2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

YOLO-VIS: Human Vision Mechanism Enhanced YOLO for Forward-Looking Sonar Images Object Detection

Ziyu Zheng[1,2] , Yuquan Wu[1,2,(✉)] , Xuewei Li[1] ,Linjuan Cheng[1] and Chenghao Hu[3]

[1] Science & Technology on Integrated Information System Laboratory, Institute of Software
Chinese Academy of Sciences,100190, Beijing, China
[2] University of Chinese Academy of Sciences, 100049,Beijing , China
{zhengziyu2024,yuquan,lixuewei,linjuan}@iscas.ac.cn
[3] Institute of Acoustics, Chinese Academy of Sciences, 450002, Beijing, China
huchenghao@@mail.ioa.ac.cn

Abstract. Forward-looking sonar images object detection plays a crucial role in marine resource exploration and national defense. Existing methods typically focus on traditional feature extraction approaches when processing sonar images, but these methods have not fully borrowed from the advanced processing mechanisms of the human brain in target recognition, leading to less than satisfactory performance in forward-looking sonar images with issues such as low resolution, dynamic changes, and noise interference. To address this, this paper proposes a brain-inspired forward-looking sonar target recognition framework named YOLO-VIS. We designed a low-level feature enhancement module based on large-kernel convolutions, which simulates the human brain's preliminary processing of images by expanding the receptive field, thereby improving the quality of feature extraction. In addition, a visual attention weighting module is proposed, which further enhances the model's focus on key features by optimizing feature selection based on the importance of neurons. Finally, through a multi-scale feature deep fusion module, the model's target recognition capability at different scales is improved. Experimental results show that YOLO-VIS significantly improves target detection accuracy over existing methods on public datasets, verifying the effectiveness of brain-inspired mechanisms in sonar image recognition.

Keywords: Underwater Object Detection, Brain-Inspired Intelligence, Attention Mechanism and Large Kernel Convolution

1 Introduction

Target recognition is an enduring hot topic in the field of computer vision. Unlike natural images captured by optical cameras, in the marine field, object detection is often achieved through sonar. However, due to the reverberation of acoustic waves in the water column and environmental noise, sonar images commonly face issues such as severe noise interference [16], non-rigid deformation of targets, and low resolution.

These problems make it challenging to construct a stable feature representation for objects in sonar images, resulting in poor detection performance [38].

To tackle this challenge, common object detection algorithms are primarily categorized into two types, one of which is the two-stage detection model. The two-stage detection model carries out object detection in two main phases: Proposal Generation and Proposal Classification. In the Proposal Generation stage, a series of candidate regions (region proposals) are first generated in the input image, which may contain certain objects. This step is usually accomplished using a Region Proposal Network (RPN). Next, the algorithm performs more precise bounding box regression and object category prediction for each candidate region generated in the first stage.

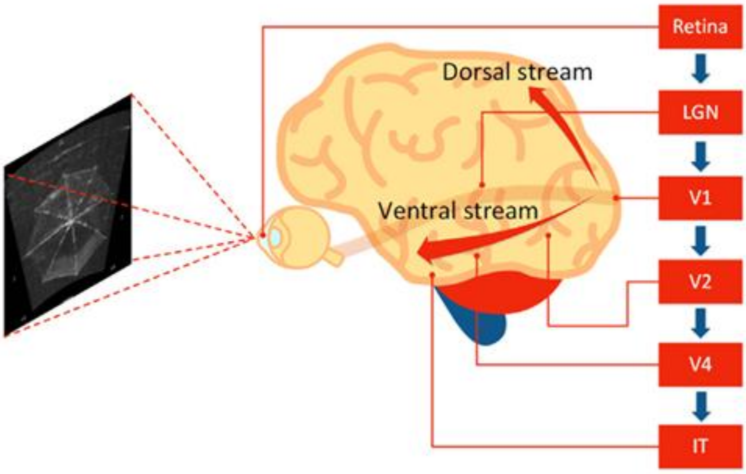


Fig. 1. Two visual pathways in the brain: the dorsal stream for spatial processing and the ventral stream for shape recognition, with the ventral stream being crucial for sonar-based underwater detection and target identification.

However, due to the complexity of the process, two-stage models do not perform well in practical applications. Currently, more commonly used are single-stage models represented by the YOLO framework, which complete the object detection task in a single forward pass without the need to generate candidate regions. They directly predict the category and position of the targets through dense grids or anchor boxes. Since they tend to produce a higher rate of false positives, there is still room for improvement in practical applications.

The brain's recognition mechanism is one of the most important means by which humans perceive the external world, primarily manifested in visual perception. Visual perception has two pathways: the ventral stream for shape perception and the dorsal stream for spatial location perception [27]. Since sonar images are primarily used for underwater detection and target recognition, their recognition performance largely relies on the ventral stream of the visual system. Therefore, we focus primarily on the ventral stream. This pathway is achieved through a series of hierarchically organized processes, as shown in Fig 1. External information reaches the retina through light reflection and is further processed by the retina and the lateral geniculate nucleus (LGN) to the striate cortex V1, then through the extra striate cortex V2 and the lateral occipital cortex V4, finally arriving at the inferotemporal cortex IT. The features extracted from the ventral pathway areas are as follows: First, neurons in the lower-level areas (V1, V2) receive visual input and represent simple features. Second, their output is integrated and processed by higher cortical levels (V4), which are specialized in representing some global representations. Finally, at further hierarchical levels (IT), their output is integrated to represent abstract forms, objects, and categories. This hierarchical structure has been proven beneficial for target recognition tasks [22], inspired by this, we hope to simulate this process with deep neural networks to achieve better recognition performance on forward-looking sonar.

In this work, we propose a YOLO-VIS framework that mimics the human brain's recognition mechanism for target identification in underwater forward-looking sonar. Specifically, the image is pre-enhanced based on large-kernel convolutions and edge detection before recognition. Building on previous research, we introduce an attention mechanism based on neuron processing in the Backbone, which can better represent features at different levels. To further obtain an abstract representation of features, a multi-scale feature deep fusion mechanism is designed at the feature fusion stage, which can achieve a deep fusion of features and thus obtain more superior recognition results. The main contributions of this work are as follows:

- A low-level feature processing method based on large-kernel convolutions and edge detection has been developed. This method achieves the enhancement of simple image features by combining the large receptive field provided by large-kernel convolutions with the handling of target edge details by edge convolutions.
- An attention mechanism based on importance of neuron has been designed. This method, without adding additional sub-networks, calculates the importance of neurons through a designed energy function and scales the corresponding elements of the feature map accordingly, achieving stronger feature extraction capabilities.
- A multi-scale feature deep fusion mechanism is proposed, which effectively integrates the semantic information of feature maps at different levels and reduces noise interference by generating initial differential features and achieving multi-scale feature fusion through large-kernel convolutions.

2 Related Work

2.1 YOLO series

YOLO (You Only Look Once) [19] is a real-time object detection system that uses a single neural network to predict bounding boxes and class probabilities for detected objects in an image. YOLO was first introduced in 2016 by Joseph Redmon, and has since undergone several updates to improve its performance. The YOLO algorithm [24] divides an input image into a grid of cells and predicts bounding boxes for objects within each cell. The predictions are made using a convolutional neural network that processes the entire image in a single forward pass, rather than scanning the image multiple times with a sliding window as some other object detection systems do.

Haoting Zhang et al. [36] proposed a object detection model for forward- looking sonar images based on an improved YOLOv5. By introducing the IoU k-means algorithm and the CoordConv method, they effectively improved the detection accuracy and speed. Jian Yang et al. [34] combined the lightweight characteristics of MobileNetV2 with the self-attention mechanism of SENet based on YOLOv5, and proposed the Fast-YOLOv5 network. They enhanced model performance through methods such as noise augmentation and K-Means label preprocessing. Liangfu Ge et al. [6] targeting the characteristics of low resolution and feature similarity in underwater sonar images, introduced the SimAM module, which realized a parameter-free attention mechanism and enhanced the feature focusing ability of CNN, verifying its generalization capability. Ken Sinkou Qin et al. [17] applied the YOLOv7 model to underwater obstacle and real-time sonar image object detection. By introducing the cross-attention mechanism and the MGGN module, they improved the model’s feature extraction ability and the detection accuracy for targets in sonar images.

2.2 Attention Mechanism

Since the introduction of the self-attention mechanism [28] by Vaswani et al.,attention mechanisms have rapidly advanced, particularly in their ability to capture long-range dependencies. Self-attention computes the similarity between elements in a sequence, yielding significant improvements in tasks such as machine translation. However, its computational complexity scales quadratically with the sequence length, leading to inefficiencies when processing long sequences or large datasets. Linformer [32] addresses this issue by reducing the attention complexity from $O(n^2)$ to $O(n)$ using low-rank approximation, significantly enhancing the ability to handle long sequences. Nevertheless, Linformer still faces challenges in capturing long-range dependencies, particularly in dealing with complex relationships.

Coordinate Attention (CA) [9] was introduced to refine the attention mechanism by incorporating spatial positional information, showing particularly strong performance in image-related tasks. CA decomposes the input feature map into two 1D encoding vectors, enabling it to capture both spatial and channel dependencies simultaneously. However, CA focuses primarily on spatial-channel relationships and overlooks the fusion of multimodal information. The Multimodal Attention Network (MAN) [10] improves performance in multimodal tasks by integrating cross-modal information. MAN not only addresses spatial and channel relationships but also establishes deep connections across different modalities. However, it incurs higher computational costs, and coordinating modalities remains a significant challenge. Swin Transformer [15] has made notable progress in image tasks by introducing sparse connections, which alleviate the limitations of traditional attention mechanisms. Brain-inspired attention mechanisms [2], which emulate selective attention in the brain, further enhance model performance on complex tasks. These mechanisms adaptively focus on the most relevant input features and establish deep connections across modalities, thereby effectively merging diverse information to improve the model’s adaptability and robustness.

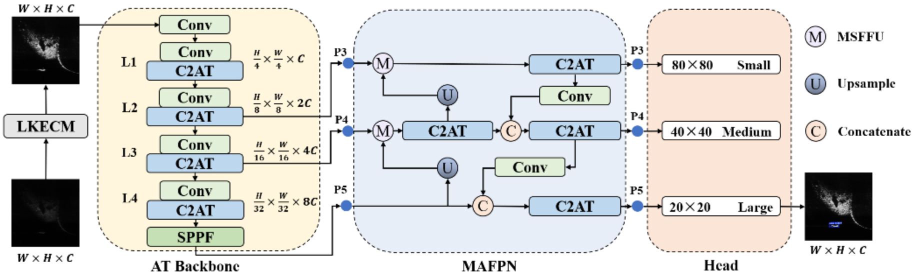


Fig. 2. Overview of the YOLO-VIS framework. This framework designs a Large Kernel Edge Convolution Module to enhance the low-level feature details in forward-looking sonar images. It introduces a C2AT into the original Backbone to enhance feature extraction capabilities. A Multi-Scale Feature Fusion Unit is designed to enhance the effectiveness of feature fusion.

3 Methods

3.1 Overall Architecture

An overview of the architecture of YOLO-VIS is depicted in Fig 2. Our underwater sonar image object detection model consists of the LKECM, AT backbone, and neck parts. Similar to the human visual cognitive process, LKECM enhances the low-level features of the image based on large-kernel convolutions. The AT backbone is an improvement on the CSPDarknet53 in the YOLO framework, using the C2AT module to replace the original C2f to achieve a richer gradient information flow. In the neck part, we adopt the MAFPN design, replace C2f with C2AT, and by using the Multi-Scale Feature Fusion Unit (MSFFU) to replace the original Concatenate operation, a more accurate feature fusion result is achieved.

3.2 Large Kernel Edge Convolution Module

Target detection in forward-looking sonar images faces a significant challenge [18]: noise and complex backgrounds in underwater environments often lead to blurred target edges, making it difficult for models to accurately capture boundary information, which in turn results in poor performance. Traditional convolution operations typically focus on extracting global features while neglecting edge details, which is particularly problematic in underwater object detection. To address this issue, we introduces a novel Large Kernel Edge Convolution Module (LKECM), as shown in Fig 3, which combines the advantages of large-kernel convolution and edge detection convolution. It emphasizes edge information while extracting conventional features. Specifically, inspired by the considerations of large-kernel convolution in [5], we introduce a 9×9 large-kernel convolution in the feature preprocessing part to obtain a larger receptive field, which is crucial for the recognition tasks in downstream processing. Additionally, it acts as a high-pass filter, achieving implicit enhancement of high-frequency information.

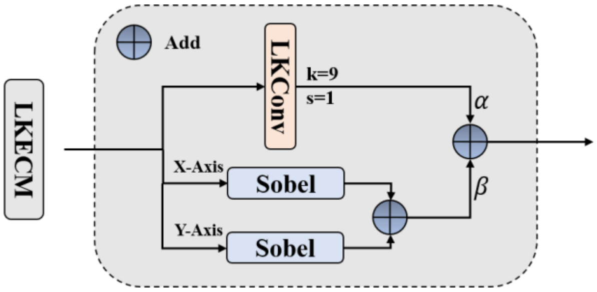


Fig. 3. The proposed Large Kernel Edge Convolution Module (LKECM) employs large-kernel convolutions to obtain a larger receptive field and multiple-dimensional edge convolutions, achieving more accurate low-level feature representation.

Edge information is processed by the lower-level regions of the human brain after receiving visual input; therefore, we simulate this process with Sobel convolution. The gradient magnitude is calculated as $\sqrt{G_x^2 + G_y^2}$ to obtain a combined edge feature map from the horizontal (G_x) and vertical (G_y) edges. The formula for the entire process is as follows:

$$Y = \alpha * LKEConv(x) + \beta * Add(Sobel_x(x), Sobel_y(y)) \tag{1}$$

where α and β are adjustable parameters that control the contribution of edge features.

3.3 C2AT

The backbone network of YOLO-VIS has been improved from the Backbone structure of YOLOv8. The original Backbone employs a series of convolutional and deconvolutional layers, using the C2f module as the basic building block to achieve good feature extraction capabilities. However, the C2f structure performs poorly in feature extraction on sonar images [3], against this backdrop, we propose a novel C2AT module, as shown in Fig 4, which resembles the intermediate feature processing part of the human brain and enhances the feature extraction capability. The C2AT module is derived from the C2F module in YOLOv8 and introduces an efficient Visual Attention (VisAT) mechanism. Traditional attention mechanisms have limitations that need to be addressed; one is the computation of attention for either positions or channels, and the other is the generation of attention weights through additional subnetworks, which does not align with the actual processing of neurons. To overcome these shortcomings, Lingxiao Yang and others have proposed an efficient attention mechanism module based on the human brain's attention mechanism. By designing an attention mechanism without additional parameters, it achieves accurate feature extraction.

To better implement attention, it is necessary to evaluate the importance of each neuron. In neuroscience, neurons rich in information typically exhibit firing patterns that differ from those of their surrounding neurons. Moreover, activated neurons usually inhibit their neighboring neurons, a phenomenon known as spatial suppression. In other words, neurons with a spatial suppression effect should be assigned higher importance. The simplest way to find these neurons is to measure the linear separability between a target neuron and other neurons. Based on these neuroscience findings, the following energy function is defined:

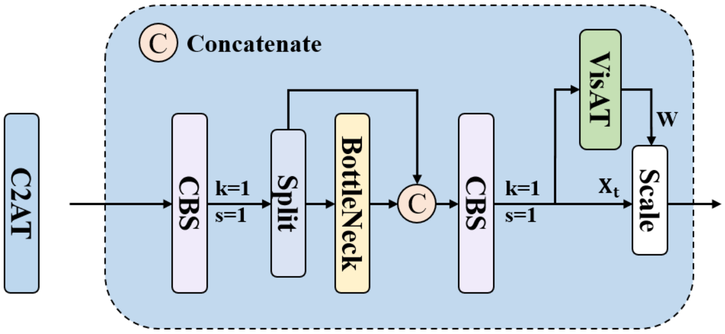


Fig. 4. The C2AT introduces the VisAT module into the original C2F module, incorporating neuron weight scaling for feature maps without designing additional sub-networks, thereby achieving feature attention functionality.

$$e_t(w_t, b_t, \boldsymbol{y}, x_i) = (y_t - \hat{t})^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}(y_o - \hat{x}_i)^2 \tag{2}$$

By introducing binary labels and adding a regularization term, the final energy function is defined as follows:

$$e_t(w_t, b_t, \boldsymbol{y}, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1 - (w_i x_i + b_t))^2 + (1 - (w_i t + b_t))^2 + \lambda w_t^2 \tag{3}$$

According to [35], an analytical solution is proposed for the equation.

$$\begin{cases} w_t = -\frac{2(t-\mu_t)}{(t-\mu_t)^2 + 2\sigma_t^2 + 2\lambda} \\ b_t = -\frac{1}{2}(t + \mu_t)w_t \end{cases} \tag{4}$$

Where

$$\begin{cases} \mu_t = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i \\ \sigma_t^2 = \frac{1}{M-1}\sum_{i=1}^{M-1}(x_i - \mu_t)^2 \end{cases} \tag{5}$$

Therefore, the goal of optimizing the energy function becomes solving the following equation:

$$e_t^* = \frac{4(\hat{\sigma}^2+\lambda)}{(t-\hat{u})^2+2\hat{\sigma}^2+2\lambda} \tag{6}$$

Eq. 6 indicates that the lower the energy, the more distinct the neuron is from its surrounding neurons, making it more important for visual processing. Studies indicate that attention regulation in the mammalian brain typically manifests as a gain (i.e., scaling) effect on neuronal responses. Therefore, we use a scaling operator instead of addition for feature refinement. The specific operation is shown in Eq. 7.

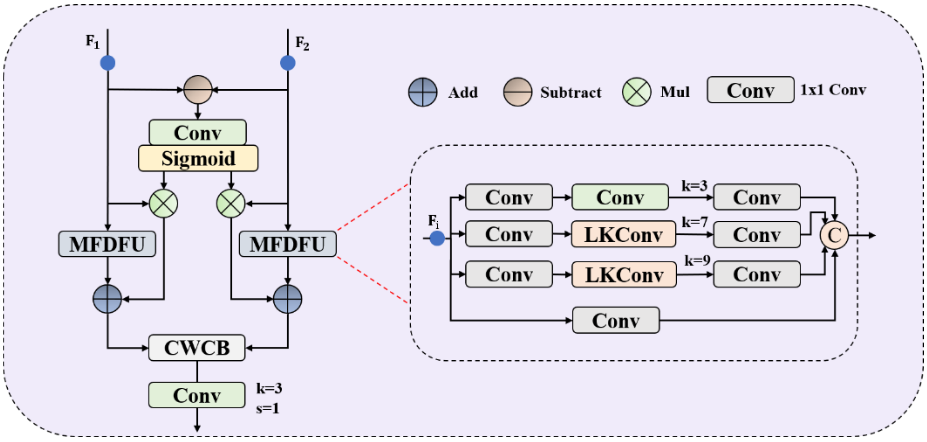


Fig. 5. The MSFFU module addresses the difficulty of multi-level feature interaction during feature fusion by designing deep fusion of multi-channel features and introducing large-kernel convolutions, achieving effective expression of multi-scale features.

$$\tilde{X} = sigmoid(\frac{1}{E}) \odot X \tag{7}$$

where E groups all inputs across channel and spatial dimensions. A sigmoid function is added to constrain excessively large values in E; it does not affect the relative importance of each neuron since the sigmoid is a unitary function.

3.4 Multi-Scale Feature Fusion Unit

Similar to how the human brain performs feature fusion in the IT area to aggregate different features into a unified representation and enhance the performance of subsequent recognition tasks, YOLOv8 employs the concept of Feature Pyramid Network (FPN) in part of its neck. However, this approach in forward-looking sonar image detection has limitations such as localized feature fusion, a tendency to use simple operations, insufficient utilization of multi-scale learning mechanisms, difficulties in multi-level feature interaction during feature aggregation, and a lack of consideration for the coupled interaction between features. Therefore, we introduce the MSFFU, which through a series of carefully designed operations, effectively integrates the semantic information from different feature maps and reduces noise interference, thereby addressing many of the issues with feature fusion in existing methods, as shown in Fig 5.

Specifically, for the two input feature maps F1 and F2, the following operations are performed: First, a pixel-wise subtraction is conducted between F1 and F2 to compute their difference. Subsequently, to ensure the non-negativity of the result, the absolute value of the subtraction outcome is taken. Next, to further extract and refine this difference information, a convolution operation is performed using a 3 × 3 convolution kernel, which aims to capture local spatial features. Finally, the result of the convolution is activated using a Sigmoid function, yielding a continuous attention map ranging between 0 and 1. This map reflects the degree of difference between the two feature maps at the pixel level, encoding complementary information between the feature maps.

Subsequently, a multi-scale feature learning mechanism is employed to enhance the effect of feature fusion, which realizes multi-scale integration based on convolutions with different kernel sizes, and is accomplished by the Multi-scale Feature Dense Fusion Unit (MFDFU). The MFDFU unit consists of convolution operations in four branches, where three branches perform specific convolution fusion operations, including one 3 × 3 convolution and two large-kernel convolutions, while the remaining branch operates with 1 × 1 convolutions. The results from these four branches are concatenated to obtain the fused feature. This process can be represented by Eq. 8.

$$Y = Concat\,(C3(X), LKC7(X), LKC9(X), C1(X)) \tag{8}$$

where LKC7 and LKC9 represent convolutional kernels of size 7 and 9, respectively.

Finally, feature fusion is performed through a Channel-wise Convolution Block (CWCB) [4], which specifically involves concatenating two feature inputs channel-wise, followed by a 3 × 3 depth-wise convolution operation [7] to obtain the fused difference feature C_i.

We applies the MSFFU to the upper and lower layer concatenation stage of the neck part in YOLOv8, aiming to address the issue of neglecting feature differences between layers during feature fusion in this part, while enhancing the multi-scale features of the neck part, thereby further improving the overall performance of the model.

4 EXPERIMENT AND RESULT ANALYSIS

4.1 Dataset and Experiments Setup

UATD (Underwater Acoustic Target Detection) dataset [33] is a multibeam forward-looking sonar dataset used for underwater target detection, collected by the Pengchen Lab using a Tritech Gemini 1200ik Multibeam Forward Looking Sonar in coastal and lake areas. The dataset comprises 10 classes, including cubes, spheres, cylinders, human bodies, cylindrical cages, rectangular cages, metal cans, tyres, aircraft, and BlueROV. It is divided into three separate sets: the training set containing 7600 images, the validation set containing 800 images, and the test set also containing 800 images.

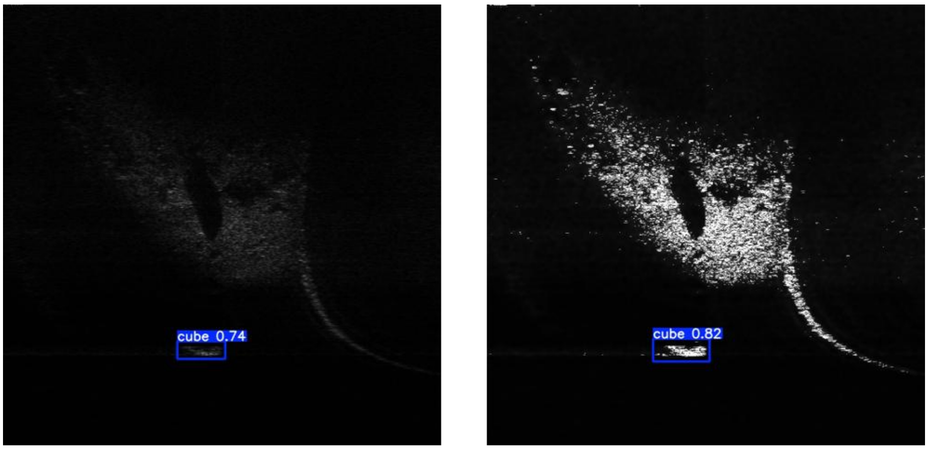
The computational platform used in this study is a work station equipped with an NVIDIA RTX 3090Ti, and the PyTorch serves as the software framework. The codebase employed is mmdetection.

4.2 Comparative experiment

In the comparison phase, we selected the state-of-the-art single-stage object recognition models represented by YOLO, including: YOLOv3, YOLOv5, YOLOv6, YOLOv7, YOLOv8, YOLOv9, YOLOv10, YOLO11 and YOLO12. Among these, the training

Table 1. Performance comparison of object detection model on the UATD dataset. P is the Params and G is GFLOPs. The bold represents the best results.

Type	Model	P	G	mAP	mAP@s	mAP@m	mAP@l
Two-Stage	Cascade R-CNN[1]	69.2	204.3	0.81	0.038	0.373	0.361
	Dynamic R-CNN[37]	68.7	198.5	0.814	0.128	0.367	0.357
	Faster R-CNN[21]	41.5	180.1	0.808	0.063	0.362	0.332
	Mask R-CNN[8]	44.1	192.6	0.783	0.117	0.35	0.336
	Sparse R-CNN[23]	38.9	175.4	0.824	0.151	0.351	0.353
Single-Stage	YOLOv3[20]	93.6	283.3	0.713	0.057	0.244	0.238
	YOLOv5[12]	9.1	24.0	0.814	0.025	0.372	0.388
	YOLOv6[14]	16.4	44.9	0.821	0.088	0.359	0.342
	YOLOv7[30]	**6.2**	**13.9**	0.802	0.04	0.32	0.318
	YOLOv8[11]	11.2	28.6	0.835	0.213	0.388	0.414
	YOLOv9[31]	7.2	26.7	0.816	0.188	0.352	0.455
	YOLOv10[29]	7.2	21.6	0.819	0.134	0.355	0.464
	YOLO11[13]	9.4	21.5	0.842	0.219	0.388	0.467
	YOLO12[25]	9.3	21.4	0.856	0.203	0.364	**0.477**
Ours	YOLO-VIS	9.6	25.7	**0.862**	**0.219**	**0.397**	0.449



(a) Without LKECM

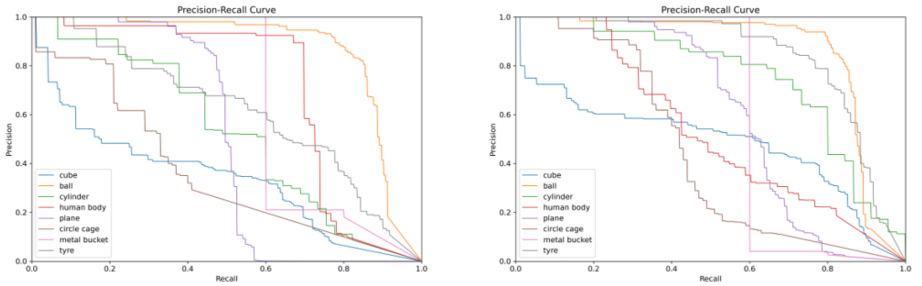
(b) With LKECM

Fig. 6. Recognition results of the improved method in real scenarios

and testing of all models were conducted based on the small size models. Meanwhile, we also compared several two-stage recognition models improved from R-CNN, including: Cascade R-CNN, Dynamic R-CNN, Faster R-CNN, Mask R-CNN, and Sparse R-CNN. As shown in Table 1, first, we can observe that two-stage models generally perform relatively well on most metrics. However, for the detection of small targets, the performance of these models is generally lower, which may be due to the fact that

Table 2. Ablation comparison of model performance improvement on the UATD dataset.

C2AT	MSFFU	LKECM	mAP	mAP@s	mAP@m	mAP@l
			0.835	0.213	0.388	0.414
√			0.848	0.221	0.383	0.468
√			0.837	0.183	0.384	**0.475**
√			0.849	0.214	0.394	0.446
√	√		0.857	0.193	0.387	0.424
√	√		0.855	0.217	0.385	0.447
√	√		0.839	0.205	0.396	0.458
√	√	√	**0.862**	**0.219**	**0.397**	0.449



(a) YOLOv8

(b)YOLO-VIS

Fig. 7. Comparison of the PR-curve before and after the improvement.

small targets are usually more difficult to detect and prone to being missed. The single-stage YOLO series models still achieve decent scores on some metrics; for instance, YOLO12 achieves 0.856 mAP, surpassing the performance of most two-stage models and attaining the best results among existing models. Notably, the YOLO series demonstrates particularly outstanding performance in large object detection.

For our proposed YOLO-VIS, the method achieved the highest scores on the vast majority of metrics listed. This indicates that YOLO-VIS not only surpasses existing methods in overall detection performance but also demonstrates outstanding effects in handling detection tasks for targets of varying sizes. This reflects that drawing inspiration from the human brain's recognition mechanism has a guiding role in improving the accuracy of underwater sonar image recognition. Fig 6 presents the performance results of our model on actual forward-looking sonar images.

Given the practical significance of model size and computational efficiency in real-world applications, we conducted a comprehensive analysis of parameter counts (Params) and computational complexity (GFLOPs) across all evaluated models. Our findings reveal that although the proposed model does not achieve optimal lightweight characteristics, it demonstrates comparable efficiency metrics to all two-stage architectures and remains competitive with the majority of single-stage counterparts. This balanced performance profile ensures robust practical applicability without compromising detection accuracy.

4.3 Ablation Experiment

To evaluate the practical effectiveness of individual components in our proposed method, we conducted systematic ablation studies on the UATD dataset. First, we present a comparison of the improved PR curves before and after the modifications, as shown in Fig. 7. The experimental results demonstrate that the enhanced YOLO-VIS achieves significant performance improvements across all detection categories compared to the baseline YOLOv8 model. The specific improvements in the results can be found in Table 2, which elucidates the impact of the proposed elements on the overall detection performance. The results indicate that the proposed LKECM, C2AT, and MSFFU modules all contribute to the enhancement of detection performance.

Specifically, LKECM significantly increases the performance of various evaluation metrics through preprocessing, suggesting that the priority given to low-level feature processing in the brain-inspired recognition mechanism is effective. The introduction of C2AT has led to a significant improvement in performance, indicating that a mechanism similar to the human brain's visual attention can be effectively applied in actual target recognition processes. The introduction of MSFFU has caused a slight decline in the recognition performance for small targets to some extent, but it has simultaneously led to a substantial improvement in the recognition performance for large targets. When multiple components are combined, the overall performance is enhanced, and we believe that this trade-off is acceptable. Through these three optimization measures, the best results on most metrics have been achieved when compared to the original YOLOv8 model.

5 Conclusion

This paper proposes a brain-inspired forward-looking sonar target recognition technology named YOLO-VIS. Firstly, in response to the characteristics of forward-looking sonar images, which have low brightness and unclear edges, the method employs large-kernel convolutions with a large receptive field and edge convolutions in different directions to achieve low-level feature enhancement. Then, without introducing additional sub-networks, an attention weighting mechanism based on neuron weights is designed and applied to the C2f module. Moreover, a multi-scale feature deep fusion method is adopted in the feature fusion stage to obtain more accurate feature representation. Extensive experimental results on the UATD dataset show that YOLO-VIS surpasses existing recognition methods in multiple performance metrics. In the future, we will explore its potential in real-world forward-looking sonar recognition tasks.

Acknowledgments. This work was supported by the Open Fund of State Key Laboratory of Acoustics under Grant SKLA202315.

References

1. Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6154–6162, 2018
2. Lichao Chen, Sudhir Singh, Thomas Kailath, and Vwani Roychowdhury. Brain-inspired automated visual object discovery and detection. Proceedings of the National Academy of Sciences, 116(1):96–105, December 2018.
3. Ruoyu Chen, Shuyue Zhan, and Ying Chen. Underwater target detection algorithm based on yolo and swin transformer for sonar images. In OCEANS 2022, Hampton Roads, pages 1–7, 2022.
4. Channel-Wise Convolutions. Channelnets: Compact and efficient convolutional neural networks via. IEEE Transactions on pattern analysis and machine intelligence43, 2021.

5. Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31×31: Revisiting large kernel design in cnns. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11953–11965, 2022.
6. Liangfu Ge, Premjeet Singh, and Ayan Sadhu. Advanced deep learning framework for underwater object detection with multibeam forward-looking sonar. Structural Health Monitoring, page 14759217241235637, 2024.
7. Yunhui Guo, Yandong Li, Liqiang Wang, and Tajana Rosing. Depthwise convolution is all you need for learning multiple visual domains. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 8368–8375, 2019.
8. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.
9. Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13708–13717, 2021.
10. Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. Sman: Stacked multi-modal attention network for cross-modal image–text retrieval. IEEE Transactions on Cybernetics, 52(2):1086–1097, 2022.
11. Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, January 2023.
12. Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, Yonghye Kwon, Kalen Michael, Jiacong Fang, Colin Wong, Zeng Yifu, Diego Montes, et al. ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearmland deci. ai integrations. Zenodo, 2022.
13. Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725, 2024.
14. Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976,2022.
15. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.
16. Huimin Lu, Yujie Li, Yudong Zhang, Min Chen, Seiichi Serikawa, and Hyoungseop Kim. Underwater optical image processing: A comprehensive review, 2017.
17. Ken Sinkou Qin, Di Liu, Fei Wang, Jingchun Zhou, Jiaxuan Yang, and Weishi Zhang. Improved yolov7 model for underwater sonar image object detection. Journal of Visual Communication and Image Representation, 100:104124, 2024.
18. Smitha Raveendran, Mukesh D Patil, and Gajanan K Birajdar. Underwater image enhancement: a comprehensive review, recent trends, challenges and applications. Artificial Intelligence Review, 54:5413–5467, 2021.
19. Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 779–788, 2016.
20. Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.
21. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 39(6):1137–1149, 2016.
22. Thomas Serre, Lior Wolf, and Tomaso Poggio. Object recognition with features inspired by visual cortex. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 2, pages 994–1000. Ieee, 2005.

23. Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14454–14463, 2021
24. Juan Terven and Diana Cordova-Esparza. A comprehensive review of yolo: From yolov1 to yolov8 and beyond. arXiv preprint arXiv:2304.00501, 2023.
25. Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524, 2025.
26. Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. arXiv preprint arXiv:2502.12524, 2025.
27. Leslie G Ungerleider and James V Haxby. ‘what’and ‘where’in the human brain. Current opinion in neurobiology, 4(2):157–165, 1994.
28. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones,Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc
29. Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458, 2024.
30. Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Train-able 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,pages 7464–7475, 2023.
31. Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. In European conference on computer vision, pages 1–21. Springer, 2025.
32. Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. ArXiv, abs/2006.04768, 2020.
33. Kaibing Xie, Jian Yang, and Kang Qiu. A dataset with multibeam forward-looking sonar for underwater object detection. Scientific Data, 9(1), December 2022.
34. Jian Yang, Jianfeng Wei, and Zhe He. A real-time underwater detector based on yolov5 using forward-looking sonar images. In 2023 8th International Conference on Robotics and Automation Engineering (ICRAE), pages 126–130. IEEE, 2023.
35. Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. Simam: A simple, parameter-free attention module for convolutional neural networks. In International conference on machine learning, pages 11863–11874. PMLR, 2021.
36. Haoting Zhang, Mei Tian, Gaoping Shao, Juan Cheng, and Jingjing Liu. Target detection of forward-looking sonar image based on improved yolov5. IEEE Access,10:18023–18034, 2022.
37. Hongkai Zhang, Hong Chang, Bingpeng Ma, Naiyan Wang, and Xilin Chen. Dynamic r-cnn: Towards high quality object detection via dynamic training. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28,2020, Proceedings, Part XV 16, pages 260–275. Springer, 2020.
38. Xu Zhao, Hongwei Dong, and Yuquan Wu. Improving class imbalanced few-shot image classification via an edge-restrained sindiffusion method. In 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 2126–2131, 2024.