Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

In-situ Volume Estimation Method for Sea Cucumbers Based on Lightweight Binocular Vision and Spatiotemporal Constraints

Yan-qiang Yang
Dalian Ocean University
Ze-jin Liu
Dalian Ocean University
He-ming Zhao
Dalian Ocean University
Rui-qian Zheng
College of Mechanical and Power Engineering, Dalian Ocean University
Xiang Zhong
Dalian Ocean University
Zhi-jun Li
Dalian Ocean University
Yuan-shan Lin
linyuanshan@dlou.edu.cn
Dalian Ocean University

Research Article

Keywords: Sea Cucumber, Binocular Vision, Lightweight Network, Three-dimensional (3D) Reconstruction, Alpha Shape

Posted Date: June 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9719182/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

In-situ Volume Estimation Method for Sea Cucumbers Based on Lightweight Binocular Vision and Spatiotemporal Constraints

Yan-qiang Yang[1,2], Ze-jin Liu[1,2], He-ming Zhao[1,2], Rui-qian Zheng[3,4], Xiang Zhong[1,2], Zhi-jun Li[1,2,*], Yuan-shan Lin[1,2,*]

Abstract

Sea cucumber body volume is a key indicator for precision management and yield estimation in marine ranching. Conventional contact-based measurement methods often induce stress-related contraction and evisceration, leading to significant measurement bias. Meanwhile, existing non-contact visual methods face two main challenges for in-situ deployment: high computational complexity that strains underwater edge device constraints, and the limitations of implicit reconstruction algorithms in characterizing non-convex papillae features. To address the trade-off between measurement accuracy and limited edge computing resources, this paper proposes a lightweight binocular vision volume estimation framework. The framework optimizes the measurement pipeline through three key modules: (1) At the perception level, an asymmetric decoupled instance segmentation network (YOLO11-ED-Head) is designed to reduce cross-channel computational redundancy while maintaining accurate mask extraction. (2) At the reconstruction level, an explicit Alpha Shape topological algorithm is introduced to help preserve non-convex manifold features and mitigate volumetric inflation. (3) At the decision level, a spatial-temporal filtering pipeline combining ROI gating and 70% quantile statistics is constructed to reduce the impact of unreliable observations caused by dynamic curling and occlusion. Experimental results on a lightweight surrogate platform demonstrate that the proposed system achieves an average processing speed of 16.8 FPS, with a Mean Absolute Percentage Error (MAPE) of 9.8% and a Root Mean Square Error (RMSE) of 6.2 cm^3. This study provides a computationally efficient framework, demonstrating applicable potential for future edge-computing deployment in the automated yield measurement of benthic organisms.

Keywords: Sea Cucumber; Binocular Vision; Lightweight Network; Three-dimensional (3D) Reconstruction; Alpha Shape

1. Introduction

As high-value benthic organisms in marine ranching, sea cucumbers require accurate body volume measurement as a core phenotypic indicator for precision aquaculture management, feed optimization, and yield estimation. For decades, sea cucumber growth monitoring has relied on manual diving capture, with volume quantified via contact-based methods such as water displacement. This approach is not only

✉ Yuanshan Lin
linyuanshan@dlou.edu.cn
✉ Zhijun Li
lizhijun@dlou.edu.cn

[1] College of Information Engineering, Dalian Ocean University, Dalian 116023, Liaoning, China;
[2] Dalian Smart Fisheries Laboratory, Dalian 116023, Liaoning, China;
[3] College of Mechanical and Power Engineering, Dalian Ocean University, Dalian 116203, Liaoning, China;
[4] Liaoning Provincial Innovation Center for Marine Fishery Equipment Technology, Dalian 116023, Liaoning, China;

labor-intensive and high-risk, but also prone to severe measurement bias: as stress-sensitive soft-bodied organisms, sea cucumbers frequently undergo body contraction or even evisceration during manual handling and air exposure (Harper et al., 2020, Liu et al., 2024a). Accordingly, developing non-contact, low-stress in-situ visual measurement methods has become an urgent demand for automated aquaculture production (Santos et al., 2025).

Early non-contact measurement studies predominantly used binocular vision to acquire target depth information, and estimated volume by simplifying aquatic organisms into regular geometric models (e.g., ellipsoids, cylinders) (Muñoz-Benavent et al., 2018). However, sea cucumbers exhibit non-convex geometry with surface papillae, which are difficult to be accurately fitted by regular models, leading to large measurement errors especially during postural changes (Jongjaraunsuk et al., 2025).

To address the morphological representation limitations of regular models, subsequent studies adopted deep learning-based instance segmentation to extract organism contours. Networks such as Mask R-CNN and early YOLO series can acquire target 2D boundaries via mask extraction and map them to 3D point clouds. However, sea cucumbers' mimicry camouflage makes them visually similar to background sediment and aquatic weeds; in low-contrast, high-noise underwater environments, conventional models frequently produce blurred boundaries and incomplete masks. Furthermore, discrete 3D point clouds from mask mapping are difficult to be directly used for closed volume calculation without surface topology reconstruction.

Recent studies have addressed background interference and closed volume calculation via cascaded large-model architectures, using the Segment Anything Model (SAM) for complex background segmentation and Poisson Reconstruction for watertight surface generation from discrete point clouds (Kazhdan et al., 2006, Kirillov et al., 2023, Zhen et al., 2024). While these schemes achieve high geometric accuracy in static tests, they face critical limitations for edge deployment in dynamic underwater video streams (Fan et al., 2024). First, at the perception stage, Transformer-based architectures like SAM have computational complexity scaling quadratically with image resolution, leading to notably high inference latency on resource-constrained edge devices (Cai et al., 2025). Second, at the reconstruction stage, implicit algorithms like Poisson Reconstruction require iterative sparse linear system solving, and produce two key errors for incomplete underwater point clouds (Ren et al., 2023): global smoothing erases surface papillae details causing volume underestimation, while implicit surface closure properties induce surface expansion at point cloud holes, both exacerbating geometric distortion. Third, most existing schemes rely on single-frame static measurement, lacking temporal constraint mechanisms to mitigate dynamic interference from occlusion and target curling (Xie et al., 2024).

To mitigate the fundamental trade-off between high-precision in-situ measurement and strict edge computing resource constraints, this paper proposes a lightweight binocular vision volume estimation framework optimized for underwater edge deployment. Unlike conventional paradigms that improve accuracy via increasing model complexity, we conduct targeted optimization of the three core pipeline stages to reduce full-link computational overhead while preserving geometric topology fidelity:

(1) To address the computational bottleneck of large models and the limited feature extraction capability of conventional lightweight networks, we design the YOLO11-ED-Head instance segmentation network with a group convolution front-end backbone and asymmetric decoupled branches. This design reduces redundant cross-channel computational redundancy, reduces inference latency, and maintains high geometric consistency mask extraction for non-convex benthic organisms.

(2) To overcome the high latency and detail loss of implicit reconstruction algorithms, we replace implicit surface fitting with an explicit Alpha Shape topological algorithm. Leveraging local simplex

connectivity constraints, this method preserves high-frequency non-convex surface features, mitigating volume underestimation caused by over-smoothing.

(3) To mitigate the asymmetric negatively skewed error distribution in underwater observations, we develop a joint pipeline of spatial ROI gating and global 70th percentile temporal filtering. This mechanism excludes truncated invalid frames at the field-of-view edge, reduces measurement errors from target curling and dynamic occlusion, and improves system stability in real underwater environments.

2. Related Work

2.1. Underwater Object Segmentation for In-situ Visual Measurement

In aquaculture and agricultural vision, deep learning-based instance segmentation has been widely adopted for biomass assessment. Representative architectures, including Mask R-CNN (He et al., 2017) and conventional YOLO-series segmentation models (Jocher et al., 2023), have achieved reliable performance in fish body length measurement and disease detection (Hu et al., 2021). Recently, Transformer-based large models (e.g., SAM) have also been introduced into complex underwater scenarios to address benthic organisms' mimicry camouflage and background interference (Niu et al., 2024), delivering high mask accuracy in offline static image processing with high-performance computing support.

However, direct deployment of these models on underwater Remotely Operated Vehicles (ROVs) for in-situ observation faces inherent latency bottlenecks in dynamic marine environments. To counteract ROV drift and target displacement caused by benthic currents, the system requires high-frame-rate continuous target capture. Existing high-performance segmentation models mostly adopt parameter-intensive symmetric standard convolution for all feature channels in the prediction head, which introduces prohibitive computational load and single-frame inference latency on resource-constrained edge devices. This latency causes spatial position calculations to lag behind the actual physical state of targets (Jiang et al., 2025, Huang et al., 2025, Yu et al., 2025). Therefore, breaking the traditional symmetric prediction structure to reduce edge computational load via feature recombination or grouped dimensionality reduction, while maintaining spatial perception capability for complex benthic organisms, is a core prerequisite for low-latency continuous underwater observation.

2.2. 3D Reconstruction Algorithms for Non-convex Biological Morphologies

3D reconstruction is the core approach for spatial geometric information acquisition in plant and animal phenotypic analysis and volume measurement. Early agricultural measurement studies mostly used bounding box approximation or voxel grid partitioning, which are suitable for volume estimation of regular rigid agricultural products (e.g., apples, potatoes) (Gongal et al., 2015). For complex organisms, recent studies have adopted point cloud-based surface reconstruction, combining instance segmentation masks and binocular stereo matching to acquire the 3D morphology of underwater organisms (Zhu et al., 2025). Representative implicit methods include Poisson Surface Reconstruction and the Marching Cubes algorithm (Lorensen and Cline, 1987), which can generate watertight smooth meshes via global implicit function solving for rigid structures or smooth-surfaced organisms.

However, these global implicit function-based algorithms face critical geometric applicability challenges for in-situ 3D mapping of sea cucumbers in turbid water. Sea cucumbers have body surfaces densely covered with non-convex papillae protrusions, and the acquired point clouds are often

incomplete with local sparsity and discrete holes, limited by underwater optical scattering and local stereo matching failures (Hao et al., 2022). When processing such data, conventional Poisson reconstruction's global smoothing constraints enforce forced surface closure across holes, which erroneously encloses surrounding water into the mesh and erases non-convex papillae features, leading to significant deviation between measured volume and true physical boundaries. Therefore, abandoning global implicit smoothing and exploring a topological reconstruction mechanism that enables explicit pure geometric extraction based on the local connectivity of underwater sparse point clouds is critical to reduce morphological distortion of flexible benthic organisms.

2.3. Target Tracking and Global Temporal Filtering in Dynamic Video Streams

In continuous dynamic visual measurement, single-frame static observations are vulnerable to ambient noise-induced fluctuations. Traditional temporal processing strategies, including multi-frame arithmetic mean filtering, Kalman filtering, and optical flow-based inter-frame smoothing, deliver reliable performance for Gaussian-distributed random noise or continuous trajectory tracking of rigid targets (Li et al., 2025).

However, in real seabed marine ranch environments, dynamic measurement errors of benthic organisms exhibit non-Gaussian and non-linear characteristics. During a tracking cycle, sea cucumbers may remain in a stress-induced curled state or suffer long-term partial occlusion by aquatic weeds for most of the observation period, generating a large number of continuous underestimated frames and forming an asymmetric negatively skewed distribution. Conversely, sporadic floating aquatic weed adhesion may trigger over-segmentation and isolated volume peak values (Hao et al., 2024, Liu et al., 2024b, Sun et al., 2019). In this scenario, traditional mean filtering is pulled down by long-term underestimation intervals, while maximum value strategies are misled by sporadic anomalies. Accordingly, there is an urgent need for a robust temporal mechanism coupling spatial gating constraints and non-parametric statistical aggregation to stably extract the true physical volume of targets from continuous data streams with highly dynamic interference and asymmetric deviations (Wang et al., 2025).

3. Methodology

3.1. Overview of the In-situ Volume Measurement System

To achieve non-contact, high-precision in-situ volume measurement of sea cucumbers in complex and dynamic marine ranch environments, this paper designs an end-to-end dynamic monitoring pipeline based on binocular vision. As shown in Fig. 1, the processing pipeline of the system consists of five steps:

(1) Image Acquisition: The raw left and right views are acquired by the ZED 2i binocular camera, which serve as the input data stream of the system.

(2) Target Segmentation & Tracking: The input images are first processed by the YOLO11-ED-Head lightweight network. This module is responsible for extracting the 2D bounding boxes and instance masks of sea cucumbers under low power consumption, to realize spatial localization and category classification of the targets.

(3) Spatial Filtering: An ROI triggering mechanism is introduced to perform spatial screening on the detection results. Only when the target center enters the measurement area will the system activate the subsequent high-load computing modules; otherwise, the system remains in a low-power cruise tracking state.

(4) 3D Reconstruction & Calculation: Once the measurement is triggered, the system uses a stereo

matching algorithm to obtain the point cloud of the masked region, and adopts the Alpha Shape algorithm to perform explicit topological reconstruction, to restore the non-convex 3D mesh with geometric fidelity from discrete data.

(5) Temporal Filtering & Final Output: The instantaneous volume is calculated according to the reconstructed mesh and stored in the temporal buffer pool. Finally, the global quantile filtering strategy is used to eliminate dynamic noise, and the smooth and robust 3D volume estimation value of the sea cucumber is output as the final result.

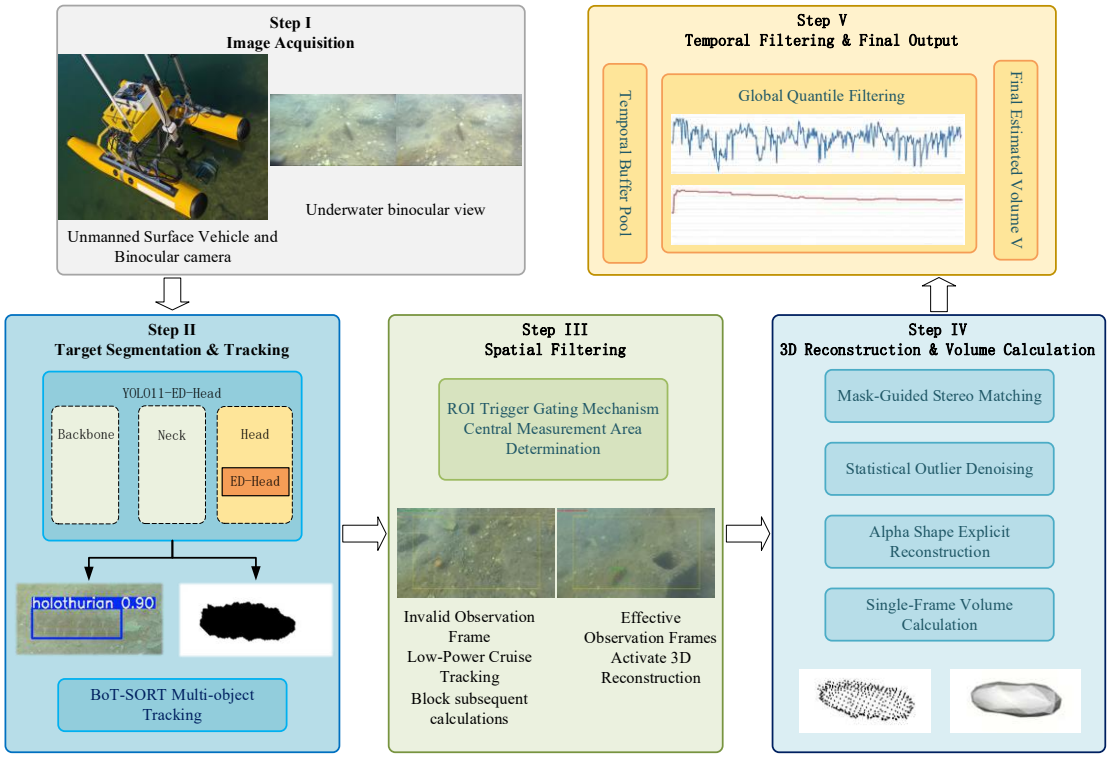


Fig. 1. Overall architecture and algorithm pipeline of the in-situ volume measurement system

3.2. Lightweight Instance Segmentation Network via ED-Head

In intelligent marine ranch monitoring, non-contact in-situ volume estimation of non-rigid benthic organisms (e.g., sea cucumbers) relies critically on high-accuracy target mask extraction from the front-end vision model. To meet the processing demands of continuous underwater video streams, this paper proposes a lightweight instance segmentation architecture with a novel Efficient-Decoupled Head (ED-Head).

3.2.1. Baseline Architecture Selection for In-situ Measurement Tasks

High-precision underwater in-situ measurement systems require continuous processing of visual data from high-resolution binocular depth cameras. For mainstream two-stage detection-segmentation large models, their underlying self-attention mechanism introduces quadratic computational complexity $O(N^2)$ with respect to the input image spatial resolution N. For high-resolution binocular video stream processing, this non-linear computational overhead causes severe video memory overflow and excessive inference latency, making these models difficult for direct deployment on resource-constrained edge platforms. To address this, this study abandons the bulky two-stage paradigm and selects the single-stage end-to-end YOLO11-seg as the baseline architecture, with the optimized overall network structure shown in Fig. 2. This architecture extracts multi-scale features via a shared fully convolutional backbone, and

outputs target spatial coordinates and prototype mask coefficients in parallel at the network head. This pure convolutional unidirectional data flow strictly constrains the overall computational complexity to linear order $O(N)$, proportional to spatial resolution. While meeting the input requirements of high-resolution binocular images, this design reserves adequate computing headroom for subsequent multi-task parallel processing.

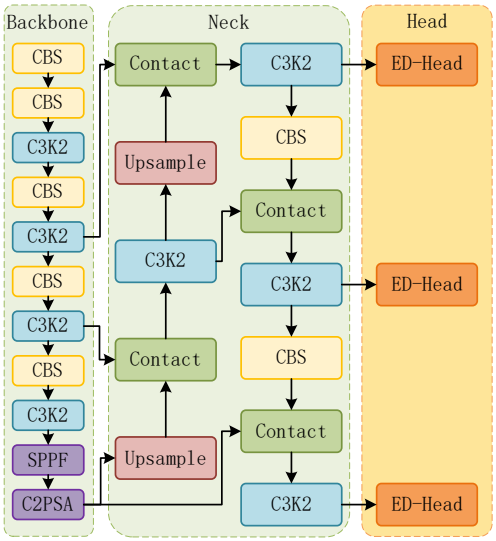


Fig. 2. Underwater lightweight instance segmentation network architecture based on YOLO11-ED-Head

3.2.2. Structural Design of the Efficient-Decoupled Head (ED-Head)

The baseline model's native multi-task decoupled head suffers from suboptimal computational resource allocation for our specific task. Its detection head adopts a symmetric convolution structure for classification, bounding box regression, and mask segmentation branches. However, sea cucumber in-situ observation is a single-class detection task with inherently low classification complexity, while mask segmentation requires high spatial feature resolution to capture non-convex papillae features in turbid underwater environments. The symmetric structure causes notably parameter redundancy in the classification branch, and independent standard convolution for feature extraction across all branches introduces unnecessary cross-channel computational overhead. To resolve this architecture-task complexity mismatch, we propose the ED-Head via asymmetric reconstruction of the native head, with two core innovations as follows:

(1) Group Convolution-based Front-end Feature Fusion Backbone

In the native decoupled head, spatial and channel mapping is performed independently by each branch after feature splitting, which increases back-end inference computational load. To address this, we front-load spatial feature extraction and design a front-end fusion stem (GroupConv Stem). As shown in the dashed box in Fig. 3, the feature map output from the Neck layer first feeds into the Stem module, which consists of two cascaded 3×3 group convolutions. The group number hyperparameter is fixed at $g = 16$ to accommodate input channel variations across different feature scales (P3, P4, P5). This module extracts local spatial texture features of sea cucumber body surfaces via group convolution, reduces redundant cross-channel interaction computation of standard convolution, and cuts the theoretical parameter count of the front-end layer.

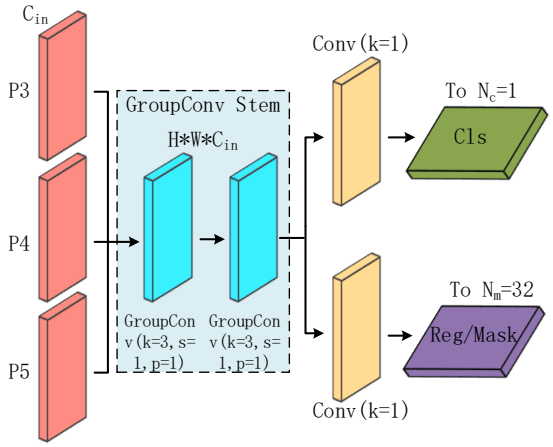


Fig. 3. ED-Head introducing the GroupConv Stem to achieve pre-feature fusion and asymmetric lightweight decoupling

(2) Asymmetric Lightweight Decoupled Branches

After Stem layer processing, features are split into classification and regression/mask branches. Since core spatial features have already been extracted by the front-end Stem, we replace the original 3×3 standard convolution in decoupled branches with 1×1 point-wise convolution for channel dimensionality reduction mapping. Specifically, an asymmetric channel configuration is adopted tailored to task characteristics: the classification branch's feature dimension is streamlined to directly map to the single target class $N_{cls} = 1$, while the regression/mask branch outputs bounding box localization parameters and a 32-dimensional prototype mask coefficient vector. Detailed tensor mapping parameters are provided in Table 1. This asymmetric structure reduces computational consumption of the low-complexity classification task, preserves the feature dimension required for high-precision mask extraction while cutting overall parameter count, and provides a reliable geometric foundation for subsequent binocular point cloud-based 3D volume measurement.

Table 1 Core structural parameter configuration of YOLO11-ED-Head at different scales.

Branch	Operator Type	Stride	Input Channel (C_{in})	Output Channel (C_{out})
P3/P4/P5 Input	-	-	64/128/256	-
GroupConv Stem	3×3 Conv	1	C_{in}	C_{in}
Cls Branch	1×1 Conv	1	C_{in}	1
Reg/Mask Branch	1×1 Conv	1	C_{in}	64/32

3.3. 3D Surface Reconstruction and Volume Estimation for Non-convex Biological Morphologies

Using the 2D mask output from the ED-Head, the system back-projects the target into 3D space via the binocular depth camera's disparity map, generating a discrete point cloud of the sea cucumber's local surface. Converting this discrete point cloud into a watertight continuous surface for volume integration is the final step to complete in-situ measurement.

3.3.1. Binocular Point Cloud Back-projection and Spatial Denoising

Limited by optical scattering in turbid underwater media and inherent deviations in binocular stereo matching, the initial back-projected 3D point cloud often contains mismatched points from suspended solids and discrete outlier points at depth map edges (Makarov et al., 2025). To ensure geometric fidelity

of subsequent topological reconstruction and reduce noise-induced volume calculation errors, we design a strict spatial cleaning pipeline prior to 3D meshing. First, we implement absolute depth filtering based on the camera's effective working range, using the physical observation boundary $[Z_{min}, Z_{max}]$ to eliminate invalid background points that are excessively close or distant. Subsequently, a Statistical Outlier Removal (SOR) filter is applied: by setting a reasonable number of neighboring evaluation points k and standard deviation multiplier threshold σ_{th}, the algorithm dynamically removes sparse discrete suspended impurity points (Rusu et al., 2008). This two-step pipeline extracts the high signal-to-noise ratio main point cloud of sea cucumbers from complex marine backgrounds, providing a reliable data foundation for subsequent explicit surface reconstruction.

3.3.2. Local Topology Extraction and Robust Estimation via Alpha Shape

Using the denoised non-convex main point cloud, we introduce the explicit Alpha Shape topological algorithm (Edelsbrunner et al., 1983). Unlike implicit methods relying on global continuity assumptions, this algorithm generates an envelope surface via local simplex constraints. In our in-situ measurement pipeline, we achieve dual geometric optimization by tuning the key control parameter α as a spatially adaptive filter:

(1) Non-convex Morphology Reconstruction with Adaptive Parameter Setting

The virtual rolling ball radius α directly determines topological reconstruction performance. Instead of a fixed empirical threshold, we establish an adaptive mathematical mapping between α and the spatial sampling density of the binocular point cloud. Let d_{avg} be the average nearest neighbor distance of the target local point cloud, and $d_{papillae}$ be the anatomical spacing of real sea cucumber papillae. We constrain α within the interval $1.5 \cdot d_{avg} \leq \alpha \leq d_{papillae}$. Under this adaptive constraint, the reconstructed surface effectively bridges basic point cloud holes while accurately fitting gaps between papillae. This explicit local connectivity strategy restores high-frequency undulations of the target's non-convex manifold, reducing volume overestimation errors common in traditional bounding box or implicit reconstruction algorithms.

(2) Topological Isolation of Underwater Sparse Noise

For free-floating impurity points in the marine ranch environment, the α threshold enables spatial topological isolation. When the spatial distance between an impurity point and the sea cucumber main body exceeds the connectivity radius α, the algorithm does not generate simplex connections between them. This mechanism effectively isolates the topological correlation between local noise and the target main body, mitigating underwater dynamic interference via a purely geometric approach during the meshing stage.

After constructing the watertight purely geometric mesh of the sea cucumber's non-convex body surface via Alpha Shape, we calculate the instantaneous single-frame volume using the discrete signed tetrahedron summation method based on the Divergence Theorem. For a closed surface composed of N triangular facets, let the spatial vectors of the three vertices of any facet i (following the right-hand rule) be V_{i1}, V_{i2}, V_{i3}, respectively. The instantaneous single-frame volume V_t of the sea cucumber is expressed as the algebraic sum of the signed volumes of all discrete tetrahedrons, with the calculation shown in Equation (1):

$$V_t = \sum_{i=1}^{N} \frac{1}{6} (V_{i1} \cdot (V_{i2} \times V_{i3}) \quad (1)$$

As single-frame V_t inevitably contains measurement noise induced by underwater dynamic interference, this time series is directly fed into the subsequent global temporal aggregation filter to

output the final robust volume estimation.

3.4. Spatiotemporal Constraints and Dynamic Filtering Strategy for Continuous Video Streams

In real marine ranch in-situ detection, single-frame-based 3D reconstruction faces two core limitations: it cannot reliably reflect the true volume of sea cucumbers under dynamic interference, and full-computation per frame easily causes computational overload on resource-constrained edge devices. To address these issues, this paper designs a joint constraint mechanism consisting of cross-frame multi-target association, spatial ROI gating, and global temporal asymmetric filtering. This mechanism dynamically allocates computing resources while stably extracting the true physical state of targets from fluctuating video streams.

3.4.1. Cross-frame Multi-target Association via Tracking-by-Detection

To handle simultaneous multi-target appearance, partial occlusion, and long-term target identity maintenance (Wang et al., 2022), the system adopts the Tracking-by-Detection paradigm downstream of the instance segmentation module, integrating the lightweight BoT-SORT tracking algorithm (Aharon et al., 2022).

In seabed environments, highly variable illumination and homogeneous apparent texture of conspecific sea cucumbers cause severe performance degradation of conventional appearance re-identification (ReID) modules for benthic organisms, while also introducing redundant computational overhead on edge devices. Accordingly, we disable the general-purpose ReID branch in the native tracking architecture, and instead rely on the Kalman filter (Kalman, 1960) for target motion trajectory prediction, combined with 2D bounding box Intersection over Union (IoU) for data association. This mechanism assigns a unique tracking ID to each independent sea cucumber in the field of view (FoV), which serves as the unique index for subsequent trajectory-level volume aggregation: all valid single-frame volumes $V_{i,t}$ belonging to the same ID_i are pushed into the same temporal buffer pool, effectively avoiding no cross-contamination of volume observation sequences between different individuals.

3.4.2. On-demand Computation Scheduling with ROI Gating and Morphological Integrity Constraints

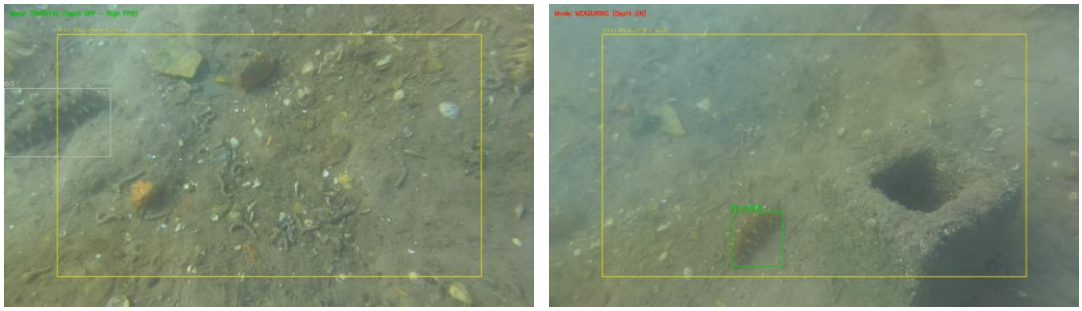
Limited by the fixed field of view of underwater cameras and the motion trajectory of unmanned surface vessels (USV), sea cucumber targets typically enter the frame from the edge before reaching the optical center. Targets at the FoV edge suffer from two critical observation defects: incomplete body truncation, and vulnerability to lens radial distortion interference. Forcibly performing high-load stereo matching and point cloud reconstruction under these suboptimal conditions not only produces geometrically distorted 3D morphologies, but also increases communication bandwidth and computational load on resource-constrained edge devices.

To effectively filters these observation defects and realize on-demand computing power scheduling, we design a strict spatial gating constraint based on a central rectangular region of interest (ROI). For a binocular camera input with original resolution $W \times H$, the system defines a safe bounding box at the FoV center, with vertex coordinates bounded as $[0.1W, 0.1H, 0.9W, 0.9H]$. We also introduce a mandatory physical truncation judgment rule: a target is classified as out-of-FoV truncated if its 2D bounding box touches the absolute physical edge of the image. During operation, the tracking module outputs the 2D bounding box coordinate set of target i in frame t, $B_{i,t} = \{x_{min}, y_{min}, x_{max}, y_{max}\}$.

Based on the inclusion relationship between $B_{i,t}$ and the central ROI, the system dynamically switches between two operating states: low-power cruise tracking and high-load precision measurement.

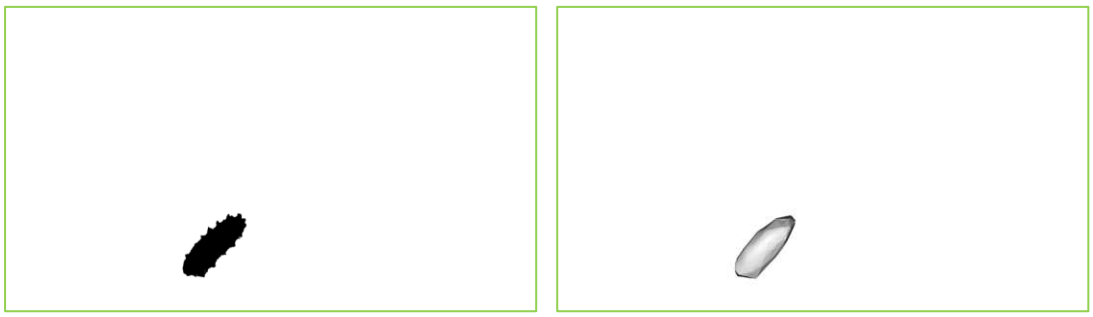
As shown in Fig. 4(a), when $B_{i,t}$ is not fully contained within the central ROI or triggers the edge truncation rule, the frame is classified as an invalid observation. The pipeline only maintains low-power 2D perception and tracking, and suspends subsequent point cloud generation and meshing processes to save computing resources. Conversely, if and only if $B_{i,t}$ is completely enclosed in the central ROI, the system identifies the optimal observation window and activates the high-load 3D reconstruction pipeline (Fig. 4(b)): it outputs a high geometric consistency 2D mask, performs stereo matching to extract the target 3D point cloud (Fig. 4(c)), and completes explicit 3D topological mesh reconstruction via the Alpha Shape algorithm (Fig. 4(d)).

This coordinate topology-based on-demand scheduling strategy isolates geometric noise from truncation and distortion at the algorithmic spatial level while acquiring high-quality 3D morphologies. It also reduce computing power congestion caused by invalid observations at the underlying architecture, ensuring smooth and stable operation of the edge detection pipeline at the target frame rate.



(a) Target partially outside the ROI (Cruise Mode: 2D bounding box only, no 3D computation)

(b) Target inside the ROI triggering the Measure Mode (2D instance mask generated)



(c) 3D point cloud of the target extracted via stereo matching

(d) 3D topological mesh model reconstructed using the Alpha Shape algorithm

Fig. 4. Visualization of the spatial gating mechanism based on ROI triggering and the subsequent 3D reconstruction process.

3.4.3. Asymmetric Aggregation Strategy for Negative-skewed Noise Mitigation

While the spatial gating mechanism guarantees single-frame reconstruction quality, the system still needs to mitigate inter-frame dynamic observation errors. Unlike the symmetric Gaussian error distribution assumed in conventional visual measurement, the actual underwater dynamic observation errors of sea cucumbers exhibit a distinct asymmetric negatively skewed distribution. Physiologically, as flexible non-rigid organisms, sea cucumbers frequently undergo body curling during seabed crawling,

while swaying aquatic weeds often cause transient visual occlusion. These physical changes lead to frequent underestimation of the instantaneous measured volume V_t, where the probability of $V_t < V_{true}$ is significantly higher than that of overestimation.

For such asymmetric long-tailed underestimation noise, conventional temporal aggregation strategies have critical limitations: global mean filtering is heavily biased downward by abundant low-value distorted frames, leading to systematic volume underestimation; while maximum filtering is overly sensitive to single-frame over-segmentation outliers caused by sporadic underwater specular highlights, resulting in amplified measurement errors. To address this distribution characteristic, we abandon conventional symmetric filters and design an adaptive temporal aggregation strategy based on global quantiles. For a valid observation sequence collected in measurement mode $V_{ROI} = \{V_1, V_2, \dots, V_n\}$, we define the q quantile of this empirical distribution as the final robust volume estimation V_{final}, with the calculation detailed in Equation (2):

$$V_{final} = \inf\{v \in V_{ROI} : F_X(v) \geq q\} \tag{2}$$

Where V_{final} (unit: cm^3) is the final robust volume estimation of the target; inf denotes the infimum operator; V_{ROI} is the valid volume observation sequence collected within the central ROI; $F_X(v) = P(X \leq v)$ is the cumulative distribution function (CDF) of the volume measurement random variable X; and q is the quantile threshold. This asymmetric temporal filtering enables the system to output a geometrically consistent volume estimation from fluctuating dynamic data streams.

The proposed spatiotemporal constraint strategy realizes on-demand computing power scheduling in the spatial dimension, alleviating bandwidth and power consumption waste caused by conventional full-time computation. In the temporal dimension, it achieves effective suppression of asymmetric negatively skewed noise, reducing systematic volume underestimation induced by dynamic occlusion.

In summary, the complete trajectory-level in-situ sea cucumber volume estimation pipeline proposed in this paper is detailed in Algorithm 1.

Algorithm 1.

Trajectory-level In-situ volume estimation pipeline for sea cucumbers.

Input: Continuous binocular video stream sequence S= $\{((I_t^L, I_t^R)\}_{t=1}^T$, binocular camera parameters, ROI boundary parameters $[0.1W, 0.1H, 0.9W, 0.9H]$, optimized topological reconstruction parameter α, quantile threshold $q = 0.7$

Output: Final volume prediction value $V_{final}^{(i)}$ or each independent sea cucumber target in the field of view

Initialize target trajectory volume buffer pool $\mathcal{B} = \{\}$
for each frame $t \in [1, T]$ **do**
  $(B_t, M_t) \leftarrow$YOLO11-ED-Head (I_t^L)
  $ID_t \leftarrow$BoT-SORT(B_t)
  for each target $i \in ID_t$ in the current frame **do**
    if ($B_{i,t}$ is completely within the ROI) **and** (does not trigger image edge physical truncation) **then**
      $P_{i,t} \leftarrow$StereoMatching $(I_t^L, I_t^R, M_{i,t})$
      $P_{i,t} \leftarrow$StatisticalOutlierRemoval $(P_{i,t}, k = 50, std = 1.0)$ //Point cloud denoising
      $Mesh_{i,t} \leftarrow$AlphaShape $(P'_{i,t}, \alpha)$ //Explicit surface reconstruction

$V_{i,t}$ ←CalculateVolume $(Mesh_{i,t})$
$\mathcal{B}[i].append(V_{i,t})$ //Store into the temporal observation buffer pool of this ID_i
end if
end for
end for
for each target ID_i in the buffer pool $\mathcal{B}$ **do**
$V_{final}^{(i)}$ ←Quantile($\mathcal{B}[i], q = 0.7$) //Extract the 70th quantile as the final volume
end for
return V_{final}

4. Results and Discussion

4.1. Experimental Setup

4.1.1. Dataset Construction and Preprocessing

To verify the robustness of the proposed model in real underwater operating scenarios, the dataset used in this study was acquired at the Xinyulong National Marine Ranch in Dalian, China. Data were collected using a ZED 2i binocular camera mounted on the hull bottom of an unmanned surface vessel, with an average water depth of 1 m during acquisition. Prior to field data collection, the camera was calibrated offline using a standard underwater checkerboard to correct radial distortion and depth scale drift. The dataset covers various typical complex underwater environmental conditions, including dynamic light spots from natural light refraction, water turbidity induced by wind waves and propeller disturbance, sediment with color characteristics similar to sea cucumbers, and aquatic weed occlusion.

The original dataset comprises 5570 wild sea cucumber images with a resolution of 1920×1080. To avoid data leakage, the dataset is split at the video sequence level into a training set (4456 images), validation set (557 images), and test set (557 images) in an 8:1:1 ratio. During the perception model training stage, the network learns features from real wild sea cucumber images, with online data augmentation strategies including Mosaic, HSV color jitter, and spatial geometric transformation applied to enhance the model's generalization ability under complex underwater working conditions.

4.1.2. Implementation Details and Experimental Platform

All algorithmic models were constructed based on the PyTorch deep learning framework and Ultralytics library. A "training-inference separation" hierarchical hardware strategy was adopted to balance training efficiency and edge deployment evaluation authenticity. All models were trained on a high-performance workstation equipped with an NVIDIA GeForce RTX 4090 (24GB) and Intel Core i7-13700KF to ensure rapid convergence on large-scale datasets. Model debugging, latency testing, and throughput analysis were migrated to a lower-resource platform with an NVIDIA GeForce GTX 1650. Although its computing power exceeds that of target embedded devices, its limited memory bandwidth and computing units can sensitively reflect the impact of model complexity on inference speed. Thus, relative latency and FLOPs changes on this platform were used as key proxy metrics for edge migration potential evaluation.

For network training, all models were uniformly trained for 100 epochs with a batch size of 32. For the model optimization process, the initial learning rate was set to 0.01, momentum factor of 0.937, and weight decay coefficient of 0.0005 to prevent overfitting. To ensure objectivity of inference speed and

latency evaluation, all speed measurement experiments excluded underlying scheduling interference: all comparative models were run at the same input tensor resolution with serial inference, simulating the real workload of underwater devices processing monocular video streams and ensuring the reported time consumption can objectively reflect the complexity of the algorithm itself.

For system-level end-to-end volume measurement evaluation, live sea cucumbers are highly prone to severe stress-induced contraction and evisceration during traditional water displacement measurement, making it difficult to obtain absolute ground-truth volumes. Accordingly, 23 high-fidelity sea cucumber models with known precise physical volumes were customized as the quantitative evaluation benchmark. During testing, these models were placed in an underwater environment with a real sediment background, and continuous binocular video streams were recorded under varying distances, viewing angles, and flow field disturbances.

The rationality of this "real-scene training, high-fidelity model testing" cross-validation strategy is based on two points. The customized bionic models are highly consistent with real sea cucumbers in color bias, texture features, and non-convex papillae morphology in underwater optical imaging, allowing the YOLO11-ED-Head trained on real datasets to directly generalize and accurately extract 2D masks. Back-end binocular stereo matching and explicit 3D topological reconstruction are pure physical-geometric mapping processes; given accurate front-end masks, back-end 3D point cloud generation and tetrahedral volume integration have mathematical equivalence between physical models and live organisms.

Thus, quantitative errors from this dynamic test set reduce physiological deviations of live organisms and rigorously reflect the real measurement accuracy of the proposed joint pipeline in actual in-situ environments.

4.1.3. Evaluation Metrics

To comprehensively evaluate the end-to-end performance of the proposed system, this study adopts three categories of evaluation metrics, covering segmentation accuracy, volume estimation precision, and edge deployment efficiency. For instance segmentation performance in turbid underwater environments, we strictly follow the COCO standard evaluation protocol (Lin et al., 2014), adopting Precision, Recall, and mean Average Precision (mAP@0.5) as core metrics to comprehensively assess the model's ability to balance suppression of underwater background artifacts and avoidance of missed detection caused by incomplete morphological capture. For in-situ sea cucumber volume measurement accuracy, Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) are adopted as the core evaluation metrics. For statistical consistency, for each test sample, the system first collects valid single-frame volume estimates fully within the spatial ROI gating from its corresponding continuous video trajectory, then outputs the final predicted volume of the independent trajectory via the global quantile temporal aggregation strategy, with the specific calculation formulas shown in Equations (3) and (4).

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{V_{gt}^{(i)} - V_{pred}^{(i)}}{V_{gt}^{(i)}}\right| \times 100\% \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(V_{gt}^{(i)} - V_{pred}^{(i)}\right)^2} \tag{4}$$

Here, N is the number of test samples, $V_{gt}^{(i)}$ is the ground-truth physical volume of the i-th

sample calibrated via the standard water displacement method, and $V_{pred}^{(i)}$ is the volume prediction of the i-th sample from the proposed system. MAPE reflects the relative error level of system measurement, while RMSE (unit: cm^3) reflects the absolute deviation scale of measurement results. To objectively reflect the system's stability in continuous observation and reduce accidental deviations from single measurements, all reported volume measurement metrics are calculated based on multiple independent observation sequences, and results are presented as "mean $\pm$ standard deviation" unless otherwise specified, to quantify the system's confidence interval and error dispersion across different individuals and dynamic underwater environments. To evaluate the lightweight characteristics and practical edge deployment potential of the algorithm under the strict computing power, memory, and power consumption constraints of marine ranch in-situ detection edge devices, three additional efficiency metrics are adopted: Params (unit: M) characterizes the overall storage overhead of network weights, directly reflecting the model's memory occupation on embedded terminals; FLOPs (unit: G) quantifies the theoretical computational complexity of a single model forward inference, serving as the core theoretical benchmark for evaluating the network's computational load, memory access cost, and underlying energy consumption; FPS quantifies the dynamic throughput of the system when processing continuous binocular video streams, providing a key physical reference for judging the deployment feasibility of the algorithm pipeline for underwater in-situ detection.

4.2. Effectiveness Analysis of the Target Detection and Segmentation Network

4.2.1. Lightweight and Segmentation Performance Validation of the ED-Head Architecture

To evaluate the migration potential of the proposed model to resource-constrained edge computing platforms, Table 2 compares the computational resource consumption (Params, FLOPs) and core instance-level mask segmentation performance (Precision, Recall, mAP@0.5) of the mainstream lightweight instance segmentation baseline YOLOv8n-seg, native YOLO11n-seg, and the proposed YOLO11-ED-Head.

Table 2 Comparison of computational load and accuracy of different target segmentation models

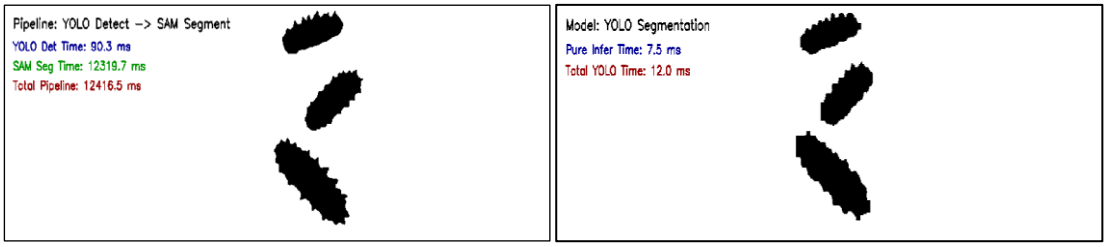
Model	Precision (%)	Recall (%)	Params (M)	FLOPs (G)	mAP@0.5 (%)
YOLOv8n-seg (Ultralytics, 2024)	93.8	76.5	3.26	12.0	85.9
YOLO11n-seg (Khanam and Hussain, 2024)	94	78.3	2.83	10.2	88.8
YOLO11-ED-Head	94.2	80	2.56	9.0	91.5

Experimental results show that the native YOLO11n-seg achieves concurrent optimization in parameter volume and segmentation accuracy compared with the earlier YOLOv8n-seg, but its mask mAP@0.5 of 88.8% still cannot meet the accuracy requirements of in-situ measurement given the challenges of underwater unstructured environments and high visual similarity between targets and backgrounds. On this basis, the proposed YOLO11-ED-Head achieves balanced optimization in computational offloading and accuracy maintenance: by introducing a group convolution-based front-end stem before multi-task splitting and adopting full 1×1 convolution for asymmetric mapping in the classification and regression/mask branches, ED-Head further compresses the model's total parameter count to 2.56M and reduces computational complexity to 9.0G, corresponding to a 9.5% and 11.8% reduction compared with the native YOLO11-seg, respectively. More critically, while achieving lightweight model design, it reaches mask Precision to 94.2% and Recall of 80.0%, with the core metric mAP@0.5 optimized to 91.5%. From the perspective of practical underwater visual measurement

requirements, the improved Precision enables the system to accurately anchor the real non-convex physical boundary of sea cucumbers and reduces false positive interference caused by misclassifying aquatic weeds or similar reefs as foreground, while the enhanced Recall means the model mitigates incomplete segmentation and missed detection of sea cucumber targets under turbid water flow and sediment occlusion. This performance improvement is attributed to the efficient aggregation of weak texture features of the sea cucumber's non-convex surface topology by the front-end GroupConv Stem, as well as the effective mitigation of feature space conflicts by the asymmetric decoupling mechanism. This quantitative result validates the previous theoretical expectation: the topological reconstruction of ED-Head not only demonstrates deployment potential adapting to the strict computing power constraints of underwater edge terminals, but also maintains stable edge extraction capability in complex turbid environments, providing a reliable data foundation for subsequent 3D point cloud reconstruction.

4.2.2. Comprehensive Evaluation of Mask Extraction Quality and Deployment Efficiency in Complex Underwater Scenarios

Following the validation of underlying computational overhead differences between models, this section conducts a combined qualitative and quantitative evaluation of mask extraction performance and inference latency between the baseline method and the proposed YOLO11-ED-Head on real underwater images, with actual mask rendering results and time pipeline monitoring of the two models processing the same sea cucumber target presented in Fig. 5. As a large-scale vision foundation model, the SAM adopted in the baseline scheme achieves high zero-shot edge fitting accuracy and pixel-level segmentation performance for the complex papillae structure on the sea cucumber body surface (Fig. 5(a)), but this high-precision detail extraction comes at the cost of severely reduced system operating speed for practical continuous dynamic monitoring on underwater unmanned surface vessels. Performance monitoring data show that the single-frame segmentation latency of SAM reaches approximately 12319.7ms, resulting in a total perception pipeline latency of 12.3s. This prohibitive computational overhead makes the baseline scheme unable to meet the processing requirements of dynamic video streams, notably restricting its practical engineering deployment potential. In contrast, the proposed YOLO11-ED-Head achieves a superior accuracy-efficiency trade-off. As shown in Fig. 5(b), benefiting from the decoupled head mechanism that enables independent parsing of geometric boundaries and semantic features, ED-Head consistently outputs high-quality masks with smooth boundaries, high foreground cohesion, and a tight fit to the topological contour of the sea cucumber main body. While the lightweight ED-Head shows a slight smooth approximation in the pixel-level depiction of individual tiny papillae compared with the large-parameter SAM model, this minor difference does not introduce significant interference to the subsequent overall 3D geometric volume calculation via the Alpha Shape algorithm. More critically, YOLO11-ED-Head compresses the total single-frame perception and inference latency to 12.0ms, achieving a notable increase in processing speed over the baseline scheme. This efficient inference capability, while maintaining high segmentation quality, effectively mitigates latency interference caused by underwater motion blur, reserves sufficient computing headroom for subsequent 3D point cloud processing, and effectively verifies the edge deployment potential of the proposed scheme for practical underwater in-situ measurement tasks.



(a) Baseline scheme YOLOv8+SAM (b) Proposed scheme YOLO11-ED-Head

Fig. 5. Comparison of time pipeline and mask extraction performance of different segmentation models on edge devices.

4.3. Effectiveness Analysis of the 3D Topology Reconstruction Algorithm

4.3.1. Quantitative Evaluation of 3D Reconstruction Algorithms

To comprehensively identify the optimal 3D topological representation scheme for unstructured underwater scenarios, this section conducts comparative experiments under a unified perception baseline. Specifically, to strictly control variables, all compared reconstruction algorithms adopt the 2D masks output by the proposed YOLO11-ED-Head as uniform input. Five representative 3D reconstruction strategies are selected for head-to-head comparison, including the conventional spatial bounding box method, implicit methods represented by Poisson Reconstruction and Voxel-based Marching Cubes (Voxel-MC), and explicit methods represented by Greedy Projection Triangulation (GPT) and Ball Pivoting Algorithm (BPA), with quantitative results summarized in Table 3.

Table 3 Comparison of measurement error and time consumption of different 3D reconstruction algorithms.

Reconstruction Algorithm	MAPE (%)	RMSE (cm^3)	耗时 (ms)
Bounding Box	16.8±10.6	10.6±6.2	0.09
Poisson Recon	27.1±28.1	20.8±15.4	224.88
Voxel-MC	18.5±14.7	13.1±9.5	24.65
GPT	15.2±14.2	9.9±8.6	16.38
BPA	21.8±14.8	14.6±9.1	13.24
Alpha Shape (Ours)	12.8±9.2	8.5±4.4	12.73

Quantitative volume measurement results reveal significant performance discrepancies of different topological reconstruction mechanisms in handling non-convex biological manifolds. As shown in Table 3, the conventional bounding box proxy model achieves a MAPE of 16.8±10.6%, demonstrating that approximating non-rigid benthic organisms with simple regular geometries are difficult to meet the accuracy requirements of complex in-situ measurement scenarios. Notably, while Poisson Reconstruction is widely recognized for superior surface fitting performance in ideal conditions, it yields a MAPE as high as 27.1±28.1% and RMSE of 20.8±15.4 cm^3 in the dynamic turbid video streams tested in this study. The large error standard deviation validates the previously mentioned bidirectional error mechanism: its global low-pass smoothing effect erases high-frequency papillae details on the sea cucumber surface, causing local volume underestimation, while the inherent closed-form property of its implicit surface easily induces spatial outward expansion distortion when processing underwater point clouds with sparse holes or outlier noise. Meanwhile, its single-frame inference latency of 224.88ms further limits its deployment potential on resource-constrained edge devices.

In addition, the lightweight implicit Voxel-MC scheme reduces inference time to 24.65ms, but introduces a "staircase effect" limited by voxel resolution that smooths out the organism's real topological

curvature. The explicit GPT algorithm, constrained by its local projection mechanism, is prone to generating geometric holes and mesh fold deformation in sparse point cloud regions. In contrast, the Alpha Shape algorithm introduced in this paper exhibits favorable geometric adaptability and robustness to dynamic underwater point clouds. Its local simplex extraction mechanism effectively isolates topological interference from outlier impurities during meshing, while allowing moderate concavity of the reconstructed surface to highly preserve sea cucumbers' real non-convex papillae contours. Ultimately, this algorithm achieves the lowest MAPE of 12.8±9.2% and RMSE of 8.5±4.4 cm^3 among all compared methods, with an ultra-low inference latency of 12.73ms, demonstrating favorable migration feasibility to underwater edge computing platforms. This result indicates that for volume measurement of unstructured underwater targets, preserving high-frequency local topology and isolating sparse noise can achieve better measurement consistency than pursuing global surface smoothness alone.

4.3.2. Visual Comparison of Reconstructed Topologies

To more intuitively analyze the geometric restoration accuracy of different algorithms when processing high-noise underwater point clouds, Fig. 6 presents a visual comparison of 3D mesh entities generated by the five algorithms under the same single-frame scattered point cloud input (Fig. 6(a)).

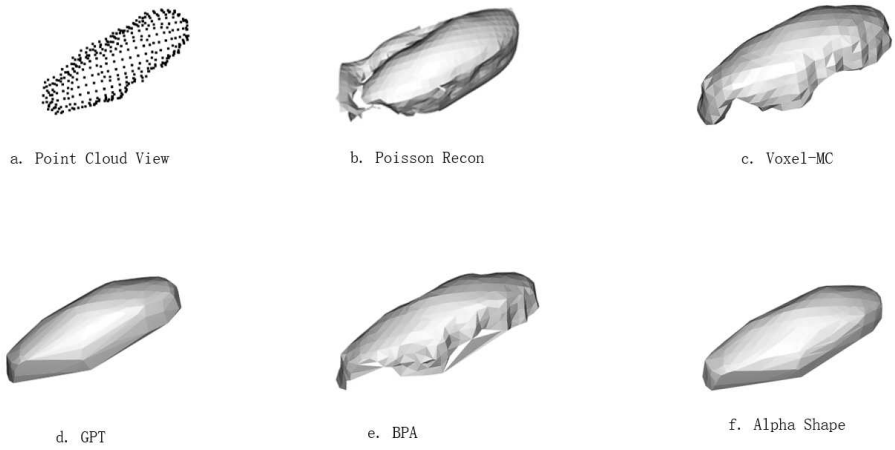


Fig. 6 Visual comparison of topological meshes from different 3D reconstruction algorithms.

As illustrated in Fig. 6, the Poisson Reconstruction algorithm—previously optimal in ideal-condition benchmark tests—exhibits significant topological distortion when applied to real dynamic sparse underwater point clouds. While its reconstructed mesh (Fig. 6(b)) has a smooth surface, it suffers from abnormal excessive expansion deformation: to enforce implicit surface closure, the algorithm encloses surrounding water noise and voids into the mesh, creating a "water expansion" effect that shows notable deviation from the natural morphological characteristics of sea cucumbers.

Other comparative algorithms also show poor adaptability to unstructured underwater environments: the implicit Voxel-based Marching Cubes (Voxel-MC, Fig. 6(c)) produces a "staircase effect" on the mesh surface due to its spatial discretization mechanism, losing the smooth transitional contours of the sea cucumber body; the explicit Greedy Projection Triangulation (GPT, Fig. 6(d)) generates overly polygonal and angular meshes, discarding the soft non-convex fold details of sea cucumbers due to excessive polygonization; and the classic Ball Pivoting Algorithm (BPA, Fig. 6(e)) is sensitive to local point cloud density distribution, leading to irregular local defects and unclosed holes at the model bottom.

In contrast, the Alpha Shape algorithm introduced in this paper demonstrates favorable geometric

adaptability. It effectively suppresses the "water expansion" phenomenon, tightly shrinks to fit the physical boundary of the input point cloud, and preserves both the overall natural trunk contour and local papillae details of sea cucumbers via its simplex-based spatial isolation mechanism. This high-quality solid mesh reconstruction mitigates the interference of high-frequency scattered underwater noise on volume integration from the perspective of explicit topological modeling, visually and quantitatively verifying the effectiveness and performance advantages of the explicit topological shrinking strategy for underwater in-situ measurement tasks.

4.4. Temporal Volume Filtering and Robustness Analysis

In dynamic underwater measurement streams, single-frame images are frequently affected by multiple disturbances including targets moving into or out of the field of view, body curling, and local occlusion. To guarantee the system's final measurement accuracy in long-duration video streams, this paper proposes a joint constraint mechanism combining spatial gating and global temporal filtering.

4.4.1. Design Principle and Morphological Integrity Constraint of Spatial Gating Mechanism

The effective perception range of an underwater binocular vision system is constrained by the camera's physical FoV. In real measurement scenarios, when a sea cucumber target moves into or out of the frame, its body structure is inevitably truncated by the image boundary. Without effective spatial boundary constraints, the system will feed these incomplete edge targets directly into the 3D reconstruction pipeline, where the Alpha Shape algorithm will forcibly close the surface on incomplete point clouds, directly causing the output instantaneous volume to be lower than the target's true physical volume. In addition, targets at the FoV edge are also vulnerable to optical interference from lens radial distortion, further aggravating geometric distortion.

To reduce these physical observation defects, this paper designs a region of interest (ROI)-based spatial gating mechanism at the front end of the measurement pipeline. The system strictly defines the measurement trigger region as a fixed rectangular bounding box covering 64% of the central FoV. During continuous observation, only when the 2D mask bounding box of a sea cucumber target is fully contained within this central ROI will the system mark it as a valid observation frame and activate volume measurement. This spatial filtering strategy effectively filters the input of truncated and notably distorted frames, guarantees the morphological integrity of the underlying point cloud fed into the 3D reconstruction module, and provides a reliable geometric benchmark for subsequent temporal aggregation.

4.4.2. Global Temporal Filtering Strategy and Asymmetric Noise Suppression

Experimental results show that system performance has obvious sensitivity to the temporal quantile threshold (Fig. 7), where the primary axis (left) denotes Mean Absolute Percentage Error (MAPE, dark blue solid line) and the secondary axis (right) denotes Root Mean Square Error (RMSE, orange solid line). Without temporal aggregation (RAW group), the system's MAPE reaches 12.8% due to underwater dynamic noise interference. After introducing quantile filtering, the overall error presents a typical nonlinear U-shaped trend with increasing threshold: low-to-medium thresholds are easily pulled down by volume-underestimated frames caused by terrain occlusion or target curling, leading to negative systematic bias; while high thresholds (e.g., 90th percentile) are easily misled by occasional over-segmentation frames induced by suspended solids or specular highlights, resulting in significant volume overestimation, with MAPE rising to 19.2% and RMSE reaching 19.5 cm^3.

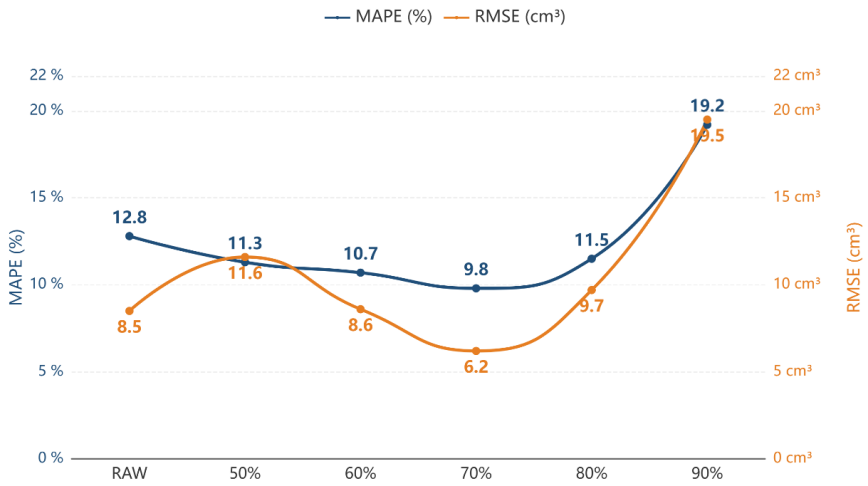


Fig. 7. Error sensitivity curve of the global temporal filtering quantile threshold.

The curve fluctuation trend clearly shows that the 70th percentile achieves an optimal statistical balance between avoiding underestimation caused by long-term curling and isolating occasional abnormal overestimation. Under this threshold constraint, the system's MAPE drops to the lowest value of 9.8% among all groups, with RMSE optimized to 6.2 cm^3.

In addition, the effectiveness of the system's temporal filtering relies on the continuous and stable observation of the front-end BoT-SORT target tracking module. In dynamic video stream verification on test models, the average continuous tracking trajectory length of a single target in measurement mode can cover the target's complete observation window within the ROI, with only a few brief ID switches caused by severe aquatic weed occlusion observed throughout the process. This indicates that the system has stable trajectory maintenance capability in conventional underwater scenarios. However, as trajectory-level volume estimation strongly depends on continuous observation results, frequent tracking interruptions will lead to incomplete observation frame sets, further affecting the accuracy of subsequent global quantile filtering and final volume aggregation. Future research can break away from the dependence on traditional appearance features, and explore lightweight multimodal association operators fusing the local geometric topology and kinematic laws of sea cucumbers, to address the challenge of long-term identity maintenance in scenarios with high-density sea cucumber aggregation or severe cross-occlusion.

4.5. Overall System Performance Comparison and Ablation Analysis

Sections 4.2 to 4.4 have independently validated the effectiveness of each core module across three dimensions: image perception, topological reconstruction, and temporal decision-making. To further quantify the synergistic effects of these optimized modules and the end-to-end performance of the full measurement pipeline, this section conducts system-level ablation experiments on a unified lightweight proxy computing platform. Starting from the mainstream cascaded baseline scheme widely used in existing studies, we sequentially embed the modules proposed in this paper with strict single-variable control, focusing on the system's final measurement accuracy (MAPE, RMSE) and operating efficiency in real dynamic underwater video streams, with quantitative results for different pipeline configurations summarized in Table 4.

Table 4 Comparison of system-level ablation experiments for dynamic underwater scenarios.

System Pipeline	MAPE (%)	RMSE (cm^3)	Avg FPS
Baseline (YOLOv8+SAM+Poisson Recon)	27.1±28.1	20.8±15.4	0.09
YOLO11-ED-Head+Poisson Recon	26.8±27.5	19.5±14.2	1.2
YOLO11-ED-Head+ Alpha Shape	12.8±9.2	8.5±4.4	9.7
Ours (YOLO11-ED-Head+Alpha Shape+ROI&70%)	9.8±4.9	6.2±3.7	16.8

The baseline cascaded scheme (YOLOv8 + SAM + Poisson Reconstruction) was first tested, with results showing that it achieved an average processing frame rate of only 0.09 FPS in dynamic underwater scenarios, accompanied by high overall measurement error due to underwater noise interference. In the first-stage ablation, replacing the large-scale visual segmentation model SAM with the proposed YOLO11-ED-Head increased the overall operating frame rate to 1.2 FPS, while maintaining roughly consistent measurement accuracy (MAPE = 26.8%, RMSE = 19.5 cm^3). This system-level performance aligns with the model-level evaluation conclusions in Section 4.2, confirming that ED-Head effectively alleviates the computational bottleneck of edge deployment without significant degradation of mask extraction quality.

Building on the speed improvement of the 2D perception module, the second-stage ablation replaced the implicit Poisson reconstruction algorithm with the explicit Alpha Shape algorithm. Results show that this module delivered a not bad optimization in system accuracy, with MAPE dropping from 26.8% to 12.8% and RMSE optimized from 19.5 cm^3 to 8.5 cm^3. This error convergence trend echoes the theoretical analysis in Section 4.3, as the explicit geometric shrinking property of Alpha Shape suppresses the "water expansion" phenomenon caused by incomplete underwater point clouds and reduces morphological distortion. Meanwhile, by alleviating the computational overhead of solving global implicit functions, the system's average frame rate was further increased to 9.7 FPS.

Following the lightweight optimization of the perception and reconstruction modules, the spatial ROI gating mechanism and global 70th percentile temporal filtering were introduced in the final ablation stage. As shown in Table 4, the full proposed system achieves a stable average frame rate of 16.8 FPS, with MAPE reduced to 9.8% and RMSE reaching 6.2 cm^3. This result validates the analysis in Section 4.4: the global 70th percentile strategy effectively filters low-value noise caused by aquatic weed occlusion and sea cucumber curling, while the ROI gating mechanism avoids unnecessary 3D reconstruction of invalid targets at the FoV edge and cuts unnecessary computational consumption.

In summary, the progressive ablation experiments quantify the independent contribution of each optimized module to system accuracy and operating efficiency, and verify the effectiveness of the proposed measurement pipeline in balancing measurement precision and edge computational cost. With a mean absolute percentage error of 9.8% and a processing speed of 16.8 FPS, the final system provides a feasible edge-side technical solution for continuous underwater in-situ observation of benthic organisms in marine ranching.

5. Conclusion

This study developed an in-situ volume estimation system for sea cucumbers, aiming to address the computational bottleneck between high-fidelity 3D measurement and processing requirements. Instead of relying on highly complex models, this research achieved an effective system balance through targeted spatiotemporal constraints and pipeline reconstruction. The systematic evaluation of the proposed framework yields the following core conclusions: First, the paradigm of asymmetric resource allocation is well-suited for single-class benthic organism detection. By reducing classification parameter

redundancy, the system substantially decreases inference latency while maintaining reliable mask boundaries in turbid environments. Second, compared to traditional implicit surface fitting, explicit topological constraints (Alpha Shape) demonstrate favorable geometric adaptability for non-convex biological targets, helping to mitigate the "water-body inflation" phenomenon caused by sparse underwater point clouds. Finally, statistical filtering based on the 70% quantile mechanism provides a practical mathematical approach to dynamic measurement noise, alleviating the asymmetric errors caused by biological stress curling and weed occlusion. Running at 16.8 FPS with an overall MAPE of 9.8%, the proposed framework serves as a computationally efficient algorithmic foundation for the automated yield measurement of benthic organisms. However, a current limitation of this trajectory-level measurement lies in its dependency on continuous observation; tracking interruptions caused by severe occlusion in high-density scenarios can still disrupt the temporal sampling distribution. Future work will explore lightweight multimodal association operators fusing targets' geometric topology and kinematic features, and introduce multi-view collaboration to improve multi-target identity maintenance and long-term measurement stability in complex seabed environments.

Author contribution All authors have read and agreed to the published version of the manuscript.

Funding The Special Fund for Basic Scientific Research Business of Provincial Universities in Liaoning (2024JBZDZ004); 2023 Central Government Subsidy Project for Fisheries in Liaoning; Liaoning Province Key Research and Development Program (2023JH26/10200015); Liaoning Province Natural Science Foundation Funding Plan (2020-KF-12-09); Liaoning Provincial Department of Education Basic Scientific Research Project (LJKZ0730, QL202016); Liaoning Province Applied Basic Research Program Project (2022JH2/101300187); Liaoning Province Science and Technology Joint Plan Project Key Technology Research and Application of Multi modal Data Fusion for Sea Cucumber Production Prediction (2024JH2/102600083)

Data availability The dataset generated and/or analyzed during this study has not yet been publicly available but can be obtained from the author upon reasonable request.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Aharon, N., Orfaig, R., Bobrovsky, B.Z., 2022. BoT-SORT: Robust associations multi-pedestrian tracking. arXiv preprint. https://doi.org/10.48550/arXiv.2206.14651.

Cai, S., Zhou, X., Cai, W., Wei, L., Mo, Y., 2025. Lightweight underwater object detection method based on multi-scale edge information selection. Sci. Rep.

Edelsbrunner, H., Kirkpatrick, D., Seidel, R., 1983. On the shape of a set of points in the plane. IEEE Trans. Inf. Theory 29 (4), 551–559. https://doi.org/10.1109/TIT.1983.1056714.

Fan, Y., Zhang, L., Li, P., 2024. A lightweight model of underwater object detection based on YOLOv8n for an edge computing platform. J. Mar. Sci. Eng. 12 (5), 697. https://doi.org/10.3390/jmse12050697.

Gongal, A., Amatya, S., Karkee, M., Zhang, Q., Lewis, K., 2015. Sensors and systems for fruit detection and localization: A review. Comput. Electron. Agric. 116, 8–19. https://doi.org/10.1016/j.compag.2015.05.021.

Hao, H., Yu, J., Yin, L., Chen, G., Zhang, S., Zhou, H., 2022. An improved PointNet++ point cloud segmentation model applied to automatic measurement method of pig body size. Comput. Electron. Agric. 205, 107560. https://doi.org/10.1016/j.compag.2022.107560.

Hao, Y., et al., 2024. Underwater swimming fish mass estimation based on binocular vision. Comput. Electron. Agric.

Harper, D.L., Bethoney, N.D., Stokesbury, K.D., Lundy, M., McLean, M.F., Stokesbury, M.J., 2020. Standard methods for the collection of morphometric data for the commercially fished sea cucumber Cucumaria frondosa in eastern Canada. J. Shellfish Res. 39 (2), 481–489. https://doi.org/10.2983/035.039.0229.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. Proc. IEEE Int. Conf. Comput. Vis., 2961–2969. https://doi.org/10.1109/ICCV.2017.322.

Hu, J., Zhao, D., Zhang, Y., Zhou, C., Chen, W., 2021. Real-time detection of uneaten feed pellets in underwater images for aquaculture using an improved YOLO-V4 network. Comput. Electron. Agric. 185, 106135. https://doi.org/10.1016/j.compag.2021.106135.

Huang, Y., Huang, J., Huang, M., 2025. A lightweight YOLO network for robotic underwater biological detection. Front. Mar. Sci. 12, 1673437. https://doi.org/10.3389/fmars.2025.1673437.

Jiang, H., Zhao, J., Ma, F., Yang, Y., Yi, R., 2025. Mobile-YOLO: A lightweight object detection algorithm for four categories of aquatic organisms. Fishes 10 (7), 348. https://doi.org/10.3390/fishes10070348.

Jocher, G., Chaurasia, A., Qiu, J., 2023. Ultralytics YOLO. https://github.com/ultralytics/ultralytics.

Jongjaraunsuk, R., Rermdumri, S., Khaodon, K., Intarachart, A., Taparhudee, W., 2025. Underwater weight estimation of three sea cucumber species in culture tanks using image analysis and ArUco markers. Animals 15 (8), 1121. https://doi.org/10.3390/ani15081121.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. J. Basic Eng. 82 (1), 35–45. https://doi.org/10.1115/1.3662552.

Kazhdan, M., Bolitho, M., Hoppe, H., 2006. Poisson surface reconstruction. Proc. Eurogr. Symp. Geom. Process. 7 (4), 61–70.

Khanam, R., Hussain, M., 2024. An overview of the key architectural enhancements in YOLOv11. arXiv preprint. https://doi.org/10.48550/arXiv.2410.17725.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Girshick, R., 2023. Segment anything. Proc. IEEE/CVF Int. Conf. Comput. Vis., 4015–4026. https://doi.org/10.1109/ICCV51070.2023.00371.

Li, M., et al., 2025. A modified adaptive Kalman filter algorithm for the distributed underwater multi-target passive tracking system. JASA Express Lett. 5 (1), 016001. https://doi.org/10.1121/10.0034764.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context. Eur. Conf. Comput. Vis., 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

Liu, S., Li, K., Ding, Y., Xu, K., Jiang, Q., Wu, Q.M.J., Song, D., 2024a. Towards the in-situ trunk identification and length measurement of sea cucumbers via Bézier curve modelling. arXiv preprint. https://doi.org/10.48550/arXiv.2406.13951.

Liu, S., et al., 2024b. WaterMask: Instance segmentation for underwater imagery. ResearchGate preprint. DOI: 10.1109/ICCV51070.2023.00126

Lorensen, W.E., Cline, H.E., 1987. Marching cubes: A high resolution 3D surface construction algorithm. ACM SIGGRAPH Comput. Graph. 21 (4), 163–169. https://doi.org/10.1145/37402.37422.

Makarov, F., et al., 2025. 3D point cloud verification and scatter removal using a dual camera/projector setup. Sensors 25. https://doi.org/10.3390/s25237250.

Muñoz-Benavent, P., Andreu-García, G., Valiente-González, J.M., Atienza-Vanacloig, V., Puig-Pons, V., Espinosa, V., 2018. Enhanced fish bending model for automatic tuna sizing using computer vision. Comput. Electron. Agric. 150, 52–61. https://doi.org/10.1016/j.compag.2018.04.005.

Niu, F., Sheng, Y., Wang, J., Li, G., 2024. Domain generalization for sea cucumber detection: Tackling background color variability in aquaculture settings. Aquacult. Int. https://doi.org/10.1007/s10499-025-02022-8.

Ren, J., Pan, L., Liu, Z., 2023. GeoUDF: Surface reconstruction from 3D point clouds via geometry-guided distance representation. Proc. IEEE/CVF Int. Conf. Comput. Vis., 14323–14333. https://doi.org/10.1109/ICCV51070.2023.01307.

Rusu, R.B., Marton, Z.C., Blodow, N., Dolha, M., Beetz, M., 2008. Towards 3D point cloud based object maps for household environments. Rob. Auton. Syst. 56 (11), 927–941. https://doi.org/10.1016/j.robot.2008.08.005.

Santos, A.E., et al., 2025. Evolving challenges and innovations in sea cucumber aquaculture. Aquacult. Int. 33 (1), 326–345. https://doi.org/10.1007/s10499-025-01988-9.

Sun, J., et al., 2019. Plasticity in the antipredator behavior of the orange-footed sea cucumber under shifting hydrodynamic forces. Behav. Ecol. https://doi.org/10.1093/cz/zoy100.

Ultralytics, 2024. Ultralytics YOLO11. https://docs.ultralytics.com/models/yolo11/ (accessed 21 October 2024).

Wang, H., Zhang, S., Zhao, S., Wang, Q., Li, D., Zhao, R., 2022. Real-time detection and tracking of fish abnormal behavior based on improved YOLOv5 and DeepSORT. Comput. Electron. Agric. 192, 106512. https://doi.org/10.1016/j.compag.2021.106512.

Wang, Z., et al., 2025. A novel fish pose estimation method based on semi-supervised temporal context network. Biomimetics. https://doi.org/10.3390/biomimetics10090566.

Xie, C., et al., 2024. UVEB: A large-scale benchmark and baseline towards real-world underwater video enhancement. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 1234–1244.

Yu, H., et al., 2025. FishKP-YOLOv11: An automatic estimation model for fish size and mass in complex underwater environments. MDPI Preprints.

Zhen, S., et al., 2024. Measurement of sea cucumber volume using binocular vision. Trans. Chin. Soc. Agric. Eng. 40 (21), 165–174 (in Chinese with English abstract). https://doi.org/10.11975/j.issn.1002-6819.202407228.

Zhu, M., Liao, W., Xu, Z., Tian, Y., 2025. Study on fish body length measurement method based on segmentation mask and binocular vision. J. Agric. Sci. Technol. 27 (2), 116–124 (in Chinese with English abstract).