frontiers in Marine Science

TYPE Original Research
PUBLISHED 26 May 2026
DOI 10.3389/fmars.2026.1824491

OPEN ACCESS

EDITED BY
Ioannis A Giantsis,
Aristotle University of Thessaloniki, Greece

REVIEWED BY
Edén V. Gaitán Morado,
Michoacán University of San Nicolás de Hidalgo, Mexico
Abhishek Bajpai,
Government Engineering College, India

*CORRESPONDENCE
Sasithradevi A.
sasithradevi.a@vit.ac.in

RECEIVED 06 March 2026
REVISED 05 May 2026
ACCEPTED 06 May 2026
PUBLISHED 26 May 2026

CITATION
M. V, A. S, Nathan S and Prakash P (2026) MorphoCal: a multi-stage deep learning framework for fish length estimation in challenging underwater pond environments.
Front. Mar. Sci. 13:1824491.
doi: 10.3389/fmars.2026.1824491

MorphoCal: a multi-stage deep learning framework for fish length estimation in challenging underwater pond environments

Vijayalakshmi M.[1], Sasithradevi A.[2]*, Sabari Nathan[3] and P. Prakash[4]

[1]School of Electronics Engineering, Vellore Institute of Technology, Chennai, India, [2]Center for Advanced Data Science, Vellore Institute of Technology, Chennai, India, [3]Deloitte Tohmatsu Deep Square Co., Ltd., Tokyo, Japan, [4]Department of Electronics Engineering, Madras Institute of Technology, Chennai, India

Introduction: Underwater imaging plays an important role in monitoring aquatic ecosystems and aquaculture environments. Accurate estimation of fish pose and body length from underwater imagery is essential for analysing fish behaviour, growth patterns, and biomass dynamics. However, underwater scenes are often affected by turbidity, uneven illumination, occlusion, and suspended particles, which reduce the accuracy of keypoint detection and metric measurements.
Methods: To address these challenges, this study proposes *MorphoCal*, a multi-stage deep learning framework for fish pose estimation and geometric length reconstruction under real-world pond conditions. The framework integrates *AquaYOLO-PoseC A*, a coordinateattention–enhanced YOLO-based keypoint detection network, with a single-shot checkerboard calibration and ray–plane projection module for centimetre-level metric reconstruction. Coordinate Attention is incorporated into the YOLO backbone to encode direction-aware spatial features and preserve positional information, improving anatomical keypoint localization for elongated fish structures. The architecture performs joint fish detection and keypoint estimation in a single-stage forward pass using PAN–FPN feature fusion.
Results: Experiments on the *DePondFi'24* dataset, consisting of multi-species fish captured under natural pond conditions, show that AquaYOLO-PoseCA achieves a bounding box mAP of 0.959, pose mAP of 0.848, and mAP_{50-95} of 0.712, while maintaining computational efficiency with 29.4 GFLOPs and 11.4M parameters. The reconstructed fish lengths show low deviation from ruler-based ground truth measurements.
Discussion: The proposed MorphoCal framework enables reliable fish pose estimation and centimetre-level length reconstruction under challenging underwater conditions, supporting noninvasive fish monitoring, growth assessment, and biomass estimation in intelligent aquaculture systems.

KEYWORDS
aquaculture, coordinate attention mechanism, deep learning, fish length estimation, keypoint detection, underwater object detection, YOLO(you look only once)

1 Introduction

Fish length is one of the first measurements taken in fisheries and aquaculture for the growth and health monitoring for production. Farmers and biologists often use it to assess growth rates, fish stock health evaluation, and estimate biomass at different stages. In routine aquaculture operations, understanding the size distribution helps determine feeding schedules, identify slow-growing groups, and plan harvest cycles. Although manual measurement is simple, it becomes impractical when fish large populations are involved (Zhao et al., 2024). The repeated handling can stress the fish, affecting both behaviour and survival. Over the past decade, there has been a gradual shift from manual or sonar-based measurements toward image- and video-based techniques. The intelligence based growth analysis has become the major trending domain in aquaculture based industry. In the initial stages of computer vision techniques (Al-Abri et al., 2025), segmentation and edge detection were used. This was found to perform fairly well in controlled environments but did not perform well in changing environments. Changes in the amount of sunlight, turbidity of water, reflections, etc., affect the performance of the algorithm. This is highly sensitive to changes in the environment. As research progressed, artificial intelligence techniques like machine learning, and later deep learning techniques (Li and Du, 2022) were used. While analyzing the image annotation, it is found that the image annotation tools like ImageJ (Andrialovanirina et al., 2020) and sonar systems provided partial automation. However, human intervention is still required. This is an area that needs to be focused on. Later, the introduction of CNN-based detectors like Faster R-CNN (Rocha et al., 2020) and YOLO (Abinaya et al., 2022) changed the way image annotation is performed. These detectors can identify the fish in the frame and estimate the length of the fish (Cao et al., 2024) using bounding boxes. However, the estimation of the length of the fish using bounding boxes is affected by the curvature of the fish, occlusion of the fish, and the angle of the fish. Moreover, the estimation of the length of the fish using bounding boxes is in terms of pixels. Due to the above-mentioned issues in the existing methods, researchers have attempted segmentation-based methods. In segmentation-based methods, the full outline of the fish is obtained using Mask R-CNN. The length of the fish is estimated using skeletonisation. Some of the YOLOv4-derived methods have divided the fish into the head, mid-body, and tail sections. However, the above-mentioned methods are highly affected by the silhouette of the fish. It is very difficult to obtain the silhouette of the fish in pond water. More recently, there has been an emphasis on the detection of keypoints or landmarks. Instead of drawing the outline of the fish, these methods locate various anatomical features such as the snout, eye, caudal peduncle, and the tip of the tail. The former performs the task of body curvature more accurately. To detect multiple landmarks, for example, the hybrid YOLOv7-HRNet (Peng et al., 2025) architecture has been applied. It uses the combination of multi-scale semantic features from the Feature Pyramid Network and high-resolution spatial detail from the HRNet. The Lite-HRNet (Li et al., 2024) variant is computationally cheaper yet comparable in accuracy. Yet again, the HRNet-based approaches are computationally expensive.

Stereo and three-dimensional estimation approaches have also been proposed. The YOLOv7-Pose stereo configuration (Feng et al., 2025) proves the effectiveness of improving the accuracy of multi-fish detection in complex environments using two-camera configurations. Other approaches utilize the attention mechanism in YOLO-based approaches for focusing on specific features in complex environments. The problem associated with stereo-based approaches is the increased hardware complexity, thereby increasing the overall cost of deployment. In the past three years, researchers have proposed various approaches that utilize the attention mechanism of the Transformer approach in YOLO-based approaches for improving the accuracy of the approach. Hybrid Attention Transformers coupled with YOLOv8 have been proposed for improving the accuracy of detecting the ripeness of fruits in complex orchard environments (Tang et al., 2025). Other works have proposed the application of YOLOv8 in lightweight real-time inference on embedded devices by incorporating attention-informed modules in terms of balancing detection accuracy and efficiency (Chen et al., 2025).

ATBHC-YOLO, as an example, has been proposed by combining Transformer blocks with bidirectional hybrid convolutional modules in terms of improving small object detection in sparse and cluttered spatial environments (Liao et al., 2025). Moreover, Transformer-YOLO hybrid architecture has been explored in terms of underwater object detection, showing robustness against various environmental noises (Neelamegam et al., 2025). Despite these advances, many attention-heavy architectures introduce additional parameters and latency, challenging real-time deployment in practical pond environments. Table 1 presents a comparison of fish detection and morphometric measurement approaches. Recent studies have also strengthened underwater vision through improved image restoration and object detection methods. For example, Tiwari et al. (2025) proposed a diffusion- and attention-enhanced framework for underwater image restoration, showing that attentionguided enhancement can improve visual clarity and support downstream marine monitoring tasks. Similarly, Bajpai et al. (2024) demonstrated the effectiveness of YOLOv8m for underwater object detection under limited visibility and inconsistent lighting, highlighting the importance of efficient deep models for subaquatic monitoring. These studies confirm the growing emphasis on robust underwater perception under challenging imaging conditions. However, these approaches are largely confined to image restoration and object detection, with limited consideration of anatomically consistent keypoint localization (Rivera et al., 2024) and monocular metric fish length estimation in real-world pond environments.

Parallel to these methodological improvements, several fish keypoint datasets have been introduced to support landmark-based morphometric analysis, such as the FishPhenoKey dataset (Liu et al., 2024) with 22 phenotype-oriented keypoints for multiple species, a keypoint dataset for Plectropomus leopardus (Liu M. et al., 2025) under underwater conditions, annotated larval zebrafish pose datasets (Scholz et al., 2025), Lightning Pose mirror-fish dataset (Whiteway et al., 2024)with multi-view keypoints, and the Robotic-Fish-Pose-Dataset (Zhang et al., 2021) for robotic fish pose estimation. However, existing datasets are typically constrained to

TABLE 1 Comparison of fish detection and morphometric measurement approaches.

Method	Architecture/ backbone	Dataset/ context	Landmark strategy	Key feature/limitation
YOLO + bbox scaling (Abinaya et al., 2022)	Single-stage detector	Controlled fish images	No landmarks	Fast and simple, but ignores curvature and occlusion
Mask R-CNN seg (Rocha et al., 2020).	Mask-based CNN	Aquatic images	Silhouette → end-points	Handles curvature; fails in turbidity
FPN-HRNet (stereo) (Li et al., 2024)	YOLOv7 + FPN-HRNet	Stereo dataset	Explicit landmarks	Accurate 3D, needs stereo rig
Lite-HRNet reg (Peng et al., 2025).	YOLOv5 + Lite-HRNet	Underwater datasets	Landmark offset regression	Efficient; may lose precision under noise
YOLOPose (Sun et al., 2024)	YOLO-based pose head	Cultured fish	Direct keypoints	Streamlined; needs calibration for metric scale
Mono 3D pose (Mei et al., 2021)	3D template + 2D keypoints	Longline videos	3D keypoint fitting	Handles deformation; complex optimization

specific species or controlled environments, restricting generalization across heterogeneous pond conditions. Table 1 summarizes and compares existing fish keypoint detection models.

Despite these developments, there are some limitations:

- Dataset Constraints: The available datasets are mostly specific to the species or the environment and are not applicable across domains.
- Environmental sensitivity: The models are mostly effective in clear underwater conditions and may not perform well in turbid or vegetation-dense pond environments.
- Computational Overhead: Models such as HRNet and Mask R-CNN are highly accurate but are computationally expensive and not feasible in embedded systems.
- Lack of metric calibration: The models are mostly effective in measuring pixel values and not in real-world metrics such as centimeters.
- Limited integration of detection and calibration: There is limited integration of fish detection and anatomical keypoint detection with geometric calibration in the models.

To address the identified gaps, the present work proposes the MorphoCal framework for length estimation using the proposed AquaYOLO-PoseCA model, a coordinate attention-based keypoint detection model with the proposed checkerboard calibration method. The proposed system includes the integration of the multi-instance decoding method for the detection of the keypoints corresponding to each fish instance (Lauer et al., 2022) and the proposed single-shot checkerboard calibration method for the estimation of the real-world conversions. The pixel coordinates of the head and tail are then projected using the ray-plane intersection method for the estimation of the fish length in centimeters without the need for multi-view or stereo estimation. This work is believed to be the first work that includes the integration of the detection, keypoint estimation, and checkerboard calibration for the estimation of the fish length using a single image.

The contributions of the present work can be summarized as follows:

- The curation and introduction of the proposed DePondFi'24 dataset, a real-world underwater pond dataset containing multi-species fish with nine anatomical keypoints for morphometric analysis.
- The proposed AquaYOLO-PoseCA model, a Coordinate Attention-enhanced YOLOv8-based keypoint detection model for underwater fish pose estimation. The model is adapted for low-contrast, turbid, and occluded pond imagery, while remaining computationally efficient with 29.4 GFLOPs and 11.4 million parameters.
- A task-specific Coordinate Attention integration within the PAN-FPN neck for improved localization of elongated fish landmarks in turbid underwater pond imagery.
- A monocular calibration and ray–plane projection framework that converts detected anatomical keypoints into real-world centimetre-scale fish length estimates using a single RGB camera. This provides an affordable alternative to stereo or multi-camera setups for practical aquaculture monitoring.
- A comparative evaluation of the proposed model with the state-of-the-art models for the estimation of the keypoints for the fish image.
- Experimental validation of the proposed MorphoCal framework using ruler-based real-world length measurements.

The remainder of this paper is organized as follows. Section 2 presents our proposed methodology, which includes the proposed AquaYOLO-PoseCA model and dimensional calibration approach. Section 3 describes the experimental design analysis, dataset characteristics and evaluation technique. Section 4 discusses model performance, ablation study and calibration results. Section 5 demonstrates the statistical results analysis. Section 6 addresses the limitations and the further improvements. Finally, Section 7 summarizes the conclusion of our proposed work.

2 Methodology

As illustrated in Figure 1, the system comprises three main stages (Zhao et al., 2024): Our proposed AquaYOLO-PoseCA keypoint detection model (Al-Abri et al., 2025); checkerboard-based camera calibration; and (Li and Du, 2022) metric length

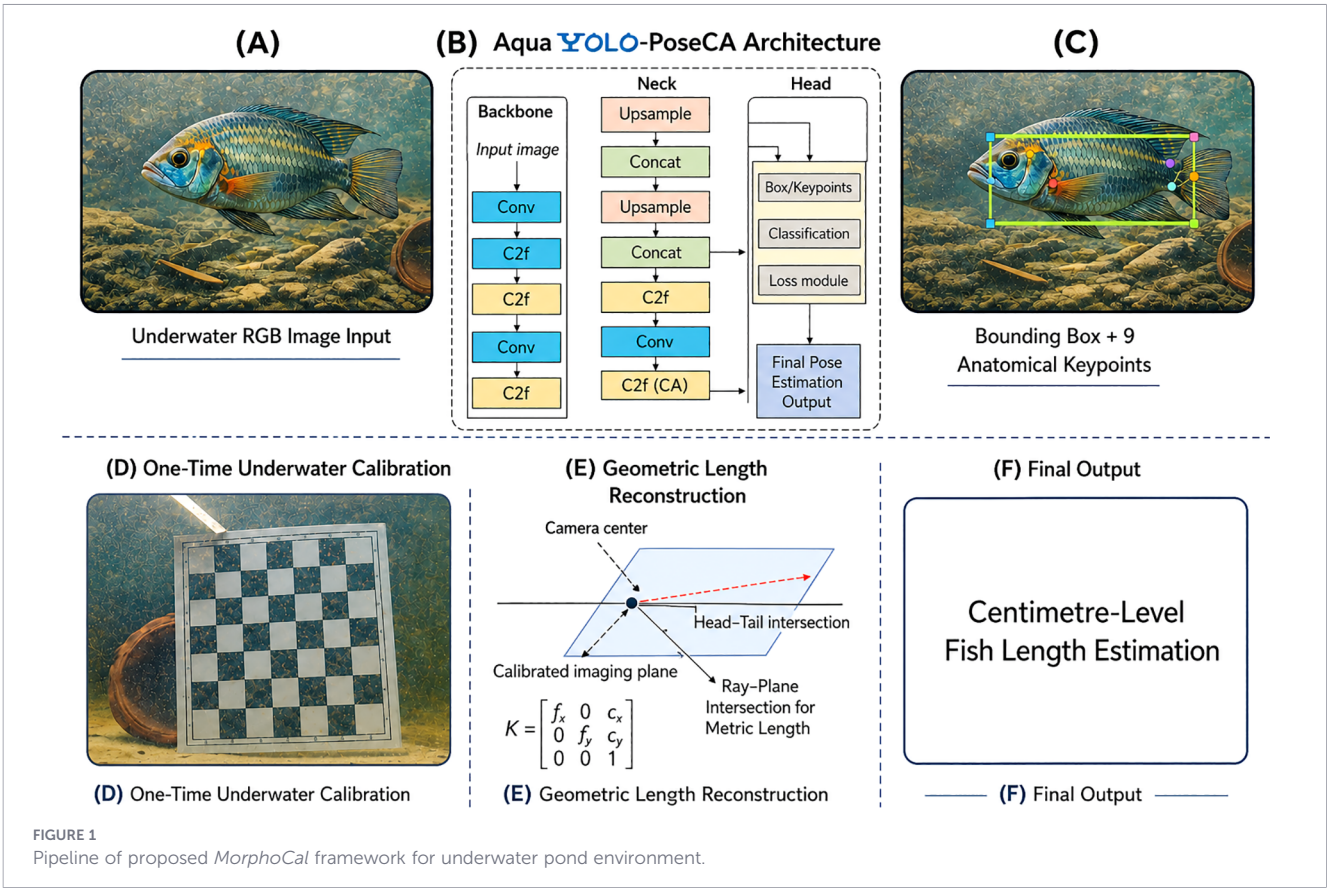


FIGURE 1
Pipeline of proposed *MorphoCal* framework for underwater pond environment.

estimation through ray–plane projection of keypoints onto the calibrated reference plane.

2.1 Proposed AquaYOLO-PoseCA architecture

AquaYOLO-PoseCA is a unified detection–keypoint estimation network that processes a 640× 640 RGB image and predicts both the fish bounding box and $K = 9$ anatomical keypoints. Attention mechanisms have become integral to improving feature discrimination in YOLO-based architectures. In this work, Coordinate Attention (CA) is adopted instead of other attention mechanisms due to its ability to preserve spatial positional information while maintaining a lightweight design. SE(Squeeze-and-Excitation) blocks collapse spatial dimensions through global pooling, which is suboptimal for anatomical keypoint detection where the relative location of landmarks along the fish body is critical. CBAM introduces both channel and spatial attention but relies on additional convolutional layers that increase computational overhead.

CA encodes long-range spatial dependencies along horizontal and vertical axes directly into channel attention, enabling the network to retain precise positional cues essential for detecting elongated and articulated fish structures under turbidity, occlusion, and pose variation. This is a characteristic that is wellsuited to fish morphometric analysis, as accurate localization of head and tail marks is a significant factor in determining the length of a fish. The ablation tests in Table 2 verify that CA improves localization robustness without compromising real-time prediction, making it well-suited for deployment in unconstrained pond environments. The model is based on a three-stage structure: backbone, neck, and double head prediction (see Figure 2).

Backbone

The backbone is based on the YOLOv8 feature extractor, with task-specific adaptation of the C2f (CrossStage Partial Fusion) (Chen et al., 2024) blocks to improve feature representation under underwater imaging constraints such as turbidity and low contrast. Given an input image $I \in \mathbb{R}^{640\times640\times3}$, the feature extraction can be expressed as shown in Equation 1:

$$F_b = \phi_{\mathrm{C2f}}(\phi_{\mathrm{Conv}}(I)), \tag{1}$$

where $\phi_{\mathrm{Conv}}(\cdot)$ denotes convolution, BN, and SiLU activation, and $\phi_{\mathrm{C2f}}(\cdot)$ performs split-transform-merge feature interactions.

The final stage of the backbone employs a Spatial Pyramid Pooling-Fast (SPPF) module (Hussein and Nemer, 2025) to increase receptive field as defined in Equation 2:

$$F_{sppf} = \mathrm{Concat}(\mathrm{MaxPool}(F_b), \mathrm{MaxPool}^2(F_b), \mathrm{MaxPool}^3(F_b)) . \tag{2}$$

Neck with Coordinate Attention (CA)

To enhance spatial localization under turbid and cluttered underwater conditions, the neck of AquaYOLOPoseCA integrates the Coordinate Attention (CA) mechanism at two feature fusion stages. Unlike conventional channel attention such as SE

TABLE 2 Hyperparameter configurations used for training the AquaYOLO-PoseCA model.

Hyperparameter	Value
Input resolution	640 × 640
Batch size	16
Number of epochs	25
Initial learning rate (lr_0)	1×10^{-4}
Optimizer	SGD with momentum
Learning rate scheduling	Cosine decay
Weight decay	0.0005
Data augmentation	Enabled (Ultralytics default pipeline)
Number of dataloader workers	4
Training device	NVIDIA Tesla T4 (16 GB VRAM)

blocks, which collapse spatial information through global pooling, CA encodes long-range dependencies along both the horizontal and vertical directions while preserving precise positional cues. For a feature map $X \in \mathbb{R}^{C\times H\times W}$, CA performs separate pooling operations along width and height dimensions, as expressed in Equation 3:

$$g_h(x) = \frac{1}{W}\sum_{i=1}^{W} X(:,:,i), \quad g_w(x) = \frac{1}{H}\sum_{j=1}^{H} X(:,j,:)\,. \tag{3}$$

The features are concatenated and fed into the shared transformation function $\delta(\cdot)$ followed by two lightweight fully connected layers as defined in Equation 4:

$$f = \delta(\text{Concat}(g_h, g_w)), \quad A_h = \sigma(W_h f), \quad A_w = \sigma(W_w f), \tag{4}$$

where σ represents the sigmoid activation function. The refined feature map is represented in Equation 5:

$$X' = X \odot A_h \odot A_w, \tag{5}$$

which adaptively highlights discriminative spatial-channel regions that are relevant for fish boundaries and anatomical landmarks. The CA blocks are inserted after the deepest and mid-level fusion layers of the PAN-FPN neck (Elmezain et al., 2025) (at layers 18 and 21), enhancing the localization sensitivity and feature propagation at multiple scales. This integration is particularly effective for underwater environments where spatial distortion and feature ambiguity are prominent. In addition to Coordinate Attention, alternative attention mechanisms such as SE and GAM were also evaluated. However, these approaches resulted in lower localization performance and higher computational cost under

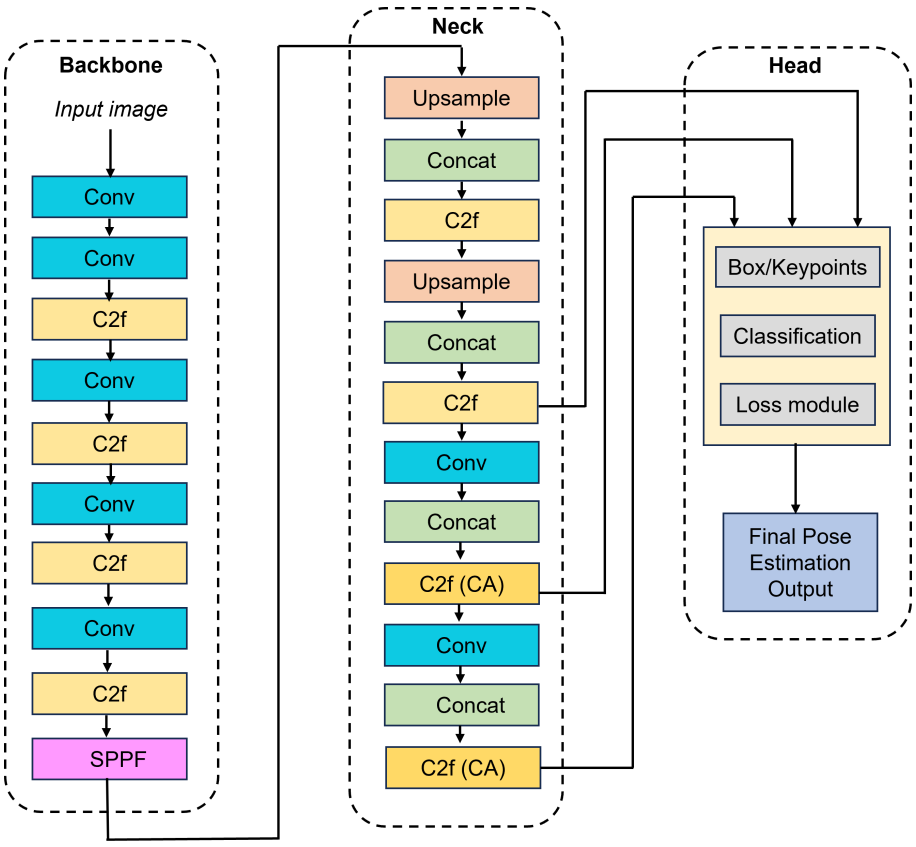


FIGURE 2
Overview of the AquaYOLO-PoseCA architecture showing the YOLOv8-based backbone, Coordinate Attention modules, Feature Pyramid aggregation, and dual-head prediction for detection and keypoint estimation.

underwater pond conditions, further supporting the suitability of CA for the proposed task.

Dual-Head Prediction

The head comprises two parallel prediction branches:

(a) Detection Head: It predicts the center coordinates (x, y), width w, height h, and objectness score $p_o$for the fish object as represented in Equation 6:

$$B = (x, y, w, h, p_o, p_c)\,. \tag{6}$$

(b) Keypoint Heatmap Head: For each anatomical keypoint $k \in K$, the network generates a Gaussian heatmap (Meng et al., 2025) at the spatial size of 64 × 64 as defined in Equation 7:

$$H_k(u, v) = \exp\left(-\frac{(u - u_k)^2 + (v - v_k)^2}{2\sigma^2}\right), \tag{7}$$

where $H_k(u, v)$ represents the heatmap value at the pixel coordinate (u, v), (u_k, v_k) represents the ground-truth keypoint location, and σ controls the spread of Gaussian response.

Training Loss

The total training loss is the weighted sum of the bounding box regression loss, the classification loss, and the keypoint heatmap regression loss, as expressed in Equation 8:

$$\mathcal{L} = \lambda_{\text{box}}\mathcal{L}_{\text{GIoU}} + \lambda_{\text{cls}}\mathcal{L}_{\text{CE}} + \lambda_{\text{kp}}\mathcal{L}_{\text{MSE}}, \tag{8}$$

with:

- $\mathcal{L}_{\text{GIoU}}$ — the Generalized Intersection-over-Union (GIoU) loss for the bounding box regression,
- $\mathcal{L}_{\text{CE}}$ — the Cross-Entropy (CE) loss for the object class probability prediction,
- $\mathcal{L}_{\text{MSE}}$ — the Mean Squared Error (MSE) loss for predicted and target keypoints

Output

In the end, the network predicts both the bounding box coordinates and the fish keypoint coordinates. For each detected fish i, the prediction is represented in Equation 9:

$$\{B_i, (x_{i,1}, y_{i,1}), \ldots, (x_{i,9}, y_{i,9})\}, \quad i = 1, \ldots, N_{\text{fish}}, \tag{9}$$

where $B_i = (x, y, w, h)$ is the bounding box coordinates.

This demonstrates that the proposed AquaYOLO-PoseCA architecture, through its underwater-specific adaptation and attention-enhanced feature learning, provides a robust solution for keypoint-based morphometric analysis in real-world aquaculture environments. The model achieves high localization accuracy even in turbid water, changing fish poses, and partial occlusions, while maintaining low computational complexity with 11.4M parameters and 29.4 GFLOPs.

Given an input underwater image $\mathbf{I} \in \mathbb{R}^{640\times640\times3}$, AquaYOLO-PoseCA carries out simultaneous fish detection and anatomical keypoint detection (Zhang et al., 2025). It predicts bounding box coordinates, as represented in Equation 10:

$$\mathbf{B} = \{(x_c, y_c, w, h)\} \tag{10}$$

in image coordinates, along with $K = 9$ keypoint coordinates, as shown in Equation 11:

$$\mathbf{H} \in \mathbb{R}^{K\times64\times64} \tag{11}$$

Each keypoint k is localized by spatial argmax.

To convert the pixel-space keypoints into real world coordinates, we utilize the calibrated camera matrix $\mathbf{K}$ and the distortion parameters θ computed in Section 3.1. For each keypoint ray, we perform back-projection into 3D space and intersect it with the checkerboard calibration plane (Pedersen et al., 2018), as described in Equation 12:

$$\mathbf{P}_k = \mathbf{C} + \lambda_k\mathbf{d}_k, \tag{12}$$

where $\mathbf{C}$ is the camera center, $\mathbf{d}_k$is the normalized direction of the ray computed from the keypoint coordinates (u_k^{px}, v_k^{px}). The value of $\lambda_k$is computed by intersecting the ray with the plane equation, as expressed in Equation 13:

$$\lambda_k = \frac{\mathbf{n}^\top(\mathbf{X}_0 - \mathbf{C})}{\mathbf{n}^\top\mathbf{d}_k}, \tag{13}$$

where $\mathbf{n}$ is the normal of the plane, $\mathbf{X}_0$ is an arbitrary known point on the plane.

The 3D coordinates of keypoints $\mathbf{P}_k$ enable centimeter-accurate morphometric measurement. Total fish length L is estimated as the Euclidean distance between the head (mouth) and caudal mid-tail keypoints, as defined in Equation 14:

$$L = \|\mathbf{P}_{\text{mouth}} - P_{\text{mid-tail}}\|\,. \tag{14}$$

This projection-based computation ensures that length estimation is independent of camera angle, fish pose, and magnification criteria.

2.2 Checkerboard camera calibration

To convert pixel-space keypoints into real-world metric measurements, a single-shot camera calibration was performed using a planar checkerboard (Huo et al., 2025) placed underwater under the same optical conditions as the fish imagery (Figure 3). The checkerboard consists of (W × H) inner-corners with a known physical square size s (in millimetres). Calibration images were captured at multiple orientations and distances to ensure adequate coverage of the image plane.

For each calibration image, inner-corner points $u_{ij} = [u_{ij}, v_{ij}]^\top$ were detected and refined to sub-pixel accuracy. The corresponding 3D object points lie on the $Z = 0$ plane, as represented in Equation 15.

$$\mathbf{X}_{ij} = [is, js, 0]^\top, \quad i = 0 \ldots W - 1, \quad j = 0 \ldots H - 1. \tag{15}$$

We adopt the pinhole projection model with intrinsic matrix $\mathbf{K}$ and per-image extrinsics ($\mathbf{R}$, $\mathbf{t}$), as expressed in Equation 16:

$$\lambda\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K}[\mathbf{R}\ \mathbf{t}]\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{16}$$

Radial–tangential lens distortion is modeled as described in Equation 17:

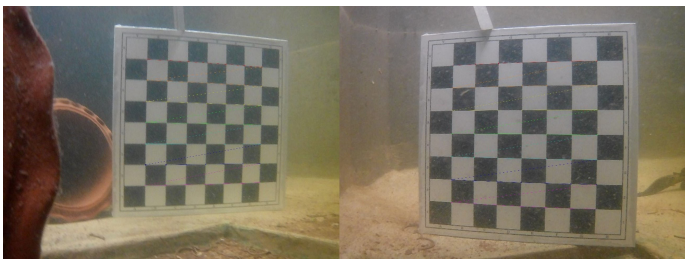


FIGURE 3
Underwater checkerboard used for camera calibration. The known square size provides a metric reference for pixel-to-centimetre scaling under the same optical conditions as fish observation.

$$x_d = x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + 2p_1 xy + p_2(r^2 + 2x^2),$$
$$y_d = y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + p_1(r^2 + 2y^2) + 2p_2 xy, \quad (17)$$

where $r^2 = x^2 + y^2$ are normalized coordinates.

The intrinsic parameters $\mathbf{K}$, distortion coefficients $\theta = \{k_1, k_2, k_3, p_1, p_2\}$, and extrinsics $(\mathbf{R}_v, \mathbf{t}_v)$ of each calibration image v are estimated by minimizing the reprojection error, as defined in Equation 18:

$$\mathcal{L}_{\text{reproj}} = \sum_v \sum_{ij} \left\| \mathbf{u}_{ij} - \hat{u}_{ij}(\mathbf{K}, \theta, R_v, t_v) \right\|^2. \quad (18)$$

The mean reprojection error (MRE) is reported as defined in Equation 19.

$$\text{MRE} = \frac{1}{\sum_{v=1}^{V} N_v} \sum_{v=1}^{V} \sum_{i=1}^{N_v} \left\| \mathbf{u}_i^{(v)} - \hat{\mathbf{u}}_i^{(v)} \right\|_2, \quad (19)$$

where $\mathbf{u}_i^{(v)}$ denotes the observed image point and $\hat{\mathbf{u}}_i^{(v)}$ represents the corresponding reprojected point obtained using the estimated camera parameters. This metric serves as the primary indicator of calibration accuracy. The mean reprojection error estimated for the setup was 0.164 pixels, indicating accurate and reliable image reconstruction. The calibrated camera matrix and distortion parameters are subsequently used to back-project detected keypoints (Arguel et al., 2025). As preprocessing all the images are resized to 640 × 640 as the input to the proposed detection model with $K = 9$ keypoint heatmaps. The peak extraction begins with the application of the sigmoid function on the heatmap logits followed by the application of the non-maximum suppression over the 3x3 region. The peaks that are higher than the confidence threshold are selected, and the fish instances are created using the multi-instance keypoint grouping. The mouth and the bottom tail of the fish, represented by the landmarks at positions k = 0 and k = 8, are selected for each fish instance i for the computation of the total length.

Let $(u_{i,k}^{px}, v_{i,k}^{px})$ denote the pixel coordinates of a predicted keypoint. MorphoCal uses the calibrated camera to project this pixel onto the checkerboard plane. First, lens distortion is removed to obtain normalized image coordinates, which are then back-projected through $\mathbf{K}^{-1}$ to form a unit ray $\mathbf{d}_{i,k}$from the camera centre $\mathbf{C}$. The 3D intersection with the calibration plane Π (defined by normal $\mathbf{n}$ and a point $\mathbf{X}_0$ on the board) is computed as described in Equation 20.

$$\mathbf{P}_{i,k} = \mathbf{C} + \lambda_{i,k} \quad \mathbf{d}_{i,k}, \quad \lambda_{i,k} = \frac{\mathbf{n}^\top(\mathbf{X}_0 - \mathbf{C})}{\mathbf{n}^\top \mathbf{d}_{i,k}}, \quad (20)$$

yielding coordinates in millimetres on the checkerboard plane. The metric total length for fish i is computed as defined in Equation 21.

$$L_i^{\text{mm}} = \left\| P_{i,\text{mouth}} - P_{i,\text{bottom-tail}} \right\|_2, \quad L_i^{\text{cm}} = \frac{L_i^{mm}}{10}. \quad (21)$$

Low-quality detections are discarded using a combined objectness and keypoint-confidence threshold together with a minimum-keypoint rule. Multi-fish scenes are handled by spatial clustering (Sattar et al., 2025) of detections prior to length computation. The same calibration (mean reprojection error 0.164 px) is applied to all images acquired from the underwater rig, ensuring a stable centimetre scale across time and across different ponds.

2.3 Ray–plane projection for metric length estimation

After keypoint localization in pixel coordinates, metric fish length is recovered by projecting each keypoint into 3D using the calibrated camera model (Yu et al., 2023). Let (u_k, v_k) denote the pixel coordinates of keypoint k and $\mathbf{K}$ be the intrinsic matrix. The normalized camera ray direction is computed as expressed in Equation 22.

$$\mathbf{d}_k = \frac{\mathbf{K}^{-1}[u_k, \quad v_k, \quad 1]^\top}{\left\| \mathbf{K}^{-1}[u_k, \quad v_k, \quad 1]^\top \right\|}. \quad (22)$$

Let $\mathbf{C}$ denote the camera center and Π be the calibration plane defined by a known point $\mathbf{X}_0$ and normal vector $\mathbf{n}$. The 3D coordinates of the projected keypoint are obtained by intersecting the ray with the plane, as defined in Equation 23.

$$\mathbf{P}_k = \mathbf{C} + \lambda_k \mathbf{d}_k, \quad (23)$$

where the scalar $\lambda_k$is defined in Equation 24:

$$\lambda_k = \frac{\mathbf{n}^\top(\mathbf{X}_0 - \mathbf{C})}{\mathbf{n}^\top \mathbf{d}_k}. \quad (24)$$

Since the reconstructed 3D point $\mathbf{P}_k$depends on the predicted 2D keypoint location, any localization uncertainty in the image plane can propagate into the metric reconstruction. Let the uncertainty in the keypoint position be represented by $(\delta u_k, \delta v_k)$. Using a

first-order approximation, the corresponding uncertainty in the reconstructed 3D point is expressed in Equation 25.

$$\delta \mathbf{P}_k \approx \frac{\partial \mathbf{P}_k}{\partial u_k} \quad \delta u_k + \frac{\partial \mathbf{P}_k}{\partial v_k} \quad \delta v_k \,. \tag{25}$$

Accordingly, the uncertainty in the estimated fish length depends on the propagated uncertainties of the selected anatomical keypoints. This highlights that the final metric accuracy is influenced by both calibration precision and keypoint localization quality.

This projection places each anatomical keypoint into a common real-world coordinate frame measured in centimetres. Fish total length is then computed as the Euclidean distance between the mouth and mid-tail keypoints, as defined in Equation 26.:

$$L = \left\| \mathbf{P}_{\text{mouth}} - \mathbf{P}_{\text{mid-tail}} \right\| . \tag{26}$$

This ray–plane projection ensures that the length measurement remains consistent regardless of camera orientation, fish pose, or distance from the camera.Calibration is done underwater environment, which incorporates the refraction effects into the camera parameters to enable accurate measurements in centimeters from a single RGB image.

From the given per-keypoint heatmaps $\mathbf{H} \in \mathbb{R}^{K\times 64\times 64}$ and the detection set $\mathcal{B} = \{B_i\}$, we assign the $K = 9$ peak locations to each detection to form instance-wise keypoint sets $\{\mathbf{k}_{i,1:9}\}$. We first gate candidate peaks that fall inside (or within a small margin of) each B_i to suppress cross-instance confusion in crowded scenes. For every detection entry, we construct a cost matrix between the K semantic keypoints and the gated peaks that balances spatial proximity and heatmap confidence, and solve a one-to-one assignment via the Hungarian algorithm. Finally, we refine each assigned keypoint by soft-argmax or quadratic peak fit to sub-pixel precision. If a semantic keypoint is missing (Li, 2025)(no confident peak), then it is excluded from length computation. This approach is applied for instance-consistent $\{\mathbf{k}_{i,1:9}\}$ for metric projection and total-length estimation, as described in Algorithm 1.

Require: RGB image(s) $\mathbf{I} \in \mathbb{R}^{640\times 640\times 3}$; calibration views $\{\mathcal{I}_v\}$; network $\mathcal{N}$; thresholds τ_{det}, τ_{kpt}; checkerboard square size s; pattern (W, H)
Ensure: Fish detections with keypoints and lengths $\{(B_i, \{\mathbf{k}_{i,1:9}\}, L_i)\}$; calibration $\mathcal{C} = \{\mathbf{K}, \boldsymbol{\theta}, \mathbf{n}, \mathbf{X}_0\}$; validation metrics
A. Fish detection and keypoint extraction
$(\{B_i\}, \mathbf{H}) \leftarrow \mathcal{N}(\mathbf{I})$3 : Filter boxes with confidence $\geq \tau_{\text{det}}$ and apply NMS
for $k = 1$ to 9 **do**
Extract peaks (u, v, s) from heatmap $\mathbf{H}_k$ where $s \geq \tau_{\text{kpt}}$
Convert to pixel coordinates $(u^{px}, v^{px}) = \frac{640}{64}(u, v)$
end for
B. Keypoint assignment
for each detection box B_i **do**
  Select candidate peaks within box
  Construct cost matrix $C_{k,p}$
  Apply Hungarian matching
  Keep matched keypoints satisfying τ_{kpt}
  Store instance keypoints $\{\mathbf{k}_{i,1:9}\}$
end for
C. Underwater calibration
if calibration $\mathcal{C}$ not available **then**
  for each calibration view $\mathcal{I}_v$ **do**
  Detect checkerboard corners
  Generate 3D grid $\mathbf{X}_{ij} = s[i, j, 0]^\top$
end for
  Estimate intrinsics $\mathbf{K}$ and distortion $\boldsymbol{\theta}$
  Compute reprojection error $\bar{e}$
  Estimate plane $(\mathbf{n}, \mathbf{X}_0)$
  Store $\mathcal{C}$
end if
D. Metric projection and fish length
for each fish instance i **do**
  for each keypoint $\mathbf{k}_{i,k}$ **do**
  Undistort (u^{px}, v^{px})
  Compute ray $\mathbf{d}_{i,k} = \text{normalize}(\mathbf{K}^{-1}[\tilde{u}, \tilde{v}, 1]^\top)$
  Intersect ray with plane Π
  Obtain world point $\mathbf{P}_{i,k}$
end for
  Compute length $L_i = \left\| \mathbf{P}_{i,\text{mouth}} - \mathbf{P}_{i,\text{bottom-tail}} \right\|$
end for
E. Length validation
Capture validation image with checkerboard or ruler
Detect validation points $\{\mathbf{u}_m^{val}\}$
Compute ground truth distances $D^*_{m_1 m_2}$
Compute predicted distances $\hat{D}_{m_1 m_2}$
Evaluate RMSE, MAPE, R^2, Bland–Altman
return Detected fish instances, calibration parameters, and validation metrics =0

Algorithm 1. MorphoCal: Keypoint detection, calibration, metric projection, and validation.

3 Experimental design and evaluation framework

The experiments were performed in Google Colab Pro (paid version) using a NVIDIA Tesla T4 GPU (16 GB GDDR6). The local development system uses an AMD Ryzen 5 5600H processor with Radeon Graphics and 12 GB of RAM. Our proposed AquaYOLO-PoseCA model was implemented using PyTorch and Ultralytics YOLO framework. The dataset of 640× 640 images was used for keypoint detection, and the hyperparameters were kept the same across model performance evaluations; the settings are tabulated in Table 2. The model was evaluated on our DePondFi'24 dataset using a batch size of 16 and 25 training epochs. The initial learning rate of 1×10^{-4} with cosine scheduling and SGD optimizer (momentum = 0.937, weight decay = 5×10^{-4}) was implemented. The validation and test set was performed on the same GPU to measure runtime performance and frame rate.

TABLE 3 DePondFi'24 dataset statistics.

Subset	No. of images	Approx. fish instances	Purpose
Training set	560	~2800	Model training
Validation set	160	~1800	Hyperparameter tuning
Test set	80	~1050	Quantitative evaluation
Total	**800**	**~ 4850**	–

Bold values represent the total count.

TABLE 4 Keypoints and their anatomical descriptions used for morphometric analysis.

ID	Region	Description
0	Mouth	Reference landmark for defining the head boundary and total length estimation.
1	Eye	Used for determining fish orientation and alignment during pose estimation.
2	Top fin	Dorsal fin apex, relevant for hydrodynamic posture and species-specific fin shape.
3	Center	Mid-body anatomical point used to assess symmetry and body curvature.
4	Bottom center	Ventral mid-body point representing lower body structure and width characteristics.
5	Start tail	Caudal peduncle location marking the transition between torso and tail fin region.
6	Top tail	Upper extremity of the caudal fin contour.
7	Mid tail	Intermediate tail joint landmark facilitating consistent tail alignment measurement.
8	Bottom tail	Lower extremity of the caudal fin defining the caudal boundary.

FIGURE 4
Underwater checkerboard used for intrinsic and extrinsic calibration showing the turbidity level.

3.1 Dataset details

The proposed study utilizes DePondFi'24 from ten aquaculture ponds located in and around Kolathur, Chennai, India. This area is chosen because it is best suited for ornamental fish production and the cultivation of enriched fish. Data collection was carried out with the support of aqua farmers and aquaculture practitioners over a period of one and a half months. Video recordings were acquired using an underwater Crosstour CT9000 camera (16 MP, 170°field of view) mounted at a depth of 4–5 m. Each video was recorded at 1920 × 1080 pixels resolution and 60 frames per second. These videos are captured under direct sunlight during 10:00–12:00 am and 04:00–06:00 pm—to capture variation and to consider all complex challenges regarding illuminations. The ponds chosen for dataset collection are highly turbid due to feed particles, soil quality, and the suspended other particles. The average turbidity, estimated during the recording sessions, ranges between 15–25 NTU. The water conditions visible in Figure 4 illustrate the real-time turbid level during image acquisition.

TABLE 5 Quantitative comparison of the proposed AquaYOLO-PoseCA, YOLO baseline models, RTMPose-s, and HRNet-W32 on the DePondFi'24 dataset.

Model	Box				Pose				Time (ms)			GFLOPs	Params (M)
	P	R	mAP_{50}	mAP^{50-95}	P	R	mAP_{50}	mAP^{50-95}	Pre	Infer	Post		
Proposed AquaYOLO-PoseCA	0.917	0.900	0.959	0.712	0.847	0.840	0.858	0.592	5.5	11.1	6.4	29.4	11.4
YOLOv8-l	0.881	0.877	0.940	0.689	0.818	0.817	0.841	0.505	7.6	37.1	5.7	168.5	44.5
YOLOv8-m	0.862	0.881	0.942	0.686	0.836	0.840	0.841	0.562	5.9	22.6	5.8	80.9	26.4
YOLOv8-n	0.909	0.869	0.952	0.699	**0.875**	0.831	**0.861**	0.576	6.3	**8.6**	6.4	8.6	3.1
YOLOv8-s	0.877	**0.913**	0.950	0.707	0.846	**0.850**	0.849	0.578	9.4	10.3	6.8	29.4	11.4
YOLOv11-l	0.881	0.890	0.932	0.683	0.839	0.849	0.829	0.518	5.5	35.7	5.8	90.3	26.1
YOLOv11-m	0.916	0.845	0.937	0.657	0.870	0.804	0.831	0.529	6.8	30.8	6.6	71.4	20.9
YOLOv11-n	0.883	0.895	0.947	0.671	0.827	0.827	0.831	0.572	6.4	8.5	4.8	**6.8**	**2.7**
YOLOv11-s	0.900	0.886	0.949	0.687	0.848	0.841	0.840	0.552	9.4	12.2	5.8	22.3	9.7
RTMPose-s	N/A	N/A	N/A	N/A	N/A	0.561	0.862	0.509	N/A	147.4†	N/A	10.74‡	5.27

N/A indicates that the metric is not directly applicable. RTMPose-s is a top-down keypoint estimator and does not output YOLO-style bounding-box precision/recall or pose precision. †For RTMPose-s, inference time is reported from the MMPose test log as the average time per test iteration. Bold values indicate the best performance for each metric.

TABLE 6 Cross-validation performance of AquaYOLO-PoseCA.

Fold	mAP@50
Fold 1	0.91
Fold 2	0.90
Fold 3	0.92
Fold 4	0.91
Fold 5	0.90
Mean ± Std	0.91 ± 0.01

The dataset consists of nearly five ornamental and freshwater fish species commonly cultivated in South-Indian regions, including Vieja melanura, Escondido cichlid, Dawkinsia filamentosa, Blood Parrot Cichlid, and several other cichlid varieties (De Ridder et al., 2025). Each frame contains an average of 5 to 7 fish with different orientations and poses, partial occlusions, and light-scattering conditions under natural lights. From the thousands of collected frames, 800 images were selected using keyframe extraction. Contrast stretching and histogram normalization techniques are applied as a preprocessing step, and the frames are resized to 640 × 640 pixels. Dataset statistics are summarized in Table 3 (DePondFi'24 dataset).

Annotations are performed manually using the Roboflow annotation software. The annotation format chosen is YOLO format (Vijayalakshmi and Sasithradevi, 2025) and also raw pixel coordinates formats for calibrations. Each fish was annotated with a bounding box and nine keypoints corresponding to key morphological points on fish. These keypoints and the descriptions are listed in Table 4. The annotation process was conducted with great attention to ensure high-quality labeling. All annotations were independently double-verified by a separate team comprising aquaculture domain experts. The dataset was split into training, validation, and test sets at 70:20:10 ratio. In addition to the fixed train–validation–test split, cross-validation was also performed to assess the stability of the proposed model across different data partitions. This was included to reduce split-specific bias and to provide a more reliable estimate of model consistency. Although the dataset size is moderate, it was curated under real-world South Indian pond conditions with substantial variability in turbidity, illumination, occlusion, and fish density, making it representative of practical aquaculture deployment scenarios.

Each annotation file (. t x t) corresponds to a single image with raw coordinate structure, where all values are recorded in pixel units corresponding to the image resolution (640× 640). Our annotation format also include yolo format, and the syntax of this format is shown below:<class_id><x_center ><y_center><width><height><kpx1><-kpy1><kpv1>…<kpx9><kpy9><kpv9>

Here, class_id indicates the fish species, and as we are not classifying the fishes, we consider 0 for allfishes, x_center, y_center, widt h, and height represent the bounding-box parameters, and each keypoint (kpx_i, kpy_i) corresponds to the pixel location of the i^{th} anatomical keypoint. The visibility flag $kpv_i \in \{0, 1, 2\}$ denotes whether the keypoint is not visible, partially visible, or fully visible, respectively. All annotation files were manually cross-verified for keypoint order sequence, totla keypoint count of 9, and geometric evaluation before being finalized. Our annotation system and the process are illustrated in Figure 5.

3.2 Evaluation metrics

The standard detection evaluation metrics (Yi et al., 2025) Precision, Recall, mAP@50, mAP@50–95 are applied for the evaluation of the model performance. For anatomical keypoint localization evaluation, pixel-based PCK@5/10/20, scale-normalized PC K $_\alpha$(e.g., PC $K_{0\cdot01}$ and PC $K_{0\cdot05}$) (Liu M. et al., 2025), and Mean Square Error (MSE) metrics are employed to quantify localization accuracy relative to bounding-box dimensions (Liu L. et al., 2025).

4 Morphocal: performance analysis

4.1 Object and keypoint detection performance

The proposed AquaYOLO-PoseCA model was benchmarked against the YOLOv8-Pose and YOLOv11-Pose families, which are

FIGURE 5
DePondFi'24 keypoint annotation process: **(a)** Keypoint topology with nine anatomical keypoints used for morphometric analysis. **(b)** Sample annotated keypoints on a fish for visualization. **(c)** Manual labeling of the Roboflow dashboard showing bounding box with keypoint fish pose estimation.

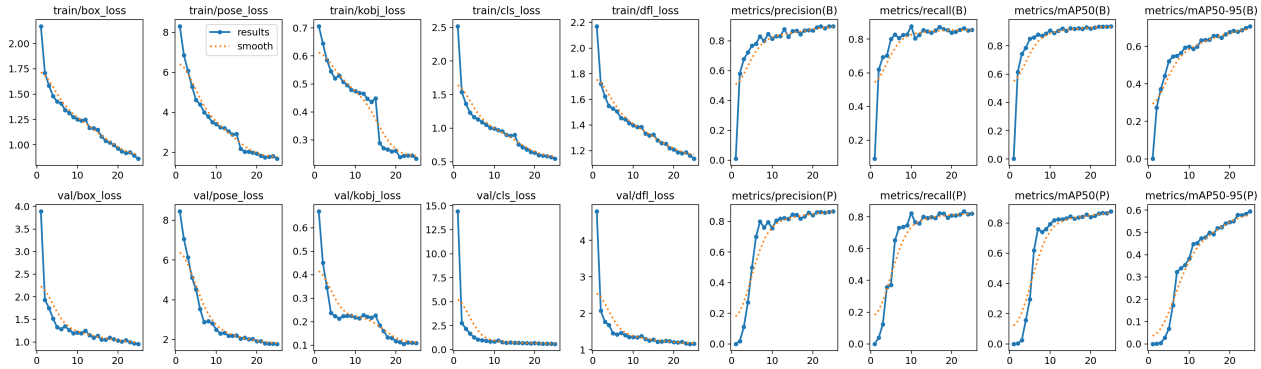


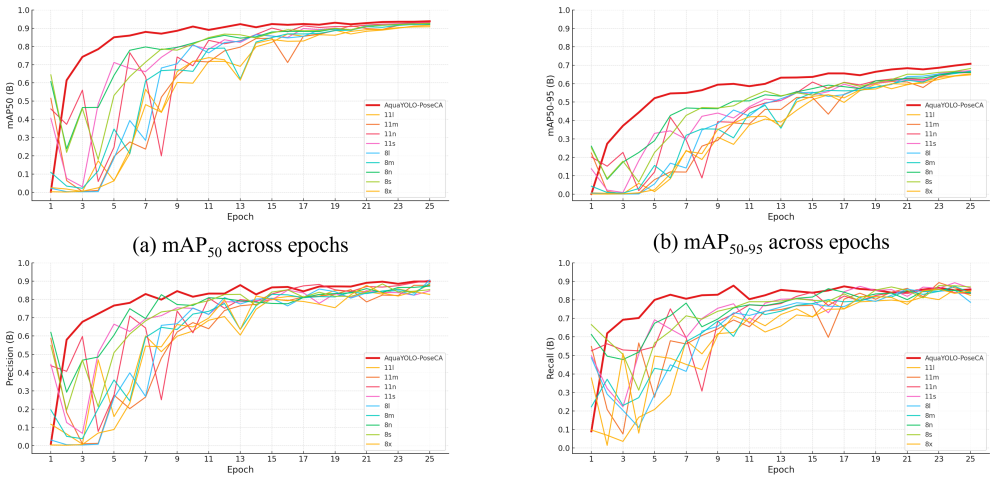
FIGURE 6
Training and validation curve of AquaYOLO-PoseCA showing the evolution of loss components (box, pose, objectness, classification) and detection metrics (Precision, Recall, mAP_{50}, mAP_{50-95}) over 25 epochs. The smooth dotted line represents the exponential moving average of the training trend.

the only standard and officially supported keypoint detection architectures in the Ultralytics framework. These baseline models are chosen as YOLOv8 is a stable and efficient anchor-free detector, while YOLOv11 adds improvements such as better feature fusion, separate detection heads, and smarter loss weighting. By adding a Coordinate Attention (CA) module (Yang et al., 2025), the spatial detail detection capacity and inter-channel relationship of YOLO can be improved, enabling better keypoint detection in challenging underwater environments. The loss curves in Figure 6 show that box loss, keypoint loss, classification loss, and distributional focal loss all decrease steadily during training and validation. And also Figure 6 with recall, and mAP metrics demonstrate a steady improvement across epochs. This shows the stable training and strong generalization of our proposed model.

The metric visualizations of baseline models, including our proposed model for detection metrics, are shown in Figure 7 (mAP_{50}, mAP_{50-95}, Precision, and Recall). The proposed model exhibits faster convergence across all evaluation criteria. Figure 8 presents qualitative validation results of AquaYOLOPoseCA, showing reliable fish localization and anatomically consistent keypoint predictions.

AquaYOLO-PoseCA achieved strong performance across both object detection and pose estimation metrics, obtaining a bounding-box mAP_{50} of 0.959 and mAP_{50-95} of 0.712, as shown in Table 5. For keypoint localization, the proposed model achieved a pose mAP_{50} of 0.858 and the highest pose mAP_{50-95} of 0.592, demonstrating improved robustness in precise landmark localization. Although YOLOv8n achieved slightly higher pose precision and mAP_{50}, AquaYOLO-PoseCA provided better overall localization quality at stricter OKS thresholds, as reflected by its superior pose mAP_{50-95}. With 11.4M parameters, 29.4 GFLOPs, and an inference time of 11.1 ms, the proposed model achieves a favourable balance between accuracy and computational efficiency compared with larger YOLOv8 (Tian et al., 2024) and YOLOv11 variants.

To address comparisons beyond YOLO-based architectures, RTMPose-s (Jiang et al., 2023) was evaluated as a representative



(a) mAP_{50} across epochs
(b) mAP_{50-95} across epochs
(c) Precision (B) across epochs
(d) Recall (B) across epochs

FIGURE 7
Performance comparison with baselines. **(a)** mAP_{50}, **(b)** mAP_{50-95}, **(c)** Precision, and **(d)** Recall metrics across training epochs. The proposed AquaYOLO-PoseCA (red curve) demonstrates faster convergence compared with baseline models.

FIGURE 8
Sample keypoint predictions on images from the validation set of the AquaYOLO-PoseCA model.

TABLE 7 **Keypoint localization error and PCK accuracy comparison of AquaYOLO-PoseCA, YOLO variants, and RTMPose-s on the DePondFi'24 test set.**

Model	Avg KPs Missed ↓	MSE ↓	MAE ↓	Fish/ Img	PCK@20px ↑	PC K @α = 0.1 ↑	PC K @α = 0.5 ↑
AquaYOLO-PoseCA-32	2.40	0.2419	0.4134	4–5	0.002885	0.072115	0.817308
AquaYOLO-PoseCA-R16	2.48	0.2568	0.4333	4–5	0.002921	0.069133	0.819864
YOLOv8-L	2.12	0.2589	0.4203	4–5	0.001923	**0.076923**	0.820192
YOLOv8-M	2.39	0.2641	0.4305	4–6	0.002885	**0.076923**	0.819231
YOLOv8-N	2.16	0.2473	0.4221	4–5	0.001923	0.074038	0.818269
YOLOv8-S	2.34	0.2665	0.4368	4–5	0.002885	0.069231	0.819231
YOLOv8-X	2.36	0.2315	0.4077	4–6	0.002885	0.067308	0.815385
YOLOv11-L	2.04	0.2566	0.4289	4–6	0.000962	0.066346	**0.828846**
YOLOv11-M	2.11	0.2958	0.4421	4–6	0.001923	0.071154	0.816346
Y OL Ov11-N	2.01	0.2831	0.4370	4–6	0.002885	0.067308	0.819231
YOLOv11-S	2.04	0.2415	0.4257	4–6	0.002885	0.067308	0.819231
RTMPose-s[†]	N/A	N/A	16.82 px	N/A	0.8138	N/A	N/A
Proposed AquaYOLO-PoseCA	2.06	**0.2221**	**0.4027**	**5–7**	**0.002923**	0.075000	0.824038

[†]RTMPose-s was evaluated using the MMPose top-down keypoint protocol. Therefore, YOLO-specific missed-keypoint count, normalized MSE/MAE, fish-per-image prediction count, and scale-normalized PCK values are not directly applicable. For RTMPose-s, absolute pixel-based evaluation produced PCK@5px = 0.2851, PCK@10px = 0.5758, and PCK@20px = 0.8138, with a mean keypoint error of 16.82 pixels and a median keypoint error of 8.31 pixels. Bold values indicate the best performance for each metric.

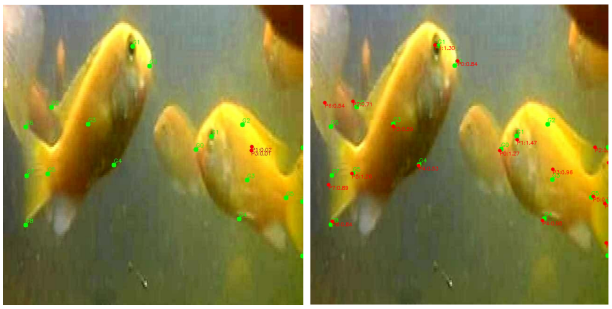


FIGURE 9
Qualitative comparison of non-YOLO top-down keypoint estimators on a sample test image from the DePondFi'24 dataset. Left: HRNet-W32 pretrained result. Right: RTMPose-s result. Green points indicate ground-truth keypoints, while red points denote predicted keypoints.

top-down keypoint estimation baseline. RTMPose-s achieved a pose mAP_{50} of 0.862; however, its $mAP_{50–95}$ decreased to 0.509, indicating reduced localization consistency under stricter OKS thresholds. In addition, HRNet-W32 (Peng et al., 2025) pretrained was examined as another non-YOLO top-down baseline. Its qualitative predictions showed weak fish landmark localization under the same 25-epoch setting. This highlights the need for task-specific adaptation when applying generic top-down pose estimators to underwater fish keypoint detection. Representative qualitative results of HRNet-W32 pretrained and RTMPose-s are shown in Figure 9.

To further examine the stability of AquaYOLO-PoseCA on the DePondFi'24 dataset, multi-fold crossvalidation was conducted, and the corresponding results are summarized in Table 6. The results show consistent performance across all folds, with low variance (± 0.01), indicating that the proposed framework is stable and not sensitive to data partitioning. This demonstrates the robustness of the model under varying in-domain conditions.

We further report the Percentage of Correct Keypoints (PCK) using both absolute and relative thresholding criteria. For an absolute pixel threshold $\tau \in \{5, 10, 20\}$, PCK represents the proportion of predicted keypoints that fall within τ pixels of the corresponding ground-truth landmark. In addition, relative PCK evaluates keypoint correctness using a scale-normalized threshold, allowing localization accuracy to be assessed with respect to object size.

Table 7 summarizes the keypoint localization error and PCK-based accuracy of AquaYOLO-PoseCA, YOLOv8–YOLOv11

TABLE 8 Extended ablation study comparing different attention mechanisms integrated into the AquaYOLO-Pose architecture.

Model	Box P	Box R	mAP_{50} (B)	mAP50 – 95 (B)	Pose P	Pose R	mAP_{50} (P)	mAP50 – 95 (P)	Pre	Infer	Post	GFLOPs	Params
YOLOv8s-Pose (Baseline)	0.896	0.856	0.938	0.707	0.865	0.820	0.877	0.594	8.4	5.3	9.2	29.4	11.41
AquaYOLO-PoseCA (r = 32)	0.907	0.900	0.949	0.712	0.847	0.840	0.848	0.592	6.1	5.7	7.4	29.4	11.42
AquaYOLO-PoseCA (r = 16)	0.879	0.889	0.933	0.690	0.852	0.856	0.880	0.585	7.2	5.6	9.5	29.4	11.41
AquaYOLO-CBAM	0.852	0.891	0.897	0.569	0.832	0.838	0.864	0.587	7.8	15.6	8.2	157.8	64.8
AquaYOLO-PoseGAM	0.861	0.798	0.872	0.459	0.815	0.857	0.832	0.542	8.2	17.7	7.4	169.9	54.2
AquaYOLO-SE	0.834	0.773	0.825	0.406	0.782	0.814	0.798	0.519	9.1	15.8	9.5	187.9	48.7
AquaYOLO-PoseCA (Proposed)	**0.917**	0.900	**0.959**	**0.712**	0.847	0.840	0.858	**0.592**	**5.5**	11.1	**6.4**	29.4	11.4

Bold values indicate the best performance for each metric.

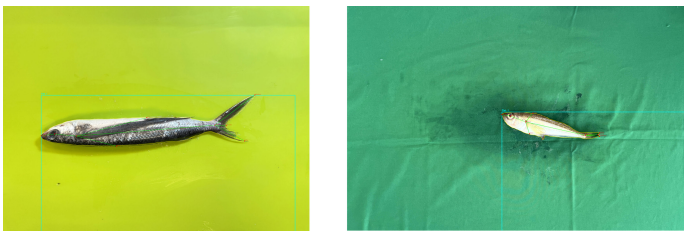


FIGURE 10
Qualitative results on the external FIB dataset showing fish detection and landmark localization by the proposed AquaYOLO-PoseCA model.

FIGURE 11
Representative failure cases in keypoint estimation under underwater pond conditions. Green points indicate ground-truth keypoints, while red points denote predicted keypoints.

baselines (Zhu et al., 2025), and RTMPose-s. For YOLO-based models, the localization errors were computed by comparing the predicted raw keypoint coordinates with the corresponding ground-truth coordinates. The proposed AquaYOLO-PoseCA achieved the lowest MSE and MAE among the YOLObased models, indicating improved landmark localization accuracy. The results suggest that the Coordinate Attention module enhances spatial consistency by preserving direction-aware positional information, which is important for localizing elongated fish structures and tail-region landmarks.

Figure 10 presents representative qualitative results on the external FIB dataset (Hasegawa and Nakano, 2024). Compared with DePondFi'24, the FIB images contain single-fish samples captured under more controlled background conditions. Despite this domain difference, AquaYOLO-PoseCA was able to identify fish regions and localize key morphological landmarks, indicating that the learned fish-structure representation can transfer beyond the original underwater pond setting.

4.2 Failure analysis of keypoint estimation

Accurate keypoint localization is essential for reliable morphometric analysis, since fish length estimation in MorphoCal depends directly on the spatial precision of anatomical landmarks. However, underwater pond environments introduce challenges such as occlusion, multi-instance overlap, body curvature, turbidity, and light scattering, which affect prediction quality.

TABLE 9 Statistical analysis of fish length estimation errors.

Metric	Value
Number of samples (N)	4810
Mean error (bias) (cm)	0.0089
Standard deviation of error (cm)	0.8638
Variance of error (cm^2)	0.7461
MAE (cm)	0.7478
MSE (cm^2)	0.7461
RMSE (cm)	0.8638
95% confidence interval of mean error (cm)	[– 0.0155, 0.0333]
Paired t-test statistic (t)	0.7168
Paired t-test p-value (p)	0.4736

As shown in Figure 11, keypoint failures mainly arise in two forms. In multi-fish scenes, overlap and occlusion lead to missing landmarks and incorrect keypoint association. In single-fish cases, curvature and underwater noise cause spatial misalignment and incomplete predictions. These errors propagate to uncertainty in metric length estimation through the calibration–projection pipeline.

4.3 Ablation study on coordinate attention integration

To evaluate the effectiveness of attention mechanisms for underwater fish detection and pose estimation, an extended ablation study was conducted by integrating multiple attention modules —Coordinate Attention (CA), CBAM, and Global Attention Mechanism (GAM)—into the AquaYOLO-Pose architecture. In addition, the internal configuration of CA was examined by varying the reduction ratio r to analyze its influence on localization accuracy and computational efficiency.

In the proposed AquaYOLO-PoseCA design, CA modules were inserted after the two deepest C2f blocks (layers 18 and 21), where high-level semantic representations are formed. Two reduction ratios were considered: the default setting of $r = 32$, which provides lightweight spatial–channel encoding, and a more aggressive configuration of $r = 16$, which increases channel emphasis and feature selectivity. As shown in Table 8, both configurations improve detection and pose performance over the YOLOv8s-Pose baseline, confirming the effectiveness of incorporating coordinate-aware attention for elongated fish structures.

Specifically, the default CA configuration with ($r = 32$) enhances box recall and results in better detection and pose mAP, all without requiring any additional computational cost over the baseline (approximately 29.4 GFLOPs and 11.4 million parameters) (Xu et al., 2024). When the ratio is reduced to $r = 16$ the responses of the features are stronger, while the precision is slightly unstable due to the sensitivity of spatial attention in challenging underwater environments.

To place CA in perspective, we compare it with SE, CBAM, and GAM in Table 8. While SE improves channel-wise feature recalibration, it does not explicitly preserve spatial positional

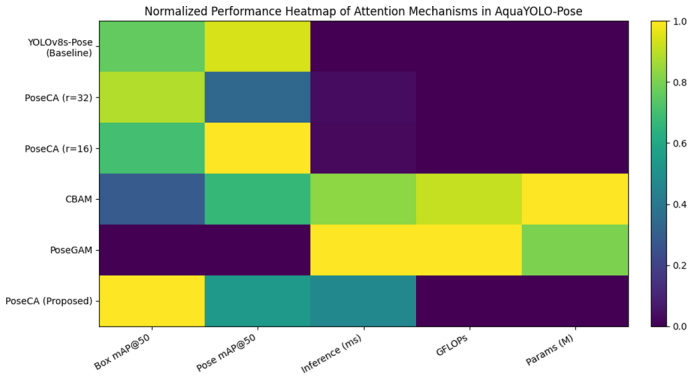


FIGURE 12
Normalized heatmap showing the accuracy–efficiency trade-off of different attention mechanisms in the AquaYOLO-Pose architecture.

information, which is important for accurate localization of elongated fish landmarks. CBAM and GAM enhance global feature representation, but this comes at a high computational cost, with substantially increased GFLOPs and slower inference times. In particular, GAM introduces dense global channel-spatial interactions that may smooth spatial gradients, which is less suitable for the precise localization of fine-grained anatomical keypoints. In contrast, CA preserves directional spatial dependencies along horizontal and vertical axes while remaining lightweight, making it more appropriate for underwater fish pose estimation under turbidity, occlusion, and pose variation. Figure 12 illustrates the trade-off between detection accuracy and computational efficiency for different attention mechanisms. The AquaYOLO-PoseCA model achieves strong box- and pose-level mAP while maintaining low computational cost, as indicated by its reduced GFLOPs and parameter count. In comparison, CBAM- and GAM-based variants require substantially higher computational resources but do not provide corresponding improvements in localization accuracy. This indicates diminishing returns from heavier global attention mechanisms.

Overall, the AquaYOLO-PoseCA proposed framework CA configuration demonstrates the most favorable balance between localization accuracy and computational efficiency. By achieving the highest box and pose detection performance while preserving real-time inference capability. These results support the use of Coordinate Attention as an effective and deployable attention mechanism for underwater fish pose estimation.

4.4 Checkerboard calibration

AquaYOLO-PoseCA and the MorphoCal length module operate on monocular RGB images acquired with the Crosstour CT9000 camera. To recover metric measurements from pixel-space keypoints, we perform a one-time underwater checkerboard calibration under the same optical conditions as the DePondFi'24 dataset (cf. Figure 3). Calibration images were captured from multiple viewpoints, tilts, and distances in order to cover a wide portion of the field of view.

For each of the 25 checkerboard views, inner corners were detected and refined to sub-pixel precision, and the camera intrinsics and distortion parameters (Khrouch et al., 2025) were estimated using a pinhole projection model with radial–tangential distortion. The recovered intrinsic matrix is

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 4594.83 & 0 & 2249.39 \\ 0 & 4595.70 & 1820.60 \\ 0 & 0 & 1 \end{bmatrix},$$

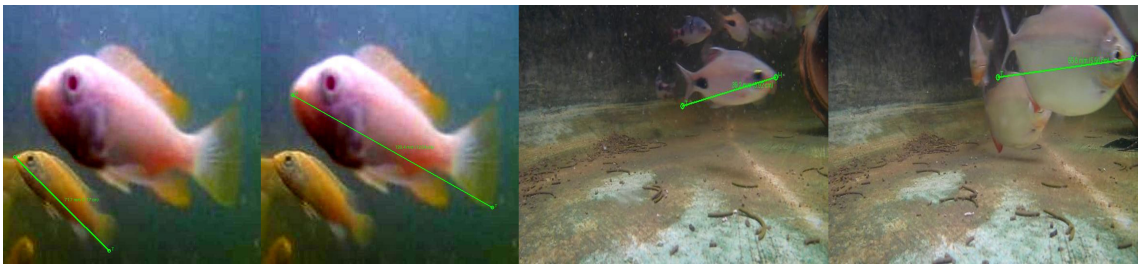


FIGURE 13
Sample outputs from the AquaYOLO-PoseCA + calibration pipeline showing metric fish length estimation across multiple underwater conditions. Each image displays the predicted mouth–tail endpoints (green points) connected by a calibrated length vector.

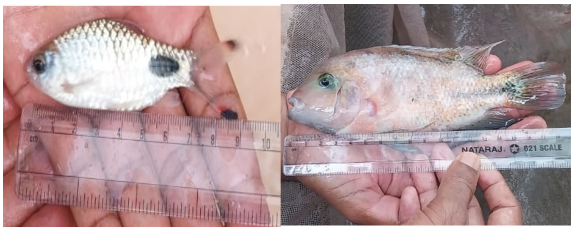


FIGURE 14
Sample ground-truth images with manual ruler measurements used to benchmark metric length estimation.

TABLE 10 **Comparison of length estimation methods.**

Method	MAE (cm)	RMSE (cm)
MiDaS-based depth estimation	1.23	1.58
Proposed MorphoCal	0.7478	0.8638

with distortion coefficients

$$(k_1, k_2, p_1, p_2, k_3) = (-0.2738,\quad 0.4506,\quad -4.21 \times 10^{-4},\quad 1.75 \times 10^{-3},\quad -0.9164)\,.$$

The mean reprojection error over all views is 0.164 px, indicating a well-conditioned calibration suitable for millimetre-scale measurements across the full image plane. The field of view is 53.58° degrees horizontally and 41.47° vertically. This calibrated tuple (**K**, θ, **R**, **t**) is saved once and reused for each inference in AquaYOLO-PoseCA + MorphoCal. This is a constant, globally consistent metric. A representative example of metric fish length estimation is shown in Figure 13.

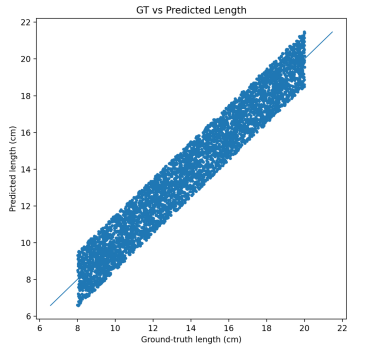
For each frame in the DePondFi'24 test split, AquaYOLO-PoseCA provides head–tail keypoints which are projected by MorphoCal into the calibrated metric plane, yielding centimetre-accurate length estimates. Ground truth lengths are obtained from ruler-referenced images collected during sampling (see Figure 14).

5 Statistical evaluation of fish length estimation

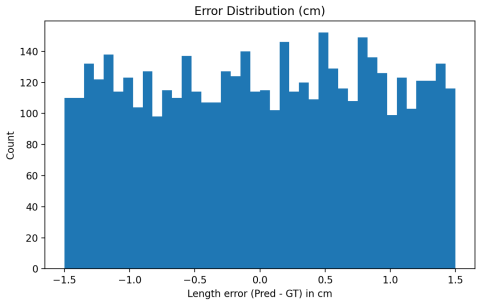
To assess the accuracy and reliability of the proposed MorphoCal framework for fish length estimation, a statistical analysis was performed by comparing predicted lengths with ruler-based ground-truth measurements. The evaluation was conducted on N = 4810 length samples obtained from monocular underwater RGB images captured in real pond environments.

5.1 Quantitative error statistics

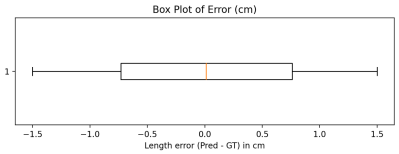
The mean error (bias) between predicted and ground-truth lengths was 0.0089 cm, indicating negligible systematic overestimation or underestimation. This is supported by the 95% confidence interval of the mean error, [− 0.0155, 0.0333] cm, which includes



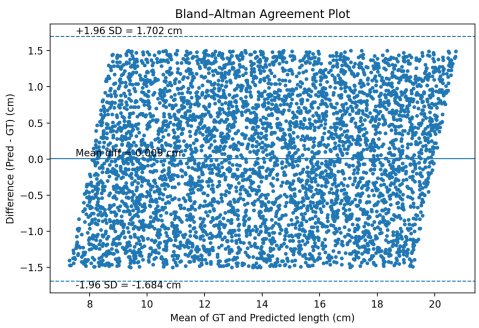
(a) Ground-truth vs predicted length



(b) Histogram of length errors



(c) Box plot of length errors



(d) Bland–Altman agreement plot

FIGURE 15
Statistical evaluation of metric fish length estimation. **(a)** Ground-truth vs predicted length correlation, **(b)** error distribution histogram, **(c)** error spread analysis using box plot, and **(d)** Bland–Altman agreement analysis.

zero. The Mean Absolute Error (MAE) was 0.7478 cm, which shows that the average error is less than a centimeter when the measurement is made in the actual pond environment. The Mean Squared Error (MSE) was 0.7461 cm^2, and the Root Mean Squared Error (RMSE) was 0.8638 cm, showing that the errors are constrained, as summarized in Table 9.

A paired *t*-test was performed to evaluate whether the predicted values are significantly different from the ground-truth measurements. The result of the paired *t*-test is $t = 0.7168$ and $p = 0.4736$ ($p > 0.05$), indicating that there is no statistically significant difference between the predicted and actual values, which reflects a high level of agreement between the proposed monocular estimation approach and manual measurements. For comparison, a MiDaS-based (Ranftl et al., 2022) depth estimation alternative was also implemented and evaluated for fish length reconstruction. MiDaS estimates relative inverse depth from a single RGB image and was originally designed for monocular depth prediction across diverse datasets.The MiDaS-based approach yielded an MAE of 1.23 cm and an RMSE of 1.58 cm, which are higher than those obtained with the proposed checkerboard-calibrated projection framework (MAE = 0.7478 cm, RMSE = 0.8638 cm), as shown in Table 10. This result indicates that the proposed MorphoCal framework provides more accurate and stable metric length estimation under underwater pond conditions.

5.2 Visual analysis and agreement assessment

Visual analysis was also performed to complement the numerical results. In the scatter plot comparing ground truths to predictions in Figure 15a, it can be observed that the plots are aligned well along the identity line. This suggests that the predictions for different sizes of fish are consistent. In Figure 15b, it can be observed that the error plots are concentrated near zero. In Figure 15c, the box plot indicates the median error, the range within which 50% of the errors lie, and the presence of any errors that lie too far from the rest to be considered representative. To further compare the predictions with the ground truths, a Bland-Altman plot was also created for the length predictions, as shown in Figure 15d. In the Bland-Altman plot, it can be observed that the majority of the plots lie within the 95% limit.

From the results, it can be observed that MorphoCal is effective in providing statistically accurate results for the length estimation of a fish from a monocular underwater image. The bias is close to zero with a confidence interval that includes zero, the mean absolute error indicates that the average error is within threshold.

6 Limitations and future work

Estimation of fish length by MorphoCal is accurate, but there are certain practical issues that need to be considered. First, it is important to understand that there is a presumption of stable geometry of the camera in relation to the reference plane. If there is any change in the elevation of the camera, its tilt, or its underwater housing, it could affect the pixel-metric relationship, making it necessary to recalculate the parameters to ensure accurate centimetre-level estimation. The dataset is specific to a certain region, and there is only limited variability in terms of pond conditions. Even though DePondFi'24 includes multiple ornamental and freshwater fish species with variability in turbidity, clutter, and illumination, it is important to validate the framework under different aquaculture conditions, seasons, water types, pond structures, and camera systems. Future work will also consider fish with broader morphological variation and disease-related body changes, such as fin erosion and skeletal deformities, which may affect landmark visibility and length reconstruction accuracy. The future of MorphoCal is focused on ensuring its robustness in different operating conditions, as well as its extension beyond a single two-dimensional reference plane, including three-dimensional reconstruction and stereo vision (Lin et al., 2025) for continuous fish monitoring in dynamic aquaculture environments.

7 Conclusion

In our work, we present the MorphoCal pipeline, which is a multi-stage pipeline that combines our proposed AquaYOLO-PoseCA model with a checkerboard-based calibration mechanism. This results in an accurate fish length estimation in challenging underwater environments. Our proposed model outperforms existing methods by achieving an Average Precision at IoU 0.50 (mAP@50) of 0.959 in bounding box detection and an mAP@50 of 0.848 in detecting poses within bounding boxes, requiring 29.4 GFLOPs and 11.4 million parameters. The calibration mechanism uses a one-time underwater checkerboard-based ray-plane projection method. With our proposed MorphoCal solution, we hope to contribute to the advancement of fish length estimation. The study serves as an early step toward AI-enabled aquaculture systems capable of combining growth analysis with environmental monitoring.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

This study collected data during fishery activities conducted by local fishers, as such, it did not require approval from an animal ethics committee.

Author contributions

VM: Writing – review & editing, Methodology, Writing – original draft, Formal analysis, Conceptualization, Data curation. SA: Validation, Visualization, Supervision, Writing – original draft, Writing – review & editing. SN: Methodology, Validation, Software,

Resources, Writing – review & editing. PP: Writing – review & editing, Project administration, Data curation, Validation.

Funding

The author(s) declared that financial support was received for this work and/or its publication. Open access funding provided by Vellore Institute of Technology.

Conflict of interest

Author SN was employed by the company Deloitte Tohmatsu Deep Square Co., Ltd.

The remaining author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Abinaya, N. S., Susan, D., and Sidharthan, R. K. (2022). Deep learning-based segmental analysis of fish for biomass estimation in an occulted environment. *Comput. Electron. Agric.* 197, 106985. doi: 10.1016/j.compag.2022.106985

Al-Abri, S., Keshvari, S., Al-Rashdi, K., Al-Hmouz, R., and Bourdoucen, H. (2025). Computer vision based approaches for fish monitoring systems: a comprehensive study. *Artif. Intell. Rev.* 58, 185. doi: 10.1007/s10462-025-11180-3

Andrialovanirina, N., Ponton, D., Behivoke, F., Mahafina, J., and Leopold´, M. (2020). A powerful method for measuring fish size of small-scale fishery catches using ImageJ. *Fish. Res.* 223, 105425. doi: 10.1016/j.fishres.2019.105425

Arguel, L., Buisson, L., Baran, P., Nemoz´, M., Blanc, F., De Sauverzac, L., et al. (2025). Relationship between hydrology and the activity of an endangered semi-aquatic mammal, the Pyrenean desman (Galemys pyrenaicus). *Mammalia* 90(2), 164–183. doi: 10.1515/mammalia-2024-0157

Bajpai, A., Tiwari, N., Yadav, A., Chaurasia, D., and Kumar, M. (2024). Enhancing underwater object detection: leveraging YOLOv8m for improved subaquatic monitoring. *SN Comput. Sci.* 5, 793. doi: 10.1007/s42979-024-03170-z

Cao, D., Guo, C., Shi, M., Liu, Y., Fang, Y., Yang, H., et al. (2024). A method for custom measurement of fish dimensions using the improved YOLOv5-keypoint framework with multi-attention mechanisms. *Water Biol. Secur.* 3, 100293. doi: 10.1016/j.watbs.2024.100293

Chen, R., Wang, P., Lin, B., Wang, L., Zeng, X., Hu, X., et al. (2025). An optimized lightweight real-time detection network model for ioT embedded devices. *Sci. Rep.* 15, 3839. doi: 10.1038/s41598-025-88439-w

Chen, Y., Liu, H., and Fang, F. (2024). A novel pansharpening method based on cross stage partial network and transformer. *Sci. Rep.* 14, 12631. doi: 10.1038/s41598-024-63336-w

De Ridder, J., Dujardin, V., Camacho Garcia, J., Sawasawa, W., Aerts, P., Svardal, H., et al. (2025). An alternative pattern of head expansion during feeding in cichlids. *Commun. Biol.* 8, 1448. doi: 10.1038/s42003-025-08851-w

Elmezain, M., Saoud, L. S., Sultan, A., Heshmat, M., Seneviratne, L., and Hussain, I. (2025). Advancing underwater vision: a survey of deep learning models for underwater object recognition and tracking. *IEEE Access*, 13, 17830–17867. doi: 10.1109/access.2025.3534098

Feng, G., Yuan, L., Wang, W., and Chen, M. (2025). Monitoring of fish fry stress behavior based on improved YOLOv8n-pose and boT-SORT. *Modern Fisheries* 52, 31–43. doi: 10.26958/j.cnki.1007-9580.2025.04.003

Hasegawa, T., and Nakano, D. (2024). Robust fish recognition using foundation models toward automatic fish resource management. *J. Mar. Sci. Eng.* 12, 488. doi: 10.3390/jmse12030488

Huo, F., Liu, K., Dong, H., Yang, C., Ren, W., and Li, X. (2025). A novel fisheye calibration device along with calibration approach in terms of a single-shot image. *Measurement* 255, 118008. doi: 10.1016/j.measurement.2025.118008

Hussein, S. K., and Nemer, Z. N. (2025). Improving YOLOv8m with neck-integrated atrous spatial pyramid pooling for enhanced detection of small fish and jellyfish. *Informatica* 49, 33. doi: 10.31449/inf.v49i33.8656

Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., et al. (2023). RTMPose: real-time multi-person pose estimation based on MMPose. *arXiv preprint*. doi: 10.48550/arXiv.2303.07399

Khrouch, H., Mahdaoui, A., Marhraoui Hsaini, A., Merras, M., Chana, I., and Bouazi, A. (2025). Improving camera parameter estimation using an adaptive genetic algorithm. *Signal. Image Video Process.* 19, 113. doi: 10.1007/s11760-024-03604-4

Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., et al. (2022). Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nat. Methods* 19, 496–504. doi: 10.1038/s41592-022-01443-0

Li, X. (2025). Improving deep image feature matching. UK: University of Oxford.

Li, D., and Du, L. (2022). Recent advances of deep learning algorithms for aquacultural machine vision systems with emphasis on fish. *Artif. Intell. Rev.* 55, 4077–4116. doi: 10.1007/s10462-021-10102-3

Li, H., Zheng, R., Jiang, W., Man, X., and Ma, X. (2024). "Fish length estimation based on stereo vision and keypoint detection", in: *2024 36th Chinese Control and Decision Conference (CCDC)* (Xi'an, China: IEEE), 1747–1752.

Liao, D., Zhang, J., Tao, Y., and Xie, J. (2025). ATBHC-YOLO: aggregate transformer and bidirectional hybrid convolution for small object detection. *Complex Intelligent Syst*, 11(1), 38. Shenzhen, China, doi: 10.1007/s40747-024-01652-4

Lin, H., Zhang, H., Huo, J., Li, J., Zhang, H., and Li, Y. (2025). High-precision 3D reconstruction of underwater concrete using integrated line structured light and stereo vision. *Autom. Constr.* 169, 105883. doi: 10.1016/j.autcon.2024.105883

Liu, M., Huang, Y., Xu, G., Zhou, J., Wei, C., Hu, J., et al. (2025). Automated underwater Plectropomus leopardus phenotype measurement through cylinder. *Sci. Rep.* 15, 24461. doi: 10.1038/s41598-025-08863-w

Liu, W., Tan, J., Lan, G., Li, A., Li, D., Zhao, L., et al. (2024). Benchmarking fish dataset and evaluation metric in keypoint detection–towards precise fish morphological assessment in aquaculture breeding. *arXiv preprint arXiv:2405.12476*. doi: 10.48550/arXiv.2405.12476

Liu, L., Wu, J., Zhao, H., Kong, H., Zheng, T., Qu, B., et al. (2025). Fish-Finder: A robust small target detection method for aquaculture fish in low-quality underwater images. *J. Fish Biol.* 106, 908–920. doi: 10.1111/jfb.15992

Mei, J., Hwang, J. N., Romain, S., Rose, C., Moore, B., and Magrane, K. (2021). "Absolute 3D pose estimation and length measurement of severely deformed fish from monocular videos in longline fishing", in: *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Toronto, ON, Canada: IEEE), 2175–2179.

Meng, F., Cong, Z., Wang, B., Guo, K., Li, Q., and Zhang, D. (2025). DeepTurbid: Underwater marker detection and pose estimation in turbid conditions. *Inf. Fusion* 126, 103568. doi: 10.1016/j.inffus.2025.103568

Neelamegam, G., Vinoth Kumar, S., Punitha, P., and Lakshmana Kumar, R. (2025). Enhancing object detection in underwater aquatic systems using a YOLO–transformer hybrid model and ioT sensor integration. *J. Hydroinf.* 27(10), 1467–1492. doi: 10.2166/hydro.2025.288

Pedersen, M., Bengtson, S. H., Gade, R., Madsen, N., and Moeslund, T. B. (2018). "Camera calibration for underwater 3D reconstruction based on ray tracing using Snell's law", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 2018, 1491–14917. doi: 10.1109/CVPRW.2018.00190

Peng, Q., Li, W., Liu, Y., and Li, Z. (2025). High-precision fish pose estimation method based on improved HRNet. *Smart Agric.* 7, 160–172. doi: 10.12133/j.smartag.SA202502001

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., and Koltun, V. (2022). Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 1623–1637. doi: 10.1109/TPAMI.2020.3019967

Rivera, M., Huamanchahua, D., and Flores, C. (2024). "Comparative analysis of state-of-the-art pre-trained human pose estimation models in underwater condition", in: *2024 IEEE Colombian Conference on Communications and Computing (COLCOM)* (Barranquilla, Colombia: IEEE), 1–6.

Rocha, W. S., da Costa Doria, C. R., and Watanabe, C. Y. V. (2020). "Fish detection and measurement based on mask R-CNN", in: Proceedings of the 33rd Conference on Graphics, Patterns and Images (SIBGRAPI), Graduate Works Workshop, Sociedade Brasileira de Computação (SBC), Porto Alegre, Brazil, 183–186.

Sattar, S., Khan, S., Khan, M. I., Akhmediyarova, A., Mamyrbayev, O., Kassymova, D., et al. (2025). Anomaly detection in encrypted network traffic using self-supervised learning. *Sci. Rep.* 15, 26585. doi: 10.1038/s41598-025-08568-0

Scholz, L. A., Mancienne, T., Stednitz, S., Scott, E., and Lee, C. (2025). *Datasets of zebrafish larvae poses – annotated frames and videos* (Melbourne, Victoria, Australia: The University of Melbourne). doi: 10.26188/29276009.v1

Sun, N., Li, Z., Luan, Y., and Du, L. (2024). "Enhanced YOLO-based multi-task network for accurate fish body length measurement", in: *2024 5th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)* (IEEE), Shenzhen, China, 334–339.

Tang, J., Yu, Z., and Shao, C. (2025). Hybrid attention transformer integrated YOLOv8 for fruit ripeness detection. *Sci. Rep.* 15, 22652. doi: 10.1038/s41598-025-04184-0

Tian, Y., Liao, W., Xu, Z., and Wang, W. (2024). "A binocular vision fish body length measurement method based on improved YOLOv8-pose", in: *2024 5th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC)* (Wuhan, China: IEEE), 622–625.

Tiwari, N. K., Bajpai, A., Yadav, S., Bilal, A., Darem, A. A., Sarwar, R., et al. (2025). DM-AECB: a diffusion and attention-enhanced convolutional block for underwater image restoration in autonomous marine systems. *Front. Mar. Sci.* 12, 1687877. doi: 10.3389/fmars.2025.1687877

Vijayalakshmi, M., and Sasithradevi, A. (2025). AquaYOLO: Advanced YOLO-based fish detection for optimized aquaculture pond monitoring. *Sci. Rep.* 15, 6151. doi: 10.1038/s41598-025-89611-y

Whiteway, M., Biderman, D., Pedraja, F., Ehrlich, D., Noone, D., and Sawtell, N. B. (2024). *Lightning Pose dataset: mirror-fish* (London, United Kingdom: figshare). doi: 10.6084/m9.figshare.24993363.v1

Xu, H., Chen, X., Wu, Y., Liao, B., Liu, L., and Zhai, Z. (2024). Using channel pruning–based YOLOv5 deep learning algorithm for accurately counting fish fry in real time. *Aquacult. Int.* 32, 9179–9200. doi: 10.1007/s10499-024-01609-x

Yang, J., Ren, G., Huan, Y., Du, W., Xue, S., Wang, B., et al. (2025). CADY-YOLOv8Pose: a keypoint-based visual inspection framework for secondary fastening detection of transmission bolts. *Measurement* 257, 118692. doi: 10.1016/j.measurement.2025.118692

Yi, J., Han, W., and Lai, F. (2025). YOLOv8n-DDSW: an efficient fish target detection network for dense underwater scenes. *PeerJ Comput. Sci.* 11, e2798. doi: 10.7717/peerj-cs.2798

Yu, Y., Zhang, H., and Yuan, F. (2023). Key point detection method for fish size measurement based on deep learning. *IET Image Proc.* 17, 4142–4158. doi: 10.1049/ipr2.12924

Zhang, H., Guo, Y., Xie, Y., and Zheng, Z. (2025). An integrated method for non-intrusive underwater fish measurement based on keypoint detection and stereo vision. *Aquacult. Int.* 33, 1–28. doi: 10.1007/s10499-025-02276-2

Zhang, T., Xiao, J., Li, L., Wang, C., and Xie, G. (2021). Toward coordination control of multiple fish-like robots: real-time vision-based pose estimation and tracking via deep neural networks. *IEEE/CAA J. Autom. Sin.* 8, 1964–1976. doi: 10.1109/JAS.2021.1004228

Zhao, Y., Qin, H., Xu, L., Yu, H., and Chen, Y. (2024). A review of deep learning-based stereo vision techniques for phenotype feature and behavioral analysis of fish in aquaculture. *Artif. Intell. Rev.* 58, 7. doi: 10.1007/s10462-024-10960-7

Zhu, T., Zhang, W., Miao, Z., Zhao, C., and Xiong, Y. (2025). YOLOv11-SKP: an enhanced model for strawberry bounding box and key point detection in harvesting scenarios. *Eng. Res. Express* 7, 045217. doi: 10.1088/2631-8695/ae0de3