

A-32 Improved YOLOv11 for Fish Detection in Real Farming Environment

Tian Ying*, Wang Sikun*, Lu Cunwei*
 (* FUKUOKA Institute of Technology)

1.Introduction

With the growing global demand for high-quality animal protein, aquaculture has become a key supporting force. It plays a crucial role in ensuring food security, reducing pressure on wild fish stocks, and promoting sustainable resource utilization. As of 2023, Japan's aquaculture production reached about 939,000 tons, ranking 12th globally. Currently, many fish farms still rely on manual observation or regular spot checks to monitor the fish population. However, this method has several problems, such as low efficiency, high error rate and high labor cost.

So improved YOLOv11 is used to realize the identification of fish and automatically count the number of fish in this paper.

2.Method

In this paper, improved YOLOv11 for fish detection in real farming environment is proposed. Firstly, make initial datasets by labeling and augmenting and train the initial dataset using YOLOv11. Next, add images containing folded fish and water ripples to the dataset and add ECA attention mechanism to YOLOv11 network structure. Finally, train the improved dataset using improved YOLOv11 and compare two results.

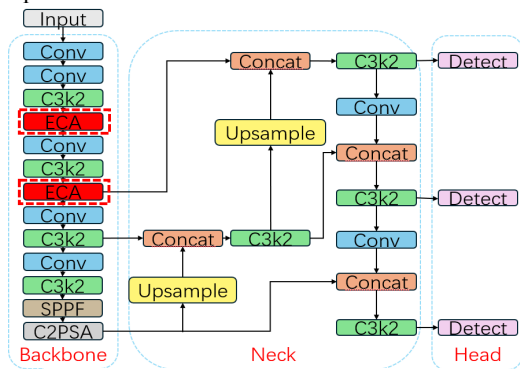


Figure 1 Improved YOLOv11 architecture

As shown in Figure 1, the YOLOv11 architecture consists of three main parts: backbone, neck, and head. Input images undergo feature extraction in the backbone through convolutional layers, C3K2(deep features), SPPF(multi-scale context), and C2PSA(cross-scale attention) modules, with resulting feature maps subsequently fused in the neck and processed by the head for detection.[1]

Two lightweight ECA attention layers are incorporated into the backbone to enhance feature representation, improving detection accuracy with minimal impact on inference speed by focusing on important channel features.

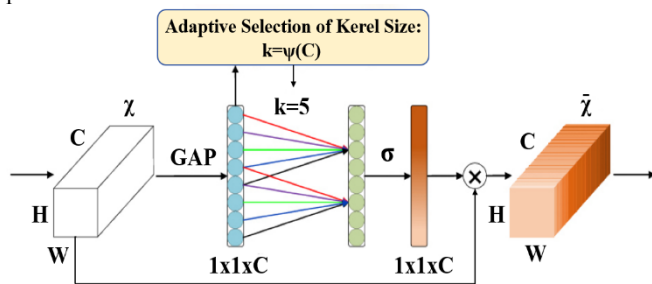


Figure 2 The Structure of ECA Attention

As shown in Figure 2, the ECA attention mechanism applies Global Average Pooling (GAP) to generate a channel descriptor. 1D convolution with kernel size 5 captures local cross-channel interactions without reducing dimensionality. The output is then passed through a Sigmoid function to generate attention weights, which are

multiplied with the original feature map for adaptive channel-wise recalibration.[2]

Initially, as shown in Figure 3, 4,800 annotated and augmented fish images formed the dataset. To improve robustness and detection accuracy, additional images containing overlapping fish and water ripples were incorporated, resulting in a final dataset of 9,344 images.



Figure 3 Initial Images

3.Result

As shown in Figure 4 and Figure 5, missed detections and false detections in areas with overlapping fish and water ripples have been improved. As shown in Table 1, improved YOLOv11 achieved an AP50-95 of 87.7%, which is a 1.6% improvement over YOLOv11. In addition, the number of parameters increased by only 7,680, and the computational cost rose by just 0.1G FLOPs.

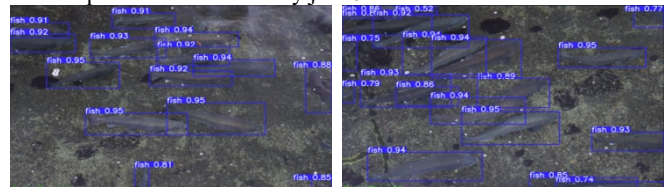


Figure 4 Testing Results of Improved YOLOv11

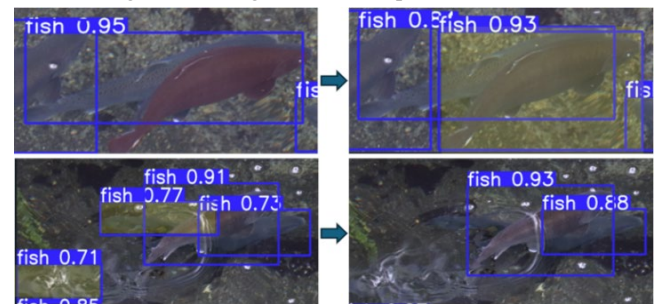


Figure 5 Results Before and After Improved

Table 1 Results of YOLOv11 and Improved YOLOv11

	AP50	AP50-95	Params	Flops
YOLOv11	99.2%	86.1%	2582355	6.3G
Improved YOLOv11	99.2%	87.7%	2590035	6.4G

To improve the detection of fish in real environments, this paper integrates the ECA attention into the YOLOv11 architecture. The results show this modification helps reduce detection errors in overlapping fish and water ripples, demonstrating its practical effectiveness.

References

- [1] Khanam R, Hussain M. Yolov11: An overview of the key architectural enhancements[J]. arXiv preprint arXiv:2410.17725, 2024.
- [2] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11534-11542.