Proceedings of 2025 IEEE
International Conference on Mechatronics and Automation
August 3 - 6, Beijing, China

HAD-Net: A Hybrid Attention Driven Dynamic Network for AUV-Based Underwater Object Detection

Xueting Liu[1], Shuxiang Guo[1,2,3]*, Chunying Li[1]*, Sihan Gao[1,2], Qirong Lei[1], Weizhi Wu[1]
1. The Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, Guangdong 518055, China
2. Advanced Institute for Ocean Research, Southern University of Science and Technology, Shenzhen, 518055, China
3. The Aerospace Center Hospital, School of Life Science and the Key Laboratory of Convergence Medical Engineering System and Healthcare Technology, Ministry of Industry and Information Technology, Beijing Institute of Technology, Beijing 100081, China
* Corresponding authors: guo.shuxiang@sustech.edu.cn, licy@sustech.edu.cn

2025 IEEE International Conference on Mechatronics and Automation (ICMA) | 979-8-3315-1426-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICMA65362.2025.11120645

***Abstract*—Underwater object detection (UOD) is critical for enabling autonomous underwater vehicles (AUVs) to carry out underwater operations incomplex marine environments. However, the underwater domain poses severe challenges for visual perception due to low visibility, color distortion, and dynamic noise. To address these issues, Hybrid Attention Driven Dynamic Network (HAD-Net), which integrated a novel dynamic convolution module named FDConv (Full-Dimensional Convolution) was proposed. Unlike existing dynamic convolution approaches that only adapt over the kernel number dimension, FDConv introduced a four-way attention mechanism across the channel, filter, spatial, and kernel dimensions, enabling precise and adaptive feature modulation. By embedding FDConv into a YOLOv11-based detection framework, HAD-Net achieved superior performance in challenging underwater scenarios. Extensive experiments on the URPC2020 and TrashCan datasets demonstrated that HAD-Net outperforms existing detectors in both accuracy and robustness, while maintaining real-time inference speed. Our method provides a lightweight and deployable solution for perception tasks for autonomous underwater vehicles (AUVs).**

***Index Terms*—Underwater object detection (UOD), Hybrid Attention, Lightweightness, Autonomous Underwater Vehicles (AUVs)**

I. Introduction

Autonomous Underwater Vehicles (AUVs) [1], [2] have become indispensable tools in advancing scientific investigations and technological endeavors within marine exploration and resource development [3]. Capable of operating at extreme depths, these robotic systems perform complex subaquatic missions in environments marked by formidable challenges and dynamic conditions [4], thus serving as reliable alternatives to human intervention in high-risk oceanic operations [5]. Underwater object detection (UOD) is a crucial method by which AUVs can delve into the mysteries of the ocean. Accurate detection allows AUVs to identify, locate and track objects or entities within the ocean, which is essential for tasks such as seabed mapping,

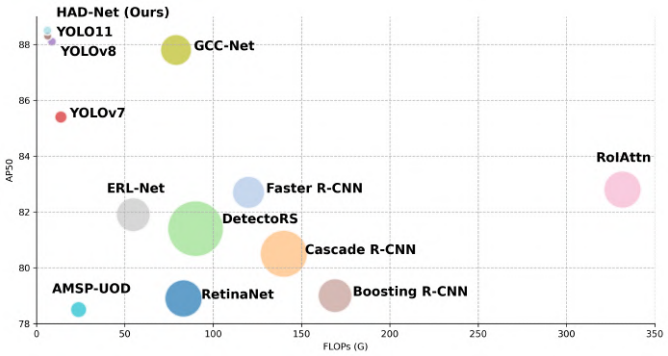

Fig. 1. Algorithm-Performance Comparison with the SOTA. The size of a circle represents its parameter count

environmental monitoring, and underwater archaeology. However, underwater environments present unique challenges for visual object detection, such as low visibility, color distortion, backscatter noise, and limited lighting conditions [6]. These factors hinder the transmission and quality of optical data, resulting in blurry, noisy, and poorly contrasted images. Traditional object detection algorithms that perform well in air, such as those used in terrestrial robotics [7], often fail in underwater settings due to drastically different lighting and visibility conditions. To overcome these challenges, various UOD approaches [8] have incorporated Underwater Image Enhancement (UIE) techniques as a preprocessing step [9]. However, UIE methods do not always translate into better detection outcomes and can occasionally reduce detection accuracy due to the introduction of domain shifts that alter image distributions. To mitigate this, Dai et al. [10] proposed GCC-Net, which integrates features from both raw and enhanced images to promote better domain adaptation. In parallel, other strategies such as contrastive learning frameworks [11] have aimed to improve generalization across

domains through advanced data augmentation. Furthermore, Zhou et al. [12] developed AMSP-UOD, a method that focuses on suppressing underwater noise during feature extraction, further enhancing the robustness of detection in underwater conditions.

Recent advances in dynamic convolution methods, such as CondConv [13] and DyConv [14], have significantly improved network adaptability by enabling input-conditioned convolution. These methods adjust the convolution kernel at runtime based on the input, allowing for better feature representation under varying conditions [15]. However, existing dynamic convolution methods primarily exploit attention along the kernel number dimension. While they improve adaptability, they fail to fully leverage other essential dimensions, such as spatial, channel, and filter dimensions, which are crucial for underwater object detection. This limitation hinders their capacity to adapt effectively to complex, dynamic, and noise-prone underwater environments. To address this issue, HAD-Net, a Hybrid Attention Driven Dynamic Network was designed specifically for underwater object detection. HAD-Net integrates FDConv into the YOLO11 backbone architecture [16], a popular real-time object detection model. This hybrid approach allows us to combine the strengths of dynamic convolutions and attention mechanisms, leveraging multi-dimensional attentions to refine feature extraction without introducing excessive computational overhead. HAD-Net aims to improve detection accuracy and robustness in underwater environments, while maintaining real-time performance for AUV-based missions.

The contributions of this paper are as follows:

1) FDConv was proposed, a novel full-dimensional dynamic convolution mechanism that integrates multi-dimensional attention across spatial, channel, filter, and kernel dimensions, ensuring both global scene understanding and fine-grained detail retention.

2) HAD-Net was introduced, a hybrid architecture that incorporates FDConv into a YOLO11-based detector, allowing the model to dynamically adapt its receptive fields and kernel responses according to varying underwater conditions and significantly improving accuracy and robustness in underwater object detection.

3) HAD-Net was validated on two underwater datasets (URPC2020 and TrashCan [17]) and demonstrated its superior performance compared to existing methods.

II. Methodology

A. Architecture of HAD-Net

HAD-Net consists of three main components: the backbone, the neck, and the head. Each of these components is designed to leverage the multi-dimensional attention mechanism to enhance the model's performance while maintaining efficiency, as shown in Fig. 2. The backbone of HAD-Net is based on CSPDarkNet, a lightweight and efficient feature extractor designed for real-time object detection. In HAD-Net, several standard convolutional layers are replaced with Full-Dimensional Convolution (FDConv) blocks to introduce dynamic feature recalibration. The FDConv layers are applied across multiple stages of the backbone to allow the model to learn spatially adaptive kernel responses and channel-specific attention. The backbone is responsible for processing the raw input image into a set of high-level features, which are subsequently refined by the neck. The backbone is responsible for extracting hierarchical feature representations from the input image. It begins with standard convolutional layers and is followed by repeated applications of the FDConv module and the C3K2 residual block. FDConv integrates three types of attention mechanisms: Kernel Attention, Spatial Attention, and Channel Attention. These attention paths operate in parallel and are adaptively fused based on their learned importance. The backbone also incorporates the Spatial Pyramid Pooling-Fast (SPPF) module to encode multi-scale context, and the C2PSA module in later stages to suppress background noise and refine spatial information. The neck module bridges the backbone and the detection head. It adopts a feature pyramid structure that utilizes upsampling and concatenation operations to merge low-level spatial features and high-level semantic features. Alternating C3K2 and FDConv modules are employed to maintain feature integrity and enhance representation across scales. The detection head consists of three output branches, each corresponding to a specific scale. Each branch includes C3K2 blocks followed by convolutional layers for object classification, bounding box regression, and objectness scoring. This multi-scale design improves the network's ability to detect objects of varying sizes.

B. Full-Dimensional Convolution

1) Attention in Spatial, Channel, Filter, and Kernel Dimensions: The FDConv module introduces an omni-directional attention mechanism that enables the network to dynamically adapt convolutional operations based on both the characteristics of the input features and the convolutional kernel parameters. Unlike traditional dynamic convolutions that typically focus on a single axis, FDConv explicitly decomposes the dynamic reweighting process into four semantically meaningful and complementary attention branches:

- *Channel-wise Attention* (α^c): Captures inter-channel dependencies within the input feature map, allowing the model to suppress redundant information and emphasize informative channels.
- *Filter-wise Attention* (α^f): Learns to scale the responses of different filters in the convolutional layer, enabling adaptive emphasis on more relevant output feature maps.
- *Spatial-wise Attention* (α^s): Assigns varying importance to different spatial positions within each convolutional kernel (e.g., each location in a 3×3 kernel), which is particularly beneficial in modeling spatially heterogeneous patterns, such as those commonly encountered in underwater visual environments.
- *Kernel-wise Attention* (α^w): Performs soft selection over a set of parallel convolutional kernels, enabling

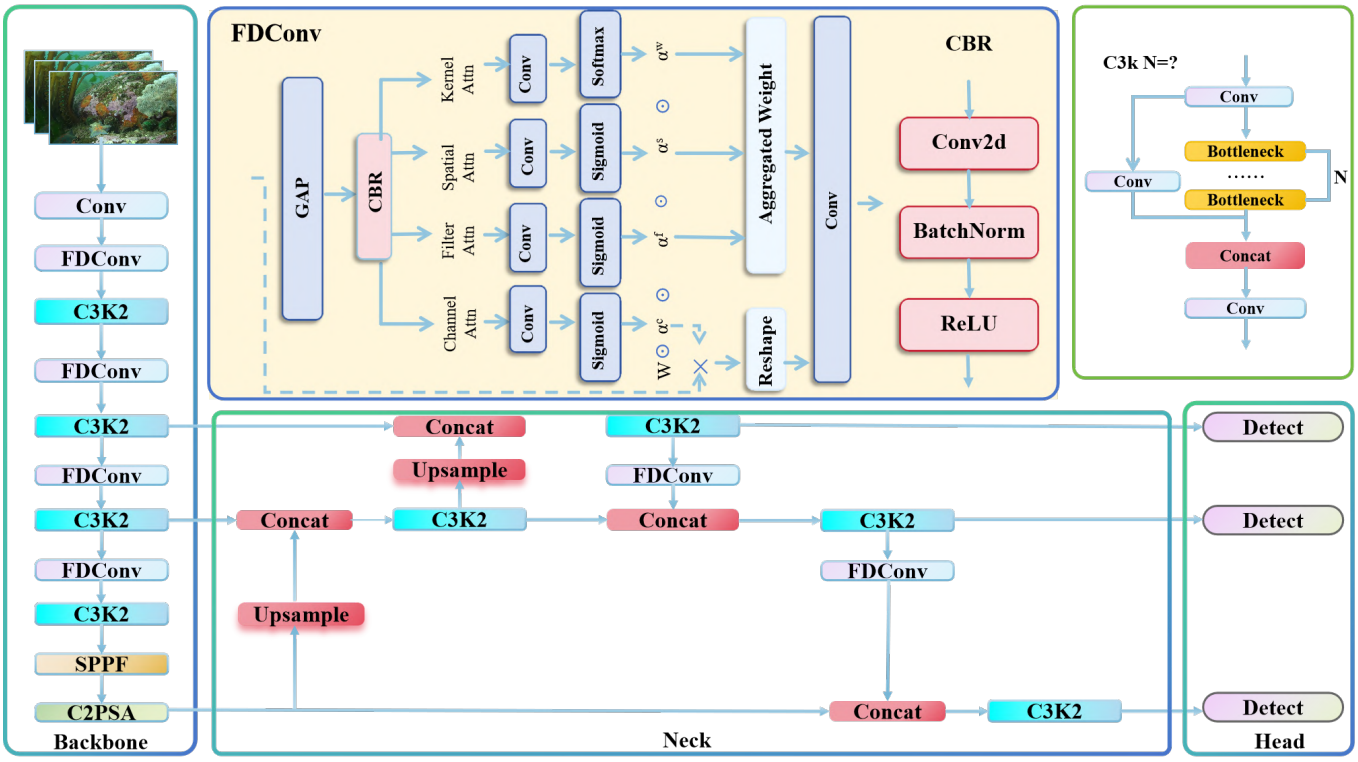

Fig. 2. The overall framework of HAD-Net

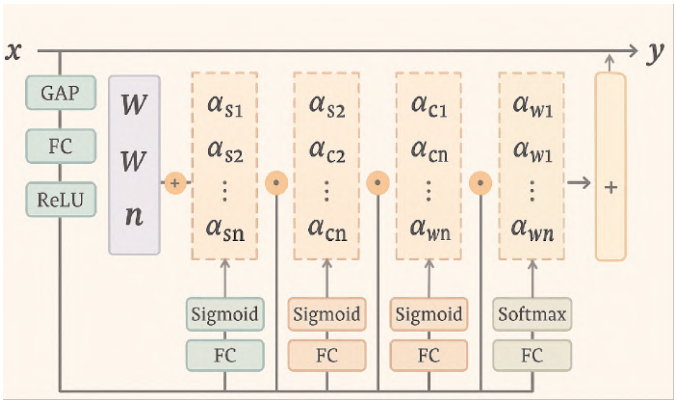

Fig. 3. Hybrid Attention-Driven Computation Flow

content-adaptive interpolation between multiple kernels for enhanced flexibility.

All four attention weights are derived from a shared global descriptor of the input feature map, which is generated via global average pooling. This descriptor is then processed through a lightweight bottleneck structure consisting of a Conv1×1, BatchNorm, and ReLU (CBR) block, before being passed to four parallel branches responsible for producing α^c, α^f, α^s, and α^w respectively.

2) Comparison with CondConv and DyConv: While prior dynamic convolution variants such as CondConv and DyConv enhance model capacity by introducing kernel-wise modulation, they are limited by their unidimensional scope—focusing only on α^w and ignoring potentially rich semantics embedded in spatial and channel dimensions.

In contrast, FDConv holistically considers multiple structural axes of convolution. Specifically, it enables: Identification of critical input channels via α^c, emphasis on informative filter responses via α^f, differentiation of spatial locations in convolutional operations via α^s, and adaptive kernel blending via α^w. This enables FDConv to realize a more fine-grained and content-aware feature transformation. The general dynamic convolution operation of FDConv can be expressed as:

$$y = \left(\sum_{i=1}^{n} \alpha_i^w \odot \alpha_i^f \odot \alpha_i^c \odot \alpha_i^s \odot W_i \right) * x \tag{1}$$

Here, W_i represents the i-th kernel in the kernel bank, and the attention coefficients perform multiplicative modulation across different axes before convolution.

3) Hybrid Attention-Driven Computation Flow: Fig. 3 illustrates the overall computation flow of FDConv. The process unfolds as follows:

Given an input feature map $x \in \mathbb{R}^{B \times C_{in} \times H \times W}$, global average pooling is applied to extract a global descriptor with shape $B \times C_{in} \times 1 \times 1$.

This descriptor is passed through a shared CBR bottleneck block to reduce dimensionality and enrich feature abstraction.

The resulting embedding is split into four branches:

- Channel attention α^c is used to recalibrate the input feature map: $x' = x \cdot \alpha^c$.
- Spatial attention α^s and kernel attention α^w are used to compute the content-aware aggregated kernel:

$$W_{\text{agg}} = \sum_{i=1}^{n} \alpha_i^w \cdot \alpha_i^s \cdot W_i \tag{2}$$

  where $\alpha_i^s \in \mathbb{R}^{k\times k}$ is the spatial weighting mask, and $W_i \in \mathbb{R}^{C_{out}\times C_{in}\times k\times k}$ is the i-th kernel in the bank.
- Convolution is performed using W_{agg}, and the result is modulated by filter attention α^f.

The final output is computed as:

$$y = \text{Conv2D}(x', W_{\text{agg}}) \cdot \alpha^f \tag{3}$$

By integrating all axes of attention in a unified framework, FDConv achieves spatially adaptive and instance-aware dynamic feature processing.

Despite the inclusion of multiple attention branches, the computational overhead is minimal due to the use of shared pooling and projection layers, compact representations (e.g., 1×1 descriptors), and efficient broadcasting mechanisms. This lightweight design ensures that FDConv remains suitable for real-time applications, particularly in scenarios such as AUVs and embedded vision systems.

III. Experimental Results

A. Datasets and Evaluation Metrics

1) Datasets: We evaluate HAD-Net on URPC2020 and TrashCan [17]. URPC dataset, employed in the Underwater Robot Professional Contest (URPC), encompasses a substantial collection of 5543 underwater images designated for training purposes. Additionally, it includes 1,200 images from its B-list answers, serving as the test set. This dataset spans across four distinct categories of underwater organisms, namely echinus, holothurian, scallop, and starfish, offering a diverse range of visual data for analysis and machine learning applications. TrashCan [17] represents a significant contribution to the field of instance segmentation annotation, specifically tailored for the challenging domain of underwater debris identification. This comprehensive dataset encompasses 16 distinct categories, not only capturing various types of garbage but also incorporating remotely operated vehicles (ROVs) and a rich diversity of underwater flora and fauna, thereby providing a robust resource for advancing research in this specialized area.

2) Evaluation Metrics: In this paper, the results adhere to the standard COCO-style Average Precision (AP) metrics, encompassing AP, AP_{50} (IoU=0.5), and AP_{75} (IoU=0.75). The AP score is calculated by averaging across various IoU thresholds, ranging from 0.5 to 0.95, with a step size of 0.05.

B. Implementation Details

The proposed model was trained on a single NVIDIA GeForce RTX 4090 GPU, employing the SGD optimizer with a weight decay factor of 0.0005 and a momentum of 0.937. The training configurations specified an input image size of 640×640 pixels, a batch size of 16, and a fixed random seed of 0 to guarantee reproducibility. Initially, the learning rate was set at 0.01, utilizing a schedule compatible with SGD, and the model underwent training for 300 epochs. Notably, no pre-trained weights were utilized during the initialization process ("Weights: None").

C. Comparisons with the SOTA

We compared HAD-Net with several SOTA methods on URPC2020 , and TrashCan [17]datasets. The results are shown in Table I and Table II.

1) Results on URPC : HAD-Net (Ours) demonstrates outstanding performance on the URPC dataset, outperforming several SOTA methods. Specifically, HAD-Net has a maximum AP_{50} of 88.5 and AP_{75} of 76.8. As shown in the tableI, HAD-Net consistently outperforms all other methods in both AP_{50} and AP_{75} metrics. Compared to the SOTA GCC-Net [10] method, our model achieved a performance gain of 0.7% in AP_{50} and 0.5% in AP_{75}. Furthermore, compared to the YOLO11 [16] model, our model demonstrates improvements of 0.2% in AP_{50} and 2.9% in AP_{75}. These results validate the effectiveness of our proposed method.

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

Moreover, as evident from the Table I and Fig. 1, HAD-Net has a relatively low number of GLOPs (Floating Point Operations), which translates to lower computational cost during inference. This makes it more efficient for deployment on AUVs with limited computational power, offering a practical solution in real-world applications compared to other larger, more computationally intensive models. HAD-Net strikes an efficient balance between accuracy and speed, making it particularly well-suited for real-time decision-making in dynamic underwater environments.

2) Results on TrashCan [17]: For TrashCan [17] datasets, we present the performance of HAD-Net and compare it with several SOTA methods. As shown in TableII, HAD-Net demonstrates superior performance across all three datasets. Specifically, it achieves the highest AP and AP_{50} scores in the TrashCan [17] datasets, outperforming other methods such as GCC-Net [10] and ERL-Net [21]. These results highlight the effectiveness of the proposed architecture in capturing both local and global features in challenging underwater environments.

In summary, HAD-Net has demonstrated significant advancements in underwater object detection, providing robust and efficient performance across a variety of underwater

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

	Methods	AP↑	AP_{50}↑	AP_{75}↑	echinus	holothurian	scallop	starfish	Params (M)↓	GFLOPs↓
GOD	RetinaNet	57.8	78.9	63.0	79.4	68.2	51.1	74.9	55.38	83.2
	Faster R-CNN	61.2	82.7	69.6	70.4	61.4	41.9	71.4	41.3	120
	Cascade R-CNN	61.6	80.5	72.4	69.0	61.9	41.9	72.0	88.15	140
	DetectoRS	60.8	81.4	69.3	69.5	60.9	41.1	70.2	123.23	90.03
	YOLOv7	49.8	85.4	64.2	73.7	66.3	50.8	74.5	6.2	13.8
	YOLOv8 [18]	59.1	88.1	72.8	95.2	**84.8**	84.7	90.9	3.2	8.7
	YOLO11 [16]	59.7	88.3	73.9	95.4	82.7	84.2	**91.2**	**2.6**	**6.3**
UOD	Boosting R-CNN [19]	63.7	79.0	72.3	70.0	64.3	46.6	74.8	45.95	169
	RoIAttn [20]	62.2	82.8	69.5	70.7	62.2	40.5	71.4	55.23	331.7
	ERL-Net [21]	63.7	81.9	72.2	70.8	66.6	45.4	73.7	45.95	54.8
	GCC-Net [10]	**69.1**	87.8	76.3	75.2	76.7	68.2	56.3	38.31	79
	AMSP-UOD [12]	40.1	78.5	–	87.5	60.6	42.5	77.5	10.4	23.8
	HAD-Net (Ours)	60.3	**88.5**	**76.8**	**95.6**	**85.5**	**85.3**	91.0	**3.4**	6.2

(a) Ground Truth **(b) YOLO11** **(c) HAD-Net (ours)**

Fig. 4. Visualization of HAD-Net

TABLE II
BENCHMARKING RESULTS BETWEEN HAD-NET AND OTHER SOTA METHODS ON TRASHCAN [17]

	Methods	AP↑	AP_{50}↑
GOD	RetinaNet	29.4	53.8
	Faster R-CNN	31.2	55.3
	Cascade R-CNN	33.6	52.7
	YOLOv7	26.1	48.8
	YOLOv8 [18]	44.2	61.5
	YOLO11 [16]	45.1	61.9
UOD	Boosting R-CNN [19]	36.8	57.6
	RoIAttn [20]	32.5	56.8
	ERL-Net [21]	37.0	58.9
	GCC-Net [10]	41.3	61.2
	HAD-Net (Ours)	**45.5**	**62.3**

environments. By integrating dynamic convolution through FDConv and leveraging a hybrid attention mechanism, HAD-Net outperforms existing state-of-the-art methods in terms of both detection accuracy and computational efficiency. Its ability to adapt to the challenging conditions of underwater environments, such as low visibility and noise, makes it a powerful tool for autonomous underwater vehicles. The extensive results on the URPC2020 and TrashCan datasets demonstrate the practical applicability of HAD-Net in real-world underwater perception tasks, offering a lightweight and deployable solution for AUVs.

IV. CONCLUSION

In this paper, HAD-Net introduced significant advancements in underwater object detection by addressing the challenges of low visibility and noise in underwater environments. By integrating FDConv and a hybrid attention mechanism, HAD-Net improved adaptability across multiple dimensions, resulting in robust feature extraction and enhanced performance in both accuracy and efficiency. Extensive

experiments conducted on the URPC2020 and TrashCan [17] datasets have validated the exceptional performance of HAD-Net, outperforming existing state-of-the-art techniques by achieving higher detection accuracy while maintaining real-time processing capabilities. This renders HAD-Net a lightweight and highly deployable solution for AUVs, pushing the boundaries of underwater perception and paving the way for advanced real-world applications. Future work will integrate underwater robotic applications, exploring multimodal data such as sonar to enhance model robustness and detection accuracy in extreme conditions.

ACKNOWLEDGEMENT

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

REFERENCES

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

[2] A. Li, S. Guo, and C. Li, "An improved motion strategy with uncertainty perception for the underwater robot based on thrust allocation model," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 64–71, 2025.

[3] X. Hou, H. Xing, S. Guo, H. Shi, and N. Yuan, "Design and implementation of a model predictive formation tracking control system for underwater multiple small spherical robots," *Applied Sciences*, vol. 14, no. 1, pp. 294, 2024.

[4] Y. Jia, X. Ye, P. Li, and S. Guo, "Contrastive adaptation on domain augmentation for generalized zero-shot side-scan sonar image classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–13, 2025.

[5] L. Qiao and W. Zhang, "Trajectory tracking control of auvs via adaptive fast nonsingular integral terminal sliding mode control," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1248–1258, 2020.

[6] C. Fabbri, M. J. Islam, and J. Sattar, "Enhancing underwater imagery using generative adversarial networks," *in Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)*, brisbane, Australia, pp. 7159–7165, 2018.

[7] X. Xu, W. Ren, G. Sun, H. Ji, Y. Gao, and H. Liu, "Grouptrack: Multi-object tracking by using group motion patterns," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, United Arab Emirates, pp. 4896–4903, 2024.

[8] R. Liu, Z. Jiang, S. Yang, and X. Fan, "Twin adversarial contrastive learning for underwater image enhancement and beyond," *IEEE Transactions on Image Processing*, vol. 31, pp. 4922–4936, 2022.

[9] O. A. Aguirre-Castro, E. E. García-Guerrero, O. R. López-Bonilla, E. Tlelo-Cuautle, D. López-Mancilla, J. R. Cárdenas-Valdez, J. E. Olguín-Tiznado, and E. Inzunza-González, "Evaluation of underwater image enhancement algorithms based on retinex and its implementation on embedded systems," *Neurocomputing*, vol. 494, pp. 148–159, 2022.

[10] L. Dai, H. Liu, P. Song, and M. Liu, "A gated cross-domain collaborative network for underwater object detection," *Pattern Recognition*, vol. 149, pp. 110222, 2024.

[11] Y. Chen, P. Song, H. Liu, L. Dai, X. Zhang, R. Ding, and S. Li, "Achieving domain generalization for underwater object detection by domain mixup and contrastive learning," *Neurocomputing*, vol. 528, pp. 20–34, 2023.

[12] J. Zhou, Z. He, K.-M. Lam, Y. Wang, W. Zhang, C. Guo, and C. Li, "Amsp-uod: When vortex convolution and stochastic perturbation meet underwater object detection, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, pp. 7659–7667, 2024.

[13] X. Wu, G. Li, S. Li, Q. Cao, S. Yang, H. Li, and M. Liu, "Enhancing deep learning model performance by integrating cbam and condconv technologies," in *Proceedings of the 2024 International Conference on New Trends in Computational Intelligence (NTCI)*. Guangzhou, China, pp. 478–482, 2024.

[14] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," *arXiv*, 2020.

[15] H. Ji, B. Chen, X. Xu, W. Ren, Z. Wang, and H. Liu, "Language-assisted skeleton action understanding for skeleton-based temporal action segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Milan, Italy, 2024.

[16] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024.

[17] J. Hong, M. Fulton, and J. Sattar, "Trashcan: A semantically-segmented dataset towards visual detection of marine debris," *CoRR*, vol. abs/2007.08097, 2020.

[18] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023.

[19] P. Song, P. Li, L. Dai, T. Wang, and Z. Chen, "Boosting r-cnn: Reweighting r-cnn samples by rpn's error for underwater object detection," *Neurocomputing*, vol. 530, pp. 150–164, 2023.

[20] X. Liang and P. Song, "Excavating roi attention for underwater object detection," in *Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP)*, Kuala Lumpur, Malaysia, 2022.

[21] L. Dai, H. Liu, P. Song, H. Tang, R. Ding, and S. Li, "Edge-guided representation learning for underwater object detection," *arXiv*, 2023.