Research on Drowning Identification Based on Improved YOLOv8

Litao Cai, Youbing Feng, Xueyun Wei, Feng Xiong

Abstract—**The frequent occurrence of drowning incidents poses a significant challenge to public safety, and timely detection is crucial for ensuring life safety. With the development of deep learning technology, some drowning detection models have been established; however, these models often suffer from low accuracy and slow response times. This paper proposes a drowning detection model based on an improved YOLOv8 algorithm, named SS-YOLOv8. The model combines the Spatial Pyramid Pooling (SPPF) module with the Efficient Channel Attention (ECA) mechanism to enhance the model's ability to capture important features while maintaining computational efficiency. Additionally, a small target detection layer is introduced to detect key features of small targets, such as the head and arms of drowning individuals, making the model more sensitive to such small target information. The experimental results demonstrate that the SS-YOLOv8 model outperforms YOLOv5, YOLOv7, YOLOv8, YOLOv9 and YOLOv10 models in terms of accuracy, recall, and mean average precision (mAP). Additionally, it exhibits superior detection speed. Therefore, the improved model meets the accuracy and real-time performance requirements essential for practical search and rescue operations.**

Index Terms—**drowning identification, YOLOv8, attention mechanism, small target detection layer.**

I. Introduction

Drowning is a neglected but deadly public health issue, with approximately 372,000 people dying from drowning accidents every year [1], posing a significant threat to public health and safety. Traditional drowning search and rescue methods rely on human resources, which are not only labor intensive, but also usually suffer from delayed response and inefficient search. Detecting drowning victims involves many challenges, such as uncertain lighting conditions and water surface fluctuations. Therefore, accurate and effective identification of drowning victims is of crucial importance for initiating timely rescue operations. In recent years, advancements in computer vision technology have driven remarkable progress in image recognition for target detection, offering innovative approaches to identifying drowning victims [2]. By processing images using computer vision techniques, autonomous detection of drowning victims is feasible, enabling rapid rescue operations. This is important for reducing drowning accidents and improving rescue efficiency.

In 2016, Salehi et al. [3] proposed an HSV color space based real-time drowning detection method for indoor swimming pool drowning detection. The color channel is optimized using video prior information, and HSV thresholding and contour detection are utilized to identify the region of interest (ROI) for every frame. In 2018, Bhaskaran [4] introduced the New Equation pool drowning detection system (NEPTUNE), where the video sequence images are merged and features are extracted using K-means clustering to generate image variables. NEPTUNE recognizes a drowning instance and triggers an alarm within 5 seconds, achieving high speed detection and a low false alarm rate. In 2019, Shiuuee et al. [5] developed the Ocean Drowning Automated Detection method, which combines image processing and background omission methods to detect drowning persons. This method uses an artificial neural network for training and testing. In 2021, He et al. [6] proposed an unsupervised video anomaly detection method for swimming pool drowning events. In this study, a new pool scene dataset was created, the data was preprocessed, video frames were reconstructed with a modified deep residual network (ResNet), and frame differences were evaluated for their effectiveness in identifying anomalous events not in the training set. In 2021, Jian et al. [7] presented an artificial intelligence-based motion analysis method that utilizes computer image processing techniques. This approach involves installing a camera at the bottom of a swimming pool, utilizing OpenPose to label joint features in the captured images, and inputting these features into a recurrent neural network to assess whether or not drowning is occurring. In 2022, Hayat et al. [8] suggested a swimmer safety alert system by augmenting the traditional Mask R-CNN framework. The convolutional backbone is greatly optimized and features of the cascade pyramid model are integrated. The system recognizes swimmers' poses in real-time in a swimming pool, detects drowning events promptly, and triggers appropriate alerts. In 2023, Dulhare et al. [9] combined the Faster Region-based Convolutional Neural Network (Faster R-CNN) with data enhancement algorithms to detect whether a human is drowning.

In recent years, with the rapid development of computer vision technology, the YOLO framework algorithms have gained widespread adoption in target detection because of their exceptional balance between recognition precision and processing speed. For example, in 2022, Lei et al. [10] presented a behavioral identification method (BR-YOLOv4) based on the YOLOv4 algorithm by analyzing the geometric association of the target's location data and designated swimming or drowning zones within the pool. The analysis

Manuscript received December 3, 2024; revised June 27, 2025.

Litao Cai is a postgraduate student of the Ocean college, Jiangsu University of Science and Technology, Zhenjiang, 212003, China. (e-mail: 221212201105@stu.just.edu.cn)

Youbing Feng is an associate professor at the Ocean college, Jiangsu University of Science and Technology, Zhenjiang, 212003, China. (corresponding author to provide phone: +86-151-8911-5635; e-mail: yzfyb@just.edu.cn)

Xueyun Wei is an associate professor at the Ocean college, Jiangsu University of Science and Technology, Zhenjiang, 212003, China. (e-mail: xywei@just.edu.cn)

Feng Xiong is a postgraduate student of the Ocean college, Jiangsu University of Science and Technology, Zhenjiang, 212003, China. (e-mail: 221712201108@stu.just.edu.cn)

assesses whether the individual is exhibiting swimming or drowning behavior. In 2022, Niu et al. [11] introduced a YOLOv4-based approach for detecting drowning risks in indoor swimming pools and utilized the MA_CBAM module, significantly enhancing precision and reliability compared to the standard YOLOv4 framework. In 2023, He et al. [12] suggested an automated and efficient method for detecting video-based surveillance for infant drowning incidents by exploiting the advantages of the Faster R-CNN and YOLOv5 model to achieve rapid and precise identification of infant submersion incidents in realistic application scenarios. In 2023, Liu et al. [13] introduced a dual-stage detection method that first utilizes the YOLOv5n network to recognize nearly vertical human bodies, considering the distinctive features of drowning individuals. This is followed by integrating a lightweight drowning detection network (DDN) with a Gaussian model to rapidly extract human body features, addressing the scarcity of drowning video datasets and the limitations of simulated video realism.

With the innovation and maturation of drowning detection methods, some drowning recognition systems are emerging. The SwimEye water safety monitoring platform developed by Norwegian researchers [14] uses distributed underwater visual sensing devices to collect real-time 3D motion data of swimmers. It performs human body recognition through a visual analysis module based on the OpenPose framework and constructs a multidimensional biomechanical parameter model using a spatiotemporal feature fusion algorithm. The Poseidon system in the United States [15] distributed visual sensing technology. High-sensitivity imaging devices are embedded and deployed on the sidewalls of the pool's deep water zone to provide underwater panoramic coverage. Wide-angle monitoring devices are installed at elevated points along the poolside to form an overhead observation plane, creating a three-dimensional dynamic monitoring space. When an object is detected to be stationary at the pool's bottom for more than the preset 10-second threshold, an automatic alarm is triggered. The Hong Kong Artificial Intelligence Drowning Detection System [16] employs a 3D skeleton keypoint tracking algorithm to construct a dynamic motion model and relies on computer vision technology to analyze swimmer behavior.

However, current mainstream drowning victim detection methods still suffer from low detection accuracy and slow inference speed, particularly in identifying small or blurred targets. To address these issues, this study presents a method for detecting drowning individuals using an optimized YOLOv8 algorithm. The main contributions of this research are summarized as follows:

[1] We constructed the SPPF_ECA module by introducing the ECA framework into the SPPF module of the original YOLOv8 to highlight the important channel features related to the detection target and ignore the unimportant ones. This strategy assists the network in selecting feature channels across various scales, enhances detection precision and improves the model's robustness when handling complicated scenarios.

[2] To strengthen the capability of the model to capture and identify small targets, we add an additional STDL for small targets such as the head and arm of a drowning person. This improves the model's adaptability across various scenarios to meet the application requirements in real-life scenarios.

[3] The improved YOLOv8 algorithm has been verified and analyzed by ablation experiments and comparison experiments, all demonstrating good detection performance and significantly improve the efficiency and adaptability of drowning recognition.

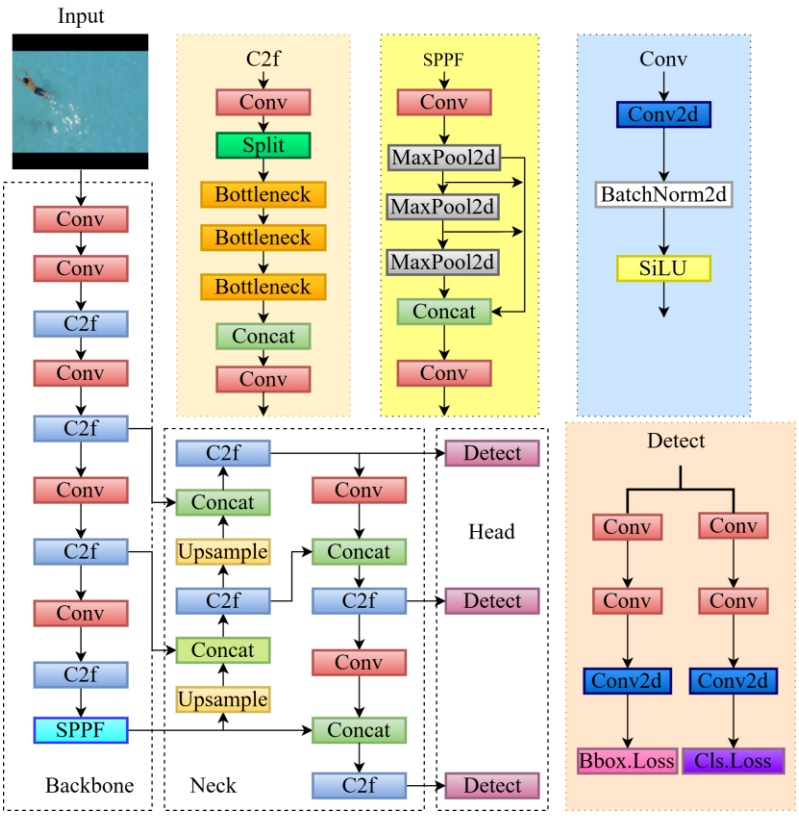


Fig. 1. YOLOv8 network structure model

II. METHOD

A. YOLOv8 Framework Synopsis

YOLOv8, developed by Ultralytics, is a highly efficient algorithm for object detection [17], which combines the network structure and modules of many mainstream target detection algorithms, and has a strong potential for secondary development. YOLOv8 has a faster detection speed and higher precision than traditional target detection algorithms. It also integrates features across multiple scales to improve detection performance by fusing different levels of feature mapping [18]. This fusion enables the model to identify objects across varying sizes and scales more effectively, improving detection precision. YOLOv8 officially provides five different sized models for users to choose from, which are YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x [19], which are the same in terms of network structure and differ only in their depth_multiple and width_multiple parameters to regulate the model size [20]. YOLOv8x is the largest model in the YOLOv8 series, having a deeper number of network layers and richer feature extraction, and it can better deal with small target detection, multi-targets, as well as targets in complex environments. Therefore, YOLOv8x is used as the main model framework for the experiments within this research. The model architecture of YOLOv8 is primarily consists of three components: backbone, Neck and Head, and the structure of its network model is presented in Fig. 1.

The Backbone module primarily extracts fundamental Characteristics of the input image, serving as the basis for the following processing steps. The Backbone of YOLOv8 adopts the C2f module, which includes two convolutional layers and multiple bottleneck layers to optimize the depth and breadth of the feature extraction through residual connectivity [21]. Furthermore, the SPPF (Spatial Pyramid Pooling Faster) framework is integrated at the end of the Backbone, enhancing the feature aggregation capability without significantly increasing the computational cost through multilevel pooling operations and residual connections.

The Neck part enhances the fusion of multiscale features obtained from the backbone to improve the recognition ability of different scale targets in the detection task [22]. The Neck layer of YOLOv8 adopts the structure of a Path Aggregation Network (PAN) and Feature Pyramid Network (FPN) for multiscale feature fusion [23]. To improve the model's generalization capabilities in complex scenarios, features from different layers are effectively integrated through a combination of up-sampling and down-sampling operations.

The Head part i is primarily responsible for the final part of target detection. A Decoupled-Head structure is used in YOLOv8 to separate the classification and regression tasks. This structure helps to capture the target's category information and positional features more accurately. In addition, the number of channels of the regression head is adjusted accordingly [24].

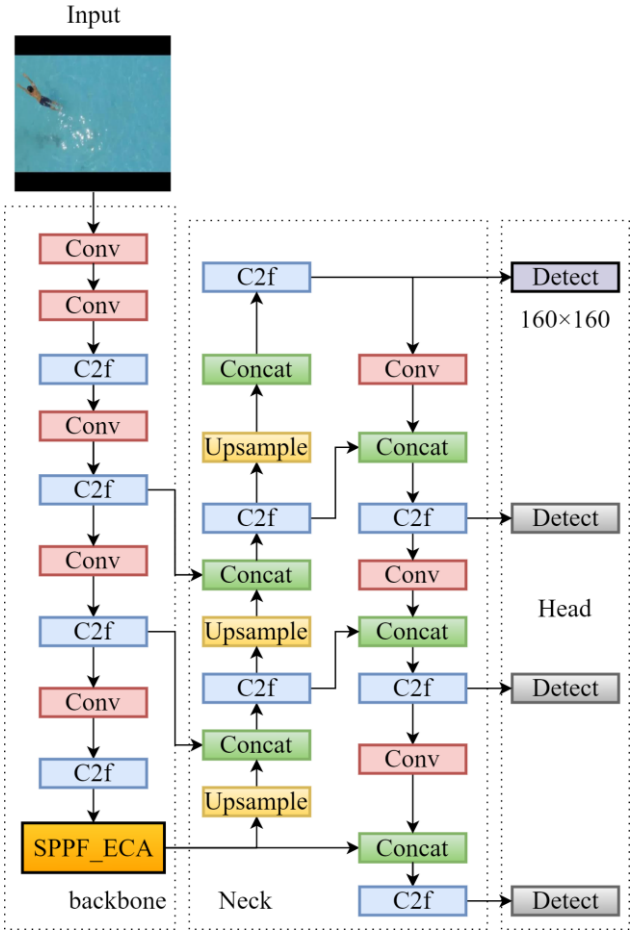


Fig. 2. SS-YOLOv8 network structure model

B. SS-YOLOv8 Model

Although YOLOv8 performs well in target detection, it still has limitations in small target detection and feature extraction efficiency in complex environments. In the drowning recognition scenarios, factors such as occlusion, water reflection, and light change will interfere with feature extraction. When YOLOv8 deals with this complex and variable environments, its feature extraction module may fail to capture the key information, leading to misdetection or omission of detection. Hence, this research designs an upgraded network structure SS-YOLOv8 based on the YOLOv8 framework, whose network structure is shown in Fig. 2.

The SPPF_ECA module is constructed at the end of the YOLOv8 backbone network and combines the ECA mechanism with the SPPF module, which ensures the lightweight of the model while enhancing the focus on key information and improving the precision and robustness of feature extraction. We add an STDL to the Head network to enhance the model's sensitivity to small targets and recognition precision, while simultaneously augmenting the original multiscale detection mechanism of YOLOv8, balancing the detection capability for different targets.

C. SPPF_ECA

The SPPF module integrated into the YOLOv8 backbone network adopts a spatial pyramid pooling structure, which is improved from the structure of SPP [25]. Its core idea is to capture characteristics across multiple scales to capture the target object's detailed information and enrich the semantic and spatial information in the feature map through maximum pooling operation and feature fusion [26]. Although SPPF can extract features from different receptive fields, it lacks information interaction between channels, prohibiting the effective distinguishing of key channel features, and lacks the ability to suppress irrelevant information. Incorporatingan attention mechanism enhances the model's ability to concentrate on the relevant input information, making it more effective in identifying and processing data critical to a specific task, thereby optimizing the model to optimize the learning process [27]. Therefore, this research adopts an ECA mechanism combined with the SPPF module to construct the SPPF_ECA module, emphasizing the channel features associated with the target, enabling the model to concentrate more effectively on the critical information, and enhancing the robustness of the model. At the same time, the lightweight design of ECA significantly improves the model's real-time performance, ensuring its suitability for applications requiring rapid processing, which is suitable for the drowning person detection task that requires fast response.

The composition of the improved SPPF_ECA module is depicted in Fig. 3. The ECA module is added after the concatenation operation of all maximum pooling layers (MaxPool2d) and before the the second convolutional layer (Conv). This setup takes full advantage of the multiscale features after the pooling operation, to enhancing the weight distribution between channels and the expression of the essential features. The feature maps after multiscale pooling contain information on different scales after splicing, and directly applying the convolution operation may lead to confusion about the integration of features during the convolution process. Thus, introducing the ECA mechanism can first carry out channel-selective enhancement of these spliced feature maps, allowing the convolution layer to accurately capture the features of the crucial channels and reduce the interference during the processing. Meanwhile, it maintains, to the greatest extent, the multiscale characteristics and lightweight design of the SPPF module.

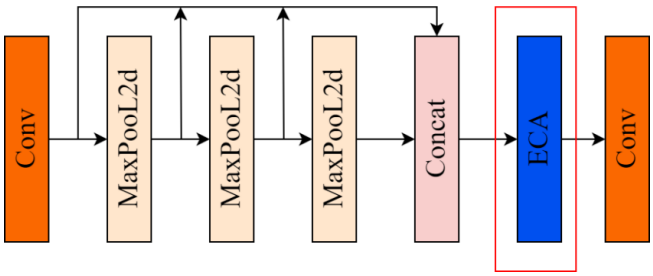


Fig. 3. SPPF_ECA overall framework, the red box shows the improved module in this research

ECA is a lightweight channel attention mechanism [28], with its module structure illustrated in Fig. 4. Traditional channel attention mechanisms typically compute the relationships between channels using fully connected layers, which introduces a significant number of parameters and computational overhead. The ECA mechanism introduces one-dimensional convolution to capture the interrelationships between channels, thus assigning adaptive weights to each channel and highlighting the channels that are more contributing to the feature expression and suppressing irrelevant channels. This strategy effectively avoids dependence on the fully-connected layer and significantly reduces computational complexity while preserving critical information between channel.

Assuming an input feature map $X \in R^{C\times H\times W}$ field，where C, H and W are the number of channels, height, and width of the feature map [29], respectively. A global average pooling operation is performed for each channel to obtain a global description vector for each channel as in (1).

$$z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_c(h,w) \quad c=1,2,\ldots,C \quad h=1,2,\ldots,H \quad w=1,2,\ldots,W \tag{1}$$

where $X_c(h, w)$ is the value of the c-th channel in the input feature map at the position with height h and width w; A channel description vector can be obtained defined as $z=[z_1, z_2, \ldots, z_c] \in R^C$, where each element z_c represents the global average information for channel c.

The ECA utilizes a one-dimensional convolution to model inter-channel relationships to minimize the significant computational burden associated with fully connected layers. A one-dimensional convolution operation using a kernel of size k is performed on the channel descriptor vector z. The output is the channel attention weight $w \in \mathbb{R}^C$ as in (2).

$$w = Conv1D(z,k) \tag{2}$$

where $Conv1D(z, k)$ denotes a one-dimensional convolution operation with a kernel size k, which is adaptively determined by the number of channels of the input features to capture the

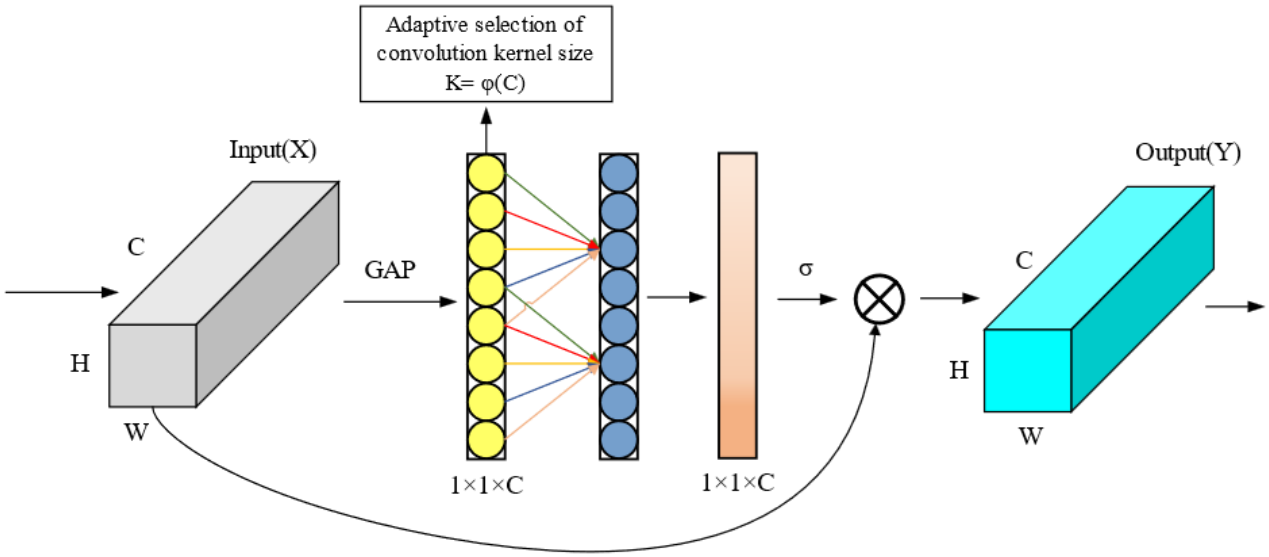


Fig. 4. ECA module

appropriate channel dependencies, and the adaptive strategy as in (3).

$$k = \psi(C) = \left| \frac{log_2(C)}{\gamma} + \beta \right|_{odd} \tag{3}$$

where γ and β are hyperparameters, which we set to $\gamma = 2$ and $\beta = 1$, $|.|_{odd}$ denotes that the result is taken to the nearest odd number to ensure that the size of the convolution kernel is odd and to avoid information loss in the convolution operation.

The channel weight vector w obtained from the 1D convolution is normalized through the Sigmoid activation function to obtain the attention weight w_c for each channel as in (4).

$$\begin{aligned} w_c &= \sigma(w_c) \\ c &= 1, 2, ..., C \end{aligned} \tag{4}$$

where σ(.) denotes the Sigmoid activation function, and the mathematical expression of σ(.) as in (5).

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{5}$$

Finally, the original input feature map X is multiplied with the generated channel weights $\in$ to achieve adaptive weighting for each channel and obtain the output feature map result $Y \in R^{C \times H \times W}$ as in (6).

$$Y_c(h,w) = w_c X_c(h,w) \tag{6}$$

Where $Y_c(h, w)$denotes the value of the c-th channel in the output feature map at a position with height h and width w , c=1, 2, …, C; h=1, 2, …, H; w=1, 2, …, W.

D. Small Target Detection layer

In real-world drowning scenarios, particularly in complex aquatic environments, there are often numerous small targets present, such as half-exposed or only a small part of the body parts on the water's surface and drowning people positioned at the edge or corner of the image or video. The small targets occupy a relatively small proportion of the image and exhibit diverse forms.

The traditional YOLOv8 algorithm performs feature fusion, it outputs 20×20, 40×40, and 80×80 feature maps for target detection. However, since small targets make up only a small portion of the image and are represented in various ways, it is challenging for the shallow network to capture their features adequately. The downsampling multiplier of the original YOLOv8 is large, so it is challenging for deep feature maps to effectively capture the feature information of small targets, which is prone to non-detection and misdetection [30].

In order to solve this problem, this research adds a new 160×160 STDL to the Head layer of YOLOv8, as in Fig. 5. The new STDL is specialized in detecting small targets, forming a 4-detection head structure corresponding to very small, small, medium, and large targets. It is connected with the shallow and deep feature maps to expand the network's detection range for very small targets in images and improve the network's ability to recognize small targets. In summary, STDL effectively reduces missed and false detections by refining target segmentation and enhancing attention to detail, thereby improving the model's adaptability.

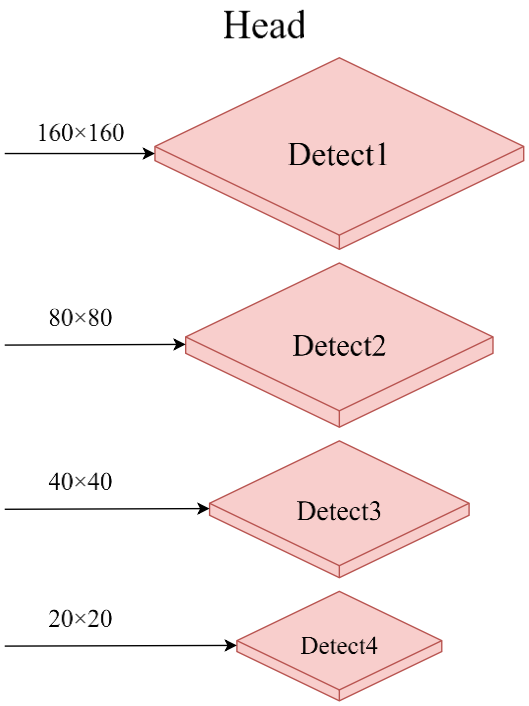


Fig. 5. Structure of small target detection head

(a) open water (b) fast-moving currents (c) swimming pool

(d) initial struggle (e) Intermediate body sinking (f) Submerged at a later stage

Fig. 6. Pictures of the dataset in different scenarios

III. Experimental Results and Evaluation

A. Dataset

The dataset utilized in this study is constructed from publicly available real-world images of drowning individuals sourced from the internet, frames extracted from drowning-related video footage, as well as images of swimmers on the water surface. This dataset encompasses a wide range of drowning scenarios: for instance, Fig. 6(a) illustrates a drowning individual in an open water area; Fig. 6(b) shows a case in fast-moving currents; and Fig. 6(c) depicts a drowning incident in a swimming pool. Moreover, the dataset captures different stages of the drowning process. As shown in Fig. 6(d), the early stage involves struggling and calling for help; Fig. 6(e) represents the intermediate stage where the body is partially submerged; and Fig. 6(f) corresponds to the late stage where the body is almost or completely submerged.

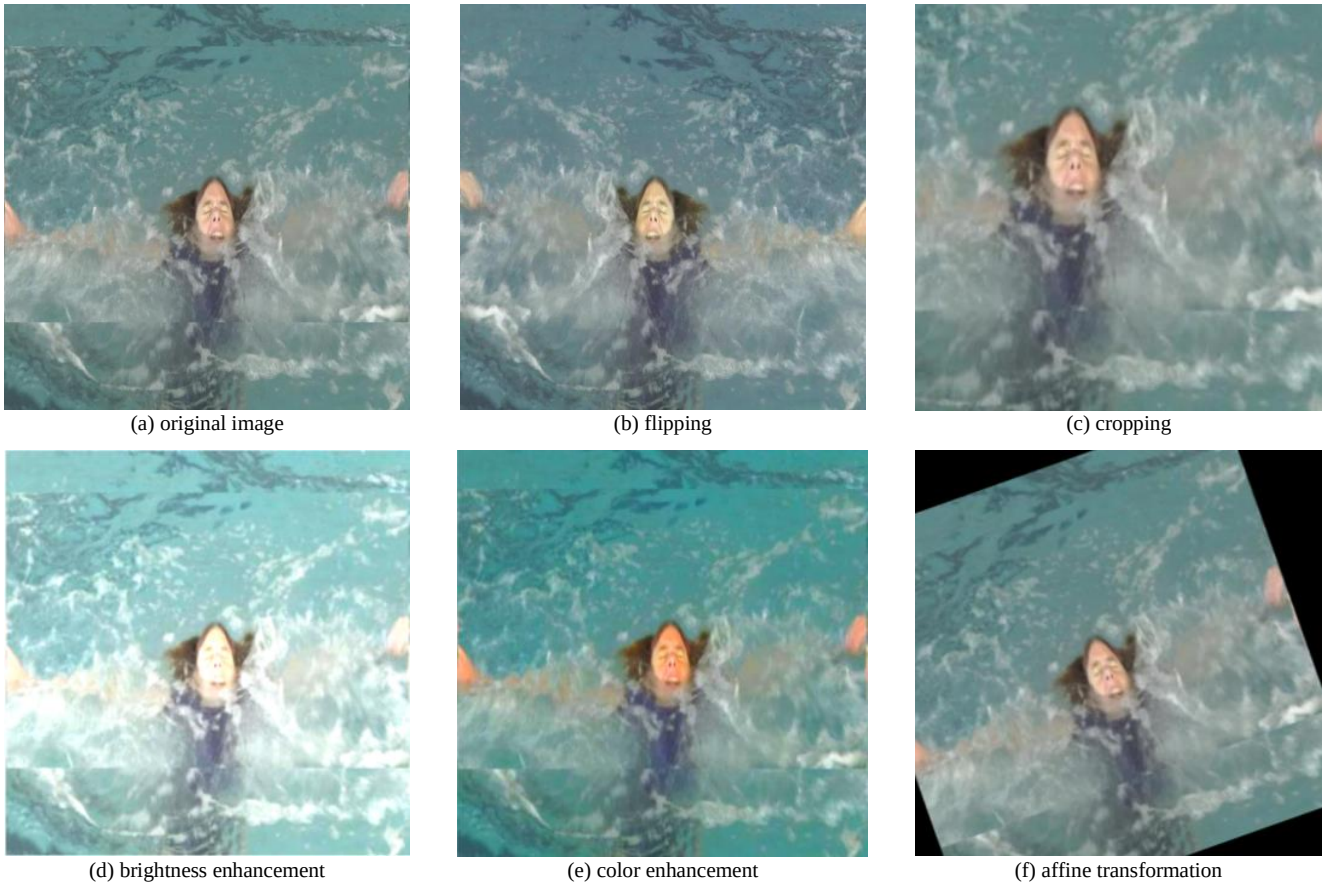
To enhance the diversity of the sample data, various data augmentation techniques were applied to the original dataset images to expand the number of training samples. As shown in Fig. 7, Fig. 7(a) presents the original image, while Fig. 7(b) shows the image after flipping, Fig. 7(c) illustrates the result of cropping, Fig. 7(d) displays the image after brightness enhancement, Fig. 7(e) presents the image after color enhancement, and Fig. 7(f) shows the image after applying an affine transformation. The above operations help to improve the generalization ability of the model as well as its applicability and usefulness in different application environments. The processed dataset has 24730 images, and dataset was divided in a 7:2:1 ratio, yielding 17,311 images for training, 4,946 for validation, and 2,473 for testing.

Fig. 8 presents the dataset's distribution. The x-axis and y-axis in Fig. 8(a) indicate the horizontal and vertical positions of the image, respectively, illustrating the distribution of the target's location in the dataset. The target distribution in this dataset is relatively decentralized, with most target objects appearing in the middle of the image. The horizontal axis in Fig. 8(b) indicates the width of the target object and the vertical axis denotes the height of the target object, demonstrating the size distribution of the targets in the dataset. The dataset demonstrates significant multi-scale characteristics in target representation, particularly with a substantial prevalence of small-scale objects.

In this study, the LabelImg software is used to annotate the images in the augmented dataset, capturing the coordinate information for each bounding box. The annotations are stored in XML label files, where each file details the class label of each object and the coordinates of its bounding box, including the top-left corner (x_{min}, y_{min}) and the bottom-right corner (x_{max}, y_{max}). Subsequently, the data needs to be converted into the TXT label format suitable for the YOLO object detection algorithm, as in (7).

$$\begin{cases} x_c = \dfrac{x_{min} + x_{max}}{2W}, y_c = \dfrac{y_{min} + y_{max}}{2H} \\ w = \dfrac{x_{max} - x_{min}}{W}, h = \dfrac{y_{max} - y_{min}}{H} \end{cases} \tag{7}$$

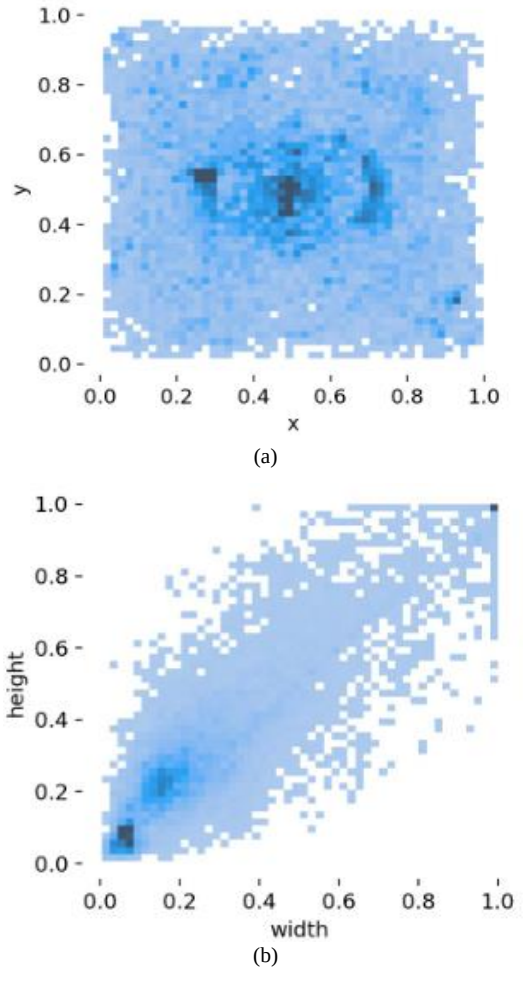
Where (x_c, y_c) denotes the coordinates of the center point of the annotation box after the normalization operation; (W, H) denotes the width and height of the original annotation box; (w, h) denotes the width and height of the annotation box after the normalization operation.



(a) original image (b) flipping (c) cropping

(d) brightness enhancement (e) color enhancement (f) affine transformation

Fig. 7. Example of a data-enhanced image of a drowning person



(a)

(b)

Fig. 8. Distribution of (a) objects center. (b) objects size

B. Experimental preparation

The experiments were conducted on an Ubuntu-18.04 64-bit Server, using an Intel Xeon with an NVIDIA 2080TI GPU with 12GB of memory. The network structure is constructed based on Pytorch1.8.1, the programming language is Python3.9, and the GPU acceleration library is CUDA11.1. The training parameters are defined below: the image size of both training and testing environments is 640×640, the training period Epoch is set to 100 rounds, and the image batch size is set to 64. The network parameters are updated using the SGD optimizer, the momentum value is configured as 0.937, the starting learning rate is defined as 0.01, and the weight decay coefficient is set to 0.0005.

C. Evaluation criteria

The evaluation indexes for the detection precision are the precision rate (P), recall rate (R) and mean average precision (mAP@ 0.5). Accordingly, the criteria for model size and real-time performance detection performance are model parameters, detection speed and frame rate. The precision rate is the percentage of correctly identified samples out of the total number of samples in the model prediction results, and the recall rate is the proportion of all actual positive samples successfully recognized by the evaluation model. P and R are formulated as in (7)-(8).

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

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

where TP represents the number of correctly identified

targets, *FP* indicates the number of incorrect target detections, and *FN* refers to the targets that were not detected. mAP is the average value used to measure the model's recognition precision for a number of categories, as in (9).

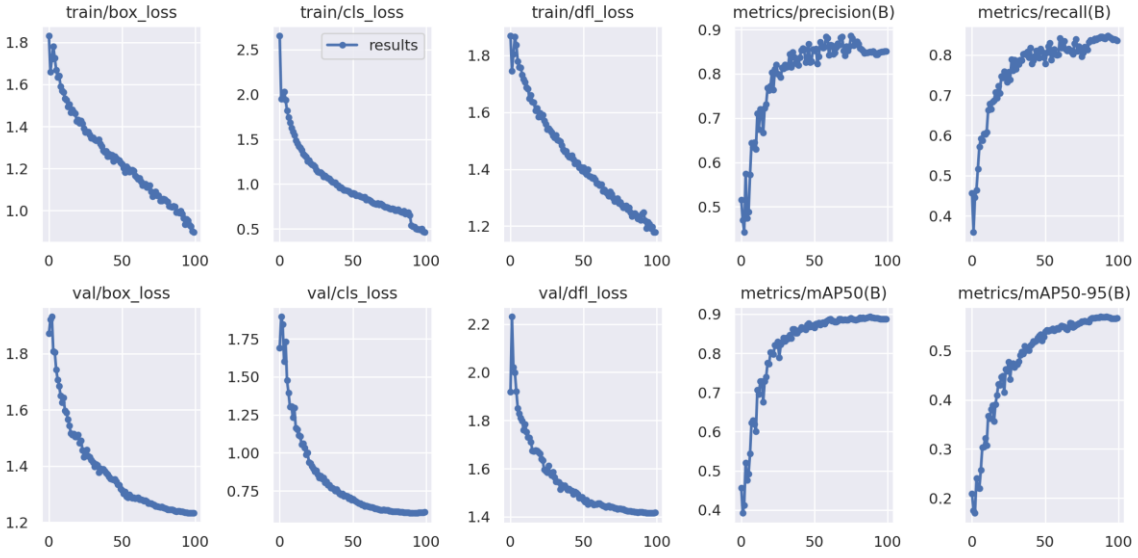
$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N}\int_{0}^{1} P(R)dR \tag{9}$$

where N is the overall count of categories and $P(R)$ denotes the precision and recall curves. In this study, mAP@0.5 is the primary evaluation metric, representing the mean average precision across all classes when the intersection over union between the predicted bounding box and the ground truth box reaches 0.5 in the object detection task[31].

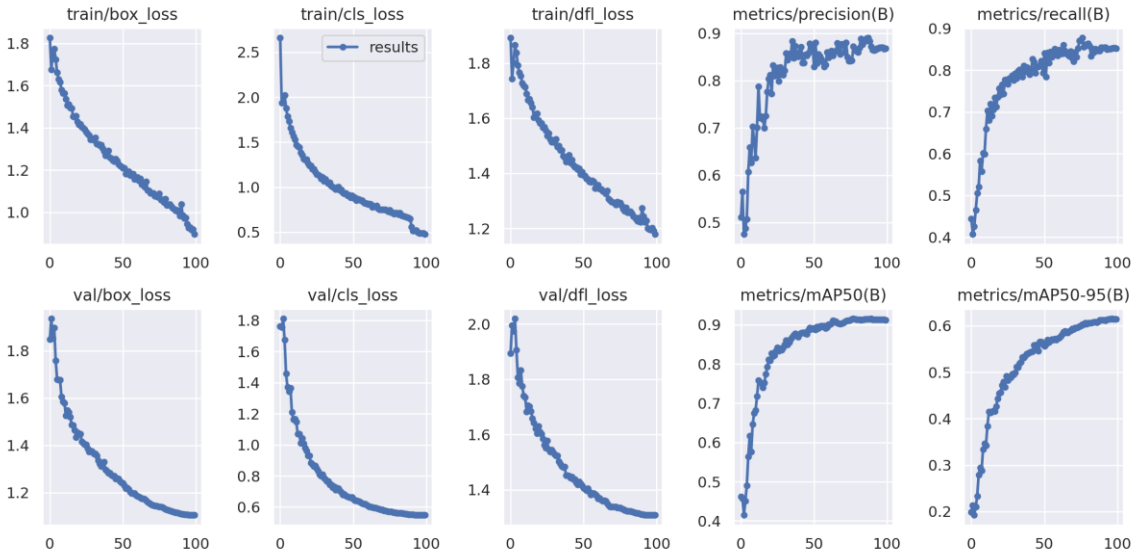
D.Experimental analysis

Fig. 9 (a) and (b) illustrate the convergence curves of the YOLOv8 and SS-YOLOv8 models. The loss curves on the training and validation sets show that the loss values of the two models are consistently low, indicating that they perform well in the boundary regression, classification accuracy and distribution prediction tasks. Comparison of the four performance evaluation metrics further demonstrates that the SS-YOLOv8 model outperforms YOLOv8 in terms of precision, recall, mAP@0.5 and map@0.5:0.95, achieving superior results across these metrics. Additionally, the convergence curves of SS-YOLOv8 exhibit greater smoothness, suggesting enhanced stability during training. Overall, the SS-YOLOv8 model demonstrates superior convergence performance compared to YOLOv8.

The P-R curve is an essential tool for evaluating the performance of a model, which visualizes the model's trade-off between precision and recall under various thresholds, especially in unbalanced classification tasks. The model's overall performance can be quantified by calculating the area under the curve, with a larger area indicating better performance. Fig. 10(a) and (b) present the P-R curves for YOLOv8 and SS-YOLOv8, respectively, demonstrating that our proposed SS-YOLOv8 algorithm achieves a mAP of 91.3% across all categories at the same IoU threshold, significantly surpassing YOLOv8's mAP of 89.4%. These results highlight the effectiveness of the SS-YOLOv8 model in improving detection accuracy.

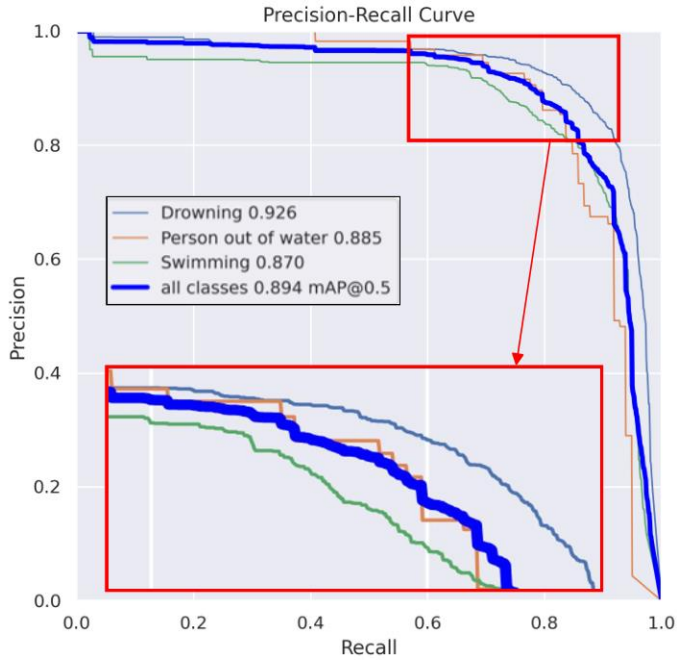


(a) YOLOv8 convergence curve

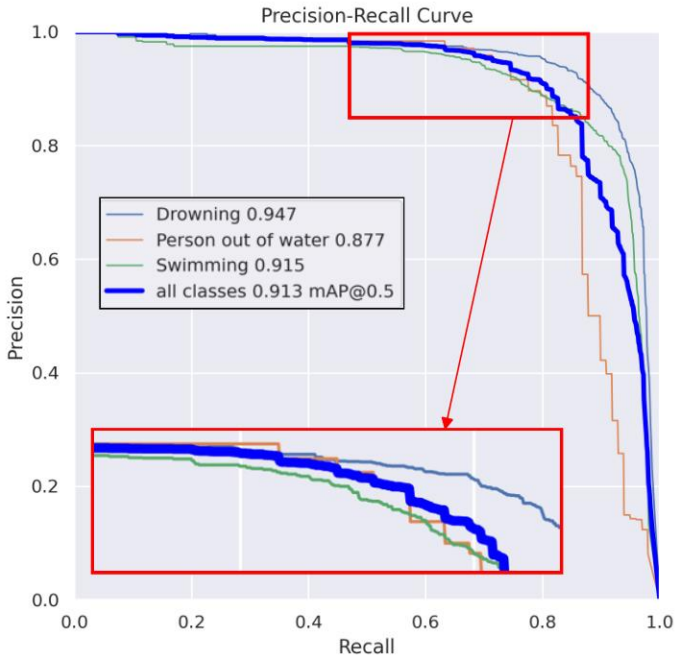


(b) SS-YOLOv8 convergence curve

Fig. 9. Comparison of YOLOv8 and SS-YOLOv8 convergence curves



(a) YOLOv8-P-R curve



(b) SS-YOLOv8-P-R curve

Fig. 10. Compare the P-R curves of YOLOv8 and SS-YOLOv8

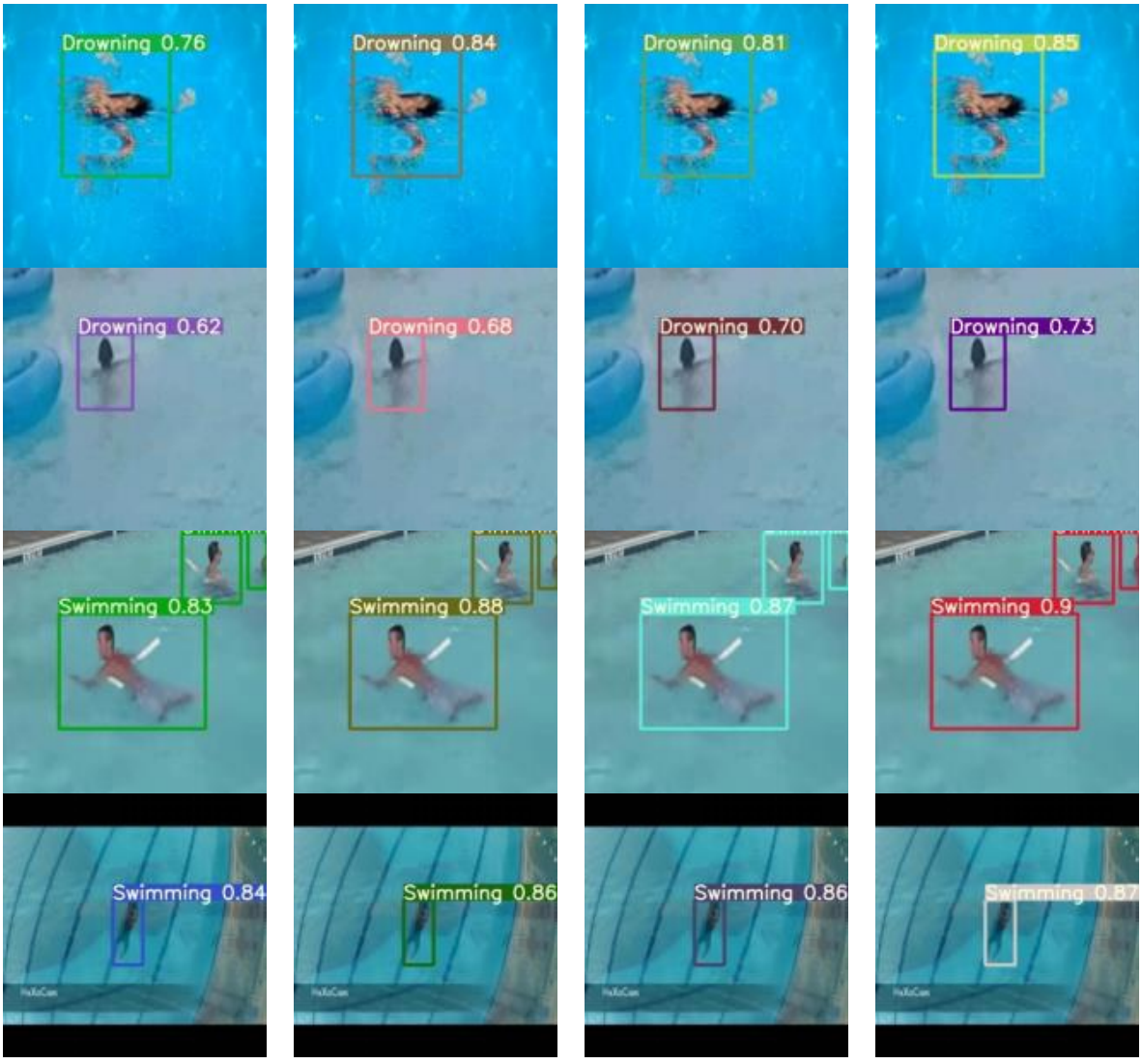
E. Ablation Experiment

To evaluate the contribution of each improvement to the overall model performance, this study conducts ablation experiments as a key component of its evaluation framework. These experiments are designed to systematically evaluate the effectiveness of individual improvements, ensuring a comprehensive understanding of their impact on the model's overall performance. Table I presents the results of ablation studies, demonstrating that the individual incorporation of either the SPPF_ECA module or the small object detection layer into YOLOv8 significantly enhances both the model's precision and recall rate. Specifically, adding the SPPF_ECA module successfully enhances the feature expression ability between channels and better extracts key features. Besides, adding the small target detection layer can effectively improve the ability to capture small targets and reduce the false alarms and omissions of the model; however, their mAP@0.5 is the same as that of YOLOv8, which suggests that using them individually is not a significant improvement to the overall model performance. The SS-YOLOv8 shows the best performance metrics, with the mean values of precision, recall and average precision reaching 87.1%, 86.5% and 91.3%, respectively, which are significantly improved by 2%, 2.1% and 1.9% compared to the YOLOv8 model. The results suggest that simultaneously adding the SPPF_ECA Attention Module and the Small Target Detection Layer contributes best to the detection effect, and significantly improves the overall detection performance of the model.

Although ECA is a relatively efficient attentional mechanism, it still introduces additional parameters and computation, especially when dealing with feature maps with many channels. Thus, the number of parameters will increase

TABLE I
RESULTS OF ABLATION EXPERIMENTS

Net	P(%)	R(%)	mAP@0.5(%)	Parameters($\times 10^6$)	Speed(ms)/Frame(fps)
YOLOv8	85.1	84.4	89.4	68.2	22.8/43.8
YOLOv8-SPPF_ECA	86.2	84.8	89.7	75.5	23.2/43.1
YOLOv8-STDL	85.9	85.7	90.2	106.4	24.5/40.8
SS-YOLOv8	87.1	86.5	91.3	111.8	25.2/39.7



(a)YOLOv8 (b)YOLOv8-SPPF_ECA (c)YOLOv8-STDL (d)SS-YOLOv8

Fig. 11. Comparison of model ablation visualization results

slowing in the detection rate. Adding a small-target detection layer will extract additional features and increase the system's computational overhead and processing time thereby increasing the model's computational bueden and detection time.

Comprehensively, the improved model proposed has the best overall detection performance, affording excellent detection precision and a minor decrease in the detection speed while maintaining a high frame rate. Therefore, the enhanced model fulfills the criteria for lightweight design, and has good practicality and balance.

To demonstrate the detection effect of the improved model more intuitively, the partial detection visualization results of each module after ablation are shown in Fig. 11. It is evident that the detection performance of the YOLOv8 model has been improved by using the SPPF_ECA module and the small object detection layer on the framework. The improved model shows excellent detection performance in the face of both multiscaleand small targets, proving the effectiveness and applicability of the algorithm improvement.

F. Comparison Experiment

To further analyze the detection effectiveness of the enhanced model, it is compared with the commonly used target detection models, YOLOv5, YOLOv7, YOLOv9, and YOLOv10 based on the metrics: P, R, mAP@0.5(%), number of parameters, detection speed and frame rate. The experimental data are summarized in Table II, highlighting that the improved model significantly outperforms competitor other models in precision, recall, and mean average precision. Compared to the latest target detection model YOLOv10, the precision, recall, and mean average

precision is improved by 0.9%, 2%, and 0.8%, respectively. The detection speed is second only to the YOLOv10 model, which better balances performance and speed, effectively meeting the requirements for both lightweight design and high precision. This demonstrates that the enhanced model generally exhibits strong detection performance, aligning well with the intended objectives.

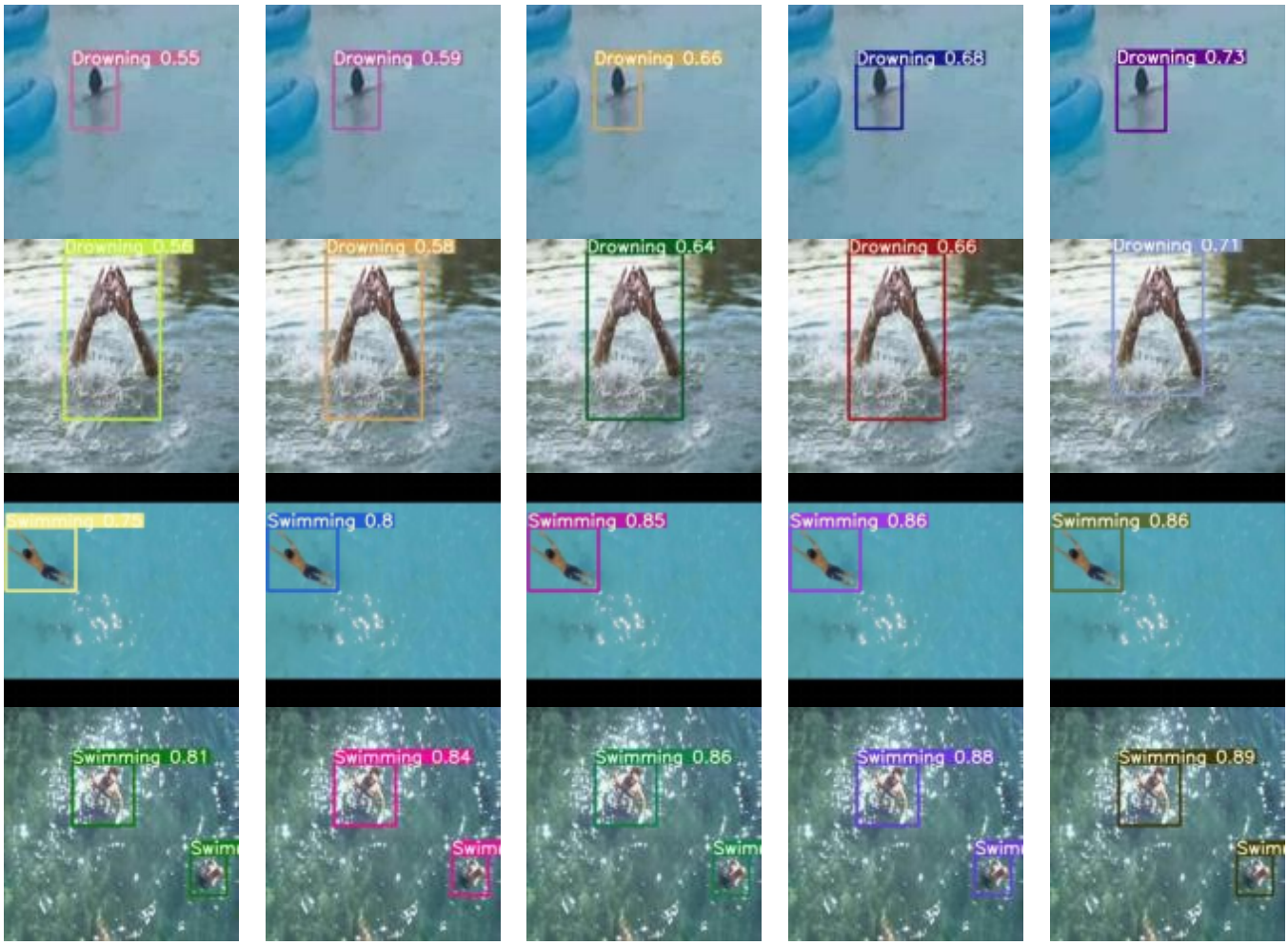
To reveal show the comparative experimental results of each model more intuitively, Fig. 12 compares the graphs between YOLOv5, YOLOv7, YOLOv9, YOLOv10, and SS-YOLOv8. The graphs infer that the proposed model fulfills the drowning person recognition task in various application scenarios. Compared with other models, the SS-YOLOv8 model outperforms other models by achieving superior confidence scores and maintaining reasonable confidence intervals. This enhanced performance indicates improved model stability, higher detection accuracy, and greater robustness, which collectively ensure reliable detection capabilities even in complex environmental conditions.

G. Visualization Results

Fig. 13 shows some representative visualization detection results evaluated in the dataset. The enhanced SS-YOLOv8 model can still be able to accurately conduct the drowning person recognition task with excellent detection performance, even in complex cases such as multi-targets, small targets, overlapping, wave overturning, image blurring, and reflections on the water surface. Indeed, even under such challenging scenarios the improved model completes the drowning person identification task relatively accurately, showing good detection performance and proving its robustness and applicability.

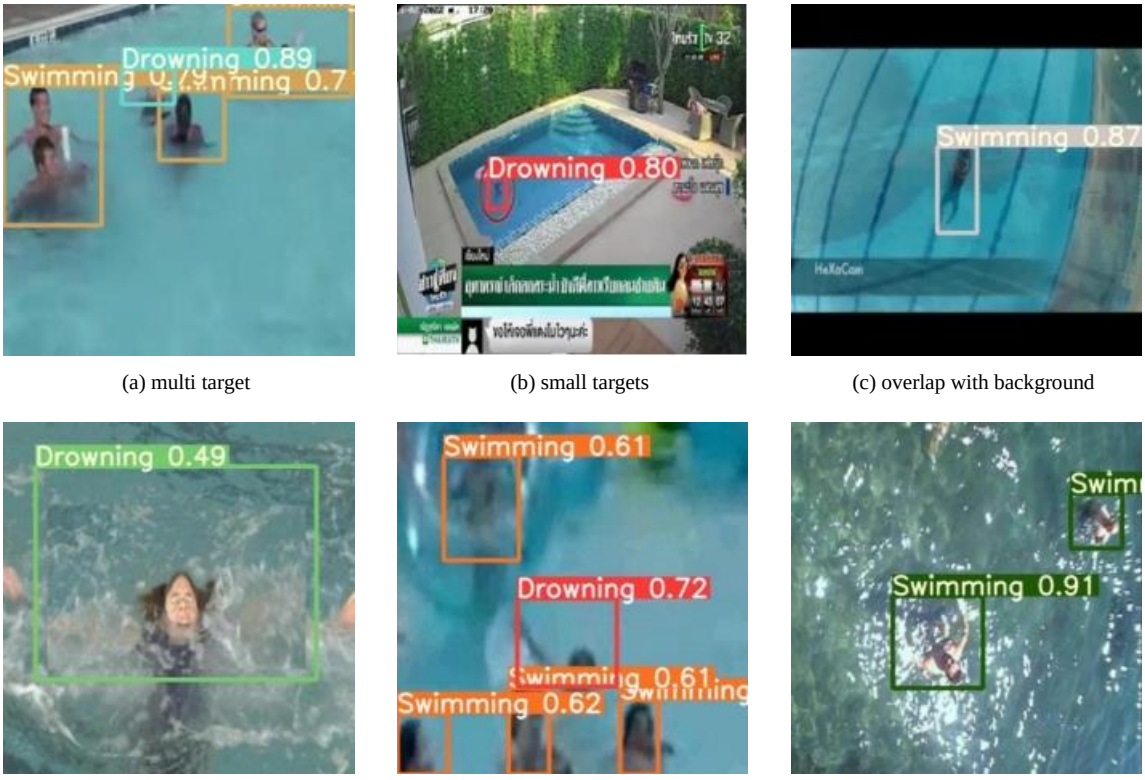
TABLE II
Comparison results of the performance of different models

Net	P(%)	R(%)	mAP@0.5(%)	Parameters($\times 10^6$)	Speed(ms)/Frame rate(fps)
YOLOv5	82.8	83.5	87.1	87.4	26.5/37.7
YOLOv7	83.6	83.6	88.5	92.3	27.7/36.1
YOLOv9	85.7	84.1	89.8	94.7	28.6/34.8
YOLOv10	86.2	84.5	90.5	86.5	23.8/42
SS-YOLOv8	87.1	86.5	91.3	111.8	25.2/39.7



(a)YOLOv5 (b)YOLOv7 (c)YOLOv9 (d)YOLOv10 (e)SS-YOLOv8

Fig. 12. Comparison of different models visualization result map



(a) multi target (b) small targets (c) overlap with background

(d) wave overturning (e) image blurring (f) reflections on the water surface

Fig. 13. Representative test results.

IV. Conclusion

This research introduces an improved YOLOv8 that focuses on drowning person recognition scenarios, focusing on poor recognition of small and complex targets in real drowning scenarios and improving recognition precision. The SPPF module in the YOLOv8 backbone network is integrated with the ECA to design the SPPF_ECA module, which augments the model's capacity to extract features across multiple scales while ensuring the detection efficiency by aggregating the multiscale context information and optimizing the weight allocation between channels. For the problem that YOLOv8is insensitive to small target detection, a new small target detection is added in the Head layer that deals explicitly with the small-scale targets in the drowning scene, significantly improving model detection of fine targets.

The experimental results demonstrate that SS-YOLOv8 achieves superior performance in drowning person detection tasks, with mean precision, recall, and average precision reaching 87.1%, 86.5%, and 91.3%, respectively. Compared to the YOLOv8 model, these metrics show improvements of 2%, 2.1%, and 1.9%, respectively. Meanwhile, when compared with the state-of-the-art object detection algorithm YOLOv10, SS-YOLOv8 exhibits enhancements of 0.9%, 2%, and 0.8% in these respective metrics. Hence, SS-YOLOv8 significantly reduces both false positives and false negatives, thereby substantially improving detection accuracy. Although the number of parameters has increased compared with the YOLOv10 model, it still maintains a fast detection speed, with a high frame rate of 39.7 fps, which meets the requirements of both lightweight design and high accuracy in drowning detection.

In the future, we will optimize the network architecture by incorporating spatial information into the attention feature maps, thereby continuously enhancing both the model's accuracy and detection efficiency. Furthermore, while current research primarily focuses on the identification of drowning individuals on the water surface, future research will expand into the area of underwater drowning detection, aiming to enhance the comprehensiveness and accuracy of aquatic rescue systems.

References

[1] WHO urges countries to invest in drowning prevention to protect children. Saudi medical journal, vol. 44, no. 8, p. 821, 2023.

[2] A. B. Amjoud and M. Amrouch, “Object Detection Using Deep Learning, CNNs and Vision Transformers: A Review,” in IEEE Access, vol. 11, pp. 35479-35516, 2023.

[3] Salehi, Nasrin et al. “An Automatic Video-based Drowning Detection System for Swimming Pools Using Active Contours.” International Journal of Image, Graphics and Signal Processing, vol. 8, pp.1-8, 2016.

[4] Prakash, Bhaskaran David, “Near-drowning Early Prediction Technique Using Novel Equations (NEPTUNE) for Swimming Pools.” 2018, arXiv : 1805,02530.

[5] Shiuuee K, Rezaei F, “A presentation of drowning detection system on coastal lines using image processing techniques and neural network.” Journal of injury and violence research, vol.11, p.18, 2019.

[6] X. He, F. Yuan, Y. Zhu, “Drowning Detection Based on Video Anomaly Detection.” International Conference on Image and Graphics, 2021, pp. 700-711.

[7] J.X. Jian, C.M. Wang, “Deep Learning Used to Recognition Swimmers Drowning,” 2021 IEEE/ACIS 22nd International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Taichung, Taiwan, 2021, pp. 111-114.

[8] M.A. Hayat, G. Yang, A. Iqbal, “Mask R-CNN based real time near drowning person detection system in swimming pools.” 2022 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), Karachi, Pakistan, 2022, pp. 1-6.

[9] Uma N. Dulhare, Mohd Hussam Ali, "Underwater human detection using faster R-CNN with data augmentation."Materials Today: Proceedings,vol. 80, pp. 1940-1945, 2023.
[10] F. Lei, H. Zhu, F. Tang, X. Wang, "Drowning behavior detection in swimming pool based on deep learning, Signal." Image and Video Processing vol. 16, pp. 1683-1690, 2022.
[11] Q. Niu, Y. Wang, S. Yuan, K. Li and X. Wang, "An Indoor Pool Drowning Risk Detection Method Based on Improved YOLOv4." 2022 IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 2022, pp. 1559-1563.
[12] Q. He, Z. Mei, H. Zhang and X. Xu, "Automatic Real-Time Detection of Infant Drowning Using YOLOv5 and Faster R-CNN Models Based on Video Surveillance." Journal of Social Computing, vol. 4, no. 1, pp. 62-73, 2023.
[13] T. Liu, X. He, L. He, F. Yuan, "A video drowning detection device based on underwater computer vision." IET Image Process. Vol. 17, pp. 1905–1918, 2023.
[14] G.D. Lunde, Bruk Av Nevrale Nettverk for Avansert Deteksjon Av Objekter under Vann, University of Stavanger, Norway, 2009.
[15] J.M. Lavest, F. Guichard, C. Rousseau, Multi-view reconstruction combining underwater and air sensors, Proceedings. International Conference on Image Processing, 22-25 Sept 3 (2002 2002) 813–816.
[16]] W.W.T. Fok, L.C.W. Chan, C. Chen, Artificial intelligence for sport actions and performance analysis using recurrent neural network (RNN) with long shortterm memory (LSTM), in: Presented at the Proceedings of the 4th International Conference on Robotics and Artificial Intelligence, Guangzhou, China, 2018.
[17] M. Hussain, "YOLOv1 to v8: Unveiling Each Variant–A Comprehensive Review of YOLO." IEEE Access, vol. 12, pp. 42816-42833, 2024.
[18] W. Jiang, D. Han, B. Han and Z. Wu, "YOLOv8-FDF: A Small Target Detection Algorithm in Complex Scenes," IEEE Access, vol. 12, pp. 119223-119237, 2024.
[19] Z. Han, Y. Cai, A. Liu et al. "MS-YOLOv8-Based Object Detection Method for Pavement Diseases," Sensors, vol. 24, no. 14, p.4569, 2024.
[20] X. Liu, Y. Wang, D. Yu and Z. Yuan, "YOLOv8-FDD: A Real-Time Vehicle Detection Method Based on Improved YOLOv8," IEEE Access, vol. 12, pp. 136280-136296, 2024.
[21] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.
[22] S. Qing, Z. Qiu, W. Wang, F. Wang, X. Jin, J. Ji, L. Zhao & Y. Shi, "Improved YOLO-FastestV2 wheat spike detection model based on a multi-stage attention mechanism with a LightFPN detection head," Frontiers in plant science, vol. 15, p.1411510, 2024.
[23] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, "Path Aggregation Network for Instance Segmentation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 8759-8768.
[24] H. Zhang, Y. Wang, F. Dayoub and N. Sünderhauf, "VarifocalNet: An IoU-aware Dense Object Detector," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 8510-8519.
[25] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 2015.
[26] A. H. Ahmed, S. M. Youssef, N. Ghatwary and M. A. Ahmed, "Myositis Detection From Muscle Ultrasound Images Using a Proposed YOLO-CSE Model," IEEE Access, vol. 11, pp.107533-107547, 2023.
[27] Le. N, Nguyen. K, Nguyen. A, and Le.H, "Global-local attention for emotion recognition," Neural Computing and Applications, vol, 34, pp.1-15, 2021.
[28] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-Excitation Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, pp. 2011-2023, 2020.
[29] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: efficient channel attention for deep convolutional neural networks," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.11531-11539, 2020.
[30] J. Lin, K. Zhang, X. Yang, X. Z. Cheng, & C. Li, "Infrared dim and small target detection based on u-transformer," Journal of Visual Communication and Image Representation, vol 89, p. 103684, 2022.
[31] L. Zhang, M. Xu, G. Wang, R. Shi, Y. Xu & R. Yan, "SiameseNet Based Fine-Grained Semantic Change Detection for High Resolution Remote Sensing Images,"Remote Sensing, vol. 15, no. 24, p. 5631, 2023.