CAAV-YOLOv11: a deep learning approach for underwater fish recognition

Honglei Wei*, Xinmin Liu

School of Mechanical Engineering and Automation, Dalian Polytechnic University, Dalian 116034, China (e-mail: weihl2005@163.com)

Abstract: The complex optical environment underwater leads to spectral selective absorption and scattering effects that significantly degrade image quality, posing challenges for fish target recognition. Additionally, current detection models face limitations in real-time applications due to their high computational complexity. To address these issues, this study introduces a two-stage fusion enhancement strategy for an underwater fish intelligent recognition framework. In image pre-processing, a collaborative enhancement mechanism combining CLAHE and Auto Levels is developed to enhance contrast while preserving image naturalness. A local adaptive processing model is established, and a dynamic color gamut compensation algorithm is integrated to correct color distortion effectively. The introduction of the ADown module optimizes the feature extraction network, reducing computational complexity. Furthermore, a VoV-GSCSP feature multiplexing mechanism is devised to enhance feature acquisition ability while maintaining precision. Experimental findings demonstrate that the model enhances detection precision and mAP@50 by 6.8% and 3.0%, respectively. Moreover, model volume and parameters are reduced by 12.7% and 11.5%. The F1 score is improved by 0.03, surpassing the baseline model in overall performance. The recognition robustness and adaptability in complex underwater environments are effectively enhanced.

Keywords: ADown, Auto Levels, CLAHE, Deep Learning, Object Detection, VoV-GSCSP

1. INTRODUCTION

Fish, being among the most widely distributed biological groups globally, plays a crucial role in assessing ecosystem health, managing fishery resources, and fostering sustainable development in aquaculture. Monitoring fish behavior not only serves as a scientific basis for safeguarding biodiversity but also underpins the optimization of aquaculture strategies and the enhancement of economic benefits. Despite the significant advancements in aquaculture, traditional manual monitoring is increasingly revealing limitations in terms of efficiency, safety, and operational conditions.

The emergence of underwater robot technology presents a promising solution owing to its automation and intelligence capabilities. However, this technology is constrained by the unique underwater environment, characterized by challenges such as the exponential decline in illumination with increasing water depth, optical scattering due to fluctuations in water temperature, and turbidity resulting from fish movements, all of which significantly compromise the reliability of target recognition systems. Consequently, the focal point of research and a critical technical challenge in this domain is the development of high-precision and robust methods for detecting underwater targets.

This paper focuses on image processing and target detection algorithms, particularly addressing the challenges posed by complex underwater environments. Scholarly research has highlighted the significant impact of factors such as underwater diffusion, uneven illumination, water absorption and scattering, noise, and low contrast on image quality. Effective image preprocessing plays a crucial role in enhancing the performance of subsequent detection tasks. (Jian et al., 2021) conducted extensive research on addressing complex underwater conditions, laying a solid foundation for further studies. (Han et al., 2020) proposed a comprehensive preprocessing workflow to address image degradation in low-light settings. (Boudhane and Nsiri, 2016) focused on denoising and restoring original images, leading to a noTable enhancement in image quality. (Ancuti et al., 2017) introduced an innovative enhancement technique based on double image fusion, which represents a significant advancement in single image processing through the strategic integration of color compensation and white balance technologies.

Object detection algorithm is an important branch of computer vision, which can identify the existence and location of specific objects in a given image or video. Target detection algorithms are usually classified into two categories according to their implementation: two-stage and single-stage algorithms. Two-stage algorithm firstly generates candidate target regions, and then classifies and fine-locates these regions, while single-stage algorithm directly predicts the category and location of targets through dense networks or anchor frames. Both methods have advantages and disadvantages and are suitable for different application scenarios. In the research of two-stage detection algorithm, scholars have made in-depth exploration and targeted improvement according to the particularity of underwater environment. Due to the complexity and variability of underwater environment and the limited sample images available, the improved Faster R-CNN algorithm developed by (Wang and Xiao, 2023) realizes accurate identification of five types of marine organisms such as sea cucumber and sea urchin. (Zeng et al., 2021) innovatively combine the occlusion resistant network (AON) with Faster R-CNN, and the detection precision in occlusion scenes is improved; (Liu and Wang, 2021) designed a detection algorithm based on Faster R-CNN to detect small and densely distributed creatures in overlapping and occluded images, which significantly improved the detection performance of small targets. In order to solve the problem of low detection precision caused by low-quality images of optical imaging for underwater target recognition, (Song et al., 2020 and Yue et al., 2023) combine Mask R-CNN with image enhancement strategy to improve the recognition precision of the model and verify the effectiveness of their improved methods. However, these methods still have shortcomings in real-time performance and computational efficiency, which seriously affect the detection efficiency.

In order to overcome these limitations, single-stage detection algorithms have gradually become a research hotspot and made remarkable progress in recent years. In view of the problems of low precision and poor real-time performance of target detection in underwater complex environment, (Qiang et al., 2020) achieved a good balance between detection speed and precision by optimizing the basic structure of SSD network; (Zhang et al., 2023) proposed a method based on improved YOLOv5 framework to solve the problem of image quality degradation and biological detection of different scales, effectively solving the problem of multi-scale detection; Considering that underwater objects are usually distributed small and dense and easily occluded, and the storage and calculation capacity of underwater embedded equipment is limited, (Zhang et al., 2024) proposed a high-precision and lightweight underwater detector based on YOLOv8 model, which is optimized for underwater scenes and achieves significant improvement; YOLOv9s-SD algorithm developed by (Zhou et al., 2025) focuses on solving the color distortion problem at low resolution, realizes real-time detection on embedded devices, and optimizes detection performance; (Mai and Wang, 2025) modify the structure based on YOLOv10 model in order to solve the problems of low precision, missed detection and false detection in complex marine environment, so as to enhance detection efficiency and improve recognition precision. Although significant progress has been made in current methods, the calculation speed of SSD network in target recognition still cannot meet the real-time requirements, the precision of early YOLO series models in complex scenes needs to be improved, and the model compression technology needs to be further optimized.

Despite advancements, Faster R-CNN and Mask R-CNN algorithms exhibit limitations in real-time performance, including high memory usage and slow computation, which impede detection efficiency. The SSD network also struggles to strike a balance between precision and computational speed in target recognition. In contrast, the YOLO series models outperform other detection algorithms in real-time detection tasks due to their streamlined design, efficient training, and high detection precision. Hence, YOLOv11 (Chen et al., 2025) serves as the foundational detection framework for underwater fish identification. Nonetheless, the underwater environment's unique characteristics result in significant image quality degradation, and the YOLOv11 base model still faces constraints in computational efficiency and feature extraction, thereby constraining the model's final recognition performance. To address these challenges, this study proposes a fusion optimization strategy. This strategy aims to enhance input data quality through advanced image enhancement algorithms and improve the network architecture to enhance the model's recognition precision and robustness.

The article is structured as follows: section 2 outlines the data collection methods and preprocessing techniques for underwater fish images; section 3 elucidates the proposed model architecture and its key module design; section 4 delineates the experimental setup, ablation studies, and comparative analysis results; and section 5 provides a summary of the work and explores future research directions.

2. IMAGE ACQUISITION AND PROCESSING

This study introduces a two-stage fusion enhancement strategy for an underwater fish intelligent recognition framework, the overall experimental process is shown in Fig. 1. Initially, fish datasets are gathered using an underwater robot, and the original images undergo enhancement through blurring and rotation to enhance the model's generalization capability. Subsequently, fish targets within the images are annotated, and image processing methods are employed to refine image quality. Following preprocessing, the dataset is partitioned, and model training ensues. Based on this, this paper constructs the CAAV-YOLOv11 detection algorithm, with main improvements including:

(1)We employ a color correction technique that integrates contrast-limited adaptive histogram equalization and an automatic color scale algorithm. This method dynamically modifies the pixel distribution of the image by creating a local adaptive processing model. It effectively addresses color distortion resulting from optical attenuation in water bodies. This process aims to generate a high-quality reference dataset for future experiments.

(2)The ADown module is utilized as a novel alternative to the conventional Conv convolutional structure within the Backbone component. This substitution is intended to achieve enhanced feature extraction at a more advanced semantic level, leading to a substantial reduction in model parameters and computational complexity. Moreover, it aims to enhance reasoning efficiency while maintaining the capability for effective feature expression.

(3)To optimize the structure of the Neck component, we first employ the VoV-GSCSP module to substitute a portion of the C3k2 structure, enhancing network characterization through optimized feature multiplexing mechanisms. Subsequently, we enhance the feature extraction capability of the model for fish targets by upgrading a segment of the convolution module to an ADown structure. This enhancement aims to boost both detection and recognition precision.

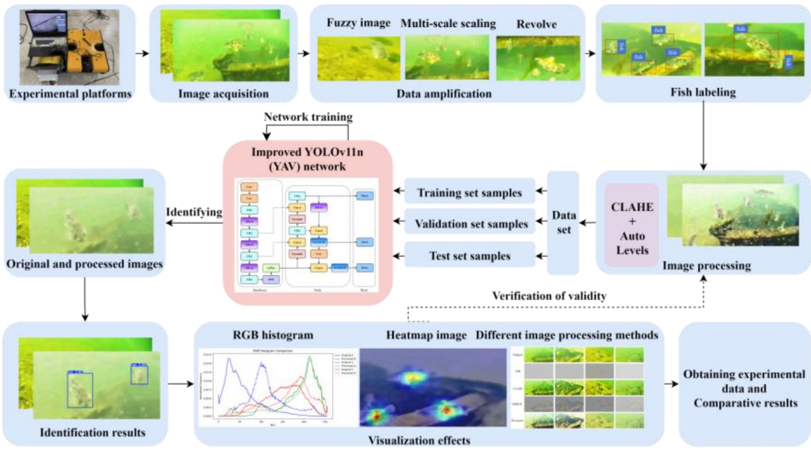

Fig. 1. Overall flow chart.

2.1 Data collection

The fish image dataset used in this study was acquired through a high-definition camera system mounted on an underwater robot platform. The robot was developed by Yantai Hailong Information Engineering Co., Ltd., located in Shandong Province, China, and is identified by the model R100. It is equipped with a 2.8mm wide-angle lens and a 2-megapixel digital HD color camera, as shown in Fig. 1 above. The image resolution is 1200 pixels × 640 pixels, which is collected under the natural lighting conditions in the aquaculture waters from 9:00 to 11:00 on July 25, 2024. In order to improve the authenticity and reliability of the imaging environment, enhance the model's generalization performance, and effectively mitigate overfitting during training, this study conducted manual selection and filtering of the original images. Invalid samples were removed, and various operations including positional translation, rotation, blurring, and multi-scale scaling were applied to the remaining images. This process resulted in a final dataset of 450 valid images.

During data annotation, we use LabelImg, a widely adopted graphical image labeling tool, to perform manual annotation (Guan et al., 2022). The tool is utilized to annotate the entire shape of the fish target by employing rectangular box labels. In cases where the target is partially occluded, only the visible parts are labeled. To maintain data quality, targets positioned at the image periphery or occupying less than 15% of the visible area are excluded from labeling. Subsequently, the annotated dataset is randomly partitioned into training, validation, and test sets at a ratio of 7:1.5:1.5. An illustration of sample labels can be observed in Fig. 1 above.

2.2 Image processing

Recognizing fish targets underwater presents several technical challenges. The complex and dynamic underwater environment introduces significant background interference in images. Suspended matter in the water leads to image quality degradation, commonly observed as a universal blur effect. Additionally, underwater optical properties result in noticeable color distortion, giving images non-uniform yellow-green tones. Moreover, light absorption and scattering in water reduce visibility, leading to a loss of texture details in fish targets. In turbid water and under uneven illumination, traditional detection methods often struggle to achieve effective target recognition.

In order to solve the above problems, this study proposes a solution based on image enhancement, which constructs a two-stage color restoration framework by fusing Contrast Limited Adaptive Histogram Equalization (Prince, 2024 and Narla, 2024) (CLAHE) and Auto Levels (Yang et al., 2023) correction algorithms. This scheme firstly improves local contrast by CLAHE algorithm, and then realizes global color balance through automatic color level adjustment, effectively eliminates fog effect in underwater image, restores natural underwater color distribution, and thus improves the sharpness and identifiability of original image. The specific flow is shown in Fig. 2.

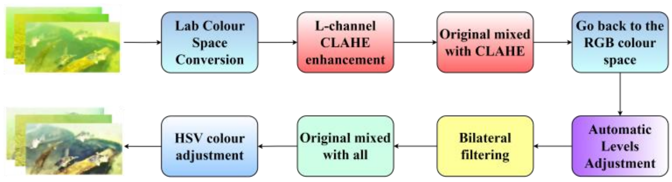

Fig. 2. Image processing flow.

CLAHE algorithm achieves fine enhancement of low contrast regions by establishing a local adaptive processing mechanism while preserving the natural characteristics of the image. The algorithm divides the image into several sub-regions, calculates gray histogram independently in each region, and uses slope limited cumulative distribution function to transform. This innovative design not only suppresses noise amplification effectively, but also retains the overall contrast characteristics of the image. Auto Levels algorithm starts from the global statistical characteristics, analyzes the probability distribution of image pixel values, intelligently identifies and eliminates abnormal extreme points (usually the first/last 0.1% pixels), and then adjusts the effective pixel values to the standard dynamic range by using the optimized mapping function. The algorithm has unique advantages in eliminating color cast and improving overall luminance distribution. It is worth noting that these two algorithms complement each other in function: CLAHE focuses on local detail restoration and enhancement, while Auto Levels is better at global tone correction and optimization. By organically combining these two algorithms, a more complete image enhancement process can be constructed to provide higher quality input data for subsequent advanced visual tasks such as target detection.

3. UNDERWATER FISH TARGET RECOGNITION METHOD

This paper investigates underwater fish target recognition using the YOLOv11 model framework. The model is structured with three key components: a feature extraction network (Backbone), a feature fusion network (Neck) and a detection output layer (Head). The model employs a modular design approach, with the innovative use of the C3k2 module for feature extraction, enhancing deep feature representation in complex scenes compared to the C2f structure of YOLOv8. The feature fusion network facilitates cross-scale fusion of features at different levels through a multi-scale feature pyramid structure, improving the model's ability to express features for multi-scale targets. The detection output layer achieves high-precision target detection and recognition by jointly optimizing localization and classification tasks. This architectural design renders the model well-suited for fish identification in intricate underwater settings.

However, there are many challenges in the underwater environment. The turbid water quality causes the image to present fuzzy yellow-green tones, suspended impurities interfere with fish target detection, and the high alertness of fish and frequent swimming easily stir the bottom sediment further increase the difficulty of target recognition. In order to effectively deal with these challenges, improve the feature extraction ability, recognition speed and reduce the computational cost under the premise of ensuring the lightweight model, this study proposes a dual improvement strategy based on YOLOv11. Firstly, ADown module is introduced in the Backbone part to replace part of the traditional convolution (Conv) structure. By optimizing the downsampling method, this module can extract image features at a higher semantic level, reducing the amount of computation and improving the efficiency of model reasoning. Secondly, double optimization is carried out in the Neck part: on the one hand, VoV-GSCSP module is used to replace part of C3k2 structure, and the performance of the network is enhanced by improving the feature multiplexing mechanism; on the other hand, part of Conv module is replaced by ADown, so as to further improve the feature characterization ability of the model for fish targets. Based on these improvements, a new target detection algorithm, namely CAAV-YOLOv11 algorithm, is constructed in this study. The optimized model architecture is shown in Fig. 3, which significantly improves the precision and efficiency of underwater fish detection while maintaining a lightweight model.

3.1 ADown

In the fish identification task, due to the large parameter quantity and high memory occupation of the traditional detection model, the calculation resource consumption is significantly increased, which seriously restricts the real-time monitoring performance in complex underwater environment.

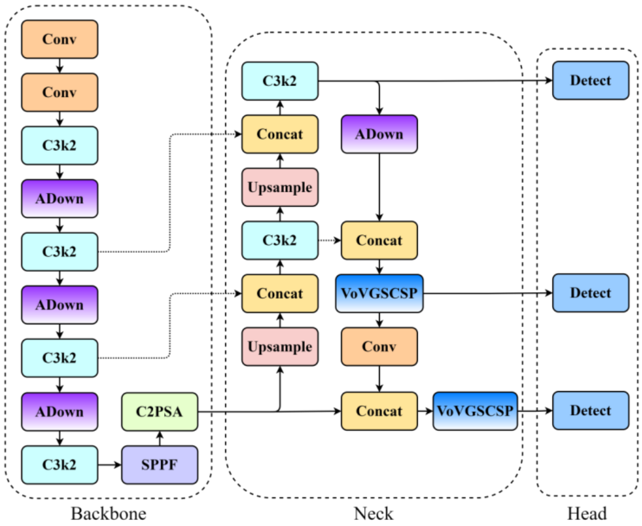

Fig. 3. Model structure of CAAV-YOLOv11 algorithm.

Aiming at this technical bottleneck, this research innovatively introduces ADown module (Li et al., 2024; Shad et al., 2025; Jia et al., 2025 and Zou et al., 2025) to optimize and reconstruct the network architecture, as shown in Fig. 4. This module mainly replaces all conventional convolution structures except the first two convolution modules in Backbone, and also replaces the first Conv structure in Neck part. ADown module, as an efficient down-sampling scheme, significantly reduces the computational complexity while maintaining the performance of the model. This module reduces the spatial dimension of feature map and retains the key visual information effectively through the carefully designed feature compression strategy, which enables the model to extract features at a higher semantic level. This improvement is particularly useful for real-time target detection tasks and has shown significant performance improvement potential in the YOLO family of models. Compared with traditional downsampling methods, ADown module shows significant technical advantages. It effectively reduces the complexity of the model through innovative parameter reduction strategy, and realizes efficient encoding of information with unique feature retention mechanism. This double optimization not only greatly improves the operation efficiency of the model in resource-constrained environment, but also improves the detection precision by enhancing the feature expression ability, achieving an ideal balance between computational efficiency and recognition performance.

In the design of module replacement strategy, this study adopts a phased structure optimization method: except for the first two convolution modules and the last convolution module of the Neck part of the Backbone network, the rest of the conventional convolution structures are replaced by ADown modules. This design decision is based on three considerations. From the viewpoint of feature extraction, the first two convolution layers undertake the key function of extracting image underlying features (such as edges, textures, etc.), and the conventional convolution operation exhibits better local feature capture ability and computational stability at this stage; From the point of view of model stability, premature introduction of downsampling operation may lead to irreversible loss or distortion of basic visual information, and then affect the extraction quality of deep features. Moreover, the second convolution module of Neck usually undertakes the task of feature compression or channel transformation, and keeping it as a conventional convolution structure helps to reduce the uncertainty and training difficulty caused by nonlinear operation. From the perspective of computational efficiency, the computational cost of conventional convolution for high-resolution feature maps processed in the first two layers is still within a controllable range, while the deployment of ADown modules in deep networks can not only make full use of its multi-scale feature extraction advantages, but also significantly reduce the overall computational complexity. This hierarchical module replacement strategy effectively balances the model representation capability with the computational resource consumption.

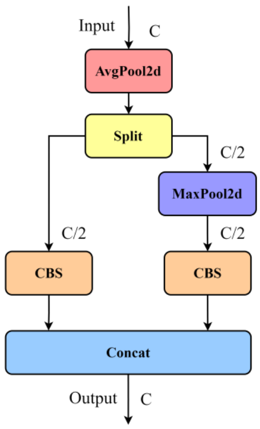

Fig. 4. ADown structure.

This module adopts multi-branch feature processing architecture to process input feature graph with channel number C. Firstly, the input features are down-sampled and smoothed by the average pooling layer (AvgPool2d). Then, the feature map is divided into two sub-feature maps with channel number C/2 by channel split operation, and the differential feature processing is carried out respectively. The main branch directly inputs the sub-feature map into CBS module, which realizes hierarchical extraction of local features through learnable convolution kernel and has multi-scale feature characterization capability. This design not only optimizes the gradient propagation path, effectively alleviates the gradient anomaly problem in deep network training, but also improves the generalization performance of the model through parameter sharing mechanism. The auxiliary branch first strengthens the key feature response through MaxPool2d, and then inputs CBS module for feature transformation. Finally, the results of dual branch processing are fused through channel concatenation (Concat) operation, and the complementary enhancement of feature information is realized while keeping the original channel number C. This module improves the detection precision by multi-scale feature fusion strategy on the basis of significantly reducing the computational complexity, and provides an efficient solution for real-time target detection system.

3.2 VoV-GSCSP

Because fish swim in disturbed water, suspended sediment and other interference factors seriously affect the image quality. As a key feature extraction component, traditional C3k2 has obvious limitations in complex underwater environment, and its ability to extract fuzzy and low-contrast features is insufficient, and the recognition precision is greatly reduced. To address this problem, this study uses VoV-GSCSP (Cross-Stage Partial Network) modules (Xiong et al., 2025; Shi et al., 2024; Wang et al., 2024) to replace the last two C3k2 structures of the Neck section. Fig. 5 (a) shows the structure of VoV-GSCSP, where the GS bottleneck module structure is shown in (b). VoV-GSCSP module can enhance feature expression capability through its unique one-time aggregation mechanism. The core advantage of this module lies in its dynamic feature selection ability. By adaptively strengthening the informative feature channels and suppressing the inefficient channels, the robustness of the model under noise interference is effectively improved. In addition, its cross-stage feature fusion mechanism promotes information interaction between different levels of features, so that shallow features such as edge texture and advanced semantic features can work together more effectively.

In terms of module replacement strategy, this study selectively replaces only the last two C3k2 modules in the Neck section. This design decision is based on two considerations. Although VoV-GSCSP can improve feature fusion ability, its computational cost is slightly higher than C3k2. Selective replacement can achieve a better balance between precision and efficiency. From the feature level characteristics, shallow features are mainly responsible for extracting detailed information and are more susceptible to underwater noise interference. If complex module design is adopted in the shallow feature extraction stage, noise may be amplified, which is not conducive to subsequent feature learning and task completion. Deep features focus on advanced semantic information and are relatively less affected by noise. Therefore, VoV-GSCSP modules are introduced in the last two places, which can better utilize their cross-stage feature fusion ability. Enhance the extraction of advanced semantic information. To sum up, this study optimizes the feature fusion ability of Neck part through the targeted introduction of VoV-GSCSP, so that it can still maintain high recognition precision in complex underwater environment.

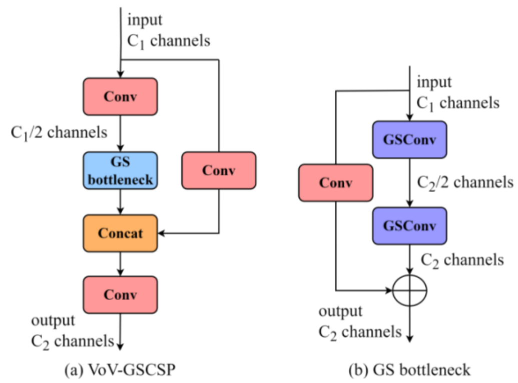

Fig. 5. VoV-GSCSP structure.

The initial convolution operation is performed on the input data with C1 channels, and the channel dimension is compressed to C1/2 by a standard convolution layer (Conv). Then the GS bottleneck structure is used for feature processing. The module enhances the characterization learning ability of the network by stacking GSConv layers. Specifically, the architecture adopts a two-path design: the main path is processed directly through the standard convolutional layer, and the branch path sequentially passes through two cascaded GSConv layers, where the first GSConv layer reduces the number of channels to C2/2 and the second GSConv layer restores the channel dimension to C2. The feature maps of the two paths are fused by adding element by element, and the final output channel number is C2 feature tensor. In order to further improve the feature multiplexing efficiency, the output of the GS bottleneck module and the result of the input data processed by the independent convolution layer are channel spliced. Finally, a standard convolution layer is used to integrate the spliced multi-scale features to generate a final output feature map with channel number C2. This optimized architecture effectively enhances the nonlinear expression ability of the network by feature reuse, and can significantly improve the mAP@50 of the model in complex underwater environments while maintaining a certain reasoning speed, especially when dealing with fish images disturbed by suspended objects. The improved model will show stronger feature discrimination ability.

4. EXPERIMENT AND RESULT ANALYSIS

4.1 Parameter setting and evaluation index

In this study, model training experiments were carried out on NVIDIA GeForce RTX 3060 Laptop GPU platform under Windows11 operating system environment. The experimental platform is configured with Python 3.9.21 programming environment, Pytorch 2.5.1 deep learning framework, and CUDA 11.8 is enabled for GPU acceleration. The model implementation is based on the ultralytics (YOLOv11) framework and trained after environmental configuration and activation. The main training parameters are set as: input image resolution 1200×640 pixels, training period 100 epochs, batch size 16. In order to evaluate the model performance comprehensively, six key indexes were selected for quantitative analysis: precision (P), mean average precision (mAP@50), model parameters (Parameters), model size (MB), frames per second (FPS) and F1 score. Precision and mAP@50 directly reflect the detection performance of the model, and the higher the value, the better the model performance. The formula is shown in (1) and (2).

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$mAP@50 = \frac{\sum_{i=1}^{N} AP}{N} \tag{2}$$

Parameters and MB represent model complexity and storage requirements, with smaller values representing lower computational resource consumption. The formula is shown in (3).

$$Parameters = C_{in} \times C_{out} \times k^2 \tag{3}$$

FPS index reflects the real-time processing ability of the model, and the higher the value, the better the calculation efficiency. The F1 score, as a harmonic average of Precision and Recall, enables a comprehensive assessment of the overall performance of the model. The formula is shown in (4) and (5), and the recall rate formula is shown in (6).

$$FPS = \frac{1}{ElapsedTime} \tag{4}$$

$$F1 = \frac{2PR}{P + R} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

where:

Formulas' parameters	Meaning
TP	the predicted value is a positive sample when the true value is a positive sample
FP	the predicted value is a positive sample when the true value is a negative sample
FN	the predicted value is a negative sample when the true value is a positive sample
N	the total number of target detection categories
C_{in}	the number of channels in the input image
C_{out}	the number of channels in the output image
k	the height of the convolution kernel (assuming the same width and height)
R	recall
ElapsedTime	the sum of the three times of image preprocessing, inference, and post-processing

4.2. Analysis of image enhancement effect

To verify the effectiveness of the two-stage color restoration method proposed in this study in image processing, Fig. 6 below presents a three-channel color histogram comparison analysis of unprocessed underwater images and processed images. From the point of view of statistical characteristics, the pixel values of the original blue channel are mainly concentrated in the lower brightness interval, and the peak value is about 20, which indicates that the original image contains a large number of dark blue pixels. In contrast, the pixel distribution of the blue channel of the processed image is significantly changed, with the number of pixels in the low brightness region decreasing and the number of pixels in the medium brightness and high brightness regions increasing. This indicates that the processing not only enhances the blue component in the image, but also makes its distribution more balanced at different brightness levels; for the green channel, the original image presents a relatively uniform but overall small number of pixels, especially in the low brightness region, almost no pixels exist, but in the processed image, the channel has a relatively increased number of pixels in the low brightness and high brightness regions; In the red channel, the original pixels are concentrated in the middle brightness region, and the peak value is located in the range of 90-110, while the processed image reduces the number of pixels in the middle and low brightness regions and increases the number of pixels in the high brightness regions.

To sum up, the processed image shows more medium to high brightness pixels in all color channels, thus improving the brightness and contrast of the overall image, making the details clearer, richer and more balanced, which is conducive to highlighting the detailed information in the image. This processing effectively improves the image quality and provides a better data set basis for subsequent target recognition tasks, thereby improving the recognition precision of fish targets and improving the recognition efficiency while reducing the calculation cost.

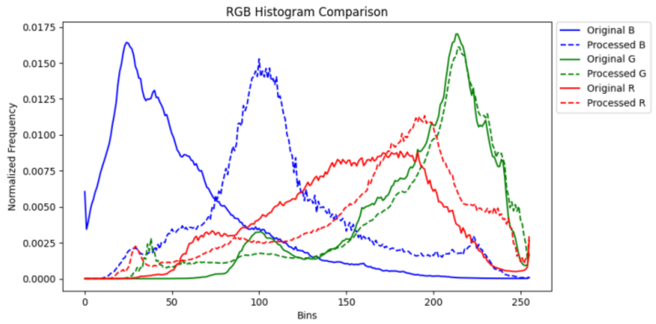

Fig. 6. RGB histogram contrast.

In addition, this paper selects three commonly used image enhancement algorithms as contrast, namely Single Scale Retinex (SSR) algorithm (Rong et al., 2024), Multi-Scale Retinex with Color Restoration (MSRCR) algorithm (Zhang et al., 2019) and Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm. Fig. 7 shows the original image and the enhancement effect after processing with four algorithms.

By contrast analysis of the original image, it can be found that there is obvious yellow-green phenomenon. Due to the interference of impurities in water and the influence of light attenuation effect, the image quality decreases significantly, and the whole image presents unbalanced hue distribution, which makes it difficult to accurately identify fish targets in the recognition process. SSR and MSRCR algorithms have serious color distortion after processing, and the image quality is obviously degraded, which is not conducive to real-time target monitoring in underwater environment. CLAHE algorithm improves the image definition to some extent, but there are still some limitations in color recovery ability and target area enhancement. In contrast, the two-stage color restoration algorithm proposed in this paper exhibits better performance in multiple evaluation dimensions. This method not only improves the detail definition and brightness distribution of the image effectively, but also realizes the color space equalization correction, which significantly weakens the yellow-green hue deviation in the original image, thus greatly enhancing the discriminability of underwater objects. Experiments show that the algorithm provides higher quality and more discriminative images for subsequent target recognition tasks, which helps to improve the overall recognition efficiency.

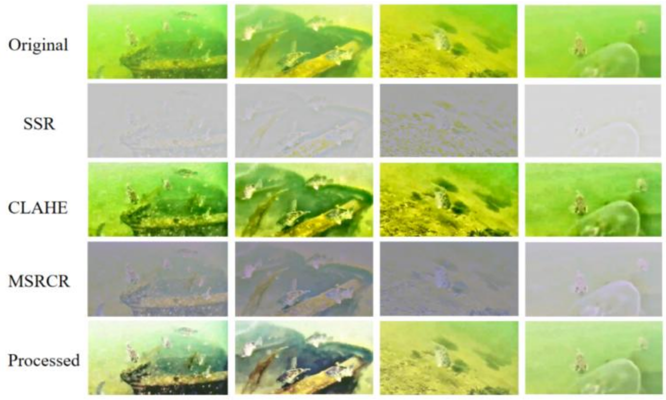

Fig. 7. Image enhancement effect of different algorithms.

To further quantify the performance of the algorithm, this study uses three objective evaluation indicators, PSNR (peak signal-to-noise ratio), MSE (mean square error) and SSIM (structural similarity), for systematic analysis (Liu et al., 2024; Xie et al., 2023), as shown in Table 1. Among them, PSNR evaluates image quality by calculating the ratio of maximum signal power to noise power, and its value is positively correlated with image quality; MSE reflects processing effect by quantifying the square mean of pixel level difference between images, and the smaller the value, the less obvious the difference; SSIM evaluates image similarity from brightness, contrast and structure, and the closer its value is to 1, the better the processing effect. Ten groups of image samples are randomly selected for testing, and the results fully verify the superiority of the algorithm.

Table 1 displays a comparative analysis of quantitative evaluation results, highlighting performance disparities among different algorithms. In objective evaluation, SSR and MSRCR algorithms perform significantly worse than our method, their PSNR values are significantly lower, and SSIM index also presents a similar trend, showing a decline range. Moreover, the MSE values of these two algorithms are higher than those of the method in this paper, indicating that there is more significant pixel-level distortion in the processed images. This result verifies the visual quality degradation phenomenon mentioned above, indicating that the traditional Retinex algorithm has limited applicability in this research scenario. CLAHE algorithm shows some competitive ability in PSNR and MSE, and its numerical value is slightly superior to that of this method, which indicates that CLAHE algorithm has certain advantages in reducing noise and maintaining basic information of image. However, the SSIM values are lower than those of the proposed method, which is especially critical because the structural similarity measured directly affects the precision of target recognition. Experimental results show that CLAHE algorithm improves the overall contrast of the image, but it is not enough to preserve the structural features of the target, which makes it difficult to meet the needs of accurate underwater target recognition in this study.

Table 1. Image quality evaluation results.

Image	SSR			CLAHE			MSRCR			Processed		
	PSNR /dB	MSE	SSIM	PSNR /dB	MSE	SSIM	PSNR /dB	MSE	SSIM	PSNR /dB	MSE	SSIM
1	9.65	0.11	0.80	21.64	0.01	0.85	12.11	0.06	0.86	13.83	0.04	0.85
2	9.59	0.11	0.69	21.64	0.01	0.77	9.62	0.11	0.70	15.66	0.03	0.80
3	9.12	0.12	0.67	22.37	0.01	0.79	9.57	0.11	0.70	14.69	0.03	0.77
4	9.23	0.12	0.83	24.44	0.003	0.93	11.15	0.08	0.87	17.74	0.02	0.94
5	10.11	0.10	0.78	20.01	0.01	0.82	12.45	0.06	0.84	14.10	0.04	0.84
6	13.88	0.04	0.92	22.73	0.01	0.93	11.31	0.07	0.90	18.33	0.01	0.96
7	10.61	0.09	0.81	18.59	0.01	0.83	12.54	0.06	0.88	14.60	0.03	0.86
8	9.19	0.12	0.84	24.35	0.004	0.93	11.16	0.08	0.87	17.37	0.02	0.94
9	8.90	0.13	0.78	19.41	0.01	0.84	12.01	0.06	0.85	15.18	0.03	0.90
10	9.07	0.12	0.70	22.45	0.01	0.77	9.79	0.10	0.68	16.16	0.02	0.79

Fig. 8 below shows the results of visual analysis (Pei et al., 2025; Huang et al., 2024) based on the thermal map of features generated by LayerCAM for some randomly selected sample images. The red area in the thermal diagram indicates the significant area of the model, and its color depth is positively correlated with the model recognition performance. The visualization results show that the model of the original image has obvious limitations in the attention region, and it can not effectively recognize the fish targets with long distance and partial occlusion, resulting in the problem of false detection and missing detection. In contrast, after two-stage image optimization processing, the feature recognition ability of the model is significantly improved, and the target object can be accurately located even in complex environmental conditions. The comparison results strongly confirm the effectiveness of the proposed method in improving fish recognition performance.

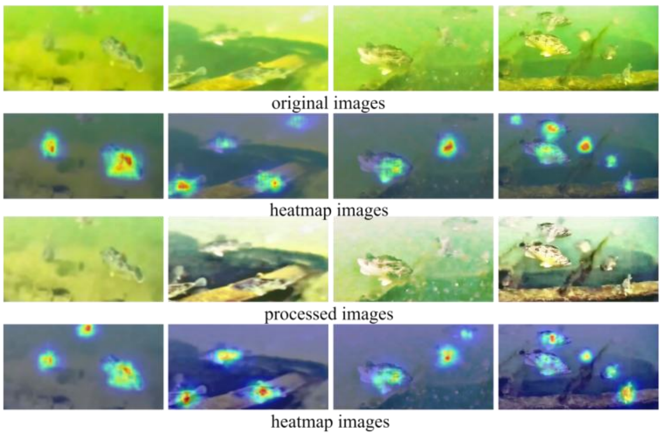

Fig. 8. Visualization.

Table 2 below shows the results of recognizing fish targets before and after image improvement using YOLOv11n basic model.

Table 2. Comparison before and after image processing.

Model	Image enhancement	P/%	mAP @50/%	Para /10^6 M	MS/MB	F1 Score	FPS
YOLOv11n	No	86.3	94.4	2.6	5.5	0.89	416.7
	Yes	94.4	93.2	2.6	5.5	0.88	416.7

Experimental results show that the proposed color restoration algorithm can optimize and balance the overall performance of the model after image preprocessing. Although the original image has reached 94.4% on the mAP@50 index, its precision is relatively low, only 86.3%. After color recovery, although the mAP@50 decreased slightly (by 1.2%), the precision was significantly improved, increasing by 8.1 percentage points to 94.4%. This improvement shows that the proposed method achieves a good balance between processing efficiency and recognition performance, which not only meets the requirements of real-time performance of the system, but also significantly improves the precision of target recognition. The results verify the effectiveness of the algorithm in improving underwater image quality from the quantitative point of view, and provide solid technical support for the follow-up underwater target recognition task.

4.3. Ablation experiments

In this paper, we propose an underwater fish target detection algorithm based on YOLOv11n. Through innovative image preprocessing scheme and model architecture optimization, the detection performance is significantly improved. In the aspect of image processing, the cooperative processing strategy of CLAHE and Auto Levels algorithm is adopted, which effectively improves the common color distortion and low contrast problems of underwater images and provides higher quality input data for subsequent detection. In terms of model improvement, two key aspects are optimized: firstly, ADown module is used to replace traditional convolution structure in BackBone, which not only enhances the extraction ability of high-level semantic features, but also realizes the lightweight design of the model; secondly, VoV-GSCSP module is introduced in Neck part and combined with the comprehensive application of ADown structure, which significantly improves the feature representation ability of underwater fish by optimizing feature multiplexing mechanism.

The experimental data are shown in Table 3. The improved model has significant improvement in multiple performance indicators. After introducing the ADown module, the model parameter quantity and size have been significantly improved, from the original 2.6×10^6 M and 5.5 MB to 2.1×10^6 M and 4.5 MB. After replacing the VoV-GSCSP module, the model detection precision and mAP@50 have been significantly improved, from 94.4% and 93.2% to 96.1% and 95.6% respectively (improved by 1.7% and 2.4%), reflecting the superiority of the improved module. Compared with the original YOLOv11n, the mAP@50 of the model is 97.4%, which is 4.2% higher; the model volume is reduced to 4.8MB, which is 12.7% lower; the parameter quantity is reduced to 2.3×10^6 M, which is 11.5% lower; and the F1 score is increased by 0.04 to 0.92. Although the precision is slightly reduced and a certain amount of computational overhead is introduced, thanks to the optimized design of the model architecture, the algorithm realizes the balance optimization of detection precision and computational efficiency while maintaining real-time performance, and provides an effective solution for fish target detection in complex underwater environments.

Table 3. Ablation experiments.

ADown	VoV-GSCSP	P/%	mAP@50/%	Para/10^6 M	MS/MB	F1 Score	FPS
Yes	No	93.2	96.3	2.1	4.5	0.94	370.4
No	Yes	96.1	95.6	2.5	5.4	0.93	357.1
Yes	Yes	93.1	97.4	2.3	4.8	0.92	370.4

4.4. Comparison of different detection algorithms

In this study, several representative models of YOLO series were selected for comparative analysis, including YOLOv3-tiny, YOLOv5, YOLOv8n, YOLOv9t, YOLOv10n, YOLOv11n and CAAV-YOLOv11 in this study. All models were evaluated in the same experimental environment using fish image datasets processed by two-stage color restoration. The experimental data are shown in Table 4. The reference model YOLOv11n achieved a recognition precision of 94.4% and an mAP@50 of 93.2%, with a model parameter size of 2.6×10^6 M, a storage volume of 5.5 MB, an F1 score of 0.88, and a processing speed of 416.7 fps. Comparative analysis shows that YOLOv3-tiny performs well in precision, mAP@50 and processing speed, but its computational complexity is high and the model volume is huge; YOLOv5 and YOLOv9t have some improvement in model compression, but their precision and processing speed are relatively low; YOLOv8n performs well in precision and processing speed, but the model compression effect is poor; YOLOv10n achieves the best image processing speed, but its precision and mAP@50 are greatly reduced. On the whole, these models have different performance limitations in key indicators such as precision, speed and model complexity, which need to be further optimized and improved.

Table 4. Comparison of results of different detection models.

Model	P/%	mAP@50/%	Para/10^6 M	MS/MB	F1 Score	FPS
YOLOv3-tiny	95.2	96.8	9.5	19.2	0.93	500.0
YOLOv5	88.7	93.5	2.2	4.6	0.90	476.2
YOLOv8n	96.7	95.6	2.7	5.6	0.94	434.8
YOLOv9t	89.7	93.3	1.7	4.2	0.88	294.1
YOLOv10n	81.8	85.0	2.3	5.7	0.80	526.3
YOLOv11n	94.4	93.2	2.6	5.5	0.88	416.7
CAAV-YOLOv11	93.1	97.4	2.3	4.8	0.92	370.4

The CAAV-YOLOv11 algorithm proposed in this study shows excellent detection performance in the comprehensive performance evaluation. Experimental data show that the algorithm achieves significant breakthroughs in key performance indicators, in which the mAP@50 is improved to 97.4%, which is 4.2% higher than the benchmark model. In terms of model efficiency optimization, through structural improvement, the model volume was successfully compressed to 4.8 MB (reduced by 12.7%), the number of parameters was reduced to 2.3×10^6 M (reduced by 11.5%), and the F1 score was also improved by 0.04. Although the precision and processing speed decreased slightly, the algorithm maintained a certain mAP@50 and excellent model compression effect while maintaining real-time performance, and each performance was well balanced.

4.5. Performance verification

In order to verify the generalization ability of the model, 7 contrast models were tested on the test set. As shown in Table 5, the CAAV-YOLOv11 model proposed in this paper exhibits significant advantages, balancing mAP@50 with computational efficiency.

Table 5. Comparison of results of model test sets.

Model	P/%	mAP@50/%	Para/10^6 M	MS/MB	F1 Score	FPS
YOLOv3-tiny	93.6	93.4	9.5	19.2	0.91	416.7
YOLOv5	96.6	91.1	2.2	4.6	0.87	400.0
YOLOv8n	90.9	91.4	2.7	5.6	0.88	434.8
YOLOv9t	84.9	88.8	1.7	4.2	0.83	312.5
YOLOv10n	78.1	81.6	2.3	5.7	0.76	625.0
YOLOv11n	84.1	88.2	2.6	5.5	0.84	416.7
CAAV-YOLOv11	94.6	95.2	2.3	4.8	0.91	384.6

In the task of target recognition in complex scenes, the model not only shows excellent environmental adaptability, but also fully verifies its generalization performance. The experimental results strongly confirm the effectiveness of the improved scheme proposed in this paper.

Fig. 9 shows the dynamic variation trend of mAP@50 of seven contrast models in the test set with increasing training period, while Fig. 10 compares the spatial positioning performance of each model in target detection task through the key indicator of bounding box regression loss, and the spatial matching degree of prediction box and real labeling box is negatively correlated. Experimental results show that CAAV-YOLOv11 model tends to be optimal gradually in mAP@50 index. Although it does not achieve the optimal value in bounding box regression loss, it achieves a good performance balance among precision, inference efficiency and model complexity. In complex underwater target recognition task, this algorithm shows better environmental adaptability and robustness than other six comparative models.

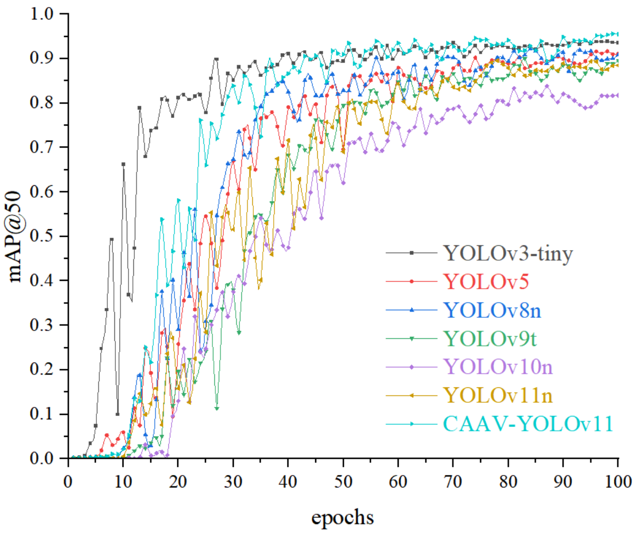

Fig. 9. mAP@50.

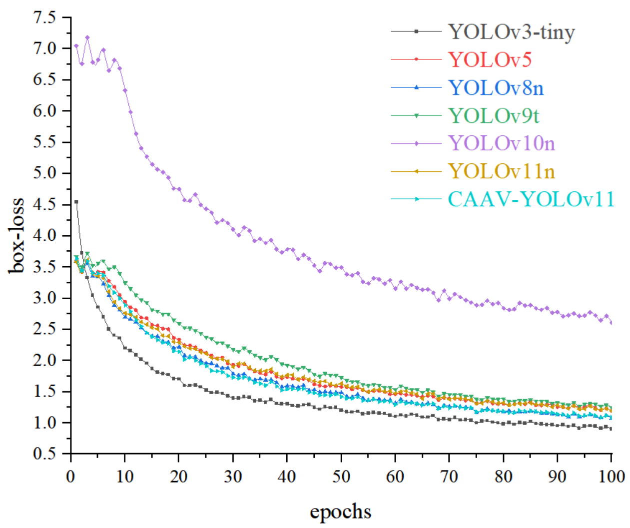

Fig. 10. Bounding box loss.

Fig. 11 and Fig. 12 show the performance of the original image input versus baseline detection model and the systematic comparison of the proposed comprehensive improvement method in terms of recognition precision and mAP@50. According to the images, the improved method achieves significant performance optimization compared with the baseline model on these two key indicators. These experimental results fully verify the effectiveness of the proposed two-stage image processing strategy and model architecture improvement scheme, and provide a better performance solution for underwater target detection tasks.

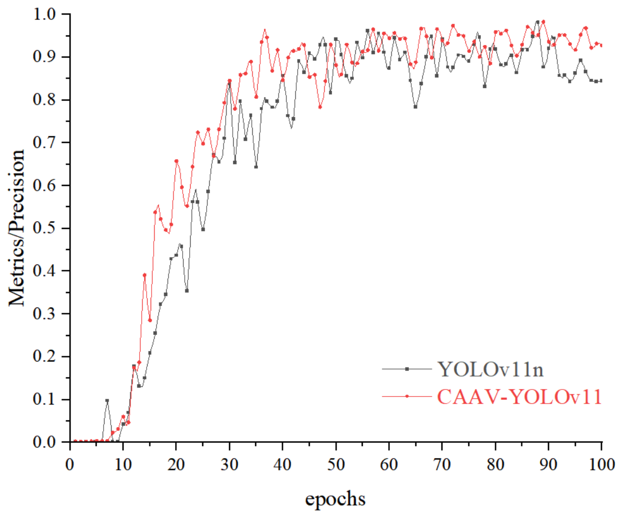

Fig. 11. Precision.

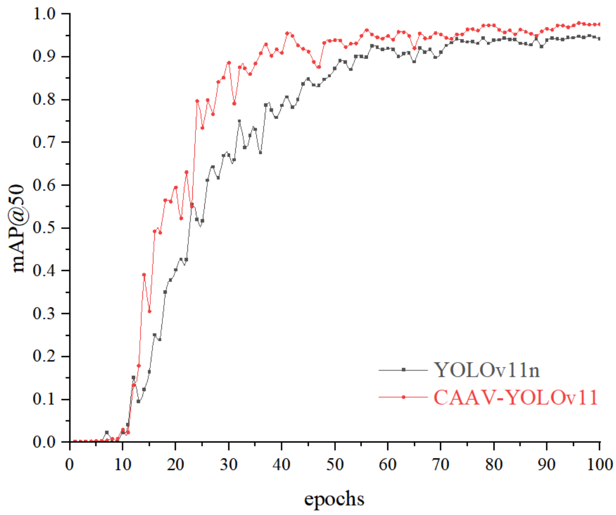

Fig. 12. mAP@50.

Fig. 13 compares the performance of YOLOv3-tiny, YOLOv5, YOLOv8n, YOLOv9t, YOLOv10n, YOLOv11n and CAAV-YOLOv11 models modified in this study on fish identification tasks. The experimental results show that these algorithms can recognize the obvious targets in the fourth group, but in the first and second group images, the environment background is complex, YOLOv3-tiny, YOLOv8n, YOLOv9t, YOLOv10n and YOLOv11n have obvious false detection and repeated detection behavior; in the third group images, YOLOv10n shows obvious missed detection phenomenon for relatively distant and fuzzy fish targets; YOLOv5 maintains low false detection rate and missed detection rate, but its confidence is insufficient. In contrast, the CAAV-YOLOv11 algorithm proposed in this study shows excellent detection performance. By using ADown module to optimize the feature extraction network and combining VoV-GSCSP feature multiplexing mechanism, the feature characterization ability of the model in complex environments is significantly enhanced. While maintaining high confidence, the false detection and missed detection problems in complex scenes are effectively overcome, and the accurate recognition of target fish is realized, which fully reflects the effectiveness, superiority and strong robustness of the improved algorithm.

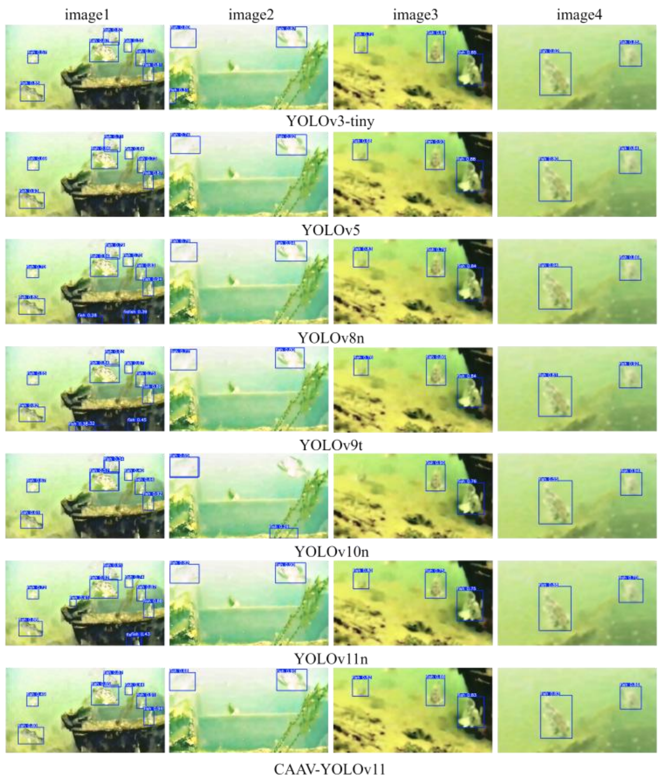

Fig. 13. Detection effect.

5. CONCLUSION

In this paper, a CAAV-YOLOv11 detection algorithm based on improved YOLOv11n framework is proposed to solve the visual detection problems caused by underwater complex optical environment, especially the image degradation caused by suspended particles and the yellow-green bias caused by spectral selective absorption. This study combines CLAHE and Auto Levels algorithms through a two-stage image enhancement strategy to effectively improve the quality of underwater images. At the model architecture level, ADown module is used to optimize feature extraction networks, and VoV-GSCSP feature multiplexing mechanism is combined to significantly enhance the model's feature representation ability in disturbed environments. Experimental results show that the proposed algorithm achieves a good balance between detection precision and computational efficiency. Compared with traditional algorithms, the proposed algorithm has obvious advantages in robustness and generalization performance. However, in view of the fact that there are still a small number of misjudgments in the model, the subsequent optimization can focus on the in-depth analysis of the false detection samples, especially the samples with complex lighting conditions or occlusion of the target, and further improve the stability and adaptability of the model by increasing the diversity of negative samples. In addition, due to the particularity and complexity of underwater environment, other feature extraction networks suitable for this kind of scene can be explored in future research, or attention mechanism can be introduced to enhance the detection ability and adaptability of fish targets in different environments.

ACKNOWLEDGEMENTS

Funding this work was supported by the Liaoning Provincial Department of Education Fund, project number (LJ212510152012).

REFERENCES

Ancuti, C. O., Ancuti, C., De Vleeschouwer, C., and Bekaert, P. (2017). Color balance and fusion for underwater image enhancement. *IEEE Transactions on image processing*, 27(1), 379-393.

Boudhane, M., and Nsiri, B. (2016). Underwater image processing method for fish localization and detection in submarine environment. *Journal of Visual Communication and Image Representation*, 39, 226-238.

Chen, R., Zhou, H., Xie, H., and Wang, B. (2025). YOLO-CE: an underwater low-visibility environment target detection algorithm based on YOLO11. *The Journal of Supercomputing*, 81(5), 1-24.

Guan, Z., Hou, C., Zhou, S., and Guo, Z. (2022). Research on underwater target recognition technology based on neural network. *Wireless Communications and Mobile Computing*, 2022.

Huang, M., and Wei, H. (2024). Faster-YOLOv5: real-time deep learning for sea cucumber defect recognition and classification. *Engineering Research Express*, 6(3), 035222.

Han, F., Yao, J., Zhu, H., and Wang, C. (2020). Underwater image processing and object detection based on deep CNN method. *Journal of Sensors*, 2020(1), 6707328.

Jian, M., Liu, X., Luo, H., Lu, X., Yu, H., and Dong, J. (2021). Underwater image processing and analysis: A review. *Signal Processing: Image Communication*, 91, 116088.

Jia, P., Sheng, H., and Jia, S. (2025). LPCF-YOLO: A YOLO-based lightweight algorithm for pedestrian anomaly detection with parallel cross-fusion. *Sensors*, 25(9), 2752-2752.

Li, J., Chen, Z., and Xu, H. (2024). A method for strawberry disease detection in real scenarios based on improved YOLOv8n. *Journal of Chinese Agricultural Mechanization*, 45(12), 267.

Liu, Z., Hong, H., and Gao, X. (2024). Underwater image enhancement based on a combination of improved gated context aggregation network and gray world algorithms. *Journal of Electronic Imaging*, 33(1), 013009-013009.

Liu, Y., and Wang, S. (2021). A quantitative detection algorithm based on improved faster R-CNN for marine benthos. *Ecological Informatics*, 61, 101228.

Mai, R., and Wang, J. (2025). UM-YOLOv10: Underwater object detection algorithm for marine environment based on YOLOv10 model. *Fishes*, 10(4), 173.

Narla, V. L., Suresh, G., Rao, C. S., Awadh, M. A., and Hasan, N. (2024). A multimodal approach with firefly based CLAHE and multiscale fusion for enhancing underwater images. *Scientific Reports*, 14(1), 27588.

Prince, H. (2024). An effective underwater image enhancement based on the fusion of CLAHE and image constancy sharpening technique. *Earth Science Informatics*, 17(2), 1359-1372.

Pei, S., Wang, W., Wu, P., Hu, C., Sun, H., Li, K., and Lan, B. (2025). Detection of hydrophobicity grade of insulators based on AHC-YOLO algorithm. *Scientific Reports*, 15(1), 9673.

Qiang, W., He, Y., Guo, Y., Li, B., and He, L. (2020). Exploring underwater target detection algorithm based on improved SSD. *Xibei Gongye Daxue Xuebao/Journal of Northwestern Polytechnical University*, 38(4), 747-754.

Rong, L., Zhang, S. H., Yin, M. F., Wang, D., Zhao, J., Wang, Y., and Lin, S. F. (2024). Reconstruction efficiency enhancement of amplitude-type holograms by using single-scale retinex algorithm. *Optics and Lasers in Engineering*, 176, 108097.

Shad, I., Zhang, Z., Asim, M., Al-Habib, M., Chelloug, S. A., and Abd El-Latif, A. (2025). Deep learning-based image processing framework for efficient surface litter detection in computer vision applications. *Journal of Radiation Research and Applied Sciences*, 18(2), 101534-101534.

Shi, P., Zhang, Y., Cao, Y., Sun, J., Chen, D., and Kuang, L. (2024). DVCW-YOLO for printed circuit board surface defect detection. *Applied Sciences*, 15(1), 327-327.

Song, S., Zhu, J., Li, X., and Huang, Q. (2020). Integrate MSRCR and mask R-CNN to recognize underwater creatures on small sample datasets. *IEEE Access*, 8, 172848-172858.

Wang, H., Li, A., Yang, Y., Zhu, X., and Song, L. (2024). Fire video intelligent monitoring method based on moving target enhancement and PRV-YOLO network. *Fire Technology*, 61(4): 1-27.

Wang, H., and Xiao, N. (2023). Underwater object detection method based on improved faster RCNN. *Applied Sciences*, 13(4), 2746.

Xiong, Q., Chen, Q., Tang, S., and Li, Y. (2025). An efficient and lightweight detection method for stranded elastic needle defects in complex industrial environments using VEE-YOLO. *Scientific Reports*, 15(1), 2879-2879.

Xie, Q., Gao, X., Liu, Z., and Huang, H. (2023). Underwater image enhancement based on zero-shot learning and level adjustment. *Heliyon*, 9(4).

Yang, T., Gao, W., Wang, P., Li, X., Lv, Z., and Di, R. (2023). Underwater target detection algorithm based on automatic color level and bidirectional feature fusion. *Laser & Optoelectronics Progress*, 60(6), 0610012-0610012-12.

Yue, Z., Yan, B., Liu, H., and Chen, Z. (2023). An effective method for underwater biological multi-target detection using mask region-based convolutional neural network. *Water*, 15(19), 3507.

Zou, L., Chen, A., Yang, X., and Sun, Y. (2025). An improved method of AUD-YOLO for surface damage detection of wind turbine blades. *Scientific reports*, 15(1), 5833.

Zhang, W., Dong, L., Pan, X., Zhou, J., Qin, L., and Xu, W. (2019). Single image defogging based on multi-channel convolutional MSRCR. *IEEE Access*, 7, 72492-72504.

Zeng, L., Sun, B., and Zhu, D. (2021). Underwater target detection based on Faster R-CNN and adversarial occlusion network. *Engineering Applications of Artificial Intelligence*, 100, 104190.

Zhou, S., Wang, L., Chen, Z., Zheng, H., Lin, Z., and He, L. (2025). An improved YOLOv9s algorithm for underwater object detection. *Journal of Marine Science and Engineering*, 13(2), 230.

Zhang, M., Wang, Z., Song, W., Zhao, D., and Zhao, H. (2024). Efficient small-object detection in underwater images using the enhanced yolov8 network. *Applied Sciences*, 14(3), 1095.

Zhang, J., Zhang, J., Zhou, K., Zhang, Y., Chen, H., and Yan, X. (2023). An improved YOLOv5-based underwater object-detection framework. *Sensors*, 23(7), 3693.