IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 April 2024, accepted 14 May 2024, date of publication 23 May 2024, date of current version 31 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3404867

RESEARCH ARTICLE

YOLO8-FASG: A High-Accuracy Fish Identification Method for Underwater Robotic System

XIANGRONG QIN[1], CHANGDONG YU[2], (Member, IEEE), BAISHENG LIU[3], AND ZHIHAO ZHANG[2]

[1]College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China
[2]College of Artificial Intelligence, Dalian Maritime University, Dalian 116026, China
[3]College of Software, Liaoning Technical University, Huludao 125000, China

Corresponding author: Changdong Yu (ycd@dlmu.edu.cn)

ABSTRACT In underwater robots, accurate identification of small fish is still a major challenge, because small fish move faster and occupy less screen space, which requires higher detection flexibility and receptive field of the model. To solve this challenge, we propose a high-precision small fish identification and tracking method named YOLO8-FASG in this paper. Specifically, the proposed method is improved in three aspects based on the YOLOv8 framework. First, Alterable Kernel Convolution (AKConv) is used in the neck network of the model to automatically adjust the shape of the convolution kernel according to the size and shape of the object. In this way, the shape and contour characteristics of rapidly changing fish can be captured more accurately and efficiently; Second, we introduce a global attention mechanism (GAM) to broaden the receptive field of the model by enhancing attention to fish features from the two dimensions of channel and space; Third, we employ Simplified Spatial Pyramid Pooling-Fast (SimSPPF) to replace the standard Spatial Pyramid Pooling-Fast (SPPF) to enhance prediction accuracy. These improvements enable the model to effectively extract image features of small, fast fish, thereby improving the robot's accuracy in identifying small fish underwater. Experiments results in the public dataset Fish4Knowledge show that YOLO8-FASG performs significantly better than traditional YOLOv8 in underwater environments. Specifically, Precision and Recall increased by 1.6% and 3.5% respectively, while mAP50 and mAP50-95 increased by 1.3% and 6.1% respectively, and our method provides an effective solution for underwater robots to identify fish schools.

INDEX TERMS YOLOv8, global attention mechanism, AkConv, SimSPPF, object detection.

I. INTRODUCTION

An underwater fishing robot is an autonomous robotic system specifically designed to perform a variety of tasks in underwater environments. Nowadays, underwater fishing missions play an important role in marine resource development [1], scientific research and rescue operations. However, challenges such as the complexity of the underwater environment and communication limitations make underwater fishing tasks extremely difficult. In order to effectively deal with these challenges, it is critical to improve the object detection and recognition capabilities of underwater fishing robot systems.

Currently, conventional object detection methods face a series of problems in underwater environments, including uneven illumination, blurred water quality [2], and diverse object shapes. Therefore, the improvement of object detection methods has become an urgent need in the field of underwater fishing robots. As an advanced object detection method, YOLOv8 [3] is fast and accurate, and has broad application prospects in underwater fishing tasks. This article aims to introduce the improved method of YOLOv8 in underwater fishing robots to improve the efficiency and accuracy of underwater fishing tasks.The diagram of identifying small objects underwater is shown in Figure 1.

Considering the complexity of the underwater environment and challenges of underwater fishing tasks, researchers are committed to improving underwater object detection and

The associate editor coordinating the review of this manuscript and approving it for publication was Sudhakar Radhakrishnan.

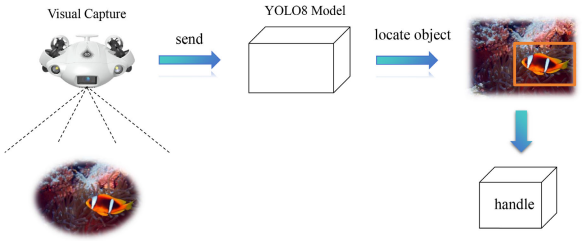

FIGURE 1. Underwater vehicle object identification process schematic.

recognition technology. Initially, Rova et al. [4] proposed a method for fish identification in underwater videos using deformable template matching, effectively combining shape context to enhance recognition accuracy. Matai et al. [5] then introduced a computer vision-based method for automatic detection and identification of fish species, significantly improving accuracy by integrating background separation and species recognition methods. Hsiao et al. [6] developed a maximum probability partial ranking method based on Sparse representation (SRC-MP), which significantly improved the accuracy of real underwater observation video data by effectively extracting feature with Eigenfaces and Fisherfaces. Palazzo and Murabito [7] demonstrated the efficient application of EMK and KDES kernel descriptors for the classification task of about 50,000 underwater images of 10 fish species under MAED 2014, significantly enhancing recognition precision.

With the gradual development of deep learning, especially the application of CNN in the field of vision, it has been proven to be an effective method for fish classification and detection. Salman et al. [8] proposed a deep learning method for fish classification in unconstrained underwater environments, significantly enhancing classification accuracy through a hierarchical feature combination of convolutional neural networks. Zhuang et al. [9] achieved a breakthrough in the SEACLEF-2017 task using advanced deep learning models, significantly improving marine life identification with pre-trained networks and BN-Inception networks. Zhao et al. [10] introduced a method combining modified motion impact graphs and Recurrent Neural Networks (RNN), significantly enhancing detection and recognition accuracy for monitoring local abnormal behaviors in densely populated aquaculture. Zhou et al. [11] proposed an automatic assessment method for fish feeding intensity based on Convolutional Neural Networks (CNN) and machine vision, significantly improving assessment accuracy through data augmentation and CNN model training.

As YOLO became prominent in the field of visual detection, more researchers innovated on its basis to make it more suitable for fish detection tasks. Cai et al. [12] put forward a novel fish detection method integrating YOLOv3 and MobileNetv1, significantly enhancing detection precision in aquaculture through optimized feature map selection. Jalal et al. [13] introduced a hybrid scheme combining optical flow and Gaussian Mixture Models with YOLO deep neural networks, effectively achieving fish detection and classification in dynamic underwater environments. Yu et al. [14] specifically proposed an improved U-YOLOv7 framework based on the YOLOv7 model for efficient and accurate underwater biological detection. Yu et al. [15] introduced an improved YOLOv5 model (termed TRH-YOLOv5) for the detection and counting of underwater fish lateral line scales, significantly enhancing recognition accuracy through transformer modules. Cai et al. [16] proposed an interesting NAM-YOLOV7 method for rapid detection for fish with SVC symptoms, which employs the NAM attention mechanism to extract features accurately and significantly improve detection efficiency.

Overall, identifying fish is an extremely challenging task in underwater environments. The environment is dark and blurry, and previous studies have performed poorly in fish recognition, with weak detection accuracy and weak image recognition capabilities. These studies have common problems: First, using square convolution kernels to process fish school images does not fully consider the outline characteristics of fish, resulting in poor recognition results; Then, the distance changes of underwater fish schools are not well captured, and spatial and channel features are not effectively extracted. In response to the above difficulties, we put forward the YOLO8-FASG method to effectively solve the above problems. In summary, our contributions in this article are described as follows:

- We embed the Alterable Kernel Convolution(AKConv) to the model to adaptively adjust the shape of the convolution kernel and capture fish-shaped outlines more effectively, thus improving the accuracy and efficiency of fish recognition.
- Global Attention Mechanism(GAM) is introduced to enhance the focus on fish characteristics from two dimensions: channel and space, further improving the accuracy and performance of the model.
- We also adopt Simplified Spatial Pyramid Pooling-Fast(simSPPF) instead of the traditional Spatial Pyramid Pooling-Fast(SPPF), which simplifies the activation function and makes the model enhance prediction accuracy.

The remaining parts of this paper are arranged as follows: Section II provides a detailed description of the improvements made to YOLOv8, including GAM, SimSPPF, and AKConv, along with the presentation of the novel framework of YOLO8-FASG. Section III presents and analyzes the experimental results of the model on open-source datasets. Finally, conclusions are drawn in Section IV.

II. OUR PROPOSED METHOD

In this section, we first introduce the structure of YOLOv8 framework and then describe the details of the improvements of our proposed method.

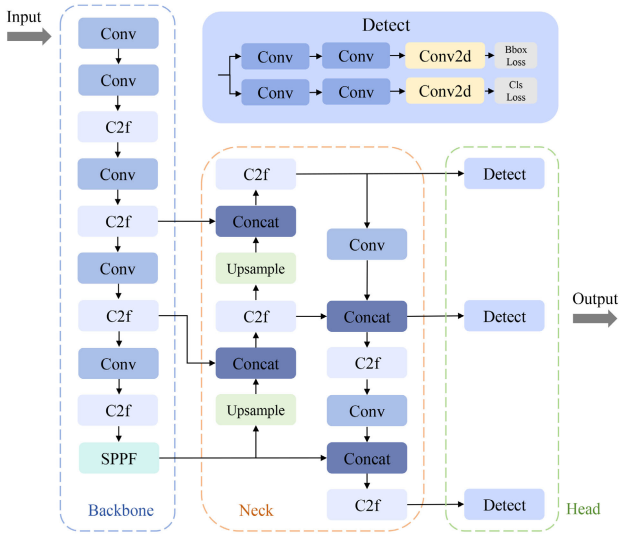

FIGURE 2. Structural diagram of YOLOv8 model.

A. YOLOv8 FRAMEWORK

YOLOv8[1] is a object detection model released by the Ultralytics team in 2023 [17], which is a detection method that can quickly and accurately identify objects. This model improves the architecture based on YOLOv5 [18] and combines the advantages of multiple object detectors. The structure diagram of the model is as shown in Figure 2.

The backbone network is the basis of the model and is responsible for extracting features from the input image. In YOLOv8, the backbone network mainly adopts the CSP Darknet [19] structure similar to YOLOv5, and introduces cross-stage connections between different stages of the network. It directly connects some feature maps with feature maps at subsequent levels to enhance the network's information transfer and feature reuse capabilities. As depicted in Figure 2, the convolution kernels Conv and C2f are repeatedly stacked with each other for feature extraction. Finally, SPPF is employed at the end of the backbone, which captures the spatial information of objects by introducing pooling windows of different scales. By this means, it is able to improve the model's detection capabilities for objects of different scales.

The neck network is located between the backbone network and head network, and is used to perform feature fusion and enhancement. This network can reduce the number of parameters while maintaining the expressive ability of the network. Specifically, it upsamples the processed backbone information through the Upsample layer to increase the size of the feature map, and then combines features of different scales through a series of concat layer splicing and C2f layers. Graphs are fused to produce richer information and diverse feature representations. The feature maps are finally fused to produce richer information and diverse feature representations.

The head network is located at the top of the entire model, and its function is to achieve the final detection task. Features extracted through the trunk and neck are fed to the head network for object feature decoding and result generation. The head layer of YOLOv8 no longer uses the coupling head of YOLOv5, but becomes the mainstream decoupling head Decoupled-Head [20], from Anchor-Based to Anchor-Free [21]. It no longer relies on predefined anchor boxes, but directly locates and detects objects on the feature map. In this way, it gets rid of the dependence on predefined anchor boxes and reduces the number of hyperparameters that need to be considered when designing the model. This makes the model simpler and easier to train. At the same time, it adopts the idea of DFL (Distributional Focal Loss) [22], focusing on difficult samples to solve long-tail classification problem.

B. IMPROVED YOLOv8 METHOD

In this paper, we improve the basic framework of YOLOv8 to make it more suitable for underwater fishing scenarios. The modified framework is shown in Figure 3 and called YOLO8-Fish AKConv SimSPPF GAM(YOLO8-FASG). Our main improvements are as follows: (1) Introduce GAM module in the backbone network; (2) Replace SPPF with SimSPPF; (3) Replace the convolution layer in the neck network with AKConv.

1) GAM MODULE

In a blurry underwater environment, the recognition accuracy will be greatly reduced. In order to enhance the model to more clearly extract the characteristics of fish species at the bottom of the water, we considered adding GAM to help to solve this problem.

As for GAM, it aims to address the problem of insufficient information retention in the channel and spatial dimensions of traditional attention mechanisms, thereby reducing information loss and amplifying the global-dimensional interactive features [23]. This mechanism adopts a sequential channel-space attention mechanism, where the channel attention sub-module employs 3D arrangement to retain information across three dimensions, and enhances the channel-space dependence across dimensions through a two-layer MLP. The internal schematic diagram of GAM is as shown in Figure 4. In the spatial attention sub-module, in order to better focus on spatial information, GAM uses two convolutional layers for spatial information fusion, while removing the maximum pooling operation that may lead to information reduction, stably improving performance. The specific details of integrating the module into our model framework are above the Figure 3. We assume that the input image feature is ϕ_{in}, the output result is ϕ_{out}, the channel function is set to M_c, and the spatial attention function is set to M_s. Then the formula of GAM is as shown in Formula 1.

$$\phi_{out} = M_s\left(M_c\left(\phi_{in}\right) \otimes \phi_{in}\right) \otimes \left(M_c\left(\phi_{in}\right) \otimes \phi_{in}\right) \quad (1)$$

[1]https://github.com/ultralytics/ultralytics

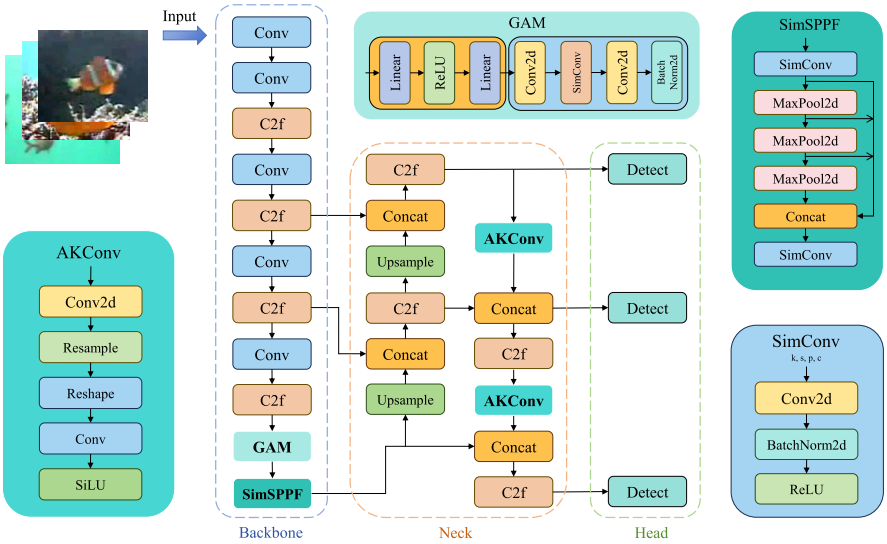

FIGURE 3. Structural diagram of our proposed YOLO8-FASG structure.

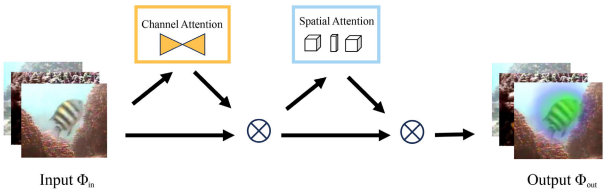

FIGURE 4. Attention effect demonstrated by GAM submodule.

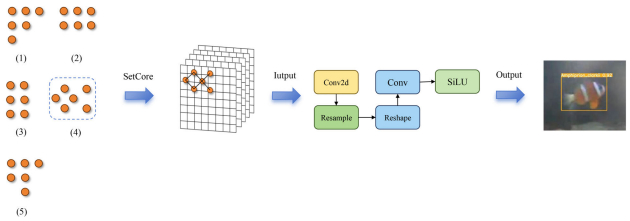

FIGURE 5. AKConv's initial convolution kernel setting and processing.

2) SimSPPF

In the original YOLOv8 model, the interior of the SPPF module uses the SiLU activation function, which is described in Formula 2. Although SiLU has the characteristics of smooth curves, the price is increased computational complexity. Hence, we replace SPPF with SimSPPF [24], that is, replace SiLU with ReLU activation function inside the network structure. This is able to achieve the effect of reducing detection time and improving model generalization ability through simple calculation. The improved hierarchical parts are shown in Figure 3. Suppose that the output of the ReLU function is σ and the input vector of the previous layer of neural network is x, and the nonlinear output result of the neurons after linear transformation is described in Formula 3.

$$SiLU\ (\alpha) = \frac{\alpha}{1 + e^{-\alpha}} \tag{2}$$

$$\sigma = max\left(0, w^T x + b\right) \tag{3}$$

3) AKConv

For the convolution operation, we adopt the novel AKConv convolution method proposed in 2023 [25], [26], and the AKConv structure is shown in Figure 5. It provides a flexible convolution mechanism that allows the convolution kernel to have any number of parameters and sampling shapes, and the size and shape can be adjusted according to actual needs to more effectively adapt to changes in different datasets and objects. Based on the characteristic contours of fish, we design the initial sampling convolution kernel (see Figure 5), and select convolution kernels similar to fish features for feature extraction. AKConv can also dynamically adjust the size and shape of the convolution kernel during the detection process, which can better extract fish features during dynamic monitoring.

For the AKConv process, we can describe it in the following steps: First, obtain the offsets corresponding to the convolutional kernel through convolutional operations, resulting in offsets of size $(B, 2N, H, W)$, denoted as offset. The formula for this process is given by Formula 4. Here, X represents the input feature map, B represents the batch size, N represents the size of the convolutional kernel. H represents the height of the input feature map, W represents the width of the input feature map, and C represents the number of channels in the input feature map. The term $2N$ arises because each position has both horizontal and vertical offsets. Next, by adding the offset P_n to the original coordinates P_0, we obtain the modified coordinates P_{new}, as shown in Formula 5. Finally, interpolation and resampling are performed based on the modified coordinates to obtain the corresponding feature values Γ, as shown in Formula 6. Here, Conv2d represents the convolution operation, and Resample

FIGURE 6. Some examples in the A dataset from Fish4Knowledge.

represents the interpolation and resampling operation.

$$offset = Conv2d\ (X) \tag{4}$$

$$P_{new} = P_0 + P_n \tag{5}$$

$$\Gamma = \text{Re}sample\ (\Gamma_{in}, P_{new}) \tag{6}$$

III. EXPERIMENTS

A. DATASETS

In this section, we use the public Fish4Knowledge dataset to verify the effectiveness of our improved method. The images in the Fish4Knowledge dataset are derived from fixed cameras capturing and monitoring marine ecosystems on Taiwan's coral reefs. The construction process of the dataset lasted more than 4 years, and more than 700000 underwater images are obtained. This dataset contains two parts, dataset A[2] and dataset B. As for dataset A, it contains 27370 images of 23 fish species, and Figure 6 presents some example images in the dataset.

The used dataset in this study is to obtain images of 23 species of fish from dataset A, and we use data augmentation to increase the number of each fish to 500, for a total of 11500 images. Then, we divide the training set, validation set and test set according to the ratio of 7:2:1.

B. EVALUATION METERICS

For testing model performance, precision and recall are indispensable evaluation indicators. Among them, precision refers to the proportion of samples predicted as positive by the model that are actually positive. Recall rate refers to the proportion of positive samples successfully identified by the model to the total positive samples. It measures how many true positive examples the model successfully predicted as positive examples. The calculation formulas of precision rate and recall rate are described as follows:

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

In addition, another evaluation index F1 is the harmonic average of precision and recall, which takes into account the precision and recall performance of the model. The higher the F1 score, the better the model achieves a balance between precision and recall. The calculation formula of $F1$ score is shown in Formula 9.

$$F1 = \frac{2 \times P \times R}{P + R} \tag{9}$$

Furthermore, mAP50 and mAP50-95 are the average accuracy calculated when the IOU threshold is set to 0.50. Specifically, for each category, we first calculate the precision-recall curve based on the IOU between the predicted box and the true box, then calculate the area under the curve, and finally average the areas across all categories. We set the average precision as AP, the precision rate at the i-th recall level as $P(i)$, and the increment of the recall rate as $dR(i)$. The formulas for AP and mAP can be obtained as follows:

$$AP = \sum_{i=1}^{n} P\ (i)\ dR\ (i) \tag{10}$$

$$mAP = \frac{1}{n} \sum_{j=1}^{n} AP_j \tag{11}$$

C. TRAINING DETAILS

This experiment is conducted on a laptop with Windows 10 operating system. The computer is equipped with a 13th generation Intel(R) Core(TM) i7-13700K processor clocked at 3.40 GHz and an NVIDIA GeForce RTX 4090Ti (24GB memory). The proposed networks are realized by the

[2]https://www.heywhale.com/mw/dataset/5e55f7960e2b66002c245df5

TABLE 1. Ablation experimental data.

Models	Precision	Recall	mAP50	mAP50-95	F1
YOLOv8n	0.950	0.933	0.978	0.817	0.941
YOLOv8-SimSPPF	0.955	0.940	0.978	0.820	0.947
YOLOv8-GAM	0.962	0.921	0.978	0.819	0.941
YOLOv8-AKConv	0.958	0.959	0.983	0.846	0.958
YOLOv8-SimSPPF+AKConv	0.973	0.961	0.989	0.865	0.967
YOLOv8-AKConv+GAM	0.969	0.964	0.986	0.860	0.966
YOLOv8-SimSPPF+GAM	**0.976**	0.963	0.988	0.865	0.969
YOLOv8-AKConv+GAM+SimSPPF	0.966	**0.971**	**0.991**	**0.878**	**0.968**

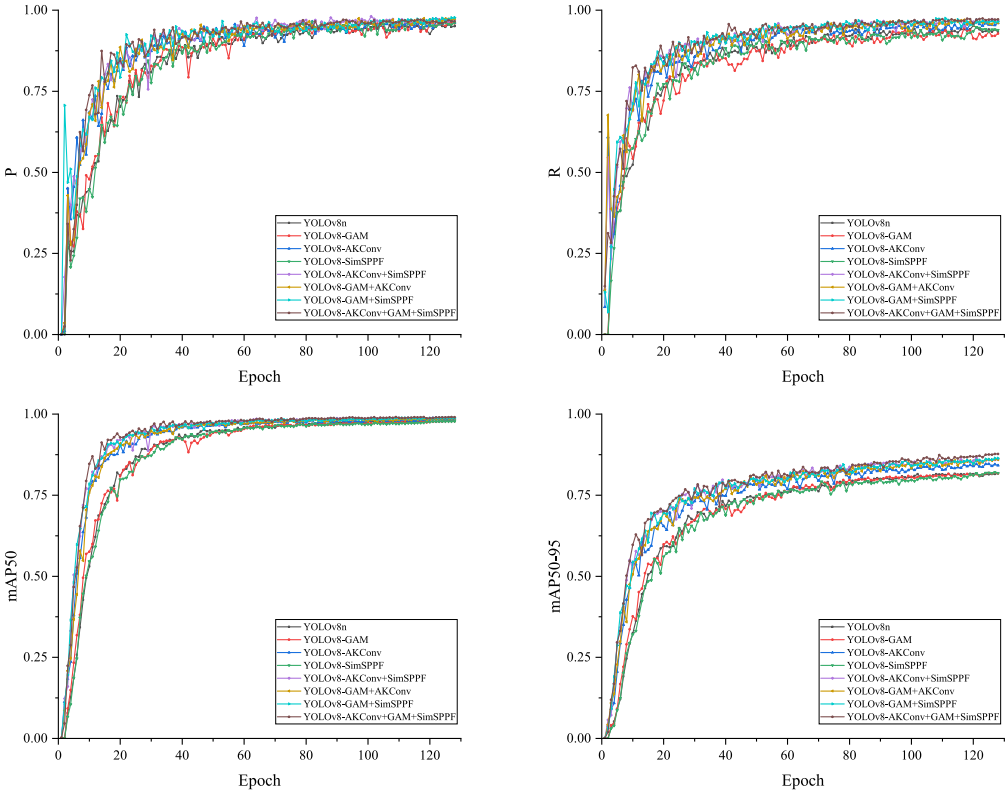

FIGURE 7. Ablation experimental data results of YOLOv8-FASG.

Python 3.8.15 and deep learning framework PyTorch. During the training process, the input image size is set to 80 × 60. We adopt the AdamW optimizer and the learning rate is initialized to 1e-3, and batch size is set to 32. We stop training the model until the trained model reaches convergence on validation set (128 epoch).

D. RESULTS AND ANALYSIS

1) ABLATION EXPERIMENT

To verify the effectiveness of the improved method, we conduct multiple ablation experiments on the public underwater fish dataset Fish4Knowledge. The relevant experimental results are shown in Table 1 and Figure 7.

Initially, it's evident that most indicators of the improved model on the underwater fish dataset surpass those of other ablation methods, especially in Recall, mAP50, and mAP50-95. Compared with the original model, the improved model has enhanced various indicators by 1.6% to 3.5% and 1.3% to 6.1%, respectively, highlighting the advantages of our improved method in enhancing detection accuracy.

Subsequently, when comparing the enhancements between single-module and dual-module approaches, it becomes evident that YOLOv8-AKConv performs the best in the underwater fish species recognition task. This superiority stems from the enhanced capability of fish-shaped dynamic convolution kernels in efficiently and accurately extracting features of small fish. Among dual-module models, YOLOv8-SimSPPF+GAM emerges as the top performer, with GAM initially aggregating the extracted spatial-channel information, followed by SimSPPF for multi-scale feature pooling. A comprehensive comparison of the data in the table reveals that GAM, AKConv, and SimSPPF all significantly contribute to the model's predictive accuracy.

Furthermore, observing the curve in Figure 7, we notice that the improved model exhibits superior stability in mAP50 compared to other models. The growth trends and

TABLE 2. Experimental data comparing multiple models.

Models	Precision	Recall	mAP50	mAP50-95	F1
YOLOv3-tiny	0.922	0.920	0.959	0.718	0.921
YOLOv5n	0.931	0.932	0.965	0.798	0.931
YOLOv6n	0.932	0.935	0.969	0.823	0.933
YOLOv8	0.950	0.933	0.978	0.817	0.941
YOLO8-FASG	**0.966**	**0.971**	**0.991**	**0.878**	**0.968**

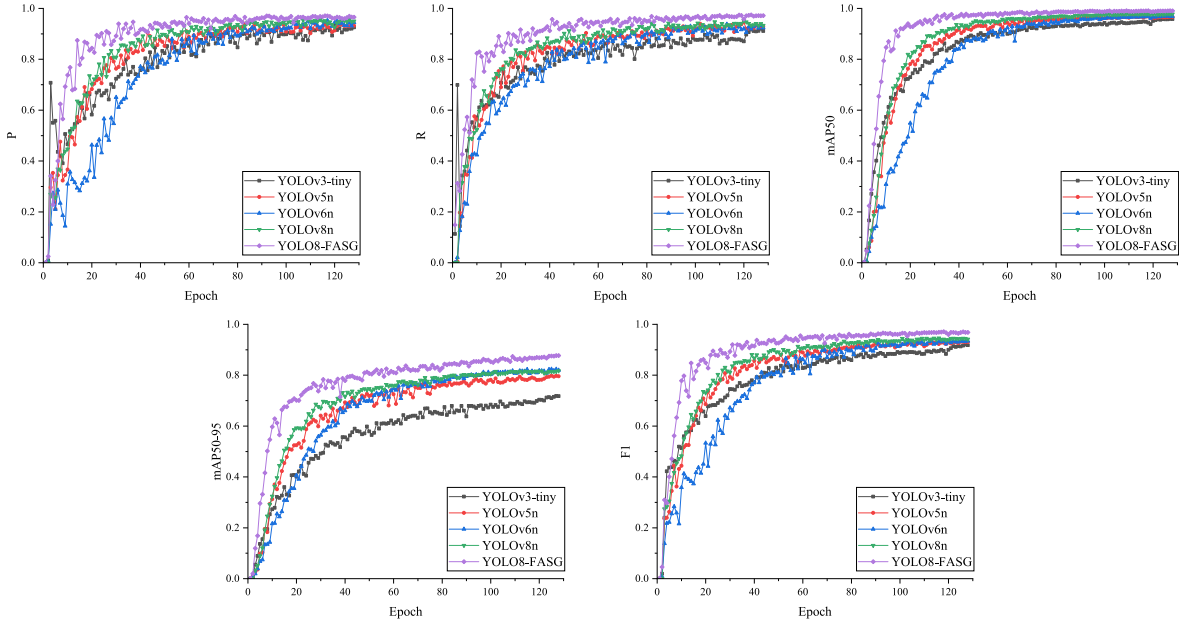

FIGURE 8. Comparison of model indicators of yolo series.

convergence speeds on other curves are faster than those of other improved submodules. In the initial growth stage of the Recall curve, YOLOv8-GAM+AKConv experiences significant fluctuations, mainly due to AKConv's adjustment of appropriate convolutional kernel shapes during early training. Additionally, in mAP50, we observe that the single-module GAM and single-module SimSPPF have similar growth convergence speeds to YOLOv8n. However, with the inclusion of AKConv in YOLOv8n-AKConv, the overall improvement benefits significantly. These experimental results indicate the pronounced effectiveness of variable convolutional kernels.

Finally, based on the aforementioned advantages, the YOLO8-FASG model integrates the global feature focusing functions of GAM and SimSPPF, as well as the spatial and channel information focusing functions of GAM, while employing the dynamic detection neck AKConv to adeptly identify small objects. This integration notably enhances detection accuracy, rendering it more suitable for underwater fish object recognition tasks.

2) COMPARISON EXPERIMENT OF TYPICAL MODELS OF YOLO SERIES

In the comparative experiments, we evaluated the performance of five different versions of the model, namely YOLOv3, YOLOv5, YOLOv6, YOLOv8, and the improved model YOLO8-FASG, on the object detection task with 128 epochs. We focused on comparing their performance across Precision (P), Recall (R), mAP50, mAP50-95, as well as F1 indicators.

Firstly, upon examining the experimental data Table 1, it is evident that the improved YOLO8-FASG model outperforms others across all indicators. This indicates significant advantages of our improvement method in underwater fish identification tasks.

Secondly, by observing the mAP50 and mAP50-95 curves in Figure 8, we notice that the YOLO8-FASG model exhibits a more pronounced upward trend in the first 10 rounds compared to other models. Furthermore, in the subsequent training process, its curve convergence speed and stability surpass those of other models. Despite the initial fluctuations in Precision and Recall for YOLOv3-tiny, its curve values are higher than those of YOLO8-FASG; however, between rounds 10 and 40, YOLO8-FASG demonstrates significantly faster curve convergence.

A comprehensive analysis reveals that the improved model surpasses others in terms of training speed and stability. This can be attributed to the effectiveness of our proposed new method in capturing fish characteristics and optimizing model structures. Notably, the utilization of AKConv adaptive convolutional kernel shape, GAM global

FIGURE 9. Comparison results of real scenarios under the test set.

FIGURE 10. The detection performance of YOLO8-FASG in real oceanic environments.

attention mechanism, and SimSPPF simplified spatial pyramid pooling method collectively provide robust support for enhancing model performance during training. Consequently, the improved model demonstrates heightened accuracy and stability in underwater fish target recognition tasks, offering a dependable solution for practical applications.

3) MODEL COMPARISON EXPERIMENT UNDER TEST SET

To validate the model's detection capabilities in real-world scenarios, we reserved a dedicated test dataset within our dataset. In this test set, we enumerated several common tropical fish species as objects for comparative experiments, as depicted in Figure 9 (from left to right: v3, v5, v6, v8, v8-FASG). The results of the comparative experiments clearly indicate that YOLO8-FASG exhibits superior performance in terms of detection accuracy, with an average detection accuracy of 0.927. Compared to the average accuracies of YOLOv8 (0.867) and YOLOv6 (0.827), YOLO8-FASG's accuracy is improved by 6% and 10%, respectively. Additionally, we subjected YOLO8-FASG to more challenging tests using real-world datasets to simulate changing water conditions and uncontrolled marine environments. In Figure 10, we observe that even in dimly lit underwater environments with blurred targets and backgrounds, YOLO8-FASG is still able to accurately locate targets. This result further validates the effectiveness of our proposed improvement method in underwater fish target recognition tasks and robustly demonstrates our model's capability to handle real-world marine small fish detection tasks.

IV. CONCLUSION

In this article, we propose a more flexible and larger receptive field deep learning model, named YOLO-FASG, to address the issue of low detection accuracy for small underwater fish species. By introducing AKConv into the neck network of the YOLOv8 model architecture, the model can adjust the shape of the convolution kernel adaptively to the fish contour, and capture the rapidly changing fish contour more accurately, thereby improving detection accuracy. Furthermore, the introduction of the GAM and the adoption of SimSPPF further enhance the model's focus on spatial-channel features of fish and receptive field, making it more effective in identifying small fish species in underwater environments.

Experimental results show that compared with the traditional YOLOv8, YOLO8-FASG has significantly improved performance indicators such as Precision, Recall, mAP50 and mAP50-95, providing a more reliable solution for underwater robots to identify fish schools. Among them, Precision and Recall increased by 1.6% and 3.5% respectively, while mAP50 and mAP50-95 increased by 1.3% and 6.1% respectively. These data show that the proposed method improves detection accuracy while maintaining good performance stability, providing important technical support for the accurate identification of small fish in underwater environments.

In summary, the YOLO8-FASG method proposed in this study has achieved remarkable results in identifying small fish by underwater robots, providing a more reliable and efficient technical means for future underwater robot applications. In the future, we will further optimize the method and explore its application in other underwater scenarios to meet a wider range of requirements.

REFERENCES

[1] E. Zereik, M. Bibuli, N. Mišković, P. Ridao, and A. Pascoal, "Challenges and future trends in marine robotics," *Annu. Rev. Control*, vol. 46, pp. 350–368, 2018.
[2] Z. Li, D. Cai, J. Wang, Y. Li, G. Gui, X. Sun, N. Wang, J. Zhang, H. Liu, and G. Wang, "Machine learning based dynamic correlation on marine environmental data using cross-recurrence strategy," *IEEE Access*, vol. 7, pp. 185121–185130, 2019.
[3] J. Glenn. (2023). *Ultralytics YOLOv8*. [Online]. Available: https://github.com/ultralytics/ultralytics
[4] A. Rova, G. Mori, and L. M. Dill, "One fish, two fish, butterfish, trumpeter: Recognizing fish in underwater video," in *Proc. MVA*, 2007, pp. 404–407.
[5] J. Matai, R. Kastner, G. R. Cutter, and D. A. Demer, "Automated techniques for detection and recognition of fishes using computer vision algorithms," in *Proc. Nat. Mar. Fisheries Service Automated Image Process. Workshop*, K. Williams, C. Rooper, and J. Harms, Eds., Sep. 2010, Paper no. NMFS-F/SPO-121.
[6] Y.-H. Hsiao, C.-C. Chen, S.-I. Lin, and F.-P. Lin, "Real-world underwater fish recognition and identification, using sparse representation," *Ecological Informat.*, vol. 23, pp. 13–21, Sep. 2014.
[7] S. Palazzo and F. Murabito, "Fish species identification in real-life underwater images," in *Proc. 3rd ACM Int. Workshop Multimedia Anal. Ecol. Data*, France, Nov. 2014, pp. 13–18.
[8] A. Salman, A. Jalal, F. Shafait, A. Mian, M. Shortis, J. Seager, and E. Harvey, "Fish species classification in unconstrained underwater environments based on deep learning," *Limnol. Oceanogr. Methods*, vol. 14, no. 9, pp. 570–585, Sep. 2016.
[9] P. Zhuang, L. Xing, Y. Liu, S. Guo, and Y. Qiao, "Marine animal detection and recognition with advanced deep learning models," in *Proc. CLEF Working Notes*, 2017, pp. 166–177.
[10] J. Zhao, W. Bao, F. Zhang, S. Zhu, Y. Liu, H. Lu, M. Shen, and Z. Ye, "Modified motion influence map and recurrent neural network-based monitoring of the local unusual behaviors for fish school in intensive aquaculture," *Aquaculture*, vol. 493, pp. 165–175, Aug. 2018.
[11] C. Zhou, D. Xu, L. Chen, S. Zhang, C. Sun, X. Yang, and Y. Wang, "Evaluation of fish feeding intensity in aquaculture using a convolutional neural network and machine vision," *Aquaculture*, vol. 507, pp. 457–465, May 2019.
[12] K. Cai, X. Miao, W. Wang, H. Pang, Y. Liu, and J. Song, "A modified YOLOv3 model for fish detection based on MobileNetv1 as backbone," *Aquacultural Eng.*, vol. 91, Nov. 2020, Art. no. 102117.
[13] A. Jalal, A. Salman, A. Mian, M. Shortis, and F. Shafait, "Fish detection and species classification in underwater environments using deep learning with temporal information," *Ecol. Informat.*, vol. 57, May 2020, Art. no. 101088.
[14] G. Yu, R. Cai, J. Su, M. Hou, and R. Deng, "U-YOLOv7: A network for underwater organism detection," *Ecol. Informat.*, vol. 75, Jul. 2023, Art. no. 102108.
[15] H. Yu, Z. Wang, H. Qin, and Y. Chen, "An automatic detection and counting method for fish lateral line scales of underwater fish based on improved YOLOv5," *IEEE Access*, vol. 11, pp. 143616–143627, 2023.
[16] Y. Cai, Z. Yao, H. Jiang, W. Qin, J. Xiao, X. Huang, J. Pan, and H. Feng, "Rapid detection of fish with SVC symptoms based on machine vision combined with a NAM-YOLO v7 hybrid model," *Aquaculture*, vol. 582, Mar. 2024, Art. no. 740558.
[17] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-time flying object detection with YOLOv8," 2023, *arXiv:2305.09972*.
[18] G. Jocher et al., 2022, "Ultralytics/yolov5: v6. 2—YOLOv5 classification models, Apple M1, reproducibility, ClearML and Deci.AI integrations," *Zenodo*, doi: 10.5281/zenodo.7002879.
[19] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1571–1580.
[20] G. Song, Y. Liu, and X. Wang, "Revisiting the sibling head in object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11560–11569.
[21] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9756–9765.
[22] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21002–21012.
[23] Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," 2021, *arXiv:2112.05561*.
[24] C. Li, L. Li, Y. Geng, H. Jiang, M. Cheng, B. Zhang, Z. Ke, X. Xu, and X. Chu, "YOLOv6 v3.0: A full-scale reloading," 2023, *arXiv:2301.05586*.
[25] X. Zhang, Y. Song, T. Song, D. Yang, Y. Ye, J. Zhou, and L. Zhang, "AKConv: Convolutional kernel with arbitrary sampled shapes and arbitrary number of parameters," 2023, *arXiv:2311.11587*.
[26] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

XIANGRONG QIN was born in Liuzhou, China, in 2003. He is currently pursuing the B.Eng. degree with Dalian Maritime University, China, in 2021. His research interests include machine learning, computer vision, and deep learning.

CHANGDONG YU (Member, IEEE) was born in Xingcheng, China, in 1996. He received the B.E. and Ph.D. degrees in information and communication engineering from Harbin Engineering University, Harbin, China, in 2018 and 2023, respectively. He is currently a Lecturer with the College of Artificial Intelligence, Dalian Maritime University, Dalian, China. His research interests include machine learning, computer vision, and swarm intelligence.

BAISHENG LIU was born in Dalian, China, in 2002. He is currently pursuing the B.Eng. degree with Liaoning Technical University, China, in 2021. His research interests include machine learning, computer vision, and deep learning.

ZHIHAO ZHANG was born in Cangzhou, China, in 2000. He received the B.E. degree from Hebei University of Science and Technology, China, in 2024. He is currently pursuing the master's degree with the College of Artificial Intelligence, Dalian Maritime University, Dalian, China. His research interests include machine learning, computer vision, and swarm intelligence.