Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.079393

ARTICLE

Robust Multi-Object Fish Tracking in Dynamic Aquatic Environments via Attention-Enhanced YOLOv8 and LSTM-Based Trajectory Prediction

Feng-Cheng Lin*, Bo-Chiao Jan and Hui-An Wu

Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan

*Corresponding Author: Feng-Cheng Lin. Email: fclin@fcu.edu.tw

Received: 20 January 2026; Accepted: 31 March 2026; Published: 08 May 2026

ABSTRACT: With the increasing refinement of ornamental fish culture, understanding fish behavioral patterns has become critical. Fish movements not only reflect daily activity ranges but also reveal responses to environmental changes such as water currents and obstacles. However, traditional manual observation is limited by manpower and time, making it difficult to record fish behaviors over long periods stably. Existing automated tracking techniques often suffer from ID switches and track interruptions caused by rapid fish movement, occlusions, or intermingling, which in turn degrade the reliability of subsequent analyses. This paper proposes a deep learning-based multi-object fish tracking system that integrates YOLOv8n for object detection and employs an IoU matching criterion to associate detections across consecutive frames, thereby maintaining object ID continuity. To further reduce ID loss under rapid motion and partial occlusion, a multiple-LSTM prediction model is introduced as a temporal compensation mechanism, thereby improving timing stability and track continuity. Moreover, considering disturbances in the experimental field (e.g., water disturbance and water current interference) that can blur fish body edges and fine details, an attention-enhanced detector, YOLOv8-CS (Convolutional Block Attention Module and Squeeze-and-Excitation), is developed by embedding CBAM and SE modules into the YOLOv8 architecture to enhance detection accuracy in dynamic waters. Experimental results demonstrate that the proposed system effectively increases the Multiple Object Tracking Accuracy (MOTA) to 77.23% and significantly reduces ID switches to 42.5, ensuring more robust and continuous trajectory tracking compared to benchmark methods. This system provides a highly reliable tool for automated behavior analysis in complex and dynamic aquatic environments.

KEYWORDS: LSTM; trajectory prediction; multi-object tracking; YOLOv8-CS; attention mechanism

1 Introduction

In recent years, ornamental fish have become an increasingly important part of family leisure and entertainment, driving steady growth in the global ornamental fish industry. As market demand expands, the industry is shifting toward refined and standardized management, which requires more reliable monitoring of fish behavior. Key indicators such as movement patterns, adaptability to environmental changes, and group behavioral dynamics directly reflect the health and growth status of fish populations.

However, traditional monitoring methods present significant challenges. Manual observation is labor-intensive and time-consuming, making long-term, continuous data collection impractical. Extended monitoring not only requires multiple observers but also yields inconsistent results owing to inter-observer variability and subjective judgment. Tagging, while useful for individual identification, can alter natural

swimming patterns and has been reported to increase stress responses, negatively affecting both animal welfare and data reliability.

These limitations highlight the need for a non-invasive, automated, and efficient fish tracking system that enables continuous monitoring with high accuracy and minimal interference. Developing such a system is therefore a critical step toward advancing precision management in modern aquaculture.

With the rapid advancement of deep learning, the YOLO series of object detection models has been widely adopted for multi-object tracking. However, several challenges persist in fish tracking tasks that can directly degrade model performance. Rapid swimming and frequent direction changes often lead to target loss, reducing detection stability in long-term tracking sequences. Similarly, overlap and occlusion among fish frequently cause identity mismatches, which reduce re-identification accuracy and compromise the reliability of subsequent behavior analyses.

In current-disturbed environments, blurred silhouettes and reflective artifacts that blend into the background can reduce detection confidence scores and increase false positives, further undermining accuracy and data quality. To address these challenges, this study proposes a multi-object fish tracking architecture that integrates YOLO-based detection, IoU-based matching, and a Long Short-Term Memory (LSTM) network. By leveraging the LSTM's temporal prediction capability, historical trajectory information is used to estimate displacement in the current frame. This approach enhances the stability of ID assignment, reduces ID switching caused by occlusion, jitter, or detection errors, and improves overall tracking accuracy and robustness.

Furthermore, experiments were conducted in a small aquarium with circulating water flow to simulate real breeding conditions. Factors such as camera limitations, lighting variations, and viewing angles were considered to validate the applicability and performance of the proposed method in practical aquaculture scenarios.

2 Background

2.1 Object Detection

In this study, YOLOv8 [1] is employed as the object detection model for fish schools, with the goal of improving the accuracy and stability of fish localization. YOLOv8 adopts an anchor-free architecture, reducing reliance on the predefined anchors used in earlier YOLO versions and thereby enhancing generalization and flexibility when tracking fast-moving fish schools with dynamic postures. For feature extraction and fusion, YOLOv8 integrates the Feature Pyramid Network (FPN) and the Path Aggregation Network (PAN), enabling more efficient handling of objects at different scales and significantly improving its ability to recognize small targets.

However, due to substantial environmental variation and the diverse morphology of fish, traditional YOLO-based models can still suffer from insufficient detection accuracy. To address this issue, many researchers have proposed improvements to enhance YOLO's adaptability and robustness. Xu et al. [2] introduced the YOLO-CTS model, which enhances YOLOv5 by integrating CBAM (Convolutional Block Attention Module), a Transformer module, and SIoU. In this framework, CBAM strengthens the extraction of salient features to improve the recognition of small fish, while the Transformer module improves adaptability to morphological changes and reduces missed detections caused by different swimming postures. Similarly, Qin et al. [3] optimized YOLOv8 by proposing the FASG (Feature Attention Selection Gate) mechanism. This approach improves feature selection to mitigate the effects of lighting variations in underwater environments and enhances recognition accuracy under occlusion.

Recent studies from 2024 and 2025 have further demonstrated the robust real-time applicability and high performance of the YOLOv8 architecture across diverse aquatic scenarios. For instance, an enhanced YOLOv8 model has been employed for precise fishery detection, introducing a 3D perception module to improve robustness against underwater noise [4]. In species-specific monitoring, the YOLOv8-TF framework integrates Transformer blocks to capture global context, thereby improving recognition accuracy for fish species while addressing class imbalance [5]. Moreover, optimized versions of YOLOv8 have been developed for water surface object detection, effectively suppressing reflection interference while maintaining real-time performance suitable for intelligent water governance [6]. Collectively, these practical applications underscore YOLOv8's ability to handle complex visual disturbances in real time, thereby justifying its selection as the baseline detector for our dynamic fish-tracking system.

2.2 Object Tracking

The core of object tracking lies in matching targets across time to establish continuous and stable trajectories. A representative example is the SORT (Simple Online and Realtime Tracking) algorithm proposed by Bewley et al. [7], which employs lightweight matching components such as a Kalman filter and the Hungarian algorithm to achieve fast and efficient tracking. SORT estimates an object's motion state through a predict–update process and maintains ID continuity using an Intersection over Union (IoU)-based matching strategy. Owing to its straightforward design and high computational efficiency, SORT has been widely adopted in many applications. However, because it relies primarily on geometric position information for matching, it is prone to mismatches or ID switches when objects overlap, experience occlusion, or move rapidly, thereby reducing tracking stability.

In response to these limitations, Wojke et al. [8] proposed the DeepSORT (Deep Simple Online and Realtime Tracking) algorithm. Building upon SORT, DeepSORT incorporates a ReID (Re-identification) module that extracts object appearance features using a convolutional neural network. By combining appearance features with positional information, DeepSORT achieves more robust matching, significantly reducing ID switches and improving tracking accuracy and stability, particularly in scenarios involving occlusion or rapid object motion.

In the field of object tracking, de Oliveira Barreiros et al. [9] proposed a zebrafish tracking method that integrates YOLOv2 with a Kalman filter. In this approach, YOLOv2 is used to detect the fish head region and calculate the center of mass, while the Kalman filter predicts position and motion direction to enable cross-frame tracking and trajectory concatenation. Their experiments were conducted primarily in a static aquarium environment, with a backlight system used to reduce reflection interference and stabilize image quality.

Zhang and Palaoag [10] proposed a fish-tracking method that combines YOLOv5 with an improved SORT algorithm. Considering the nonlinear and irregular motion of underwater fish, they noted that the Kalman filter in SORT could introduce prediction errors that accumulate over time and undermine tracking stability. To address this issue, they replaced Kalman-based prediction with IoU-based bounding box matching and introduced the CBAM attention mechanism to enhance recognition robustness during tracking.

Furthermore, Liu et al. [11] developed the Fish Track algorithm, which integrates the YOLOX detector with an appearance-based re-identification mechanism. By incorporating visual feature similarity into object matching, Fish Track effectively reduces ID-switching errors caused by long-term occlusion and appearance variations, thereby improving long-term tracking accuracy and stability.

2.3 Trajectory Prediction

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) introduced by Hochreiter and Schmidhuber [12] in 1997 to address the vanishing- and exploding-gradient problems. Traditional RNNs often struggle to preserve long-range sequence information, which makes them less effective for modeling long sequences. In contrast, LSTM incorporates a memory cell and gating mechanisms that enable the network to selectively retain or discard information. As a result, LSTM can learn long-term dependencies more effectively, supporting robust modeling of sequential data.

Wang et al. [13] proposed an LSTM-based three-dimensional fish-tracking method that combines a master–slave camera architecture with multi-view image fusion to improve the accuracy of zebrafish trajectory tracking. LSTM-based prediction is used to assist with view-angle conversion and to maintain continuity in cross-view matching, thereby enhancing the stability and accuracy of 3D dynamic tracking. In addition, Palconit et al. [14] demonstrated that, in turbid waters, LSTM can effectively reduce trajectory prediction errors and adapt well to the movement patterns of underwater fish. These studies highlight the broad applicability and potential of LSTM models for trajectory prediction, providing a strong foundation for their further application in multi-object fish tracking.

Overall, these studies demonstrate that intelligent sequence modeling and learning-based optimization methods have broad applicability in dynamic IoT systems. They also suggest that integrating LSTM-based trajectory modeling with IoT-enabled sensing and control mechanisms could be a promising direction for future intelligent aquarium management, particularly in scenarios requiring adaptive monitoring, anomaly detection, and automated decision support.

This paper summarizes the studies reviewed above and notes that most approaches adopt multicomponent designs for underwater fish detection, tracking, and behavior prediction. Such designs typically enhance detection features using attention mechanisms, apply tracking compensation via Kalman filters or related methods, and integrate time-series models for trajectory prediction. However, most experimental data in the current literature are collected in static or low-disturbance water environments, where current fluctuations, reflectance changes, and background texture variations are relatively limited. Consequently, most studies primarily address episodic occlusions and transient appearance changes, rather than the more complex challenges encountered under highly dynamic aquatic conditions.

In this study, we focus on scenarios characterized by prolonged strong currents and high background interference, in which fish often exhibit blurred silhouettes and merge with the background during detection. To address this issue, we introduce an attention module to enhance feature extraction at the detection stage and integrate it with time-series prediction for subsequent tracking. This approach represents a distinct research direction, differing from existing methods in both its processing logic and its emphasis on challenging environmental conditions.

3 Methodology

3.1 Data Collection

The dataset used in this study was collected from fish-movement images captured in a real aquarium environment. To effectively train the YOLOv8 and LSTM models, we designed a comprehensive data-processing pipeline that includes image collection and acquisition, as well as annotation, to ensure that the models are trained on high-quality data.

3.1.1 Video Shooting

To obtain high-quality images of swimming fish, we mounted a mobile phone on a stand above the aquarium to capture video frames. This setup ensured that the entire aquarium area was within the camera's field of view, allowing fish movements to be fully recorded from a top-down perspective. To maintain consistent image quality, we also considered the potential impact of lighting variations on detection performance. Therefore, all images were captured under stable lighting conditions to minimize flicker and shadow changes that could interfere with fish detection.

3.1.2 Image Capture

Because the video recordings are relatively long, we converted the videos into continuous image frames for subsequent annotation and model training. Frames were extracted at a fixed rate of 30 FPS to preserve fish-movement information and to avoid discontinuous trajectories that could result from a lower frame rate.

3.1.3 Image Annotation

During the image annotation stage, we used CVAT (Computer Vision Annotation Tool) to ensure both the accuracy of the fish detection labels and the efficiency of the annotation process. CVAT is an open-source platform developed by Intel that supports a wide range of tasks, including object detection, segmentation, and classification. In addition, its temporal annotation features make it particularly well suited for the frame-by-frame labeling required for object tracking in continuous video sequences.

3.2 System Flow

As shown in Fig. 1, this study first applies YOLOv8 for object detection to locate fish in each frame. Intersection over Union (IoU) is then used as the matching criterion to associate detections across consecutive frames and maintain object ID continuity. However, ID loss may still occur when the YOLOv8 bounding box fails to align with that of the previous frame due to rapid fish movement, partial occlusion, water-current interference, lighting variations, or detection errors.

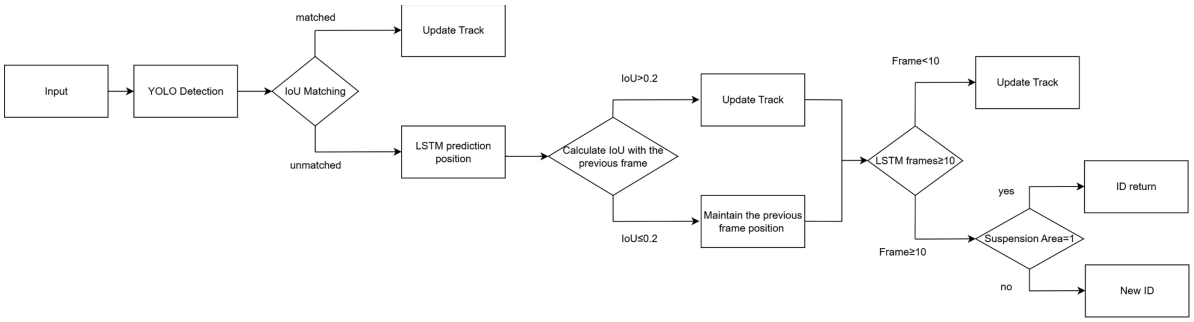


Figure 1: The proposed system flow chart.

To improve tracking stability in such cases, this study introduces a Long Short-Term Memory (LSTM) model as a temporal compensation mechanism. When an object ID cannot be matched to any detection in the current frame, the system uses the positions of that ID in the previous eight frames as input to the LSTM model to predict its likely position in the current frame. The predicted bounding box is then compared with the previous frame's bounding box using IoU. If the overlap exceeds a threshold (IoU $\geq$ 0.2), the prediction is regarded as valid and the track is updated; otherwise, the previous position is retained to avoid abnormal jumps.

If YOLOv8 fails to reacquire an object after 10 consecutive predictions, the object is considered temporarily untrackable, marked as missing, and moved to the missing area.

For each new YOLOv8 detection in the current frame that has not been matched, the system first checks whether the missing area contains exactly one object. If so, the missing ID is reassigned to the new detection to restore continuity and avoid frequent creation of new IDs due to temporary occlusion. If restoration is not possible, the detection is treated as a new object and assigned a new ID for subsequent tracking. This process effectively improves tracking stability and ID continuity for fish in high-disturbance environments.

3.3 YOLOv8 Model Modification

In this study, we selected YOLOv8n as the baseline detection architecture to balance model lightweightness with the required detection accuracy, thereby meeting the requirements of fish tracking in small aquarium environments. However, during our experiments, we found that water currents introduced substantial interference in fish detection. Dynamic background changes, such as ripples and suspended particles, often caused the fish's body contour to blend with the background, producing blurred outlines that degraded detection accuracy and stability. This effect was particularly evident when fish crossed regions with strong water flow, where detections were sometimes missed despite the strong overall performance of YOLOv8.

To address the challenges posed by current interference and turbulent flow in fish detection, this study proposes an improved model called YOLOv8-CS (CBAM + SE). Two attention mechanisms, CBAM (Convolutional Block Attention Module) and SE (Squeeze-and-Excitation), are integrated into the YOLOv8 architecture. The CBAM module guides the model to focus on key spatial regions and channel features where fish are present, helping to mitigate issues such as blurred contours and background blending caused by turbulent flow. Meanwhile, the SE module strengthens the weighting of important features and suppresses background noise such as water ripples and air bubbles. Specifically, the CBAM and SE modules are embedded into three critical layers of YOLOv8 (as shown in Fig. 2): the P2 (shallow features) and P3 (intermediate features) layers of the backbone, and the P5 layer of the head, which produces deeper semantic representations. This design highlights salient cues such as fish-body edges and textures across multi-scale feature maps, thereby improving recognition performance in dynamic water environments and occluded scenes while preserving the lightweight and efficient characteristics of the original YOLOv8 architecture.

In terms of module embedding, the attention module is inserted after each C2f layer. As the primary feature fusion module in YOLOv8, C2f performs initial information integration and feature extraction. The subsequent attention module then refines these features and guides the model to focus on regions with high recognition value by redistributing weights across channels and spatial dimensions. This configuration not only enhances the model's feature representation capability but also avoids disrupting early feature learning, thereby promoting stable training and effective convergence of the overall framework.

Regarding the order of the attention modules, this study places the CBAM [15] module before the SE [16] module. Since CBAM incorporates both channel and spatial attention, it enables the model to automatically identify the most informative feature channels and spatial regions. This helps filter background noise, such as water ripples and floating particles, while emphasizing the contours and textures of the fish body.

From CBAM [15] module. First, the channel attention module learns the weights of different channels in the input feature maps. It generates two one-dimensional vectors via global max pooling and global average pooling, which are fused through a fully connected layer with shared weights. The output is then passed through a sigmoid activation function to produce a channel attention weight map, which is used to reweight

each channel. This process enables the model to emphasize semantic information relevant to the fish body (e.g., edges and textures) while suppressing background features unrelated to the task.

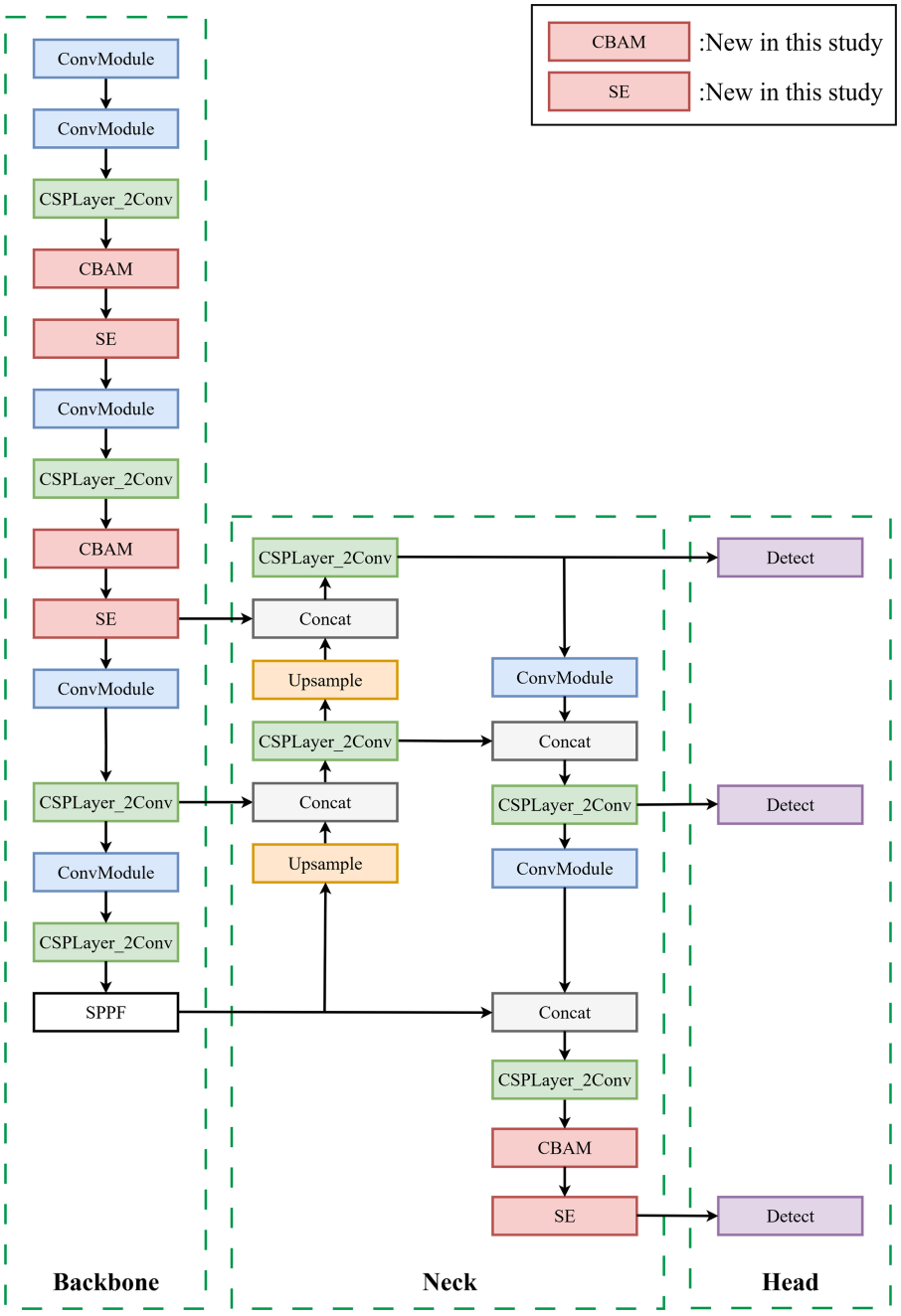


Figure 2: The proposed model based on YOLOv8 with integrated CBAM and SE attention mechanisms.

Next, the spatial attention module enhances the spatial representation of the feature maps. It first applies max pooling and average pooling along the channel dimension to the channel-refined feature maps, and then generates a spatial attention map through a 7×7 convolution. This map guides the model to focus on more discriminative regions in the image, such as fish body locations and silhouette edges, while reducing the influence of water disturbances and background noise.

After CBAM suppresses irrelevant features, the SE [16] module further redistributes channel weights, allowing the model to strengthen or weaken features based on their importance adaptively. This refinement improves both classification and localization accuracy. Together, CBAM and SE act as complementary modules: CBAM provides global-to-local attention across spatial and channel dimensions, while SE fine-tunes inter-channel relationships. Their combination enhances semantic discrimination in dynamic underwater environments. The SE module architecture, which performs channel reweighting through squeeze-and-excitation operations to strengthen feature representation and further improve detection accuracy.

Through this strategic embedding design, YOLOv8-CS effectively strengthens the model's focus on fish body regions, improves noise suppression under dynamic water conditions, and ultimately achieves stable, high-accuracy fish detection and tracking performance despite interference induced by water currents.

3.4 LSTM Training Framework

Bridging the detection and tracking stages, the refined spatial and channel features enhanced by the attention mechanisms ensure accurate object localization, which in turn provides reliable coordinate inputs for the LSTM-based trajectory prediction model (Fig. 3), enabling it to compensate for potential ID loss. In this study, we adopt this LSTM architecture, which learns fish movement patterns from the previous eight frames and predicts the fish's position in the current frame. By incorporating angle information, the model improves prediction accuracy and compensates for ID loss and mismatches caused by occlusion, detection failures, or environmental changes.

The LSTM model designed in this study is illustrated in Fig. 3. It consists of an input layer, an LSTM hidden layer, a fully connected layer, and an output layer. The input layer receives fish movement data from the previous eight frames, with each frame containing center coordinates and angle information, enabling the model to capture both spatial and directional changes. The LSTM hidden layer uses a two-layer structure: the first layer contains 128 hidden units, and the second layer contains 64 hidden units. A dropout rate of 0.1 is applied between the layers to prevent overfitting. The fully connected layer reduces the dimensionality of the time-series features produced by the LSTM and applies a ReLU (Rectified Linear Unit) activation function to enhance nonlinear representation. Finally, the output layer generates the predicted coordinates and angles of the fish in the current frame, and the results are normalized using a sigmoid function to constrain the values to a reasonable range and ensure stable outputs.

3.4.1 Assessment Indicators

Since the goal of the LSTM model is to predict the fish's position in the current frame based on the previous eight frames, this task can be formulated as a regression problem. For regression tasks, MSE (mean squared error) and MAE (mean absolute error) are commonly used evaluation metrics because they measure the error between predicted values and the ground truth. Therefore, in this study, MSE and MAE are selected as the primary evaluation metrics to capture overall accuracy and control extreme errors, ensuring that the LSTM model can predict fish movement trajectories both stably and accurately.

MSE and MAE represent mean squared error and mean absolute error, respectively. MSE evaluates the squared differences between predicted and actual values, making it more sensitive to large deviations and suitable for assessing the model's ability to handle extreme errors. In contrast, MAE measures the mean absolute difference, providing a more intuitive reflection of prediction bias and being less affected by small fluctuations. By jointly analyzing these two metrics, the model's accuracy and robustness can be assessed

more comprehensively. The formulas for MSE and MAE are provided in Eqs. (1) and (2).

$$MSE = \frac{\sum_{t=1}^{N} (real_t - pred_t)^2}{N} \tag{1}$$

$$MAE = \frac{\sum_{t=1}^{N} |real_t - pred_t|}{N} \tag{2}$$

where N is the number of samples, $real_t$ denotes the ground truth at time t, and $pred_t$ denotes the value predicted by the LSTM model at time t.

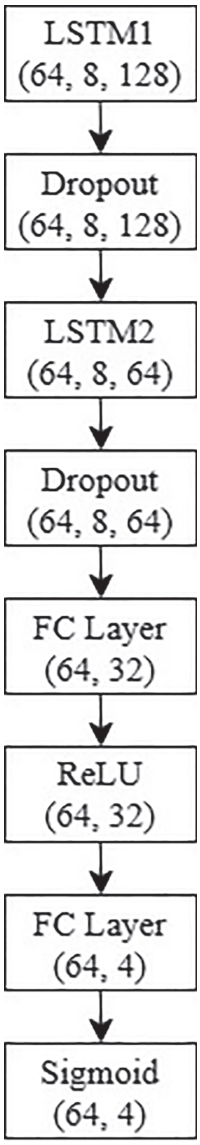


Figure 3: A multiple-LSTM model architecture.

3.4.2 Trajectory Comparison Plot

After training and testing the model, we further visualized and analyzed the LSTM prediction results by comparing the predicted trajectories with the ground-truth trajectories. During the testing phase, the trajectories predicted by the LSTM were plotted alongside the corresponding ground-truth trajectories for direct comparison.

This visualization approach helps analyze the model's prediction error. When the predicted trajectory closely matches the ground-truth trajectory, it indicates that the LSTM has effectively learned the fish's locomotion patterns. Conversely, significant deviations suggest that the model has not yet fully captured the specific movement characteristics. By examining these trajectories, we can intuitively evaluate the applicability of the LSTM model in different scenarios and identify opportunities to optimize model parameters, thereby improving both prediction accuracy and stability.

4 Experiments

This section provides a detailed description of the experimental procedures and evaluation results. In these experiments, the performance and stability of the proposed fish detection and tracking method are validated through comparisons with various benchmark approaches.

4.1 Training Data

The training data for this study were obtained from aquarium videos of schooling fish captured from a top-down viewpoint, covering real conditions with current disturbances and involving three fish species (zebrafish, brickfish, and sailfish). To preserve sufficient temporal information, the original videos were sampled frame-by-frame at 30 frames per second (FPS = 30) to generate continuous inputs for the model. All images were manually annotated, and each fish was assigned a consistent ID to ensure accurate feature learning during detection and tracking. The training dataset, illustrated in Fig. 4, comprises 1500 images.

Figure 4: Top-down aquarium training image with labeled fish instances.

The dataset was partitioned into training, validation, and test sets in an 8:1:1 ratio for model training, hyperparameter tuning, and final performance evaluation, respectively. This partitioning strategy enhances the model's generalization capability and helps ensure the representativeness and consistency of each stage, thereby improving the overall stability of the detection and tracking tasks.

4.2 Experimental Environment and Model Setup

The experimental environment for this study consisted of a well-equipped local machine used for training and testing, with detailed hardware and software specifications provided in Table 1.

Table 1: Hardware and software specifications of the experimental environment.

Category	Specification
CPU	Intel(R) Core (TM) i5-14600K
GPU	NVIDIA GeForce RTX 4060 8 GB
Operating system	Windows 11
Deep learning framework	PyTorch 2.4.1 + cu118
Programming language	Python 3.8

For the object detection task, the proposed modified YOLOv8 model, termed YOLOv8-CS, was employed as the primary training framework. The model incorporates the CBAM and SE attention mechanisms into the YOLOv8 backbone and feature output layers, thereby enhancing its ability to recognize fish regions under current disturbances. Training was conducted for 200 epochs with a batch size of 16 and a learning rate of 0.01. Model parameters were updated using the stochastic gradient descent (SGD) optimizer, and the complete training configuration is summarized in Table 2.

Table 2: Training configuration for the YOLOv8-CS model.

Parameter Item	Setting
Epochs	200
Batch Size	16
Optimizer	SGD
Image Size	$640 * 640$
Learning Rate	0.01

For the object tracking task, the designed LSTM model predicts the current position of each fish based on the preceding eight frames, and the training parameters are summarized in Table 3. Both the input and output dimensions are set to 4, representing the center coordinates and angle information. The hidden layer consists of two layers with 128 and 64 neurons, respectively, and a dropout rate of 0.1 is applied to mitigate overfitting. The model was trained for 500 epochs with a batch size of 16 and a learning rate of 0.0001, ensuring stable convergence and strong predictive performance.

Table 3: LSTM model training parameters for fish trajectory prediction.

Parameter Item	Setting
Input/Output	4/4
Hidden Size	[128, 64]
Epochs	500
Batch Size	16

(Continued)

Table 3 (continued)

Parameter Item	Setting
Optimizer	Adam
Dropout	0.1
Learning Rate	0.0001

4.3 Comparison of Training Results and Analysis

This section compares the performance of the proposed YOLOv8-CS model with the baseline YOLOv8 and presents the training results for the LSTM model.

4.3.1 YOLOv8 and YOLOv8-CS Training Results

Fig. 5 illustrates the changes in loss and evaluation metrics for the YOLOv8-CS model during training. The first column shows the metrics for the training set (box loss, classification loss, distribution focal loss, precision, and recall), while the second column presents the corresponding trends on the validation set. All loss metrics decrease steadily, particularly within the first 50 epochs, indicating that the model effectively learns fish detection features. Meanwhile, precision, recall, and mAP consistently improve before converging. Ultimately, both precision and recall approach 1.0, and mAP50 also nears 1.0, demonstrating that the model achieves high detection accuracy and strong recall.

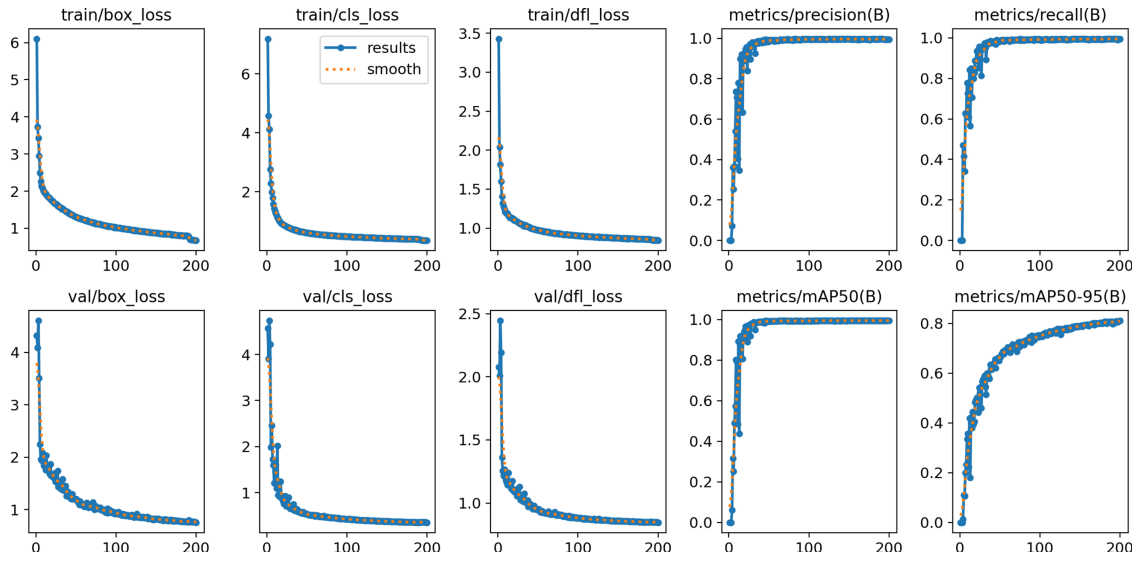


Figure 5: YOLOv8-CS training progress on loss reduction and metric convergence.

Fig. 6 presents the confusion matrix obtained after training. The confusion matrix provides insight into classification accuracy and the distribution of errors across categories. Overall, the model demonstrated stable performance in classifying all fish categories, with most classes (0 to 6) achieving over 99% accuracy and some reaching 100% correct classification. This result indicates strong feature learning and discriminative capability. Misclassifications were observed only in the background category, likely due to blurred fish-body edges in certain images or color similarity between the fish and the background. These results confirm that the

proposed framework can effectively discriminate among different fish species with high stability in multiclass recognition tasks, thereby providing a solid foundation for subsequent tracking.

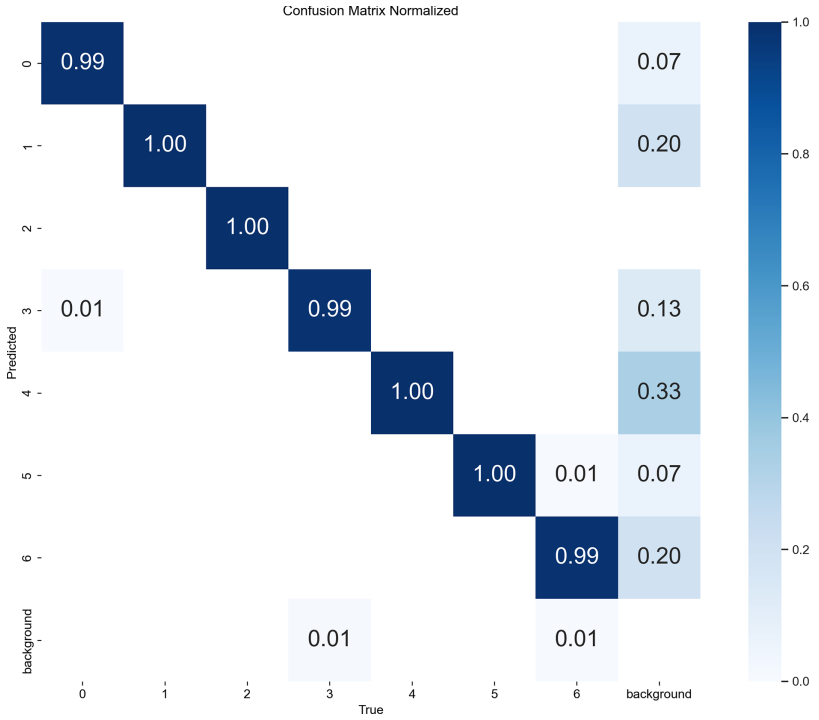


Figure 6: Confusion matrix of YOLOv8-CS showing class-wise fish classification performance.

Table 4 presents the ablation study results comparing the original YOLOv8 model with the incremental addition of attention modules. This analysis aims to verify the specific contribution of CBAM and SE modules to the detector's performance before integrating it into the tracking system.

Table 4: Comparison of YOLOv8 and YOLOv8-CS models.

Model	F1-Score	Precision	Recall	mAP50	mAP50-95
YOLOv8n(Baseline)	99.66	99.86	99.46	99.41	81.14
YOLOv8 + CBAM	99.51	99.47	99.57	99.45	81.07
YOLOv8 + SE	99.51	99.70	99.34	99.39	80.80
YOLOv8-CS	99.52	99.55	99.49	99.41	81.15

The $F1 - Score$, defined as the harmonic mean of precision and recall, is used to reflect the balance between detection accuracy and completeness. This is particularly suitable for the underwater imaging scenario in this study, where fish morphology is diverse. The formula for the $F1 - Score$ is provided in Eq. (3):

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

As shown in the table, the baseline YOLOv8n model already achieves high precision and recall. The addition of the CBAM module (YOLOv8 + CBAM) focuses on spatial and channel importance, helping the model maintain stability in feature extraction. Similarly, the independent addition of the SE

module (YOLOv8 + SE) adaptively recalibrates channel-wise feature responses. Finally, by incorporating both CBAM and SE (the proposed YOLOv8-CS), the model achieves the highest recall and a balanced mAP50-95 of 81.15. While the overall numerical metrics remain high across all versions, the YOLOv8-CS configuration demonstrates superior feature discrimination in the subsequent qualitative analysis under turbulent water flow.

To empirically support the practical significance of YOLOv8-CS, we conducted a targeted analysis of the missed-detection rate in specific high-interference video sequences (Video 1 and Video 2). As shown in Fig. 7, the missed-detection rates for both models are maintained at a low level (below 4%). In Video 1, which is characterized by significant water current disturbances, YOLOv8-CS reduced the missed-detection rate from 3.84% to 3.49%. In Video 2, the rate dropped from 1.85% to 1.77%. While these numerical improvements appear marginal, in continuous multi-object tracking, preventing even a single-frame detection failure is critical for avoiding trajectory fragmentation and ID switches.

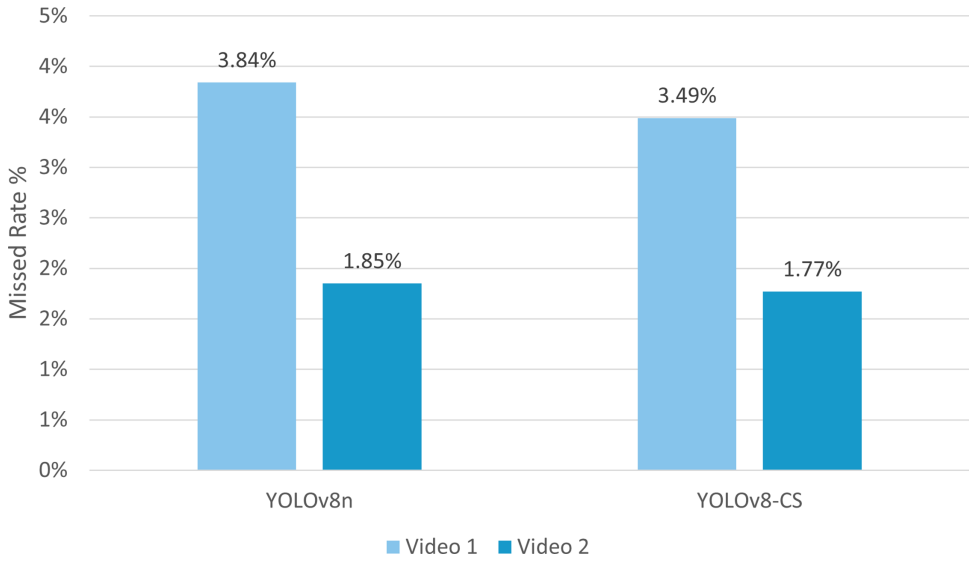


Figure 7: Comparison of missed-detection rates between the baseline YOLOv8 and the proposed YOLOv8-CS in high-interference aquatic environments (Video 1 and Video 2).

To further demonstrate the qualitative superiority of YOLOv8-CS, the original single-case comparison is expanded into a multi-case comparative visualization, as shown in Fig. 8. We carefully selected three representative "hard-case" scenarios that frequently lead to detection failure in dynamic aquatic environments: (1) Strong water ripple interference, where dynamic surface reflections distort the fish's appearance; (2) Bubble occlusion, where aeration systems create visual barriers that break the target's continuity; and (3) Blurred fish edges, where the target's color and texture blend with the complex background under rapid movement.

In these challenging scenarios, the baseline YOLOv8n model frequently suffers from missed detections or unstable localization, as indicated by the absence of bounding boxes in high-interference regions. In contrast, the YOLOv8-CS model, enhanced by the synergy of CBAM and SE attention modules, successfully suppresses background noise and recalibrates feature weights to focus on valid biological cues. This robust detection performance ensures that the target's bounding box remains continuous across frames, which is critical for the subsequent tracking phase to prevent ID switches and trajectory fragmentation caused by intermittent detection loss.

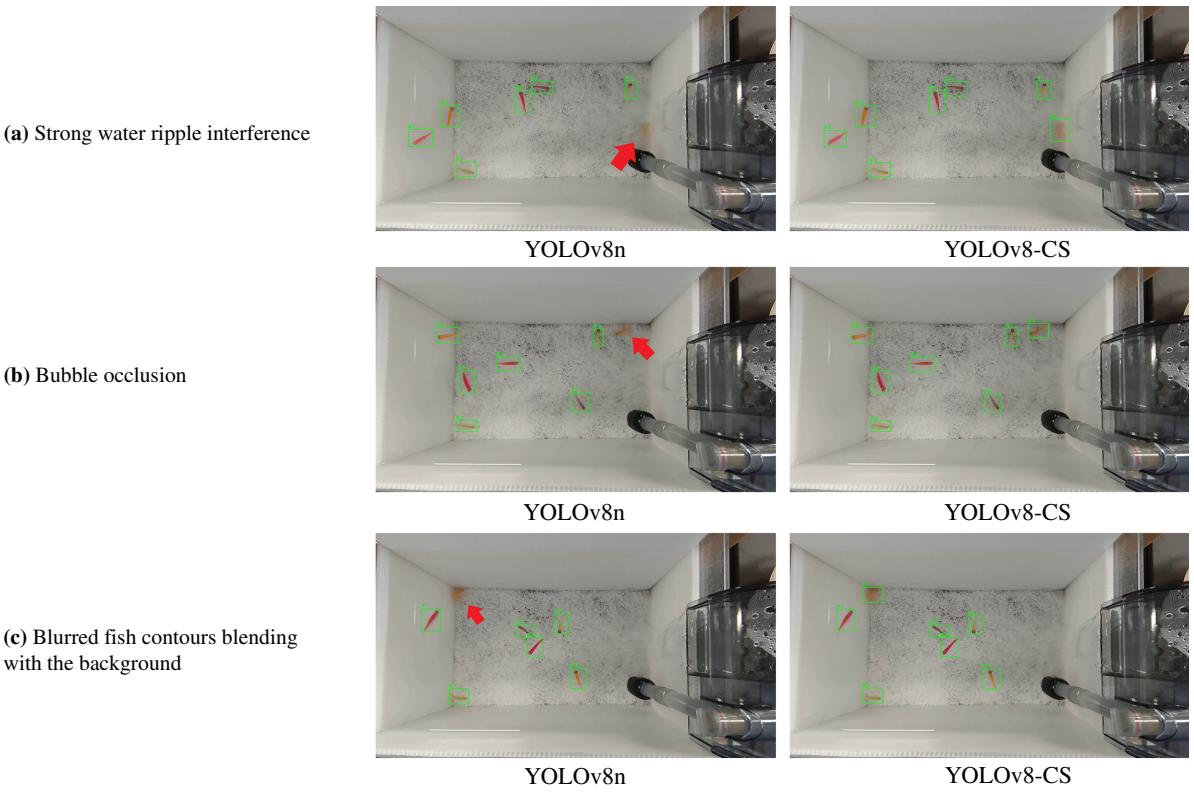


Figure 8: Qualitative detection comparison between the baseline YOLOv8n and the proposed YOLOv8-CS under three challenging scenarios: (**a**) Strong water ripple interference, (**b**) Bubble occlusion, and (**c**) Blurred fish contours blending with the background. Red arrows or dashed boxes highlight instances where the baseline model fails to detect the targets.

4.3.2 LSTM Training Results

In this study, an LSTM model was employed to predict the movement trajectory of the fish school. The model demonstrated stable performance during training and achieved strong predictive accuracy on the test set, as summarized in Table 5. With a dropout rate of 0.1, the test loss was 0.28, the center error was 0.00, and the angle error was 0.61. The overall prediction accuracy reached 1.00, while the regression errors were 0.31 (MAE) and 0.28 (MSE), indicating that the trained model can accurately capture both the position and orientation of the fish school's center point.

Table 5: LSTM test performance under different dropout rates.

Dropout	Test Loss	Center Error	Angle Error	Accuracy	MAE	MSE
0.1	0.28	0.00	0.61	1.00	0.31	0.28
0.3	0.30	0.01	0.65	1.00	0.32	0.30
0.5	0.30	0.01	0.60	1.00	0.31	0.30

To further examine the impact of the dropout rate on model performance, additional experiments were conducted under different settings. The results indicate that although dropout rates of 0.3 and 0.5 also achieved an accuracy of 1.00, they produced slightly higher loss, MAE, and MSE values than a dropout rate

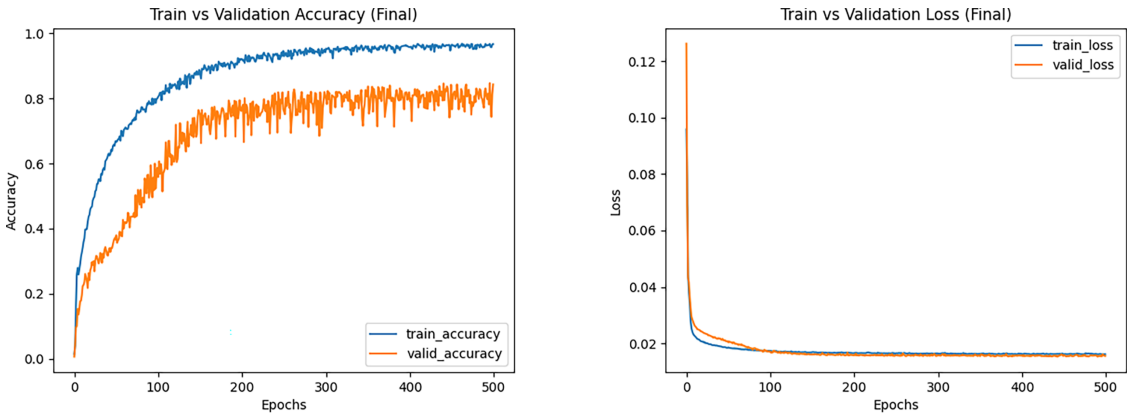
of 0.1. Based on these findings, the final model configuration in this study adopts a dropout rate of 0.1, which provides the best balance between stability and error minimization.

To provide empirical justification for the system configuration, we analyzed the impact of LSTM input frame length. We compared sequence lengths of 4, 8, and 12 frames. As shown in Table 6, while 12 frames achieve slightly lower MSE, 8 frames maintain a zero center error and provide high accuracy (1.00) with balanced regression errors, making it the most stable choice for trajectory prediction in dynamic environments.

Table 6: Performance comparison of different LSTM input frame lengths.

Input Frames	Test Loss	Center Error	Angle Error	Accuracy	MAE	MSE
4	0.30	0.00	0.63	1.00	0.32	0.30
8	0.28	0.00	0.61	1.00	0.31	0.28
12	0.27	0.01	0.59	1.00	0.29	0.27

Fig. 9 shows the training and validation accuracy and loss curves, respectively. In the accuracy plot, performance increases rapidly within the first 100 epochs and stabilizes after approximately 300 epochs, ultimately reaching about 97% on the training set and 83% on the validation set. The loss plot also shows a clear convergence trend, with only a small gap between the training and validation losses and no apparent signs of overfitting.



(a) Accuracy curves for training and validation

(b) Loss curves for training and validation

Figure 9: Accuracy (**a**) and Loss (**b**) Curves for training and validation.

As illustrated by the training and validation accuracy curves in Fig. 9a, both metrics show a steady upward trend before stabilizing. It should be noted that the ~14% gap between training accuracy (~97%) and validation accuracy (~83%) does not indicate overfitting in the conventional sense. The "accuracy" here is a threshold-based metric that counts a prediction as correct only when the predicted fish position falls within a fixed tolerance of the ground truth; this metric is inherently more sensitive to distributional variation between the training and validation sets than continuous metrics such as MAE and MSE. Critically, Fig. 9b shows that both training and validation losses converge smoothly and in parallel with only a small gap at convergence and no upward divergence in the validation loss—the characteristic signature of overfitting is absent. Consistent with this, Table 5 reports low and stable MAE and MSE values on the held-out test set. The existing dropout regularization (rate = 0.1) was selected through ablation (Table 5), which showed that higher

rates degraded regression performance, suggesting the model is appropriately regularized for the available training data. Taken together, the evidence supports the conclusion that the model has learned generalizable trajectory features rather than memorized training patterns.

Fig. 10 compares the predicted and actual trajectories produced by the LSTM model across different test samples. In the figure, the blue line represents the actual trajectory, and the orange line represents the predicted trajectory. The two lines overlap closely in both overall trajectory trends and coordinate positions, demonstrating that the model has strong temporal feature-learning capability and can effectively reproduce the dynamic behavior of fish schools. Overall, these results indicate that the LSTM model developed in this study performs well in predicting fish movement trajectories, exhibiting not only high accuracy but also robust learning and generalization capabilities, thereby providing a solid foundation for subsequent multi-object tracking.

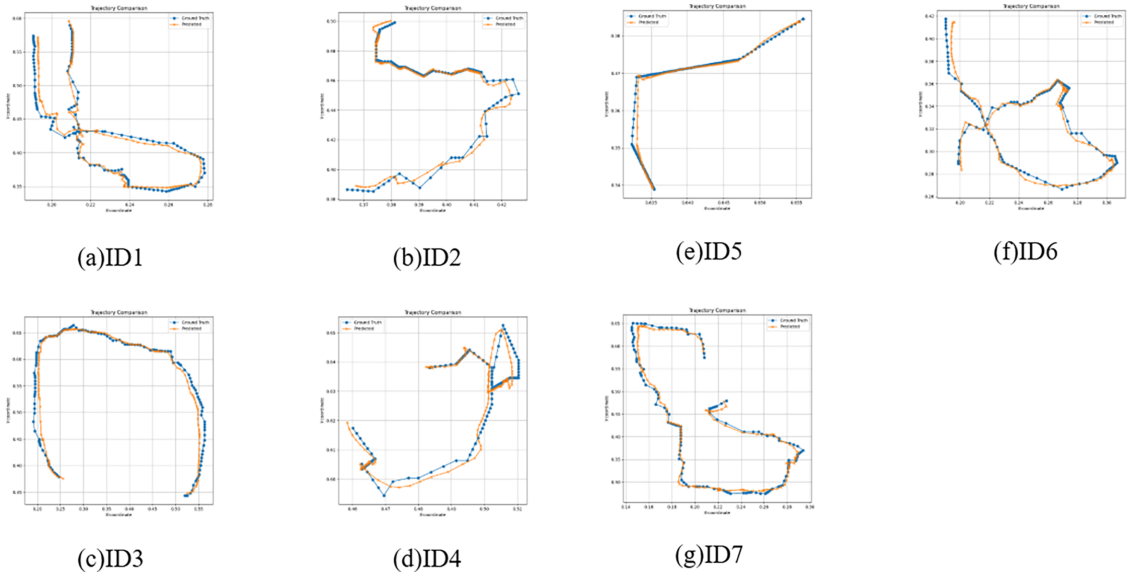


Figure 10: Comparison of predicted and ground-truth trajectories for seven individual fish: (**a**) ID1, (**b**) ID2, (**c**) ID3, (**d**) ID4, (**e**) ID5, (**f**) ID6, and (**g**) ID7. The orange lines represent LSTM predictions, while blue lines represent actual ground-truth paths.

4.4 Experimental Results

This section presents a comprehensive analysis of tracking performance, including evaluation metrics, stability assessment, and behavioral discussion.

4.4.1 Tracking Assessment Indicators

In this study, Multi-Object Tracking (MOT) performance was evaluated using the standardized indicator system proposed by MOTChallenge, enabling a comprehensive comparison of different tracking methods in the context of fish tracking. The evaluation included MOTA, ID switches, IDP, IDR, precision, recall, IDF1, and MOTP, and the metrics were computed using an open-source tool to ensure accuracy and comparability.

MOTA (Multiple Object Tracking Accuracy) is the most widely used overall performance metric, and its formula is shown in Eq. (4). It integrates false positives, missed detections, and ID switch errors, thereby providing a comprehensive measure of tracking system stability. Precision and recall, whose formulas are given in Eqs. (5) and (6), are fundamental indicators of detection quality. Precision reflects the proportion of

correctly detected objects among all positive predictions, primarily measuring false alarms, whereas recall represents the proportion of real objects correctly detected by the model, mainly reflecting missed detections.

IDP (ID precision) and IDR (ID recall), defined in Eqs. (7) and (8), measure identification accuracy. IDP denotes the proportion of correctly matched IDs among all assigned IDs, while IDR denotes the proportion of ground-truth IDs that are successfully matched by the tracker. IDF1 (identification F1), shown in Eq. (9), is the harmonic mean of IDP and IDR.

To measure the localization accuracy of the predicted bounding boxes, we adopt the MOTP (Multiple Object Tracking Precision) metric to compute the average IoU error over successfully matched pairs, as shown in Eq. (10), where $IoU_{t,i}$ is the IoU of the ith matched pair in frame t, and c_t is the number of successful matches in that frame. A smaller MOTP value indicates a higher overlap between the predicted and ground-truth boxes. Finally, this study also counts the number of identity switches (IDs) occurring during the tracking process. This value represents the number of times an object is assigned an incorrect identity during tracking; a lower value indicates more stable ID assignment.

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

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

$$IDP = \frac{ID_{TP}}{ID_{TP} + ID_{FP}} \tag{7}$$

$$IDR = \frac{ID_{TP}}{ID_{TP} + ID_{FN}} \tag{8}$$

$$IDF1 = \frac{2}{\frac{1}{IDP} + \frac{1}{IDR}} = \frac{2ID_{TP}}{2ID_{TP} + ID_{FP} + ID_{FN}} \tag{9}$$

$$MOTP = \frac{\sum_{t,i} (1 - IoU_{t,i})}{\sum_t c_t} \tag{10}$$

Through a comprehensive comparison of evaluation metrics, this study analyzes the performance of multiple object tracking methods, using the modified YOLOv8-CS as a common object detection model to assess algorithm performance in an actual fish-tracking task. The fish videos used in this study were self-recorded, and two 30-s videos (1800 frames in total) were selected for comparison, considering the cost of manual annotation and the need for experimental controllability. Although the video length was limited, the videos included interference factors such as fish intermingling, occlusion, rapid swimming, and water-current disturbances, which were sufficient to validate the tracking stability of the proposed method in a highly disturbed environment.

Before comparing our approach with benchmark algorithms, we conducted a sensitivity analysis on the IoU matching threshold to determine the optimal criterion for trajectory association. As shown in Table 7, a threshold of 0.3 yielded the highest MOTA (77.23%) and IDF1 (47.18%), while minimizing identity switches (42.50). This optimal threshold effectively balances precision and recall by filtering out weak spatial associations caused by water ripples, and was thus selected for the final system evaluation.

Table 7: Sensitivity analysis of the IoU matching threshold.

IoU	MOTA↑	IDs↓	IDP↑	IDR↑	Precision↑	Recall↑	IDF1↑	MOTP↓
0.1	75.54 ± 4.30	43.00 ± 8.49	45.28 ± 12.93	44.17 ± 10.97	88.84 ± 0.76	87.17 ± 4.12	44.71 ± 11.93	0.36 ± 0.02
0.2	77.08 ± 2.08	46.00 ± 7.07	44.46 ± 11.73	44.61 ± 11.56	88.75 ± 0.91	89.11 ± 1.35	44.53 ± 11.65	0.36 ± 0.02
0.3	77.23 ± 1.92	42.50 ± 12.02	47.10 ± 15.47	47.25 ± 15.29	88.80 ± 0.86	89.15 ± 1.32	47.18 ± 15.38	0.36 ± 0.02

To evaluate the robustness of the proposed framework, we compared our method with classic trackers (SORT, DeepSORT, and ByteTrack) and the recent state-of-the-art (SOTA) algorithm (BoostTrack and BoT-SORT). Furthermore, we conducted a control experiment by replacing the LSTM module with a traditional Kalman Filter to evaluate the impact of different position estimation mechanisms. All tracking results are summarized in Table 8, presenting the mean performance with standard deviations across test sequences.

Table 8: Comparison of results for the combined YOLOv8-CS tracking algorithm (note: ↑ indicates that higher values are better, while ↓ indicates that lower values are better).

Method	MOTA↑	IDs↓	IDP↑	IDR↑	Precision↑	Recall↑	IDF1↑	MOTP↓
SORT	72.54% ± 3.92%	139.50 ± 4.95	25.99% ± 2.36%	25.04% ± 1.96%	88.80% ± 2.49%	85.59% ± 1.35%	25.51% ± 2.15%	0.38 ± 0.02
DeepSORT	66.37% ± 8.34%	186.50 ± 6.36	22.75% ± 4.69%	24.01% ± 4.30%	82.83% ± 4.82%	87.61% ± 2.68%	23.36% ± 4.51%	0.38 ± 0.02
ByteTrack	72.12% ± 4.53%	62.50 ± 10.61	36.29% ± 1.34%	35.85% ± 1.06%	87.02% ± 2.67%	85.96% ± 1.98%	36.07% ± 1.20%	0.38 ± 0.02
BoostTrack	71.05% ± 8.53%	130.00 ± 53.74	41.10% ± 1.25%	44.69% ± 0.30%	83.69% ± 5.13%	90.96% ± 3.43%	42.82% ± 0.82%	0.37 ± 0.02
BoT-SORT	74.89% ± 6.31%	82.00 ± 5.66	34.24% ± 3.04%	34.79% ± 2.41%	87.51% ± 3.89%	88.95% ± 2.22%	34.52% ± 2.73%	0.38 ± 0.02
Kalman Filter	78.25% ± 5.47%	82.00 ± 66.47	44.82% ± 8.28%	44.73% ± 7.91%	89.82% ± 2.97%	89.73% ± 3.68%	44.77% ± 8.09%	0.38 ± 0.02
Ours (YOLOv8-CS + LSTM)	77.23% ± 1.92%	42.50 ± 12.02	47.10% ± 15.47%	47.25% ± 15.29%	88.80% ± 0.86%	89.15% ± 1.32%	47.18% ± 15.38%	0.36 ± 0.02

The experimental results demonstrate that the proposed method (YOLOv8-CS + LSTM) achieves the best balance between localization accuracy and identity consistency. Notably, while the Kalman Filter variant achieved a slightly higher MOTA (78.25%), it suffered from nearly double the identity switches (82.00 vs. 42.50) and lower IDF1 (44.77% vs. 47.18%) compared to our LSTM-based approach. This highlights the limitation of Kalman filtering, which relies on linear motion assumptions that frequently fail during the sudden direction changes and "burst-and-coast" swimming patterns typical of fish. In contrast, the LSTM's ability to capture long-term temporal dependencies ensures significantly more stable identity assignment in dynamic aquatic environments.

Compared with the recent BoostTrack, our system demonstrates substantial improvements across most identification metrics, particularly by reducing the number of ID switches by 87.5 (from 130 to 42.5). These results confirm that the synergy between the attention-enhanced YOLOv8-CS detector and LSTM trajectory prediction provides a highly reliable solution for long-term fish behavior monitoring, outperforming both traditional filters and modern SOTA trackers in maintaining identity continuity.

To demonstrate the quantitative benefit of LSTM-based trajectory prediction more clearly over detection-based association alone, Table 8 presents a three-level ablation analysis. First, in the Detection + IoU matching only (SORT) setting, tracking relies exclusively on frame-to-frame IoU overlap without any motion prediction, which serves as the closest approximation to detection-only association. Under this setting, the MOTA is 72.54% and the IDF1 is only 25.51%, with 139.50 ID switches per sequence. Second, in the Detection + Kalman filter prediction (linear motion model) setting, the incorporation of Kalman filter-based trajectory prediction improves the MOTA to 78.25% and the IDF1 to 44.77%, while reducing the number of ID switches to 82.00. Nevertheless, the relatively large standard deviation (±66.47) suggests unstable performance across sequences, which may be attributed to the mismatch between linear motion

assumptions and the highly erratic swimming behavior of ornamental fish. Third, in the Detection + LSTM prediction (our full system) setting, replacing the Kalman filter with an LSTM further reduces the number of ID switches to 42.50, corresponding to a 48.2% reduction relative to the Kalman filter variant and a 69.5% reduction relative to SORT, while the IDF1 further increases to 47.18%. Notably, the standard deviation of ID switches decreases substantially from ±66.47 to ±12.02, indicating that the LSTM not only improves tracking accuracy but also enhances robustness and consistency across diverse motion patterns. Overall, this three-tier comparison among IoU-only matching (SORT), linear motion prediction (Kalman filter), and learned trajectory prediction (LSTM) effectively isolates the contribution of each module and confirms that LSTM-based trajectory prediction provides a substantial improvement in identity continuity beyond what can be achieved through detection-based association or linear motion modeling alone.

As shown in Table 8, a direct before-and-after comparison is already available through the ablation setting in which the LSTM module is replaced with a traditional Kalman filter while the YOLOv8-CS detector is kept unchanged. This setting effectively isolates the contribution of the LSTM to identity consistency. The results indicate that the Kalman filter variant yields 82.00 ± 66.47 identity switches per sequence, whereas the proposed LSTM-based method reduces this value to 42.50 ± 12.02, corresponding to a 48.2% reduction attributable specifically to the LSTM trajectory prediction module. In addition, compared with the baseline SORT tracker (139.50 ± 4.95 IDs), the reduction reaches 69.6%, while a 67.3% reduction is observed relative to BoostTrack (130.00 ± 53.74).

Moreover, the Kalman filter variant exhibits a much larger standard deviation (±66.47) than the LSTM-based method (±12.02), suggesting that the proposed approach not only lowers the average number of ID switches but also improves robustness and consistency across different test sequences.

Regarding trajectory fragmentation, we note that this effect is partially captured by the IDF1 metric, which evaluates the proportion of correctly identified detections over the complete lifetime of a trajectory. When a trajectory is fragmented, that is, interrupted and subsequently reassigned a new identity, the IDF1 score is directly penalized due to the false identity reassignment at the restart point. In this regard, the proposed method achieves the highest IDF1 score (47.18%), outperforming both the Kalman filter variant (44.77%) and SORT (25.51%). This provides indirect yet principled evidence that trajectory fragmentation is reduced.

4.4.2 Behavioral Observation

Photographs taken from an overhead viewing angle provide a comprehensive representation of fish planar movement in the aquarium, which is valuable for assessing the potential effects of environmental variables on behavior. In this study, to investigate whether the installation of environmental devices, such as the filtration system, restricted the fish's movement range, the track coordinates stored in the tracking system were visualized and plotted as two-dimensional trajectory maps for analysis and observation. As shown in Fig. 11, despite the addition of a new filtered water-flow system on the right side of the aquarium, the fish continued to pass through the area naturally, and the trajectory distribution did not exhibit any clear concentration or avoidance patterns. These results indicate that the introduction of water flow did not significantly affect the fish's activity space or restrict their natural behavior under the experimental conditions.

The controlled aquarium environment with a fixed overhead camera used in this study represents an idealized setting, and deployment in realistic aquaculture scenarios introduces additional challenges that warrant explicit discussion. We consider three key dimensions of variability in more complex environments: lighting, turbidity, and viewpoint.

Figure 11: Visualization of fish movement paths in the aquarium (overhead view).

On lighting variability: The current YOLOv8-CS detector was trained under stable indoor illumination. In outdoor or semi-outdoor aquaculture ponds, illumination can vary substantially due to solar angle, cloud cover, and surface reflections. Although data augmentation strategies, such as random brightness and contrast perturbations, were applied during training to provide partial robustness, a more systematic solution would involve domain adaptation techniques or retraining on datasets collected under diverse lighting conditions. We have identified this as a concrete direction for future work.

On water turbidity: In recirculating aquaculture systems (RAS) and outdoor ponds, elevated turbidity, suspended particles, and algal blooms can significantly reduce the visual contrast between fish and the background. Our current framework relies on RGB imagery, which is directly affected by such conditions. Infrared imaging or multispectral sensing could help mitigate this limitation, and we discuss these modalities as hardware-level extensions that would complement the software framework proposed here.

On viewpoint variation: The fixed top-down perspective used in this study eliminates perspective distortion and occlusion caused by tank walls, thereby simplifying both detection and trajectory analysis. In contrast, real-world aquaculture facilities often employ side-view or oblique-angle cameras, and in net-pen or pond environments, fish occupy a three-dimensional volume rather than a near-planar layer. Extending the framework to address perspective distortion, multi-camera fusion, or three-dimensional reconstruction from stereo image pairs represents a meaningful, though nontrivial, engineering challenge that we identify as a priority for future investigation.

We appreciate that these are genuine limitations of the current study. The primary contribution of this work lies in demonstrating that attention-enhanced detection combined with LSTM-based trajectory prediction yields measurable improvements in tracking stability under challenging conditions involving high-density, visually similar fish in an unstructured environment. Validating this architecture under the broader conditions described above constitutes an important direction for future work.

5 Conclusion

We propose a multi-object tracking system for aquarium fish that improves tracking stability and accuracy in challenging scenarios such as fast swimming, occlusion, and water-current interference. For object detection, we incorporated attention mechanisms (CBAM and SE) to mitigate motion-induced

boundary distortion; for tracking, we combined IoU matching with an LSTM model for temporal prediction. The integration significantly reduces identity switches and ensures trajectory continuity in dynamic aquatic environments.

Despite the successful validation of the framework, we acknowledge certain limitations in the current experimental setup. The dataset used in this study (1500 images and two 30-s videos) was primarily restricted by the prohibitively high cost of manual frame-by-frame annotation for dense fish schools. Consequently, this research serves as a preliminary validation of the framework's performance within a specific controlled environment characterized by a stable top-down view and constant illumination.

Beyond its technical contributions, the proposed framework also has important practical implications for aquaculture monitoring. Maintaining stable individual trajectories over time is a fundamental prerequisite for automated behavioral biomarker extraction. Abnormal locomotor patterns, such as reduced swimming velocity, irregular trajectory curvature, and increased inter-individual avoidance distance, are widely recognized as early indicators of physiological stress, oxygen deficiency, and infectious disease in fish populations. Current monitoring practices in aquaculture facilities still rely largely on periodic manual observation, which is labor-intensive and often delays detection. By reliably preserving individual identities across occlusion events and abrupt directional changes, as reflected by the low ID-switch rate and high IDF1 score achieved in this study, the proposed tracking system can generate the continuous trajectory data required for automated health surveillance pipelines. This is especially relevant in recirculating aquaculture systems (RAS), where high stocking densities and controlled environments make continuous automated monitoring both feasible and economically necessary.

From a behavioral analysis perspective, the LSTM module's ability to learn species-specific and environment-specific movement patterns provides a promising basis for individual behavioral profiling beyond simple counting or density estimation. Trained on historical trajectory sequences, the LSTM can implicitly capture characteristic locomotor tendencies, including typical turning radii, preferred swimming zones, and inter-individual spacing. Deviations from these learned patterns at the individual level may serve as data-driven anomaly signals, enabling early-warning systems without the need for predefined rule-based thresholds. In addition, the overhead-view trajectory data generated by the proposed system are directly compatible with established ethological metrics for analyzing schooling behavior, such as nearest-neighbor distance, polarization order parameters, and spatial occupancy entropy. This broadens the framework's utility for behavioral ecology research, particularly in studies of group cohesion and social hierarchy in ornamental fish species. Future work may extend the framework to multi-camera configurations and three-dimensional trajectory reconstruction, further enhancing the behavioral resolution available to both aquaculture operators and researchers.

While the proposed framework demonstrates strong performance in a controlled aquarium environment with a fixed overhead camera, its generalizability to real-world aquaculture settings remains constrained by the experimental setup. Three major sources of variability deserve explicit discussion. First, outdoor and semi-outdoor aquaculture ponds are exposed to substantial lighting fluctuations caused by solar angle, cloud cover, and surface specular reflections. Although random brightness and contrast augmentation were incorporated during training to improve partial robustness, systematic domain adaptation or retraining on multi-condition datasets would still be required for deployment in uncontrolled lighting environments. Second, elevated suspended particulate matter and algal blooms in recirculating aquaculture systems (RAS) and earthen ponds can reduce visual contrast between fish and the background, thereby lowering detection confidence in RGB-based systems. This limitation may be alleviated through complementary sensing modalities, such as near-infrared or multispectral imaging, which are compatible with the proposed software architecture. Third, the fixed top-down perspective adopted in this study eliminates perspective distortion

and lateral occlusion, conditions that are not preserved in side-view or oblique-angle installations commonly used in commercial aquaculture facilities. Future extensions incorporating perspective-aware detection, multi-camera fusion, or stereo-based three-dimensional trajectory reconstruction would substantially broaden the framework's applicability to net-pen, pond, and raceway environments.

Future work will focus on enhancing the model's robustness and generalizability. We plan to incorporate larger-scale public datasets to further verify the method's scalability. Moreover, we aim to test the system under more complex scenarios commonly encountered in intensive aquaculture, such as low light, strong surface reflections, dense occlusions, and turbid water conditions. Integrating multiple viewpoints (e.g., side-view imaging) and advanced Re-ID modules will also be explored to compensate for information loss in highly cluttered environments. By addressing these challenges, the proposed system could eventually provide a comprehensive tool for large-scale fish health monitoring and social behavioral studies in industrial aquaculture applications. In addition, to facilitate subsequent validation and extension by other researchers, we are willing to provide the training code used in this study.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the National Science and Technology Council (NSTC), Taiwan (NSTC 114-2221-E-035-076- and 114-2420-H-006-004-).

Author Contributions: The first draft was written by Feng-Cheng Lin and Bo-Chiao Jan. The authors confirm contribution to the paper as follows: Conceptualization, methodology, writing—original draft preparation, Feng-Cheng Lin and Bo-Chiao Jan; software, investigation, Bo-Chiao Jan and Feng-Cheng Lin; data curation, validation, Feng-Cheng Lin and Hui-An Wu; funding acquisition, supervision, Feng-Cheng Lin. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: 1. Data availability: The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request. 2. Materials availability: Not applicable. 3. Code availability: The source code for this study is available from the corresponding author on reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Jocher G, Chaurasia A, Qiu J. YOLO by ultralytics [Internet]. 2023 [cited 2026 Mar 30]. Available from: https://github.com/ultralytics/ultralytics.
2. Xu X, Hu J, Yang J, Ran Y, Tan Z. A fish detection and tracking method based on improved interframe difference and YOLO-CTS. IEEE Trans Instrum Meas. 2024;73:2532913. doi:10.1109/TIM.2024.3476529.
3. Qin X, Yu C, Liu B, Zhang Z. YOLO8-FASG: a high-accuracy fish identification method for underwater robotic system. IEEE Access. 2024;12:73354–62. doi:10.1109/ACCESS.2024.3404867.
4. Jiang H, Zhong J, Ma F, Wang C, Yi R. Utilizing an enhanced YOLOv8 model for fishery detection. Fishes. 2025;10(2):81. doi:10.3390/fishes10020081.
5. Shah C, Nabi MM, Alaba SY, Ebu IA, Prior J, Campbell MD, et al. YOLOv8-TF: transformer-enhanced YOLOv8 for underwater fish species recognition with class imbalance handling. Sensors. 2025;25(6):1846. doi:10.3390/s25061846.
6. Zhu W, Xu R. Research on an improved YOLOv8 algorithm for water surface object detection. Electronics. 2025;14(18):3615. doi:10.3390/electronics14183615.

7. Bewley A, Ge Z, Ott L, Ramos F, Upcroft B. Simple online and realtime tracking. In: Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP); 2016 Sep 25–28; Phoenix, AZ, USA. doi:10.1109/ICIP.2016.7533003.
8. Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric. In: Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP); 2017 Sep 17–20; Beijing, China. doi:10.1109/ICIP.2017.8296962.
9. de Oliveira Barreiros M, de Oliveira Dantas D, de Oliveira Silva LC, Ribeiro S, Barros AK. Zebrafish tracking using YOLOv2 and Kalman filter. Sci Rep. 2021;11(1):3219. doi:10.1038/s41598-021-81997-9.
10. Zhang A, Palaoag TD. A real-time tracking algorithm for underwater fish based on track by detection framework. In: Proceedings of the 2024 7th International Conference on Communication Engineering and Technology (ICCET); 2024 Feb 22–24; Tokyo, Japan. doi:10.1109/ICCET62255.2024.00019.
11. Liu Y, Chen Z, Hu H, Wan S, Gao X. An identity based method for tracking fish. In: Proceedings of the 2023 16th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI); 2023 Oct 28–30; Taizhou, China. doi:10.1109/CISP-BMEI60920.2023.10373346.
12. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80. doi:10.1162/neco.1997.9.8.1735.
13. Wang C, Wu Z, Chen Y, Zhang W, Ke W, Xiong Z. Improving 3-D zebrafish tracking with multiview data fusion and global association. IEEE Sens J. 2023;23(15):17245–59. doi:10.1109/JSEN.2023.3288729.
14. Palconit MGB, Almero VJD, Rosales MA, Sybingco E, Bandala AA, Vicerra RRP, et al. Towards tracking: investigation of genetic algorithm and LSTM as fish trajectory predictors in turbid water. In: Proceedings of the 2020 IEEE Region 10 Conference (TENCON); 2020 Nov 16–19; Osaka, Japan. p. 744–9. doi:10.1109/tencon50793.2020.9293730.
15. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Proceedings of the Computer Vision–ECCV 2018; 2018 Sep 8–14; Munich, Germany. doi:10.1007/978-3-030-01234-2_1.
16. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell. 2020;42(8):2011–23. doi:10.1109/TPAMI.2019.2913372.