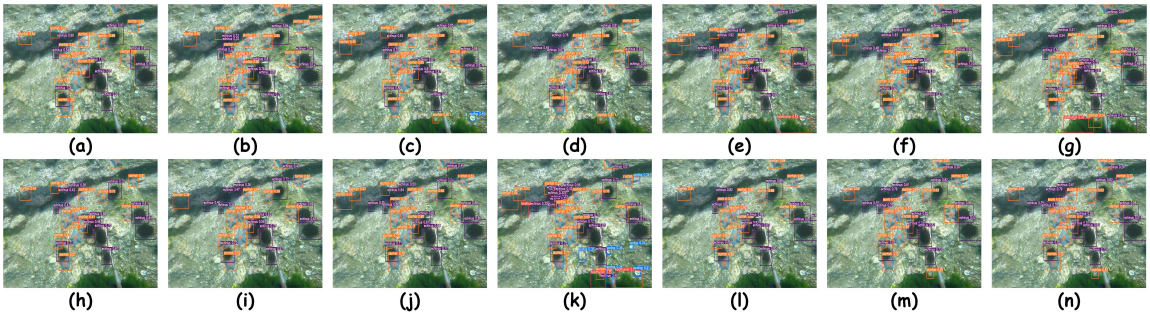

Figure 5: Visualization of object detection results of different object detection methods on URPC (Zhanjiang). (a) YOLOv3 (Redmon and Farhadi 2018), (b) YOLOv5s (Jocher 2020), (c) YOLOv6s(Li et al. 2022), (d) YOLOv7-tiny (Wang, Bochkovskiy, and Liao 2023), (e) Faster R-CNN (Girshick 2015), (f) Cascade R-CNN (Cai and Vasconcelos 2018), (g) RetinaNet (Lin et al. 2017), (h) FCOS (Tian et al. 2019), (i) ATSS (Zhang et al. 2020b), (j) TOOD (Feng et al. 2021), (k) PAA (Kim and Lee 2020), (l) Ours-Standard, (m) Ours-AMSP-VConv, (n) Ours-AMSP-VConv + NMS Similar.

Baselines	Time	Memory	P	R	$mAP^{0.95}_{0.5}$
DSC	15s	**4.61G**	0.824	0.510	0.371
GC	52s	5.10G	0.796	0.637	0.396
VC (w/o SW)	14s	5.36G	0.730	**0.694**	0.386
VC (w/o AMSP)	14s	4.65G	0.833	0.631	0.397
AMSP-VConv	**14s**	4.65G	**0.845**	0.612	**0.398**

Table 2: Ablation of AMSP-VConv. Time: Inference Time (per epoch), DSC: Depthwise Separable Conv, GC: Ghost Conv, SW: Shared Weight, VC: AMSP-VConv, P: Precision, R: Recall.

This observation is also substantiated by subsequent ablation studies. We believe this significant improvement can be attributed to the noise suppression capability of the VConv design combined with the outstanding feature perception ability of FAD-CSP. Especially in intricate underwater environments, our method adeptly boosts the recognition accuracy of waterweeds, which are treated as a small-sample target, to a remarkable 99.3%. Furthermore, the integration of the NMS-Similar strategy imparts a clear enhancement in detection rates for the Vortex version. This strategy efficiently curtails false positives and misses, thus ensuring the integrity and accuracy of object detection. In comparison with the series of YOLO models and other leading detection techniques, our method consistently manifests marked superiority on a foundation of high precision. In conclusion, our method exhibits exemplary efficiency and adaptability in UOD, underscoring its profound potential for real-world underwater applications.

Ablation Studies

In order to verify the impact of the proposed module on the network performance, we conducted a series of ablation experiments.

Ablation of AMSP-VConv: In Table 2, we find that the combination of VConv with the AMSP strategy provides an optimal balance, in terms of precision, recall, and mAP, while maintaining reasonable inference time and memory usage. Compared to Depthwise Separable Convolution (DSC) and Ghost Convolution (GC), AMSP-VConv demonstrates superior performance in complex object detection tasks, particularly in intricate scenarios that need to balance multiple performance metrics. The ablation experiments further reveal the importance of shared parameters and the AMSP strategy for enhancing both accuracy and efficiency. Ultimately, the integration of VConv and the AMSP strategy proves its potential in improving object detection tasks, providing robust support for real-world applications.

Underwater scenarios are susceptible to noise interference, and noise robustness is a crucial metric for evaluating UOD methods. Ensuring all operations that influence network metrics are equivalent, Gaussian noise was used to simulate the underwater noise environment, creating multiple noise levels (i.e., the original scenario augmented with Gaussian noise of varying standard deviations). We trained our network using the URPC dataset. As shown in Figure 6, our network's mAP score remains stable under the influence of noise level 4. In contrast, the mAP@0.5 of YOLOv5s, serving as the Baseline, decreased by 16.3%. In high-noise scenarios, our AMSP-VConv demonstrates superior noise robustness, while the accuracy of Our-Standard, which merely replaces the AMSP-VConv module with standard convolutions, aligns closely with that of the Baseline. This indicates that AMSP-VConv in the Backbone network provides AMSP-UOD with strong noise robustness, validating the effectiveness of AMSP-VConv. It offers an excellent solution for denoising and complex underwater scenarios.

Ablation of NMS-Similar: From Table 4, it is evident that NMS-Similar achieves a commendable balance between accuracy and efficiency. Compared to Soft-NMS, NMS-Similar retains similar detection accuracy while significantly reducing computation time. Especially in challenging underwater detection scenarios, where closely located or overlapping objects are frequent, the performance of NMS-Similar stands out, underscoring its immense value in real-world applications.

Method	URPC		URPC Categories AP50					RUOD	
	AP↑	AP_{50}↑	Ho↑	Ec↑	St↑	Sc↑	Wa↑	AP↑	AP_{50}↑
YOLOv3	29.7	58.9	63.5	83.1	68.1	46.4	33.2	49.1	80.3
YOLOv5s	38.6	66.2	67.3	84.7	76.7	57.2	43.0	53.8	81.4
YOLOv6s	36.1	62.8	61.4	85.2	68.1	49.0	50.1	60.1	84.9
YOLOv7-tiny	35.9	62.2	57.9	84.9	72.3	50.1	66.3	57.9	84.3
Faster R-CNN	31.0	59.0	66.9	85.9	72.1	55.4	14.7	49.1	80.3
Cascade R-CNN	31.6	59.1	67.1	86.0	71.3	56.2	14.7	53.8	81.4
RetinaNet	26.3	51.1	61.3	81.8	66.2	46.2	0.00	48.0	77.8
FCOS	29.2	58.1	61.8	83.5	68.8	53.9	22.3	49.1	80.3
ATSS	29.0	55.6	64.0	84.8	71.4	55.8	2.20	53.9	82.2
TOOD	30.1	56.7	65.0	86.1	72.7	58.3	1.30	55.3	83.1
PAA	34.2	62.3	65.1	85.2	70.9	55.9	34.6	53.5	82.2
Ours (Standard)	**45.0**	73.4	**69.1**	86.6	75.3	53.1	83.0	62.1	85.9
Ours (AMSP-VConv)	36.6	74.8	62.9	87.1	72.9	51.6	99.3	61.4	85.3
Ours (AMSP-VConv + NMS-Similar)	40.1	**78.5**	67.3	**87.5**	**77.5**	**60.6**	**99.5**	**65.2**	**86.1**

Table 3: Comparison with existing methods on the URPC and RUOD datasets. Ho: holothurian's AP50, Ec: echinus's AP50, St: starfish's AP50, Sc: scallop's AP50, Wa: waterweeds's AP50. AP: AP@[0.5:0.05:0.95], AP50: AP@0.5. Bolding and underlining is highest, underlining only is second-highest.

Baselines	Time(ms)	$mAP_{0.5}$	$mAP_{0.5}^{0.95}$	$AP_{echinus}$
NMS	**14.20**	0.748	0.366	0.477
Soft-NMS	337.3	0.785	0.400	0.509
NMS-Similar	46.90	**0.785**	**0.401**	**0.509**

Table 4: Ablation of NMS-Similar

Baselines	P	R	$mAP_{0.5}$	$mAP_{0.5}^{0.95}$
a	0.836	0.610	0.675	**0.397**
b	0.734	0.625	0.640	0.370
c	0.720	0.658	0.679	0.377
d	**0.858**	0.612	0.681	0.369
All	0.844	**0.681**	**0.748**	0.366

Table 5: Ablation of FAD-CSP

Ablation of FAD-CSP: In Table 5, we evaluated the contributions of various components in the FAD-CSP. Four configurations are tested: a) Without the GFA module. b) Replacing the Pooling Groups with Individual Pooling Layers. c) Removing the AMSP strategy from the GFA module in FAD-CSP. d) Replacing Repbottleneck with Bottleneck. Among the tested configurations, using the FAD-CSP method achieves the best results, with the highest mAP of 0.748 and an improved recall rate of 0.681. This underscores the importance of each component in enhancing the detection performance. In particular, removing the GFA module (a) or the AMSP strategy from GFD (c) leads to a decrease in performance, highlighting their critical roles in the framework. Additionally, using Repbottleneck (as opposed to the standard Bottleneck) further bolsters the detection results, emphasizing its effectiveness in the context of the FAD-CSP method.

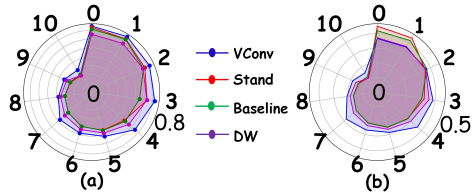

Figure 6: Noise robustness ablation for AMSP-VConv. (a) and (b) show mAP@0.5 and mAP@0.5:0.95 under varied noise levels. Numbers 0-10 represent noise (0 + Gaussian noise standard deviation). Level 0 represents original underwater scene. Methods are not pre-trained on noisy images. Blue is AMSP-VConv, red is Standard Conv, green is YOLOv5s model, and purple is depthwise-separated Conv.

Conclusion

In this work, we proposed AMSP-UOD, a novel network for underwater object detection, addressing non-ideal imaging factors in complex underwater environments. With our innovative AMSP Vortex Convolution, we enhance feature extraction and network robustness, while our FAD-CSP module improves performance in intricate underwater scenarios. Our method optimizes detection in object-dense areas and outperforms existing state-of-the-art methods on the URPC and RUOD datasets. The practical evaluations highlight the potential applicability of AMSP-UOD to real-world underwater tasks, making it a promising contribution to UOD.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No.62301105), the 2022 National Undergraduate Innovation and Entrepreneurship Training Program Project (No.202210577003), National Key Research and Development Program of China (No.2018AAA0100400), China Postdoctoral Science Foundation (No.2021M701780), the High Performance Computing Center of Dalian Maritime University, and the Supercomputing Center of Nankai University. We are also sponsored by CAAI-Huawei MindSpore Open Fund.

References

Bodla, N.; Singh, B.; Chellappa, R.; and Davis, L. S. 2017. Soft-NMS – improving object detection with one line of code. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 5562–5570. IEEE.

Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6154–6162.

Chen, L.; Zhou, F.; Wang, S.; Dong, J.; Li, N.; Ma, H.; Wang, X.; and Zhou, H. 2022. SWIPENET: Object detection in noisy underwater scenes. *Pattern Recognition*, 132: 108926.

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 1251–1258. IEEE.

Chou, E.; Southall, B. L.; Robards, M.; and Rosenbaum, H. C. 2021. International policy, recommendations, actions and mitigation efforts of anthropogenic underwater noise. *Ocean & Coastal Management*, 202: 105427.

Feng, C.; Zhong, Y.; Gao, Y.; Scott, M. R.; and Huang, W. 2021. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3490–3499. IEEE Computer Society.

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

Ghiasi, G.; Lin, T.-Y.; Pang, R.; and Le, Q. V. 2019. NAS-FPN: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7036–7045.

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

Guo, C.; Wu, R.; Jin, X.; Han, L.; Chai, Z.; Zhang, W.; and Li, C. 2022. Underwater ranker: learn which is better and how to be better. In *AAAI Conference on Artificial Intelligence (AAAI)-Oral*.

Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; and Xu, C. 2020. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1580–1589.

Jocher, G. 2020. Ultralytics YOLOv5.

Kim, K.; and Lee, H. S. 2020. Probabilistic anchor assignment with iou prediction for object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, 355–371. Springer.

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, C.; Guo, C.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; and Tao, D. 2019. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29: 4376–4389.

Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. 2022. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

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. *arXiv e-prints*, arXiv:2106.05681.

Liu, J.; Wu, G.; Luan, J.; Jiang, Z.; Liu, R.; and Fan, X. 2023. HoLoCo: Holistic and local contrastive learning network for multi-exposure image fusion. *Information Fusion*, 95: 237–249.

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.

Liu, W.; Liao, S.; Ren, W.; Hu, W.; and Yu, Y. 2019. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5187–5196.

Ma, L.; Jin, D.; An, N.; Liu, J.; Fan, X.; Luo, Z.; and Liu, R. 2023. Bilevel fast scene adaptation for low-light image enhancement. *International Journal of Computer Vision*, 1–19.

Qiao, S.; Chen, L.-C.; and Yuille, A. 2020. DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *European Conference on Computer Vision*, 145–161. Springer.

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.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 91–99.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9627–9636.

Wang, C.-Y.; Bochkovskiy, A.; and Liao, H.-Y. M. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7464–7475.

Xu, S.; Zhang, M.; Song, W.; Mei, H.; He, Q.; and Liotta, A. 2023. A systematic review and analysis of deep learning-based underwater object detection. *Neurocomputing*, 527: 204–232.

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.

Zhang, H.; Chang, H.; Ma, B.; Wang, N.; and Chen, X. 2020a. Dynamic R-CNN: Towards high quality object detection via dynamic training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6729–6738.

Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; and Li, S. Z. 2020b. 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.

Zhang, W.; Zhuang, P.; Sun, H.-H.; Li, G.; Kwong, S.; and Li, C. 2022. Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement. *IEEE Transactions on Image Processing*, 31: 3997–4010.

Zhou, J.; Li, B.; Zhang, D.; Yuan, J.; Zhang, W.; Cai, Z.; and Shi, J. 2023a. UGIF-Net: an efficient fully guided information flow network for underwater image enhancement. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–17.

Zhou, J.; Liu, Q.; Jiang, Q.; Ren, W.; Lam, K.-M.; and Zhang, W. 2023b. Underwater camera: improving visual perception via adaptive dark pixel prior and color correction. *International Journal of Computer Vision*, 1–19.

Zhou, X.; Koltun, V.; and Krähenbühl, P. 2020. Tracking objects as points. In *European Conference on Computer Vision*, 474–490. Springer.

Zhou, X.; Zhuo, J.; and Kr"ahenb"uhl, P. 2019. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 850–859.

Zhu, C.; He, Y.; and Savvides, M. 2019. Feature selective anchor-free module for single-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 840–849.

Zhuang, P.; Wu, J.; Porikli, F.; and Li, C. 2022. Underwater image enhancement with hyper-laplacian reflectance priors. *IEEE Transactions on Image Processing*, 31: 5442–5455.