Artificial Intelligence Review (2025) 58:7
https://doi.org/10.1007/s10462-024-10960-7

A review of deep learning-based stereo vision techniques for phenotype feature and behavioral analysis of fish in aquaculture

Yaxuan Zhao[1,2] · Hanxiang Qin[2,3,4,5] · Ling Xu[2,3,4,5] · Huihui Yu[1,2] · Yingyi Chen[2,3,4,5]

Accepted: 16 September 2024 / Published online: 7 November 2024

Abstract
The industrialization, high-density, and greener aquaculture requires a more precise and intelligent aquaculture management. Phenotypic and behavioral information of fish, which can reflect fish growth and welfare status, play a crucial role in aquaculture management. Stereo vision technology, which simulates parallax perception of the human eye, can obtain the three-dimensional phenotypic characteristics and movement trajectories of fish through different types of sensors. It can overcome the limitations in dealing with fish deformation, frequent occlusions and understanding three-dimension scenes compared to the traditional two-dimensional computer vision techniques. With the deep learning development and application in aquaculture, stereo vision has become a super computer vision technology that can provide more precise and interpretable information for intelligent aquaculture management, such as size estimation, counting and behavioral analysis of fish. Hence, it is very beneficial for researchers, managers, and entrepreneurs to possess a thorough comprehension about the fast-developing stereo vision technology for modern aquaculture. This study provides a critical review of relevant topics, including the four-layer application structure of stereo vision technology in aquaculture, various deep learning-based technologies used, and specific application scenarios. The review contributes to research development by identifying the current challenges and provide valuable suggestions for future research directions. This review can serve as a useful resource for developing future studies and applications of stereo vision technology in smart aquaculture, focusing on phenotype feature extraction and behavioral analysis of fish.

Keywords Stereo vision · Aquaculture · Phenotype feature extraction · Behavioral analysis · Deep learning

Yingyi Chen is the principal leader of the research team in which the author's group is working. He provided the significant intellectual and financial support for this research.

Extended author information available on the last page of the article

1 Introduction

The demand for aquatic animal-source foods has surged in recent years, driving the need for efficient and high-quality fish farming (FAO 2024). The phenotype feature and behavior of fish are the basic and most important characteristics for cultured fish and plays a key role in precision aquaculture management (Harvey 2003; Shi et al. 2020; Zhou et al. 2018b; Yang et al. 2021a). Automatic, accurate and real-time monitoring of these information is therefore essential for improving aquaculture production. Emerging technologies such as artificial intelligence (AI), computer vision, robot navigation, and the Internet of Things (IoT) are transforming the traditional aquaculture towards smart digital aquaculture. Currently, there are already intelligent aquaculture solutions available, including fish disease monitoring, fry sorting, and automated feeding (Wang et al. 2021; Ahmed et al. 2022; Zhao et al. 2021). Compared to the acoustic systems, biosensor technology, and traditional two-dimensional (2D) computer vision technology, the stereo vision technology is a more precise and intelligent non-invasive method with the abilities of understanding 3D scenes and producing more interpretable data (Li and Du 2022a). Hence, it has become a typical, fast-developing and widely used technology for fish phenotype extraction and behavior analysis that begun to replace manual work in precision aquaculture management (Strachan 1993; Israeli and Kimmel 1996).

Early studies on fish phenotype and behavior are mainly conducted using 2D computer vision due to its simplicity and low cost. However, 2D images or videos naturally have limitations in providing sufficient information when capturing fish in a real 3D scene and handling the challenges of fish occlusion and deformation. In some studies, fish have been compelled to swim within some fixture at a known distance from the camera, such as a fixed tube (Hao et al. 2016), in order to obtain more accurate information regarding their size or position (Zion 2012). These limitations hinder the further application of 2D computer vision technology in aquaculture.

Stereo vision technology, on the other hand, enables the precise acquisition of fish phenotypes and the motion of free-swimming fish in three-dimensional (3D) coordinate systems. The trend towards automation, intelligence, and precision in aquaculture management has led to the adoption of stereo vision as an important tool for constructing intelligent aquaculture models and acquiring information. Harvey and Shortis (1995) developed an early stereo-video system along with manual image processing software for underwater fish measurements. Subsequently, various deep learning-based stereo vision models have been developed. These models often exhibit a "multi-stage" pattern. By extracting features of fish from individual RGB images and leveraging the 3D perception capability of stereo vision technology, traditional tasks such as fish detection, key-point detection, instance segmentation, and tracking can be elevated to the 3D spatial level. These models can provide a more accurate and intelligent solution for fish size measurement (Perez et al. 2018; Muñoz-Benavent et al. 2018; Huang et al. 2020; Ubina et al. 2022; Hsieh and Lee 2023), biomass estimation (Serna and Ollero 2001; Tonachella et al. 2022; Shi et al. 2022), behavior analysis (Somerton et al. 2017; Zhou et al. 2018a; Bao et al. 2018). Additionally, there are also studies attempting to apply more cutting-edge techniques, such as 3D human pose recognition networks (Hsieh and Lee 2023) and radiation-based 3D reconstruction techniques (Sethuraman et al. 2023; Wang et al. 2024), to exploit the deep features provided by stereo vision The aim is to achieve end-to-end multimodal analysis models for fish stereo data in

aquaculture. Moreover, applying stereo vision technology to fish disease diagnosis (Li et al. 2022), such as identifying abnormal behavior or measuring surface damage rates (Tran et al. 2018), is also an application direction. However, there is still substantial room for research and development in this area.

Stereo vision technology involves vision capturing, image processing, and understanding of 3D information in a real 3D scene provided by various sensors. The application of stereo vision in aquaculture faces different challenges compared to its use in ground-based environments, such as robot navigation and autonomous driving. One challenge is the color deviation, blurring and low contrast of images due to the absorption and scattering of light by the water body. This can impede the accuracy of 3D reconstruction and the effectiveness of feature learning by deep learning models (see Fig. 1a). Laser-based devices, such as LiDAR, which can obtain highly accurate 3D point cloud data, are limited in their use underwater and on live fish (Risholm et al. 2022; Maccarone et al. 2023; Li et al. 2020b; Dubrovinskaya et al. 2018). Due to the non-rigid deformation and high interclass similarity of fish, and frequent occlusion caused by high density aquaculture, the accuracy of stereo matching is prone to be affected (see Fig. 1b), while also posing challenges to real-time 3D detection, segmentation and tracking of fish (see Fig. 1c). In addition, the need of over- or underwater camera calibration and stereo rectification will also introduce extra complexity and potential errors (see Fig. 1d).

There have been several works investigating the application of computer vision techniques in aquaculture (Li et al. 2020b; Yang et al. 2021b; Zhao et al. 2021; Li and Du 2022; Liu et al. 2023), but few of them focus on the deep learning-based stereo vision technology. Therefore, this review aims to address this knowledge gap by providing a comprehensive survey of the current literature and a critical analysis of the state-of-the-art research development on the phenotypic and behavioral analysis of fish in aquaculture. Particularly, it discusses the applications of the deep learning-based technological solutions in stereo vision to address the emerging challenges of smart aquaculture management.

This paper presents the findings in a systematic and structured way. As shown in Fig. 2, the paper analyses and summarizes stereo vision applications in aquaculture based on four layers: stereo data acquisition, stereo images preprocessing, stereo vision model, and further analysis and applications. As there are few public stereo datasets of fish available, the first step is to consider stereo data acquisition. The collected underwater stereo images may suffer from degradation and the lack of quantity, which requires the application of image enhancement and augmentation to eliminate their impact on subsequent data processing.

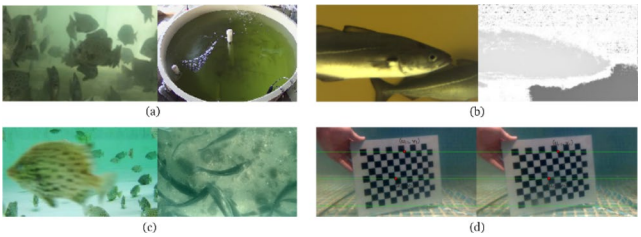

Fig. 1 Challenges of stereo vision applications in aquaculture. **a** Color deviation and low contrast of image (Hsieh and Lee 2023); **b** Insufficiency of stereo matching due to fish occlusion (Garcia et al. 2020); c Scale variation, deformation and similarity of fish (Silva et al. 2023); d Additional complexity of underwater camera calibration (Huang et al. 2020)

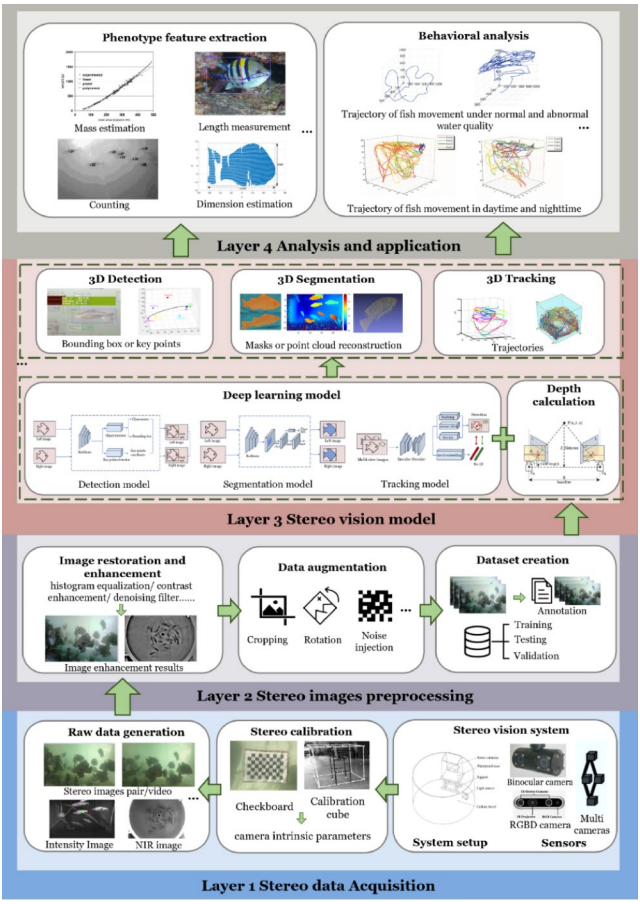

Fig. 2 The analytical framework of stereo vision technology applications in fish phenotype feature extraction and behavioral analysis

In the third layer, the collected data will be processed by stereo vision models constructed based on deep learning, including 3D fish detection, segmentation, tracking and 3D understanding models. Finally, the output of the models needs further analysis to serve the actual needs, such as size measurement, mass estimation, fish counting and behavior analysis. By identifying existing contributions and unresolved issues, the future research directions are clearly delineated. This survey can be contributed to the further development of smart aquaculture based on stereo vision technology, which is a current research hotspot.

The content of the paper is structured as follow: Sect. 2 provides an overview of the scope and criteria used for literature retrieval in this paper. Section 3 describes the relevant techniques for setting up a stereo vision system for data acquisition, and stereo data preprocessing methods. Section 4 illustrates common methods of processing data used in stereo vision models. This includes various network structures used for detection, segmentation, tracking in stereo vision and 3D reconstruction models. Section 5 summarizes and analyzes the application of stereo vision technology in fish size measurement, biomass estimation,

and behavior analysis. Section 6 discusses the challenges and gives a future outlook. Finally, Sect. 7 concludes with a summary.

2 Methodology

The study focuses on reviewing the most relevant articles published between January 2015 to April 2024 using a combination of keywords, including "stereo vision," "three-dimensional," "RGB-D," "binocular," "fish," "aquaculture," "size," "measurement," "weight estimation," and "behavior." The articles are retrieved from the WoS and Google Scholar databases. Since research on fish phenotyping and behavioral analysis using deep learning-based stereo vision technology in aquaculture is still in its early stages, the number of relevant papers is not large. To encompass all research in the field, including new explorations and trends presented at conferences, both journal articles and conference papers were considered. The retrieval results were manually screened based on their relevance and quality, following four criteria: (1) the study focused on underwater free-swimming fish rather than fish in anesthetized, frozen, or post-capture states; (2) the research utilized stereo vision techniques, including but not limited to the use of binocular cameras, multi-view camera arrays, RGB-D cameras, or laser-based sensors to acquire stereo vision data, excluding studies using monocular cameras or other non-computer vision sensors; (3) the study content was related to fish phenotype or behavioral analysis, excluding physiological analysis or review articles. (4) the paper was indexed in SCI or EI databases. A total of 45 papers were retrieved and analyzed in this study. To illustrate the early development of stereo vision in aquaculture, an additional 13 journal papers published before 2015 are also briefly introduced in the article.

The following sections present the analysis and summary of the stereo vision technology applications in fish phenotype feature extraction and behavioral analysis using the analytical framework presented in Fig. 2.

3 Stereo image acquisition and preprocessing of fish in aquaculture

3.1 Construction of fish stereo image acquisition system

3.1.1 Equipment and layout for fish stereo image system acquisition

Binocular and multicamera systems are more commonly used for underwater fish monitoring than laser-based systems. The latter require careful consideration of effects of backscattering from the water column (Dubrovinskaya et al. 2018). For binocular camera systems, there are three main types of orientations and relevant arrangements. Most studies use a side-by-side orientation of two cameras (see Fig. 3a) in water to capture the lateral view of swimming fish (see Fig. 3e). This allows for a more typical morphological characterization of the fish, yet requires consideration of the underwater calibration of the camera. It can also be positioned perpendicular to the water's surface to capture views of the dorsal or ventral side of fish, which allow for clearer visibility of the fish's bending (see Fig. 3d) (Muñoz-Benavent et al. 2018), yet requires consideration of the refraction of light through the dif-

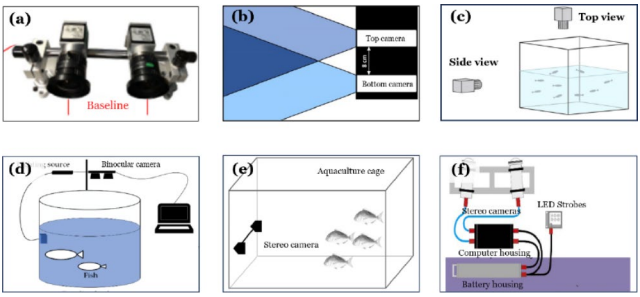

Fig. 3 Common stereo vision system composition. **a** Side by side orientation(Shi et al. 2020); **b** Vertical orientation (Tonachella et al. 2022); **c** orthogonal orientation (Cheng et al. 2018); **d** Top view arrangement (Muñoz-Benavent et al. 2018); **e** Side view arrangement (Komeyama et al. 2018); **f** Other facilities (Chuang et al. 2015)

ferent media of air and water. Some researchers have also attempted to employ two cameras positioned one above another to form an in-line stereo-vision system in a vertical orientation (see Fig. 3b) (Tonachella et al. 2022). Such system only requires an in-situ calibration, and is insensitive to changes in the angular alignment of cameras (Dunbrack 2006). Orthogonal placement mode requires 2–3 cameras to capture a top view and one or more side views of the fish body (see Fig. 3c). This enables the acquisition of the 3D spatial position of the fish. Due to the specificity of aquaculture scenes, regardless of the arrangement method chosen above, factors such as camera waterproofing, selection of fill lights, and data transmission methods are additional considerations that need to be taken into account (see Fig. 3f).

Another factor needs to be taken into consideration is the determination of the baseline and convergence angle. The baseline refers to the distance between two cameras and is closely related to the imaging range of a multi cameras system. Wide-baseline systems have a larger field of view (FOV) and are more accurate (Shi et al. 2020). However, they may have difficulty in capturing objects in close proximity (Cai et al. 2010; Shen et al. 2014). Some studies have chosen to converge the cameras inward at an angle to optimize stereo overlap in the FOV (Shafait et al. 2017). According to the study of Aguiar et al. (2016), the measurement accuracy of the same object at the same baseline and distance generally increases with the convergence angle.

3.1.2 Stereo calibration for fish stereo image

The calibration of the stereo camera system involves acquiring the intrinsic and extrinsic parameters of the camera system and is of importance for the modelling of the imaging process and for the elimination of errors introduced by perspective projection. The current calibration methods are performed by photographing an object of known size from different angles. One such method involves the utilization of a 3D cuboid lattice frame, known as a calibration cube, which is marked with multiple points. Zhang (2000) proposed a more flexible method that requires only a 2D checkerboard for calibration. This method is easy to use and can provide modest accuracy when the 2D calibration fixture is roughly similar in proportion to the field of view (FOV) of the camera system (Boutros et al. 2015; Shortis 2019). This method is easy to use and can provide moderate accuracy for short-range situations. Once the initial calibration is completed, it is important to verify the calibration

accuracy. A commonly used method is to calculate the root-mean-square error (RMSE) or evaluate the proportional error between true length and measured length of diagonal lines of the checkerboard.

P_l and P_r: actual imaging points in the left and right imaging planes; X_l and X_r: ideal imaging points in the left and right imaging planes; d_1 and d_2: distance from X_l and X_r respectively to the left of the imaging planes.

Ideally, the stereo camera's imaging planes should be coplanar, so that the epipolar lines of the two images are parallel. Yet the two actual imaging planes always exist at a certain angle. Stereo rectification is an important process used to reproject image planes onto a plane parallel to the camera's optical centerline. Considering a horizontally aligned binocular cameras, as shown in the Fig. 4, the imaging process of a point P in space in the left and right cameras with optical centers O_L and O_R is depicted (Brown et al. 2003). A simple and general method of rectification is using camera parameters obtained from calibration (Fusiello et al. 2000). The stereo matching algorithm can then be used to find the disparity between the imaging points X_l and X_r of point P in the left and right images (As shown in the Fig. 4, the disparity $D = |\, d_1 - d_2 \,|$). The focal length of the camera is represented by f, and the distance Z of point P from the baseline B can be obtained using the principle of similar triangles Eqs. (1 and 2). The 3D coordinates of the point P can then be calculated using the camera calibration parameters and its position in the image pairs.

$$\frac{Z}{B} = \frac{Z - f}{X_l X_r} \tag{1}$$

$$Z = \frac{f \cdot B}{B - X_l X_r} = \frac{f \cdot B}{D} \tag{2}$$

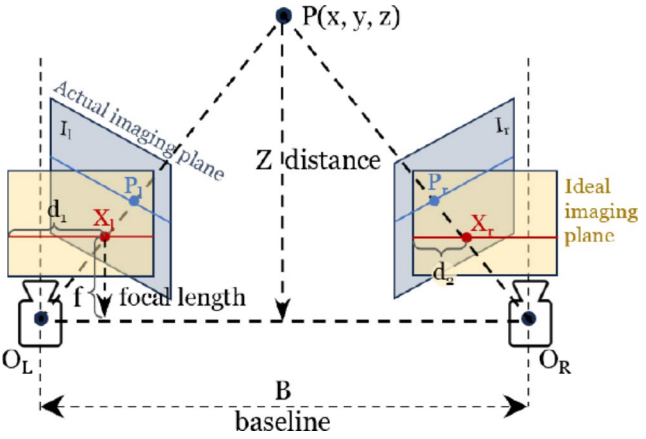

Fig. 4 The schematic representation of stereoscopic imaging geometry

3.2 Stereo image preprocessing for fish model

Stereo vision data is primarily composed of multi-view images. The current preprocessing method for this type of images involves separating them into individual image and then applying the single image-based method for preprocessing. The preprocessed individual images can then be matched to form multi-view stereo image pairs.

3.2.1 Stereo image restoration and enhancement for fish model

Due to factors such as light absorption, scattering, and potential issues like eutrophication and insufficient illumination in aquaculture environments, underwater fish images captured often face challenges of image degradation and noise interferences (as shown in Fig. 5). These issues can cause disruptions and difficulties in subsequent feature learning using deep learning-based model. It is necessary to restore and enhance the image with dehazing, contrast enhancement, denoising, and other techniques. The commonly used methods can be categorized into physical model-based, non-physical model-based, and deep learning model-based methods.

The physical model-based approach aims to restore images based on underwater imaging process. Akkaynak and Treibitz (2019) developed the sea-thru method for restoring the color of RGBD image. Inspired by the dark channel prior, the method estimates backscatter using dark pixels and relevant prior information. (as shown in Fig. 5a). It is more practical to enhance image visual effects directly by using non-physical methods, which can be categorized into spatial-domain and frequency-domain methods. Histogram stretching is one of the most commonly used spatial domain methods. Hsieh and Lee (2023) addressed the color deviation issue by simply conducting histogram equalization (HE) (as shown in Fig. 5b). The contrast-limited adaptive histogram equalization (CLAHE) can further improve HE in preserving local details while avoiding amplifying noise (Reza 2004). Another popular method is Retinex, which is based on the theory of color constancy (Li 2013). Zhou et al. (2017a) proposed an adaptive image contrast enhancement algorithm for recirculating aquaculture systems (RAS) based on Multi-Scale Retinex and adaptive image contrast enhancement algorithm (as shown in Fig. 5c). Noise interference caused by low quality of the image can be another problem. To address this issue, denoising filters are commonly used, such as Gaussian filtering (Salman et al. 2016) and median filtering (Jin and Liang 2017) (as shown in Fig. 5d). Frequency-domain methods aim to transform image pixels to the corresponding frequency domain and perform image enhancement or denoising based

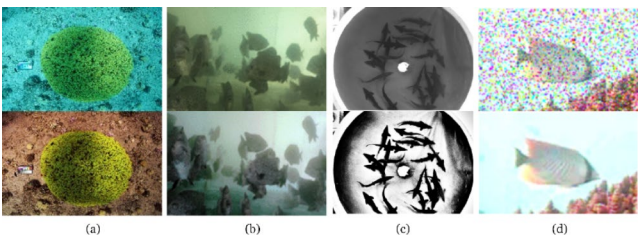

Fig. 5 Effects of underwater image restoration and enhancement. **a** Color restoration (Akkaynak and Treibitz 2019); **b** Histogram equalization and edge enhancement (Hsieh and Lee 2023); **c** Contrast enhancement (Zhou et al. 2017a); **d** Denoising (Jin and Liang 2017)

on the rate of change of the pixels. For instance, Stationary Wavelet Transform (SWT) can be used for improving image contrast (Priyadharsini et al. 2018) and denoising (Prabhakar and Praveen Kumar 2010).

Deep learning-based methods for enhancing underwater images have been developed, with convolutional neural network (CNN) and generative adversarial network (GAN) being two commonly used structures. Perez et al. (2017) built the first CNN-based underwater image enhancement model, which was used for removing haze of underwater images by learning the relationship between degraded and clear image pairs. Current CNN-based methods aim to achieve lightweight and functional diversity (Fu and Cao 2020; Li et al. 2020a; Liu et al. 2020), but they still rely on learning the mapping relationship between degraded and normal images. However, underwater images naturally lack normal images. To address the lack of paired data, GAN are being utilized. Ye et al. (2018) proposed a stacked GANs for detecting haze and correcting color jointly. However, deep learning-based methods need to be weighed against the balance between cost and benefit when applied to aquaculture due to their more complex training processes.

3.2.2 Stereo image augmentation for fish model

Deep learning-based models are a powerful tool for processing fish images and are currently trending in aquaculture applications. Most current studies use self-collected data for training. However, these self-built datasets often suffer from issues such as sample imbalance and limited environmental diversity, which hinder the improvement of model accuracy and generalization performance. The image augmentation algorithms can be utilized on the basis of existing data to increase the training samples. The simplest augmentation techniques involve flipping, rotating, cropping, scaling and transforming the image (Huang et al. 2019a; Mujtaba and Mahapatra 2021; Ben Tamou et al. 2022; Tonachella et al. 2022). These methods can simulate new image samples with the fish body presents different positions, sizes, and swimming directions. Additionally, there are methods to simulate underwater images of varying quality by injecting noise and changing the relevant parameters of the color space (Salman et al. 2016; Wei et al. 2018; Shorten and Khoshgoftaar 2019). These methods can help the model learn more robust features and increase its generalization performance.

Ben Tamou et al. (2022) used a pretrained ResNeXt101 network for transfer learning, and proposed a new criterion of targeted data augmentation techniques based on training and validation loss curves. This could alleviate the problem of insufficient datasets effectively. The employment of GANs for image augmentation has recently gained popularity. Wang et al. (2020a) used a GAN-based image generation model to expand 2D fish images. This method was found to be more effective in fitting subsequent semantic segmentation networks compared to traditional augmentation methods.

3.3 Current datasets of stereo image of fish

The dataset is the foundation for training stereo vision models for fish detection, recognition, tracking and other studies. According to relevant literature (Li and Du 2022b), publicly available fish image datasets are mainly two-dimensional, such as the Fish 4 Knowledge (http://groups.inf.ed.ac.uk/f4k/, 2015) and LifeCLEF-15 (http://www.imageclef.org/).

There are few publicly available stereo datasets for fish. 3D-ZeF (Pedersen et al. 2020), one of the few publicly available stereo datasets for fish, was constructed for multi-object tracking of zebrafish, containing eight video sequences with durations between 15 and 120 s (http://vap.aau.dk/3d-zef). The data collection scenario was a small self-constructed glass fish tank with 1–10 zebrafish swimming freely. The size and diversity of the dataset are not yet sufficient to support studies on fish in aquaculture. The OzFish (Australian Institute Of Marine Science 2020) dataset contains more than 3,000 videos, acquired with stereo baited remote underwater video systems (stereo-BRUVS). This dataset comprised of 507 species of fish, with annotations of bounding boxes and fish nose and tail pixel locations. However, the lack of fish head and tail labels and the unordered annotations create difficulties in training for more tasks such as fish key-points detection. The majority of current studies are not dependent on any public datasets. Rather, they all collect their own data, analyses it and build models based on the application requirements. In order to provide a clear demonstration of the datasets utilized in different studies, we present the dataset descriptions, including their collecting scenes, size, annotation, and preprocessing, in Tables 1 and 2, and 3. The equipment used for dataset collection in various studies predominantly consists of stereo camera systems. The scenes are mostly self-constructed experimental platforms, and some studies have deployed camera systems in sea cages or underwater trawls. The size of these self-built datasets is approximately 2000 to 8000, and the fish species are relatively homogeneous, which is not conductive to the generalization of the models due to the limitations of the labor burden of manual annotation and the collection scenarios. The construction of publicly available large-scale underwater three-dimensional fish datasets is a necessary step to promote the further development of aquaculture research applications towards mechanization and intelligence. The specific path and outlook will be discussed in Chap. 5.

3.4 Commonly used stereo image labeling techniques and tools in aquaculture

When establishing a stereo dataset for fish in aquaculture, concisely labeling fish objects, key points on the fish body or fish masks is a critical challenge to address. There are many open-source annotation tools available that support various forms of annotation, such as rectangles, key points, and polygons on images, such as Labelme, VGG Image Annotator (VIA), LabelImg, and DarkLabel. These tools can be applied in aquaculture to support tasks like fish detection, mask segmentation, and tracking. However, there are still several issues that need to be resolved. First, manual labeling may introduce errors. For instance, in the annotation of fish key points, different annotators may introduce offsets at the same key point location, especially for locations like 'fish tail connection point' that are not as easily identifiable as 'fish eyes' (Yu et al. 2023). These discrepancies can lead to errors in subsequent tasks such as dimension measurement and stereo matching. Additionally, the labor-intensive burden of manual annotation poses obstacles to constructing large-scale datasets. To alleviate the labor-intensive nature of manual annotation, Marrable et al. (2023) developed a semi-automated annotation tool where users can click anywhere on the fish body in left and right images, enabling automated localization of fish head and tail key points in stereo image pairs with the assistance of deep learning-based models. Fernandes et al. (2020) constructed a fish segmentation dataset by crowdsourcing annotation tasks on the Amazon Mechanical Turk (MTurk) platform, which also required secondary reviewer checks. Moving forward, the development of unsupervised or semi-supervised deep learn-

ing models based on autoencoders (AE) and generative adversarial networks (GAN) for feature learning from unlabeled or limited labeled data may represent future directions.

4 Stereo vision models based on deep learning in aquaculture

Locating the three-dimensional position of fish in images or video frames forms the basis for extracting phenotypic features such as key points, dimensions, or contour segmentation. Furthermore, three-dimensional tracking across frames enables extraction of behavioral characteristics of fish including swimming speed and acceleration. These are crucial capabilities for using stereo vision technology to meet the needs of aquaculture applications. Therefore, this chapter focuses on the fundamental techniques of four categories of models: three-dimensional detection, segmentation, tracking, and reconstruction based on deep learning technologies, and their development trends applied to fish.

4.1 Detection models

4.1.1 3D object detection

3D object detection in aquaculture aims at determining the 3D location and species of the interested fish, which is a prerequisite for fish phenotypic feature extraction and behavioral analysis. Computer vision-based 3D object detection is somewhat reliant on 2D image processing tasks. While, facing the challenges of inter-class similarity, scale variation, deformation and frequent occlusion of fish in the image, 2D-based detection cannot further meet the needs of the development of precise aquaculture. Stereo data can facilitate the generation of more realistic results. The current 3D object detection methods include LiDAR-based, monocular-based, and multi-view image-based detection. LiDAR-based 3D detection methods are able to exploit the geometric features of high-precision dense point cloud data to provide more robust detection results against illumination variations and texture loss (Shi et al. 2019). Monocular-based methods require additional auxiliary information, such as matching with a known 3D template (Mei et al. 2021), or estimating depth based on deep learning techniques (Koh et al. 2023). Multi-view images can provide more interpretable depth information and are more commonly used in aquaculture applications. Depending on the stage at which the depth information is utilized, it can be classified into three categories: "pre-detection fusion", "detection fusion" and "post-detection fusion". The "pre-detection fusion" refers to the conversion of the stereo-image pairs into a sparse point cloud, which is then used as the basis for detection. The idea of "fusion in detection" is to infer the 3D bounding box of an object end-to-end during the detection process by using cues such as parallax relationships between stereo image pairs and texture information in the images (Li et al. 2019). The "post-detection fusion" approach uses a "two-stage" strategy (see Fig. 6) that first relies on a common 2D image-based detection network to detect the 2D position or key-points of the object detection in the left and right images. Afterwards, the 3D coordinates of the fish body or key-points are reconstructed by relying on stereo matching or depth information. This relies on precise camera calibration and sufficient texture information to accomplish stereo matching, but is truly more common in current aquaculture applications.

Table 1 Summary of various fish detection methods proposed, the datasets used and their evaluation results in stereo vision

Challenge	Strategy	References	Algorithm	Data set				Evaluation results
				Description	Annotation	Size	Preprocessing	
High inter-class similarity	Integrated the backbone with Convolutional Block Attention Module (CBAM), focusing on key area.	Liu et al. (2022)	YOLOv5	Self-built dataset, including RGB images and infrared images captured by an underwater depth camera	Annotation of fewer mask data sets compared to labeled box data sets	N/A	N/A	Measurement accuracy = 96.9%
		Deng et al. (2022)	Keypoint RCNN	Self-built dataset, including images of five different types of fish captured by a binocular camera in a culture pool	Annotated according to the human key-point annotation format of the COCO dataset	Training/validation/test = 7200/1800/900	Enhancement: random two attributes from saturation, brightness, contrast, and sharpness were adjusted stochastically	The mAP for fish detection increased by 2.4% after integrating with CBAM
		Deng et al. (2023)	RetineNet CenterNet	Self-built dataset, captured by a binocular camera above the water surface, reserved only one side of the image pairs. The collection scenes included both outdoor and indoor lighting environments	Only the straight and uncovered fish was annotated with the key-points of head and tail	Training/validation = 6400/1600 for fish detection, 19,200/4800 for key points detection	Enhancement: saturation, brightness, contrast and sharpness were adjusted stochastically	The pixel distance errors decreased about 0.215 pixels after integrating with CBAM

Table 1 (continued)

Challenge	Strategy	References	Algorithm	Data set				Evaluation results
				Description	Annotation	Size	Preprocessing	
	Integrated the backbone of Deep Layer Aggregation with Transformer	Yu et al. (2023)	CenterNet	Including RGB images collected from Internet and field, and most were captured in clear water bodies	Each fish was labeled with 9 key-points	Training/validation/test = 1260/158/158	Enhancement: turbid underwater image enhancement based on parameter-tuned stochastic resonance	The AP for key-points detection increased by 1.8 after integrated with transformer only
Multi-scale variations	Increased the skip connection through deformable convolution for a better feature aggregation	Yu et al. (2023)	Already mentioned above					The AP for key-points detection increased by 3.9 after integrated with Aggregation only
	Replacing Feature Pyramid Networks with the Improved-Path Aggregation Network	Deng et al. (2022)						The mAP for fish detection increased by 1.4% after replacing with I-PANet
Multi-scale variations	Replacing FPN with ASFF structure to improve the scale invariance of the features in an adaptive way	Deng et al. (2023)	Already mentioned above					Already mentioned above

Table 1 (continued)

Challenge	Strategy	References	Algorithm	Data set				Evaluation results
				Description	Annotation	Size	Preprocessing	
Misdetection of key-points due to the high density of fish	An intermediate supervision scheme of Stacked Hourglass-based network can avoid the false detection	Suo et al. (2020)	Faster R-CNN Stacked Hourglass	Self-built stereo dataset, captured by a binocular camera in a culture pool.	Annotation of bounding box and 7 key points	Training/validation=1117/124 for fish detection, and 551/61 for key points detection	N/A	The mAP for fish detection is 0.905 and the Averaged Object Keypoint Similarity (OKS) is 0.667
	Using the bottom-up method designed for multi people key-points detection and tracking, using PAFs for detecting at key part of the same body	Hsieh and Lee (2023)	OpenPose ArtTrack	Self-built stereo dataset, including image pairs of Oplegnathus punctatus captured by a binocular camera in a culture pool	Each fish was labeled with 9 key-points and 9 bones	1000 images for training	Enhancement: histogram equalization (white balance) and edge enhancement	Measurement relative error=4.49%
Real-time requirements	Pre-trained the model with open datasets	Tonachella et al. (2022)	YOLOv4 RESNET-101	Open Image datasets for fish detection Self-built stereo dataset for key-points detection, captured by a stereo camera in a sea cage	Cropped fish images with snout tip and the base of the middle caudal rays labelled	Training/test=1120/280 for fish detection, 8960/3840 for key points detection	Augmentation: scale, noise, rotation, translation, and brightness	The mAP for fish detection is 87%, and the MSE for landmark detection is 0.23
	Using a lightweight model YOLO v5 small with transfer learning strategy	Marrable et al. (2023)	YOLO v5 small	Open dataset from the OzFish stereo-BRUVS imagery	Cropped fish images with head and tail labelled	Training/validation/test=5348/2292/4154	N/A	Deep learning precision for key-points detection is 77.40%

Table 1 (continued)

Challenge	Strategy	References	Algorithm	Data set				Evaluation results
				Description	Annotation	Size	Preprocessing	
	Using a lightweight and pre-trained model Deep Layer Aggregation network DLA-X-60-C	Deng et al. (2023)	Already mentioned above					The number of parameters decreased 72.207 M after using pre-trained lightweight model
Real-time requirements	Using the bottleneck and group convolution can effectively improve the training efficiency.	Deng et al. (2022)	Already mentioned above					The model size only increased 0.811 M, and the mAP of fish detection increased by 4.55%

AP, *mAP* and *MSE* correspond to average precision, mean average precision and mean square error

N/A not available

Table 2 Summary of various fish segmentation methods proposed, the datasets used and their evaluation results in stereo vision

Challenge	Strategy	References	Algorithm	Data set				Evaluation results
				Description	Annotation	Size	Preprocessing	
Overlapping	Each fish is tracked across frames and a series of 3D models are computed to reduce interference from occlusion	Ubina et al. (2022)	Mask R-CNN	Self-built dataset, including 8 videos collected by stereo camera from aquatic pond and offshore fish cage	Manually labeled mask with 'front view' and 'side view'	Training/test 700/1300	N/A	Model accuracy in the range of 84-95% in different scenarios
	Using multi-label expansion to refine the segmentation Proposing a new IoU* that defines the concurrency ratio threshold in the presence of occlusion	Garcia et al. (2020)	Mask R-CNN	Self-built dataset, including stereo image pairs of fish captured in the trawl using the Deep Vision system	Manually labeled mask, which was divided into overlapping fish and non-overlapping fish	Training/validation/test 1284/321/200	Enhancement: linearization and correction of uneven illumination Augmentation: image translations, horizontal and vertical reflections, rotations, and shear transformations	Accuracy of 0.994 for single fish and 0.984 for overlapping fish
	Using SOLO v2 with better segmentation accuracy. Discard heavily obscured fish by fish integrity discrimination.	Liu et al. (2022)	SOLO v2	Already mentioned in Table 1				Measurement accuracy=96.9%

Table 2 (continued)

Challenge	Strategy	References	Algorithm	Data set				Evaluation results
				Description	Annotation	Size	Preprocessing	
	Replacing the FPN with an improved Path Aggregation Network and incorporating a channel attention mechanism into the head of the decoupled-SOLO structure Classifying occluded fish into negative categories based on binocular vision-based fish length measurements	Yu et al. (2022)	CAM-Decoupled-SOLO	Self-built dataset, collecting from recirculating aquaculture system using an underwater binocular system	Manually labeled mask of positive sample (unoccluded fish) and negative sample (occluded fish)	Training/test 633/177	N/A	The improved network achieved a segmentation result with a 1.2% increase in mAP compared to SOLOv2.
	Only label the uncovered fish Using Grabcut to refine the segmentation	Huang et al. (2020)	Mask R-CNN	Self-built dataset, using an underwater binocular system	Only uncovered fish were labelled	N/A	N/A	The average error in length measurement is around 5.5 mm

Table 2 (continued)

Challenge	Strategy	References	Algorithm	Data set				Evaluation results
				Description	Annotation	Size	Preprocessing	
Overlapping	Acquiring high-resolution underwater depth images using a range-gated 3D camera	Risholm et al. (2022)	DBSCAN	Self-built dataset, including high-resolution underwater intensity and depth images of salmon, captured by a range-gated 3D camera	N/A	3000	Applying a median filter to filter out points in the depth map that have a low confidence / weak signal	Fish length measurement error of 1%
High background noise and severe deformation	Combining background subtraction and disparity maps to classify background planes with clustering and scoring strategies and refine segmentation using Conditional Random Fields with color and geometric features	Huang et al. (2019a)	Background plane clustering, CRFs	Self-built dataset, including stereo videos collected from fishing vessels using stereo cameras. The collection scenes included early morning, morning and afternoon of different sunlight environment	Manually labeled bounding boxes of each fish in the left images	10,000	Augmentation: horizontal flip, random crop, expansion and color distortion	Fish length measurement error of 4%

mAP correspond to mean average precision

N/A not available

Table 3 Summary of various fish tracking methods proposed, the datasets used and their evaluation results in stereo vision

Challenge	Strategy	References	Algorithm	Data set				Evaluation results
				Description	Annotation	Size	Preprocessing	
Inter-frame fish deformation, occlusion, frequent entries and exits leading to target loss	Tracking using the Viterbi data association algorithm based on dynamic programming that exploits temporal relationships throughout the target's lifecycle	Chuang et al. (2015)	Detection and segmentation: based on double local thresholding and histograms Matching: based on 4 cues, including Euclidean distance, area difference, motion direction and histogram distance Association: based on the proposed multiple-target VDA algorithm, which create a separate trellis for every target to track. Path in different trellises can share the same nodes.	Self-collected dataset including several 8-bit grayscale video clips	N/A	N/A	N/A	Tracking success rate=0.88
	Projecting the 2D image into 3D space and re-scoring multiple proposals using 3D spatial detection and tracking information.	Huang et al. (2019b)	Detection: SSD YOLO v2 Prediction: Kalman filter Association: greedily matching; Unmatched: applying the NMS with IoU threshold 0.5	Self-collected dataset from two fishing ship	Bounding boxes of each fish in the left images of stereo image pairs were labeled.	10,000 frames	Augmentation: horizontal flip, random crop, expansion and color distortion	IDS=110 Frag=149 MOTA=96.3
	Fish motion prediction using LSTM with strong learning ability of long-term dependencies	Wang et al. (2017)	Detection: DoH blob detection method for fish head in master view and Gabor filter as well as Max-Min Distance clustering for fish eye in slave views. Prediction: Kalman filter and LSTM network Association: implement top-view cross-frame association by motion continuity and appearance coherency and using motion continuity and Epipolar constraint for cross-view association	Self-collected dataset from one master view from top and one or two slave view from side	N/A	50,000 sequences	N/A	P=0.961 IDS=0.9 Frag=3.9

Table 3 (continued)

Challenge	Strategy	References	Algorithm	Data set				Evaluation results
				Description	Annotation	Size	Preprocessing	
	Top-view tracks less prone to occlusion are prioritized; multiple views increase accuracy Recognize occlusions based on skeleton representation of fish body, correlate track segments before and after occlusions	Qian et al. (2017); Liu et al. (2019) Qian and Chen (2017);	Detection: select head and tail point based on the skeleton extracted by augmented fast marching method (AFMM) as representation of fish Association: using a greedy algorithm based on the association cost function for the top-view tracking and implement cross-view association using the epipolar constraint and motion consistency constraint	Self-collected dataset from one master view from top and one or two slave view from side	N/A	2000 frames 2000 frames 2000 frames	N/A	IDS=14 TF=1.8 IDS=1.2 Frag=1.7 P=98.1% IDS=4 Frag=6
Inter-frame fish deformation, occlusion, frequent entries and exits leading to target loss	Remove occluded fish targets based on brightness using the IREF system	Pautsina et al. (2015)	Detection: based on background subtraction Association: based on the closest object selection Overlapping objects were handled by simply separating based on different brightness. Noise was filtered based on the pixel size	NIR images	N/A	N/A	Model the background using the first 100 images to complete the background subtraction	Mean depth estimation error of 1.6 ± 1.3 (SD) cm

Table 3 (continued)

Challenge	Strategy	References	Algorithm	Data set				Evaluation results
				Description	Annotation	Size	Preprocessing	
	Predicting fish movement states using nonlinear models	Palconit et al. (2021)	Detection and segmentation: threshold Prediction: MLR, ANFIS, GPR and MGGP Association: KNN algorithm	Stereo video pairs of 3 individual fishes	N/A	40 frames selected from 313 frames	N/A	GPR: $P=89.11$ accuracy=91.83 RMSE=20.00
	Tracking using the DeepSORT network containing the matching cascade and ReID network	Saad et al. (2024)	Detection: YOLO v7 Prediction: Kalman filter Association: Hungarian algorithm based on both Mahalanobis distance and cosine distance	Self-collected dataset	Bounding boxes	332 images randomly selected from 18,000 frames	N/A	$P=87.7\%$ Recall=88.2%

P, *TF*, *IDS*, *Frag*, *MOTA* and *RMSE* correspond to precision, trajectory fragmentations, ID Switches, fragments, multiple object tracking accuracy and root-mean-square error

SD corresponds to standard deviation

N/A not available

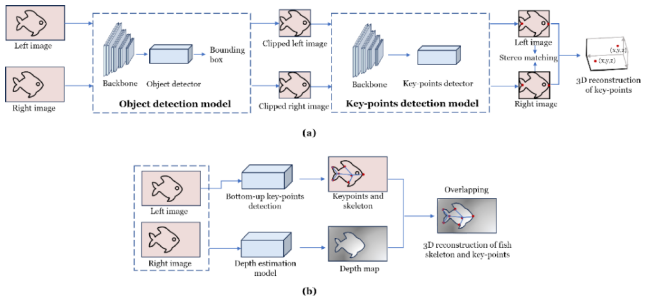

Fig. 6 3D reconstruction of fish key-points in two different post-detection fusion modes. **a** Two-step key-points detection and depth fusion after detection based on stereo matching; **b** Direct key-points detection and a dual-branch 3D key-points reconstruction model

4.1.2 Fish detection in stereo vision

Fish detection in aquaculture often adopts the "two-stage" strategy of "post-detection fusion". This mainly includes two modes: one is a single-threaded two-stage mode where fish are first detected, followed by key points detection, and then utilizing the key points detected in the left and right images for stereo matching (Fig. 6a). The other mode borrows the bottom-up key points detection method from human pose recognition, employing a dual-branch model to simultaneously perform 2D key points detection and 3D reconstruction (Fig. 6b). However, both methods involve integrating depth information after completing 2D key points detection.

It is necessary to first pay attention to different methods employed in stereo vision to detect fish in each single 2D image. For fish detection, the attention mechanism is widely used to cope with the high inter-class similarity of fish in images. Liu et al. (2022), Deng et al. (2022) and Deng et al. (2023) integrated Convolutional Block Attention Module (CBAM) into the backbone of Keypoint RCNN, YOLO v5, and CenterNet separately to improve their detection performance. CBAM can learn useful information from both the channel and spatial dimensions and can focus on specific areas that play a key role in identifying fish. Yu et al. (2023) proposed the CenterFishNet for fish key-point detection in stereo vision. This network used Deep Layer Aggregation-Transformer (DLAT) as its backbone, which introduced the self-attention mechanism to capture long-range dependencies. By integrating above attention mechanism, all the proposed models demonstrate improved accuracy in either fish object detection or dimensional measurement tasks combined with stereo vision (see Table 1 for Evaluation results). For the challenge of multi-scale variations exhibited by free-swimming fish in images, some studies have proposed different feature fusion strategies. Considering that the classical top-down fusion order, such as Feature Pyramid Networks (FPN), results in inadequate fusion of bottom layer features with top layer features, Deng et al. (2022) proposed an improved double direction features fusion network called I-PANet (Improved-Path Aggregation Network). The network performs the initial feature fusion by employing bi-Cubic interpolation for up-sampling, followed by refining the fused features through a convolution process utilizing an inverted bottleneck structure. Subsequently, the feature map is down-sampled to accomplish the second feature fusion. Deng et al. (2023) replaced FPN with adaptively spatial feature fusion (ASFF). ASFF structure can learn fused spatial weight for each scale feature adaptively, thereby improve the scale

invariance of the features. Yu et al. (2023) increased the skip connection through deformable convolution in an Aggregation module, which can fuse spatial features more effectively. Compared to the fixed feature fusion approach of FPN, these methods can better handle fish targets of different scales.

For free-swimming fish with frequent non-rigid deformations, a common solution is to represent the fish as a set of key points. Some studies adopted a two-step strategy of detecting the fish before feeding the results to the key-points detection network (as shown in Fig. 6a). Due to the high density of fish in aquaculture, it is easy to have false results of key points belonging to different fish fall into the same fish bounding box. Suo et al. (2020) employed the stacked hourglass structure as the key-points detection network. An intermediate supervision scheme was used during the refinement of the detection process, effectively avoiding the false detection of key-points belonging to non-primary fish within the bounding box. In the study by Marrable et al. (2023), it is believed that the use of detection network that can automatically rotate the detection box can reduce redundant areas and to minimize the probability of different fish key-points falling into one box. Besides, advances in the field of human pose estimation have inspired aquaculture researchers. Hsieh and Lee (2023) used the OpenPose network designed for 2D human pose detection. In contrast to the two-stage approach previously described, this type of bottom-up structure employs a direct approach to key-point detection and establish fish skeleton by encoding the relationship between different parts of fish body using Part Affinity Fields (PAFs). The anchor-free network can perform end-to-end regression of key points and bounding boxes with great robustness to the different fish postures.

Considering the real-time requirements of aquaculture applications, strategies such as transfer learning using public datasets for pre-training (Tonachella et al. 2022; Marrable et al. 2023), selection of lightweight network structures (Marrable et al. 2023; Deng et al. 2023) and exploiting bottleneck or group convolution, among others, can be used to reduce the number of parameters and computational requirements.

In 3D reconstruction based on 2D detection, the most common mode of utilizing depth information is “post-detection fusion”. Most studies use the calibration parameters of the camera to calculate the 3D coordinates of key points by matching them between stereo image pairs (Williams et al. 2016; Tanaka et al. 2019; Tonachella et al. 2022; Deng et al. 2022; Zhou et al. 2023). When the stereo camera system is underwater, the stereo matching process is the only thing that needs to be considered to achieve the 3D reconstruction. Deng et al. (2022) used a binocular camera located underwater to complete stereo matching of key-points in the left and right images based on IoU (Intersection over Union) and OKS (Object Keypoint Similarity). Additionally, the offset of the horizontal coordinates of the key-points in the right image is taken into account and compensated, thus enabling the completion of the 3D reconstruction. This stereo matching method only for key-points can avoid the high computational effort of global matching and cost aggregation-based stereo matching for underwater fish image applications. When capturing underwater fish with a camera outside the water, the process of stereo matching needs to consider the relationship between the actual fish and its imaging due to refraction at different medium interfaces. Deng et al. (2023) proposed a geometric model-based stereo reconstruction algorithm for capturing fish from an overhead binocular camera. The algorithm incorporated Snell’s law, the camera projection matrix, the refraction coefficients of light rays in air and water, and the distance between the camera and the refraction plane. By calculating the coordinates of the

refraction point, the algorithm established and corrected the beam vector of light, enabling the crucial 3D reconstruction of key points within the two media. In the "fusion after detection" paradigm, several studies have explored the use of multimodal data fusion for 3D reconstruction. Hsieh and Lee (2023) proposed a dual-branch 3D key-points reconstruction model (as shown in Fig. 6b). One branch of the model predicted depth maps from stereo image pairs, while the other branch detected nine key points and the skeleton of fish bodies. Furthermore, the authors overlaid the key-points and skeleton information from RGB images onto the depth maps to obtain the 3D coordinates of the key points. Such methods effectively use the semantic information from RGB images and the depth information from stereo image pairs.

There are also some studies that utilize monocular images for 3D fish detection. Mei et al. (2021) employed a relative deformable 3D fish template specific to flatfish and leveraged the fixed distance between fish and the camera in longline fishing to predict the 3D absolute pose and length of flatfish based on 2D segmentation masks. Koh et al. (2023) proposed the Aqua3DNet model, which integrates YOLOv3, SORT, and Udepth to create 3D density heatmaps of fish motion from monocular video clips. Udepth is used to predict the relative depth of the detected fish. However, monocular 3D detection is an ill-posed problem, requiring specific prior knowledge or complex 3D ground truth to train monocular depth perception models, thus limiting its scalability. Due to the costs and complexity of deployment of laser sensors, as well as the impact of underwater scattering and considerations for fish welfare, the application of laser-based sensors in aquaculture is not commonly observed (Risholm et al. 2022). Currently, some research efforts have explored the use of LiDAR for underwater robot navigation, such as in biomimetic fish. However, real-time data processing remains a challenge in these applications (Maccarone et al. 2023).

The challenges in applying stereo vision to fish detection and their corresponding solutions are summarized in Table 1 for a clearer illustration. Additionally, the table provides evaluation results of each study and descriptions of the datasets built and used, including annotations, size and preprocessing methods.

4.2 Segmentation models

4.2.1 3D object segmentation

The purpose of 3D fish segmentation is to provide fine-grained labels for fish bodies in real-world scenarios, enabling richer information extraction for fish phenotyping. Currently, the main approaches for 3D object segmentation can be categorized into methods based on multi-view images, RGB-D data, and point cloud data. Methods based on multi-view stereo images fuse the segmentation features extracted from different 2D views to form a global representation of the 3D object. Some traditional segmentation methods, such as Stereo GrabCut (Ju et al. 2013), estimate foreground and background by significance analysis of depth and achieve consistent segmentation between left and right images within the framework of graph cut theory. Segmentation networks designed for 2D images, such as Mask R-CNN and FCN, can also be employed to perform individual image segmentation. However, the challenge lies in how to fuse features from multiple single-view images. Much of the research based on multi-view stereo images focuses on 3D reconstruction of objects, such as MVSNet and Neural Radiance Fields (NeRF). MVSNet constructs a cost volume for

each pixel through Cost Volume Regularization and reconstructs the depth map of the scene through 3D convolutions (Yao et al. 2018). NeRF, on the other hand, can reconstruct 3D scene by mapping the position and viewing direction of a scene point to its color and density information using an MLP network, taking multi-view 2D images as input (Mildenhall et al. 2021). However, these methods have not been widely studied and applied in aquaculture.

Segmentation methods based on RGB-D data focus on the fusion of features from the two modalities. Early methods utilized a dual-branch architecture to independently process RGB images and depth maps. To use the correlation between the two modalities, some studies have employed various feature fusion approaches. Models based on RNNs and Transformers have also gained attention as they can overcome the challenges faced by CNN-based models in capturing global and long-range semantic information (Wu et al. 2022c; Yang et al. 2022).

Methods based on dense point cloud data enable direct and accurate 3D segmentation and can be categorized into three types. The first type involves performing convolutional operations directly on the point cloud data, such as the classic PointNet series (Qi et al. 2017a). The second type involves transforming point cloud data into other representations, such as voxels or meshes. The third type involves converting point clouds into multi-view images for processing. For example, SnapNet renders 3D point clouds as 2D images and applies CNNs to process the images, followed by inverse projection back to the 3D point cloud space to complete the segmentation (Boulch et al. 2018). In addition, there have been studies that project pixels from 2D images into 3D space and perform segmentation based on the constructed k-nearest neighbor graphs on the 3D point cloud (Qi et al. 2017b). These studies provide an inspiration for feature fusion in segmentation based on multi-view stereo images. For instance, some research matches the masks extracted from multi-view 2D images to form a sparse 3D point cloud representation of the objects. These methods, which combine 2D appearance and 3D geometric information, offer a new perspective on feature fusion in multi-view segmentation.

4.2.2 Fish segmentation in stereo vision

There are two main approaches of fish segmentation in stereo vision. One approach focuses on extracting 2D fish masks within individual images to serve as a precursor for further extracting 3D information such as fish key-points and skeletons (Fig. 7a). The other utilizes multi-view images to generate sparse point clouds, often combining image-based semantic information with depth maps or geometric information from point clouds to achieve "stereo" segmentation (Fig. 7b).

The main challenge in fish segmentation in images is the frequent occlusion between fish bodies, which can lead to less accurate segmentation of fish boundaries. One potential solution is to distinguish between occluded and unoccluded fish when preparing the segmentation dataset. Huang et al. (2020) addressed this issue by annotating unoccluded fish with complete visibility when building their training dataset for Mask RCNN, thus avoiding the problem of occlusions. Ubina et al. (2022) addressed this issue by tracking each fish in a video stream and calculating a series of 3D fish models to reduce the impact of occlusions across different frames. During the training of Mask R-CNN, they annotated the fish masks as either lateral or frontal views, enabling the segmentation of fish masks with different poses and providing semantic information from different perspectives for subsequent gen-

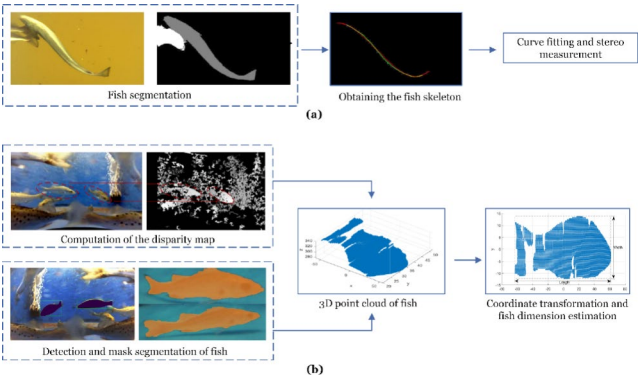

Fig. 7 Two main purposes of fish segmentation in stereo vision. **a** Segmentation of the fish mask to extract the fish skeleton and measurement of fish morphological size using stereo cues (Garcia et al. 2020). **b** Combining image-based segmented masks with disparity map to generate 3D point clouds of fish (Huang et al. 2020)

eration of fish point cloud data. In the study of Garcia et al. (2020) the training data were artificially divided into cases with non-overlapping fish and cases with overlapping fish. To distinguish between these two types of fish, a new Intersection over Union (IoU) threshold was set, calculated as the ratio of the area of the target fish to the combined area of the target fish and the overlapping fish. After the initial segmentation by Mask R-CNN, a multi-label expansion operation was applied to refine the segmentation further. When the expansion encountered resistance (indicating the presence of overlapping fish), the expansion was stopped, and the fish mask was filled to improve its accuracy. The achieved segmentation accuracy for occluded fish was 0.984. In addition to training the network to recognize heavily occluded fish targets, some studies have considered improving the accuracy of the segmentation model. In the work by Liu et al. (2022), a fish integrity assessment module was incorporated before fish segmentation, which discarded severely deformed or incomplete fish instances. Meanwhile, the researchers employed the instance segmentation network SOLO v2, which is based on the dynamic convolution position and information of instance, to address the issue of fish instances being adhered or overlapping more effectively. Yu et al. (2022) proposed a fish instance segmentation model called CAM-Decoupled-SOLO to improve segmentation accuracy and efficiency. The segmentation network based on SOLO simplifies segmentation by discarding bounding boxes, and the decoupled-SOLO structure further reduces the output dimension, thereby improving segmentation efficiency. Meanwhile, the authors replaced the FPN with an improved PANet and incorporated a channel attention mechanism into the head to enhance the representation of low-level features and suppress irrelevant features. In addition, the segmentation of fish in aquaculture applications may encounter challenges posed by dynamic background variations. Huang et al. (2019b)addressed this issue by leveraging the geometric relationships from depth maps and applying clustering and scoring strategies to segment fish from the dynamic background without the need of prior modeling. Furthermore, Conditional Random Fields (CRFs) were employed to refine the segmentation. This approach, which combines the geometric relationships from stereo vision and the color relationships from RGB images, demonstrates advantages in mitigating background noise interference.

Most studies adopt stereo matching-based methods to calculate and reconstruct sparse fish point clouds based on the multi-view masks. Since many studies only acquire binocular views, the generated point clouds primarily represent part of the visible surfaces of fish. In this process, various approaches leveraging image segmentation and point cloud clustering are employed to refine the segmentation outcomes. For instance, Huang et al. (2020) employed GrabCut to further refine the masks extracted by Mask R-CNN. Liu et al. (2022) utilized density-based clustering to denoise and refine the sparse point cloud data. It is important to note that this method relies on the assumption of coplanarity among fish instances and may exhibit limitations in terms of scalability when applied in complex aquaculture environments or for severely deformed fish. Nevertheless, this approach presents an economical solution that integrates both image semantic information and depth information, effectively meeting the application requirements of aquaculture. Several studies have also explored techniques to generate higher-precision point cloud data. Risholm et al. (2022) employed a range-gated 3D system to capture high-resolution underwater intensity and depth images. Following calibration and conversion into point cloud data, they employed the density-based clustering algorithm, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), to segment distinct fish individuals, achieving an accuracy of measurement of 99%.

The challenges in applying stereo vision to fish segmentation and their corresponding solutions, are summarized in Table 2 for a clearer illustration. Additionally, the table provides evaluation results of each study and descriptions of the datasets built and used, including annotations, size and preprocessing methods.

4.3 Tracking models

4.3.1 3D object tracking

The development of object tracking has witnessed a transition from single-object to multi-object tracking and from 2D tracking to 3D tracking. Single-object tracking focuses on selecting the candidate box with the highest confidence for the next frame, relying on three types of cues: appearance features, motion models, and location search. Deep learning-based multi-object tracking frameworks can be divided into two major types: Tracking by Detection (TBD) and Joint Detection and Embedding (JDE). The former separates object tracking into independent detection and association modules, linking detections across frames for tracking. Classical association strategies include greedy matching based on Intersection over Union (IoU), nearest neighbor algorithms, and the Hungarian algorithm. To bridge the information gap between detection and association, JDE has been proposed (Wang et al. 2020b). This paradigm trains a shared neural network to simultaneously perform detection and association, improving computational efficiency and tracking accuracy.

The implementation of 3D tracking is mostly built upon the tracking by detection paradigm, with a focus on incorporating the 3D position or motion cues of the objects, while also considering the utility of appearance features as complementary information. The key to the transition from 2D to 3D lies in the use of depth information. One category of methods uses depth information primarily at the detection stage, and designs different association strategies based on the different forms of 3D detection results. For example, AB3D MOT extends the Kalman filter into 3D space after obtaining 3D bounding boxes from LiDAR-based 3D

detection, thus enabling motion estimation for objects. The similarity calculation and association between objects and trajectories are based on the stereo IoU between 3D bounding boxes, combined with the Hungarian algorithm (Weng et al. 2020). On the other hand, 3D tracking methods that treat objects as key-points often use simple greedy algorithms based on the distance for association. For example, CenterTrack, which relies on CenterNet for detection, performs greedy matching based on the Euclidean distance between key-points and can be easily extended to 3D tracking applications (Zhou et al. 2020b). Another category of methods based on multi-view 3D object tracking uses depth information primarily at the association stage. These methods usually perform 2D-based tracking in a master view, and obtain 3D tracking results by associating the detections or tracking results from the other views (Wang et al. 2017; Cheng et al. 2018). The information from multiple views can effectively avoid the tracking errors caused by occlusion, but the multi-dimensional input can incur a significant computational cost.

4.3.2 Fish tracking in stereo vision

Multi-view 3D object tracking is more common in current aquaculture applications. Early attempts at tracking fish in stereo vision have mostly utilized a mirror and a single camera. By incorporating additional mirror imaging, it is possible to obtain the 3D coordinates of the target fish as well as recognize and avoid occlusions (Xiao et al. 2016; Mao et al. 2016). However, this method is challenging to apply to large-scale aquaculture due to the limited range of mirror imaging and the difficulty in tracking fish as their numbers increase.

Currently, most research utilizes multiple orthogonally placed cameras to capture top-view and side-view perspectives. This approach enables obtaining broader imaging coverage and more accurate 3D trajectories of fish. The top view typically serves as the master view for tracking, while other side views are used for stereo matching with the tracking results of the top view, as shown in Fig. 8 (Qian et al. 2017; Wang et al. 2017; Qian and Chen 2017; Cheng et al. 2018; Liu et al. 2019). This model involves not only cross-frame detection and association, but also cross-view association, which is usually achieved by leveraging epipolar constraints and motion consistency constraints. In order to improve the accuracy of association and tracking efficiency, many researchers have simplified the representation of the fish bodies by using feature points. To address tracking failures caused by fish deformation, occlusion, and frequent entry and exit from the field of view, Huang et al. (2019b) projected the detected 2D object pixels into 3D space and rescored multiple

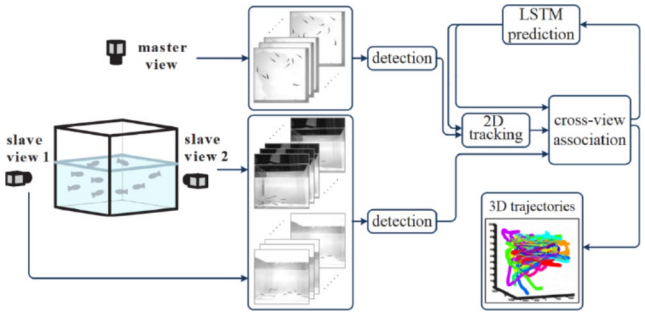

Fig. 8 Example of a workflow that uses master-slave strategy for tracking Wang et al. (2017)

proposals using Kalman filter. They also used a segmentation method based on clustering and fully-connected Conditional Random Field (CRF) to support tracking. The fish positions were updated based on the segmentation results. By using the 3D spatial detection and tracking information, they overcame the problem of low detection scores caused by fish deformations between frames. The proposed method outperformed the baseline method solely based on SSD. It achieved an improvement of 25.7 in Multiple Object Tracking Accuracy (MOTA) and reduced the number of identity switches (IDS) and fragments (Frag) by 604 and 929 respectively. Wang et al. (2017) applied the Kalman filter for initial tracking and then used LSTM, which is more effective at learning long-term dependency relationships, to predict fish motion states. For the tracking of 10 fish, the proposed method showed an improvement of 0.128 in Precision compared to idTracker, with IDS and Frag decreasing by 6.4 and 33 respectively. Cross-frame association in the master view was performed considering motion continuity and phenotype similarity cues, while association between the master and slave views was constrained by epipolar and motion continuity conditions. This approach with the top-down view as the primary tracking view better mitigated occlusion, as occlusion was more likely to occur in the side views (Qian et al. 2017; Qian and Chen 2017; Liu et al. 2019).

Some studies also employ binocular cameras to perform 3D tracking of fish from one specific angle. Chuang et al. (2015) proposed an improved Viterbi multi-object association strategy to address the frequent object loss issue in fish tracking using low frame-rate camera systems. In their approach, 8-bit low frame-rate video frames collected from underwater trawling were processed using a threshold-based automatic segmentation method to localize the fish and perform stereo matching. Temporal matching was then accomplished based on four cues: Euclidean distance between targets, area differences, motion direction, and histogram differences. The proposed multi-object Viterbi association strategy, based on dynamic programming, established a tree for each target, allowing different paths to share nodes across different trees and effectively addressing occlusion issues. Palconit et al. (2021) placed binocular cameras above the fish tank to capture stereo image pairs of the fish in a top-down view. After calculating the three-dimensional coordinates of the fish centroids, tracking was performed using the K-Nearest Neighbors (KNN) algorithm (see Fig. 9). To evaluate which method could better predict the fish's motion state, both linear and nonlinear algorithms were employed in the experiments, including Multiple Linear

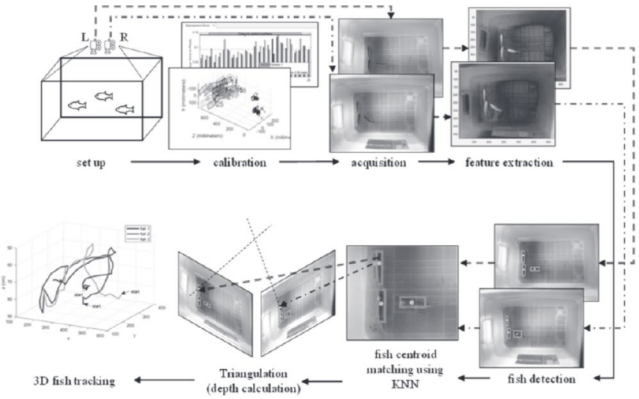

Fig. 9 Example of a workflow that uses binocular camera for tracking Palconit et al. (2021)

Regression (MLR), Adaptive Neuro-Fuzzy Inference System (ANFIS), Gaussian Process Regression (GPR), and Multigene Genetic Programming (MGGP). The results indicated that the nonlinear methods, MGGP and GPR, achieved higher accuracy and better prediction of the motion of freely swimming fish. Among them, GPR had shorter computation time per iteration, making it more suitable for aquaculture applications. Saad et al. (2024) placed a binocular camera system inside a marine aquaculture tank to capture side-view images of fish. YOLO v7 was employed to detect fish targets in the left and right views, and stereo matching was used to compute their three-dimensional coordinates, providing 3D bounding boxes for the detected fish. To overcome the challenge of frequent entry and exit of fish targets, the Deep SORT network was used for tracking based on the 3D detection results of the fish. DeepSORT's matching strategy, based on Mahalanobis distance and cosine distance, together with the matching cascade, provided robust tracking by providing multiple confirmations for lost ID targets.

In addition, there are applications using infrared reflection systems and RGB-D cameras. Pautsina et al. (2015) used an Infrared Reflection (IREF) system, which separates overlapping fish vertically based on the brightness differences in fish images. After calculating the fish centroids, tracking is accomplished by selecting the nearest neighboring target. Compared to binocular cameras, this method has lower hardware costs and does not require stereo matching or precise alignment. However, it cannot detect and track fish at the bottom of the fish tank, and the estimation accuracy is relatively lower. Saberioon and Cisar (2016) adopted a structured light sensor, Kinect I, and utilized idTracker to track four Nile tilapia fish based on their individual feature fingerprints. Compared to binocular cameras, it achieved a lower false detection rate. However, there are limitations on the distance between the sensor and the tracked objects.

The challenges in applying stereo vision to fish tracking and their corresponding solutions are summarized in Table 3 for a clearer illustration. Additionally, the table provides evaluation results of each study and descriptions of the datasets built and used, including annotations, size and preprocessing methods.

4.4 3D reconstruction models

4.4.1 Basic theory of stereo matching

Stereo matching is a common method in stereo vision for obtaining depth information (as discussed in Chap. 3). The pipeline for the generic stereo matching method consists of four steps: cost computation, cost aggregation, disparity computation, and disparity refinement. Traditional methods can be categorized as global methods, local methods and semiglobal methods. Global matching methods are used to obtain more accurate and denser disparity maps. This is achieved by constructing a global energy function which is then minimized using optimization methods. Commonly used methods include dynamic programming (Brown et al. 2003), belief propagation (Sun et al. 2003), and graph cuts (Wang and Lim 2011). However, these methods are computationally intensive and may not be effective in detecting target edges. Local matching methods are computationally efficient and construct a cost function based on the information around each pixel to compute its local disparity. Commonly used methods, such as the block matching methods, gradient-based methods, and feature matching methods (Kumari and Kaur 2016), may be sensitive to noise and may

not perform well in textureless or repetitive regions. To balance accuracy and computational cost, semi-global methods optimize the energy function in the global matching algorithm. They approximate the two-dimensional optimal solution by aggregating costs along multiple one-dimensional paths, improving efficiency while maintaining comparable accuracy to global matching methods. The semi-global block matching method (SGBM), encapsulated in the OpenCV library, introduces the idea of semi-global matching on the basis of block matching method, which can obtain satisfactory disparity at a faster speed. This is one of the most popular methods in simple real-time applications (Shi et al. 2020; Huang et al. 2020; Cheng et al. 2020). Due to noise interference or textureless regions in stereo images, the disparity map may contain errors or noise. Methods such as left-right consistency check, median filtering, and bilateral filtering can be employed to assess and improve accuracy (Zbontar and LeCun 2015).

Deep learning-based methods can further enhance the quality of stereo matching. Initially, CNNs were used to replace manual computation of image similarity for learning the cost function (Zbontar and LeCun 2015). Later, end-to-end networks were introduced (Mayer et al. 2016). Siamese network architectures, encoder-decoder structures, and 3D regularization structures are three popular architectures used for disparity estimation (Zhou et al. 2020a).

4.4.2 3D reconstruction models based on deep learning in aquaculture

Striking a balance between the accuracy of obtaining three-dimensional information of fish and the computational cost of 3D reconstruction is a crucial consideration. To improve the efficiency of stereo matching, some studies have employed object-based matching strategies, which involve matching the detected objects rather than the entire image. This approach is more robust than traditional pixel-based methods (Saad et al. 2024). Ubina et al. (2022) used a video interpolation CNN (VICNN) to synthesize intermediate objects and establish pixel correspondences between the left and right objects. The integration of interpolated signals helps reduce matching errors in targets. Due to the complex light field conditions and image degradation in underwater environments, traditional stereo matching methods based on similarity detection are prone to errors. Therefore, it is meaningful to employ CNN networks with powerful feature extraction capabilities for matching. Several classic stereo matching networks have been proposed, such as GC-Net and PSMNet, which utilize 3D convolution and 4D cost volume for matching (Kendall et al. 2017; Chang and Chen 2018). However, their large computational requirements hinder practical applications. Hsieh and Lee (2023) employed the Adaptive Aggregation Network (AANet) proposed by Xu and Zhang (2020) for 3D reconstruction. The network consists of an Intra-scale Aggregation module (ISA) and a Cross-scale Aggregation (CSA) module. The ISA efficiently performs local cost aggregation and enhances fish body edges using a mechanism similar to deformable convolution, addressing common issues of overlapping occlusions. The CSA combines the coarse discriminative features extracted from weak texture regions through down sampling with local high-resolution details, alleviating the impact of underwater image haze caused by light attenuation and scattering. Addressing the challenge of acquiring ground truth data for establishing stereo matching training datasets from underwater images, Wang et al. (2024) propose a method based on NeRF rendering to generate underwater images from various viewpoints for creating such training datasets. They integrate a

residual network with deformable convolutions into their NeRF-supervised stereo matching network, enhancing the model's capability to learn texture features and mitigate errors in 3D reconstruction at fish body edges. The network also incorporates FPN and a lightweight Triplet Attention module for multiscale feature fusion and improved adaptability in disparity search. The proposed stereo matching model achieves 92.35% in the similarity metric.

Moreover, utilizing binocular images to compute sparse point clouds for reconstructing the visible surfaces of fish bodies is an effective method to meet practical requirements (Huang et al. 2020; Liu et al. 2022). However, three-dimensional reconstruction techniques commonly used for terrestrial targets have not yet been widely applied in underwater fish studies. This aspect will be discussed in the future prospect section.

5 Applications of stereo vision techniques in aquaculture

In aquaculture production, assessing the health status of fish is a primary consideration for ensuring fish welfare and economic benefits. Estimating fish size and biomass in real-time allows evaluation of growth stages and conditions. Monitoring abnormal fish behavior enables timely detection of potential hazards, while recognizing feeding behavior assists farmers in optimizing feeding strategy and enhancing conversion rates. Stereo vision technology, a precise method for extracting fish phenotypic and behavioral characteristics, has been explored and developed in these applications. This chapter provides an overview of size measurement, biomass analysis, and behavior recognition in the intersection of computer vision and aquaculture.

5.1 Morphological size measurement

Fish body size is crucial for assessing the growth cycle of fish in aquaculture. Stereo vision technology can assist farmers in non-contact, rapid, and accurate measurement of fish body dimensions. There are three major types of modes used in fish size measurements (as shown in Fig. 10). The first is to mark or detect landmarks (usually the snout and tail fork of the fish) and convert them into 3D coordinates to find distances. Early methods still relied on human labor. Harvey et al. (2003) developed a computer interface that utilizes cursor positioning and mouse clicks to locate the snout and caudal fin of fish, enabling the calculation of fish length based on the principle of stereoscopic intersection. The average error achieved was 1.72 millimeters. Some researchers have employed Direct Linear Transformation (DLT) to estimate the three-dimensional positions of manually labeled landmarks, yielding error

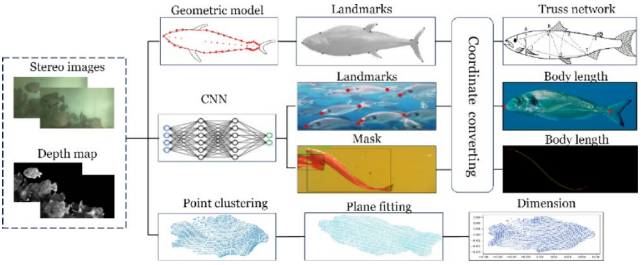

Fig. 10 Three main routes for morphological size measurements of fish in stereo vision

rates of approximately 2.3% (Tanaka et al. 2019) and 5% (Torisawa et al. 2011). Later some algorithms such as pattern recognition (Rosen et al. 2013) and template matching (Shafait et al. 2017) have been applied to reduce the workload. Users only need to deal with the fish in a single image, and the corresponding stereo pairs and subsequent images from the same video can be automatically matched based on inter-frame differences and epipolar lines.

With the advancement of AI, the trend is shifting to using deep learning models to automatically detect key points on fish. Some studies on fish length measurement focus on detecting only two key points, such as the fish's head and tail (or snout and caudal fin), and then calculate the fish's length based on the distance between these key points. Tonachella et al. (2022) developed a low-cost prototype by placing a stereo camera inside a smart buoy and tested it in a commercial mariculture cage. They captured side-view images of sea bream and European sea bass. Using YOLOv4 and RESNET-101 for fish detection and key point detection, they marked the tip of the nose and the midpoint of the tail as landmarks. They established a micron-to-pixel conversion model based on a polynomial regression algorithm and stereo camera calibration information. Length prediction can be automatically performed based on the number of pixels between the landmarks. The accuracy of the length estimation reached 97% compared to the ground truth, with an error rate of ± 1.15 cm. By using transfer learning strategies and a developed low-cost prototype, they presented an affordable and easy-to-use fish length and weight prediction tool that can meet the needs of modern sustainable aquaculture. Deng et al. (2022) constructed an experimental setup with underwater stereo cameras to capture side-view images of five fish species for fish body measurement. An improved Keypoint R-CNN network was utilized to identify fish species and detect key points such as the fish head and tail. The stereo camera calibration parameters were used to perform 3D reconstruction of the key points, obtaining their 3D coordinates. Fish length estimation was calculated based on the Euclidean distance between these coordinates, resulting in an average relative error of less than 10% compared to ground truth. This self-built culture tank differs significantly from actual factory farming scenarios, and further research is needed to explore its practicality in high-density and large-scale aquaculture environments. Deng et al. (2023) deployed a stereo camera setup on the water surface to capture an overhead view of Micropterus salmonids fish, also in an experimental environment. The model employed RetinaNet and CenterNet for object detection and key point detection, respectively, yielding an average relative error of $1.05 \pm 3.30\%$. These two-stage methods offer greater flexibility in selection. However, the aforementioned approaches, which focus solely on the fish head and tail key points, fail to account for measurement errors resulting from the deformation of the fish body between the head and tail.

Many studies have started focusing on the detection of multiple key points to calculate fish body size. Suo et al. (2020) defined seven key points for fish body: the head, eyes, the start and end points of the dorsal fin, pectoral fin, gluteal in, and caudal fin. Underwater stereo cameras were used to capture side-view images of cultured fish, and the Stacked Hourglass network was employed for key point detection, achieving an average error rate of 5.58%. Garcia et al. (2020) utilized a commercial deep vision system deployed in an underwater trawl net to capture stereo images of cultured fish in a side-view. Landmarks were defined as the intersections between the fish skeleton and the fish body patches. The fish body patches were segmented using Mask-RCNN, and the skeleton was estimated as a cubic polynomial based on morphological operations. This approach has an advantage in handling cluttered images containing overlapping fish, but the validation of the size mea-

surement has not been provided yet. Hsieh and Lee (2023) identified nine key points on the fish body, including the mouth, pectoral fin, pelvic fin, start and end points of the dorsal fin, and the connecting point between the tail fin and the tail. They employed a bottom-up approach to directly detect these key points on the spotted sea bass and encoded the relationships between different parts based on part affinity fields to reconstruct the fish skeleton. The length and height of fish were calculated by considering the distances between different key points, resulting in relative errors of 4.49% and 10.49%, respectively. This method, inspired by human key point detection, can better encode the skeletal relationships of fish bodies (Voskakis et al. 2021). However, due to the diverse morphological characteristics of different fish species, the generalizability of such methods based on human key point detection for different fish species deserves further research. Yu et al. (2023) utilized an improved keypoint R-CNN and CREStereo model to detect and obtain three-dimensional information of seven key points on the body of *Oplegnathus punctatus*. The fish body length was fitted as a spatial curve passing through four key points. This method overcomes the limitation of using the Euclidean distance between just the fish mouth and tail base points as a size prediction, which fails to accurately reflect the curved length of the fish body, thus demonstrating high practical value. The authors developed fish length prediction software based on the model and integrated it into the National Digital Fisheries Platform (China) for real-time monitoring of fish growth in recirculating aquaculture systems. In practical production scenarios, the average accuracy of size estimation reaches 93.18%.

The second approach involves utilizing the geometric contour of the fish. Tillett et al. (2000) manually fitted a three-dimensional point distribution model (PDM) to capture the fish body boundaries and measure the skeletal networks. However, the fitting performance is easily affected by fish body curvature and occlusions. Muñoz-Benavent et al. (2018) developed three novel deformable geometric models to fit the contour of bluefin tuna. To better initialize the model parameters, they also integrated feature extraction algorithms, resulting in an error rate of less than 3%. However, the applicability of these models to other fish species needs further exploration. Shi et al. (2020) employed geometric calculations based on background subtraction and convex hull. The two farthest points on the contour were selected as landmarks, and the 3D distance was computed to obtain the fish length. This yielded an average relative error of less than 2.55%. Zhou et al. (2023) employed contour tracing algorithms and the minimum bounding rectangle to distinguish between linear and curved fish based on binary fish patches. The fish skeleton was extracted using the Zhang-Suen thinning algorithm and then straightened to calculate the length, resulting in an average relative error of 0.91%. However, the two aforementioned methods, which employ simple binary processing for fish localization, may not be suitable for complex background scenarios. Additionally, geometric contour-based methods usually require intricate computational processes, may be applicable only to specific fish species, or require the fish to maintain a specific pose during image capture. Further validation is required to assess the generalizability of these methods.

The third approach is based on sparse 3D point clouds, which can be generated from stereo images or depth maps. Risholm et al. (2022) used a range gating 3D system UTOFIA, which can capture high-resolution depth information even in turbid water conditions. Depth maps were converted into point clouds through camera intrinsic parameters, and an unsupervised clustering technique, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), was used to segment fish and extract their centerlines directly from the

depth maps. In addition, this study used a tracker to follow the fish and filtered the length estimates for each frame in the same trajectory to obtain more robust length measurements. However, the cost and complexity of such equipment need to be considered for practical aquaculture applications. More research has been conducted to obtain fish body masks from RGB image pairs and combine them with generated point cloud data, which involves a two-step process. Liu et al. (2022) utilized a stereo camera with infrared functionality to capture underwater fish images. The added infrared texture alleviates the weak texture issue in the scene. The original images underwent YOLOv5 detection for object classes and SOLOv2 for fish segmentation masks. By combining the image and point cloud data, the corresponding point cloud of the fish mask was filtered. DBSCAN was applied to process the fish point cloud, followed by plane fitting and PCA to calculate the fish dimensions. The resulting accuracy of length estimation reached 96.9%. Huang et al. (2020) combined the segmented image mask and 3D point cloud to obtain the fish point cloud. They estimated the length and width of the fish in different positions and orientations through coordinate transformation and fitting. The average errors were 5.5 mm and 2.7 mm, respectively. To address the depth calculation errors at fish body edges caused by occlusions in one view of stereo images, Li et al. (2022) proposed a Cross-modal Feature Fusion Mask R-CNN (CMFF Mask R-CNN). By integrating RGB-disparity cross-fusion modules into Mask R-CNN, they enhanced the integrity and accuracy of three-dimensional reconstruction of fish body mask edges. To more accurately fit the posture of curved fish bodies, the authors introduced a fish length estimation algorithm based on point cloud normalization segmentation transformation. This method divides the fish body outline into three parts and computes their sum to derive the fish length. The average relative error in estimating salmon length is less than 5%, and for other freshwater fish species, it is also less than 5%. While RGB image pairs cannot generate high-density and high-precision point cloud data, they can be applied to underwater fish size measurement at a lower cost.

In order to provide a comprehensive overview of the applications of stereo vision for fish size measurement, a table was constructed to summarize the relevant literature. This table presents a detailed overview of the various measurement modes, specific methods, comments, and evaluations, which can be seen in Table 4.

5.2 Biomass estimation

Fish biomass refers to the total mass of fish in a specific area of water. It is typically estimated by multiplying the number of fish in the culture nets by their average mass (Li et al. 2020b). Therefore, fish biomass is closely related to both mass and quantity. The mass can be estimated from size data through computer vision technology and mathematical models. There are two types of mass estimation models: single-factor and multi-factor models, which can be fitted using linear, power, logarithmic, or second-order polynomial equations.

In the single-factor model, the most commonly used relationships are the length-mass and area-mass relationships. The power function model, $W = aL^b$ (TW 1904), is the most classical length-mass relationship model. The model determines the mass of a fish (in grams) based on its total length (in centimeters) and two experiential coefficients, a and b. The coefficients are influenced by factors including species type or growth variation (Hile 1940; Cren 1951). This simple length-mass relationship has been widely used in early fish mass estimations (De Verdal et al. 2014; Viazzi et al. 2015). However, the majority of the

studies rely on 2D imagery, which requires the fish to be anesthetized or guided through a restricted artificial channel, resulting in certain limitations (Hufschmied et al. 2011). Methods have been developed to measure fish length using stereo vision technology, which revitalizes mass prediction based on the straightforward length-mass relationship (Da Silva Vale et al. 2020; Tonachella et al. 2022; Shi et al. 2022). The correlation between area and mass of fish has been demonstrated to be stronger. Liang and Chiou (2009) used manual measurement and 2D image processing techniques to conclude that the correlation between tilapia's body mass and its projected area was the strongest among various single-factor models ($R^2=0.9303$). Gümüş and Balaban (2010) utilizing 2D images concluded that the optimal fit was achieved with a power equation ($Y=A\cdot X^B$) having an R-squared value of 0.99. Here, Y represents the estimated mass, and X denotes the field of view's area, typically calculated by pixel conversion to centimeters. Compared to multi-factor models, the area-mass relationship shows similar promise (Balaban et al. 2010b; Viazzi et al. 2015). Thus, the single-factor model is preferred and utilized as it requires fewer parameters for prediction and its accuracy satisfies practical needs (Hufschmied et al. 2011; De Verdal et al. 2014). Studies based on 3D images also show the accuracy of the area-mass relationship. Shi et al. (2022) employed binocular stereo vision technology to establish the mass-area and mass-length relationships, along with six fitting formulas between mass and fish body area. The results indicate that the accuracy of the area-mass relationship surpasses that of the length-mass relationship, and the linear model is the most appropriate fitting formula. Yu et al. (2022) utilized underwater binocular cameras to capture lateral views of *Oplegnathus punctatus*. They employed an improved segmentation model called CAM-Decoupled-SOLO to extract the contour perimeter of the fish body. Linear, power-law, and square root models were fitted to establish the perimeter-weight relationship. The results demonstrated a significant correlation between fish body perimeter and weight, with the square root model being the most suitable predictive model.

Multifactorial estimation models are frequently analyzed using multivariate linear regression analysis. Beddow et al. (1996) utilized a digital stereoscopic camera system to capture lateral images of fish. They measured specific combinations of several parameters, including truss distances between different key points such as fins and eyes, and depth of the fish body in different locations. From these measurements, they established a series of multifactorial regression equations. They selected twelve applicable measurements, and estimated the individual mass with an error of -0.1. Costa et al. (2013) used a combination of contour morphometry (elliptic Fourier analysis) and multivariate techniques (based on partial least squares modeling) to model and estimate fish body masses. The study utilized 2D image processing to extract fish body area, lengths of the long and short axes, perimeter, and size of the center of mass parameters. The correlation coefficient was found to be $r=0.9875$. Da Silva Vale et al. (2020) utilized fish body length, width, and other parameters to calculate fish body volume. They generated a point cloud of the fish body surface using a stereo camera and used the scattering theorem to predict fish body mass. The error rate was approximately 4.55%. Hsieh and Lee (2023) used stereoscopic vision to automatically extract nine key points on the fish body and generate the fish body skeleton using a deep-learning-based pose prediction network model. They then established a linear regression model between the length and height of the fish body and its mass, with an error of 5.035%. Wang et al. (2024) proposed a framework for three-dimensional reconstruction of fish body contours integrating NeRF-supervised stereo matching networks and MiVOS instance seg-

Table 4 Summary of applications of stereo vision in morphological size measurement of fish

Mode	References	Facility	Environment	Method	Comments	Results and evaluations
Manually	Harvey et al. (2003)	stereo–video system	Sea cage	Manually mark key points and converted to 3D coordinates based on triangulation.	Time-consuming, and labor-intensive.	Snout to fork length: relative error of 0.16%; Maximum body depth: relative error of 0.51%
	Dunbrack (2006)	in-line stereo camera system	Ocean	Manually mark key points and converted to 3D coordinates based on triangulation.	Has less stringent alignment requirements than other stereo techniques	Fork Length: average absolute error of 1.6%
	Torisawa et al. (2011)	stereo-video system	Sea cage	Manually mark key points and converted to 3D coordinates using direct linear transformation (DLT).	Estimate robust fish lengths by taking the mean of 20 sequential frames of data.	Fork length: error ratios less than 5%.
	Tanaka et al. (2019)	Multi-camera system	Aquaculture cages	Manually mark key points and converted to 3D coordinates using DLT.	More accurate than other stereo techniques	Fork length: accuracy approximately of 2.3%.
Semi-automatically	Rosen et al. (2013)	Deep Vision system	Coastal trawling	Semi-automatically marked key points based on pattern recognition;	Construct a model relating each landmark distance to fork length for fish bending and occlusion.	Body lengths: error less than 1% for the majority of individuals.
	Shafait et al. (2017a)	AM100 stereo vision system	Sea cage	Semi-automatically marked key points based on template matching.	Inability to handle fish bending and occlusion.	Body lengths: error of less than 1%.
	Tillett et al. (2000)	Stereo cameras system oriented vertically	Experimental platform	An improved 3D point distribution model (PDM) that combines edge strength and proximity were used to measure fish truss networks by capturing fish body boundaries.	Contour boundary delimitation was independent of the size and orientation of fish, but the fitting algorithm needs to be initialed manually.	Fish length: average error of 5% for well-presented fish

Table 4 (continued)

Mode	References	Facility	Environment	Method	Comments	Results and evaluations
	Marrable et al. (2023)	stereo-BRUVS	Dataset	Click on the fish body in the image arbitrarily, and the fish body could be cropped from images based on YOLOv5. A YOLOv5 small model was trained using fish image data with head and tail labels to detect head and tail key points.	Inability to handle fish body at an angle of 30° to 45° to the camera plane.	Landmarks: precision of 77.40%.
Automatically	Ravanbakhsh et al. (2015)	Stereo-video system	Sea cage	Automatic landmarks marking using haar classifier and fish outline captured based on shape-based level-sets framework	Ability to overcome limitations such as poor contrast boundaries, background clutter, and occlusions.	Fish detection: accuracy is 94.3% with a buffer width of 5 pixels.
	Shi et al. (2020)	Stereo camera system.	Closed recirculating aquaculture system	Fish body contours were obtained through background subtraction and morphological operations. Landmarks were then localized based on convex hull approximation.	Lower accuracy when the angle between the fish and the optical axis of the camera was not orthogonal.	Body length: mean relative error less than 2.55%.
	Zhou et al. (2023)	RealSense D435	Experimental platform	Landmarks were identified using a contour finding algorithm and a minimum external rectangle algorithm.	Skeletons were extracted and straightened to handle fish bending. Easier than deep learning methods that require large datasets.	Body length: mean relative error of 0.91%, the maximum relative error less than 4%.
	Garcia et al. (2020)	Deep Vision system	Cruise trawling	Fish were segmented using Mask RCNN and improved with localized gradients. The fish skeletal curves were approximated using RANSAC-estimated cubic polynomials. The endpoints of fish were obtained by identifying the intersections of the skeleton with the fish body masks.	Ability to successfully process cluttered images containing overlapping fish species.	Mask RCNN: IoU* average values of 0.89 and 0.79 for non-overlapping and overlapping fish respectively.
	Costa et al. (2006)	Stereo-video system	Sea cage	Fish contours were obtained by using elliptical Fourier analysis and post-processed with ANNs to return corrected 3D positions.	Suitable for specific fish species	Fish length: error about 5%.

Table 4 (continued)

Mode	References	Facility	Environment	Method	Comments	Results and evaluations
	Muñoz-Benavent et al. (2018)	AM100 stereo vision system	Sea cage	Fitting the ventral silhouette of tuna with a deformable geometric model. Development of snout and fork feature extraction algorithms to better initialize model parameters	Identifying straight fish bodies using the bending angle of the geometric model and discarding curved fish samples	Up to 90% of the samples bounded in a 3% error margin
	Williams et al. (2016)	CamTrawl system	Ocean	Fish were segmented based on background subtraction and thresholding. The endpoints of fish were obtained by identifying the extreme points.	Define two more points along the midline to account for the curvature of the fish body.	Coefficient of variation (CV) of 3%
	Risholm et al. (2022)	Range gating 3D system	Fishery cage	Fish were segmented using DBSCAN based on depth map. Fish lengths were estimated based on extracted centerline.	Estimate robust fish lengths by taking the mean of per-frame estimates.	Fish length: relative error less than 2%.
Automatically	Wu et al. (2019a)	Binocular camera in an AUV platform	Sea	Fish were detected using YOLO v2 and edges were extracted with binarization process and Canny operator. Fish height and length were calculated with the depth value obtained through SGBM-based stereo matching	Morphological-based edge extraction methods face challenges in handling the non-rigid deformations of fish.	Fish length: error in 0.2–1.2 cm
	Huang et al. (2020)	Binocular camera	Not mentioned	Fish were segmented using Mask-RCNN and refined with GrabCut and morphological processes. 3D point cloud was obtained after coordinate transformation, and fish dimensions were estimated through a fitting process.	It was assumed that the fish outline is coplanar, and the curvature of the fish body was not taken into account.	Fish length: error of 4%. Fish width: error around 2.7 mm.
	Ubina et al. (2022)	Stereo-video system	Aquaculture pond and open-sea cage	Fish were segmented based on Mask-RCNN. The disparity computation was refined with VICNN based on SGBM. The converted coordinate points estimated the body dimensions based on the fish posture recognition result.	Low-cost stereo camera system with accuracy distance limitation.	Fish length: error rate less than 5%.

Table 4 (continued)

Mode	References	Facility	Environment	Method	Comments	Results and evaluations
	Deng et al. (2022)	Stereo-video system	Experimental platform	Species identification and keypoint detection using an improved key point R-CNN.	Integration of the attention module into the backbone network and replacement of FPN with I-PANet to reach a higher precision.	Fish length: relative errors less than 10%.
	Tonachella et al. (2022)	Smart buoy with stereo camera	Sea cage	Fish were detected using YOLO v4 and the results were fed into CNN RESNET-101 to automatically identify landmarks.	A low-cost prototype and highly automated tools for use in marine ranching	Fish length: accuracy reached 97%.
	Voskakis et al. (2021)	Stereo camera	Sea cage	Fish skeletons were reconstructed based on part affinity fields (PAFs) and Openpose that directly detects6 key points. Fish length was then calculated based on LabVIEW.	Set an acceptable threshold which is at 40% to reject the unaccepted values.	Length estimation with mean relative error of Gilthead seabream and the European seabass of 3.15% and 7.4%, respectively.
	Hsieh and Lee (2023)	Stereo camera	Aquaculture pond	Fish skeletons were reconstructed based on part affinity fields (PAFs) and Openpose that directly detects 9 key points. Fish body parameters were then calculated with the depth values obtained from the depth prediction model that based on adaptive aggregation network.	Adoption of methods designed for human pose recognition which can better encode the skeletal relationships of fish bodies	Fish length: relative error of 4.49% Fish height: relative error of 10.49%
Automatically	Liu et al. (2022)	RealSense D415	Not mentioned	Fish body detection was implemented using YOLOv5, and the results were fed into SOLOv2 for segmentation. 3D point cloud was obtained after coordinate transformation and got fish dimensional estimations through DBSCAN and PCA.	Avoiding repetitive measurements for the same fish by adding a SORT-based target tracking algorithm. While there are certain requirements for the distance between the fish and the camera.	Fish length: accuracy of 96.9%. Fish width: accuracy of 91.8%.

Table 4 (continued)

Mode	References	Facility	Environment	Method	Comments	Results and evaluations
	Deng et al. (2023)	Stereo camera	Experimental platform	Fish body detection was implemented using RetinaNet and then key points detection was performed on the cropped fish body image using CenterNet.	A modified 3D reconstruction method was proposed to eliminate the effect of light refraction when using surface stereo vision.	Fish length: mean relative error 1.05±3.30%.
	Yu et al. (2023)	Stereo camera	Aquaculture pond	Seven key points of fish were detected based on improved Keypoint RCNN (SE-D-KP-RCNN). The disparity was calculated using CREStereo. Length and height of multi-pose variant fish body were calculated based on curve fitting.	Using improved keypoint RCNN and CREStereo, the accuracy of fish keypoint detection and 3D reconstruction has been enhanced. Additionally, fitting multi-segmented 3D curves can further improve the accuracy of measuring curved fish body dimensions.	Fish length: accuracy of 96%. Fish height: accuracy of 96.29%.
	Li et al. (2022)	Stereo camera	Aquaculture pond	Fish were segmented using a cross -modal feature fusion Mask R-CNN. The 3D point cloud of the fish body contour is obtained through SGBM computation and subsequent normalization and segmentation transformation. The fish body length is calculated by summing the measurements of three parts of the contour separately.	Combining features from RGB images and depth images enables more comprehensive extraction of fish body contours. Using multiple segments for measurement can better fit the shape of curved fish bodies, thereby enhancing the accuracy of length measurements.	Fish length: mean relative error less than 5%

mentation. They computed the fish perimeter by summing distances between 3D coordinate points of the contour. By establishing linear, power, and logarithmic relationships between perimeter and fish weight, the power relationship demonstrated more stable and accurate weight predictions with an R value of 0.98. The RMSE and mean absolute error (MAE) were 18.13 and 14.72, respectively. Hao et al. (2016) employed Mask R-CNN and SGBM stereo matching for fish segmentation and three-dimensional reconstruction, constructing models for mass and length relationships excluding the caudal fin using least squares regression. Specifically, they estimated the angle between the fish's major axis and the optical axis, discarding predictions outside the 60° to 120° range. Evaluation of weight accuracy on the test dataset resulted in $R^2 = 0.89$, RMSE = 6.58 g, and MAE = 8.02 g. These highlight a growing emphasis on accurate size prediction and the research trend towards multi-factor estimation model of weight.

Researchers have proposed using machine learning methods to model the relationship between multiple morphological parameters and mass. Odone et al. (2001) utilized an SVM model to determine the correlation between fish mass and shape parameters, thus predicting fish mass. Saberioon and Císař (2018) utilized an RGBD camera to extract eight dorsal geometric features, such as the total area, maximum width, and length of the fish's back. Mass prediction was performed using Random Forest and Support Vector Machines, both of which demonstrated significant predictive power. Support Vector Machines exhibited a higher coefficient of determination and superior performance. Zhang et al. (2000) equipped an underwater remotely operated vehicle (ROV) with a binocular vision system and utilized an enhanced YOLOv5n model to measure the length and height of fish bodies. They developed a fully automated biomass estimation system using CatBoost to establish relationships between length, height, and weight. Compared to single-factor models linking length-weight and height-weight, this multi-factor weight prediction model combining deep learning and machine learning achieves higher accuracy, with an average relative error (MRE) of 2.87%.

Some studies have used IREF systems to acquire infrared images of fish bodies containing depth information to predict fish mass. Other studies have combined stereo vision with acoustic techniques to obtain information on fish size and density (Boldt et al. 2018). Additionally, GANs have been used to convert sonar images into realistic daytime images for fish morphometric measurement and mass estimation.

Among the various mass prediction models, the area-mass single-factor model and the multifactor model that considers the inclusion of area and more factors correlate better, but there are four issues to be considered for mass prediction. Firstly, fins and tails contribute significantly to the area but negligibly to the mass, which may result in a certain estimation error. Balaban et al. (2010a) removed the fins and tails and found that there was no significant improvement in the matched R^2 compared to the whole fish image. In contrast, Viazzi et al. (2015) used perch shapes without fin tails, and the area-mass model predicted better results, with little change in R^2 and reduced root-mean-square error. Secondly, the appropriate camera angle, either dorsal or side view, needs to be determined based on the shapes of the fish species being tested. Most studies photographed the side of the fish to obtain morphological features. However, for flat-shaped fish, photographing from the top was more suitable to capture the overall contour. Thirdly, the main components affecting mass vary in different growth stages. In the early stages, growth is likely to occur in the length and area of the fish, while in later stages, it may occur in the thickness of the fish. In mature fish populations, the presence of both male and female individuals may impact the

predicted outcome, as females are likely to carry eggs during certain seasons. Additionally, underwater fish occlusion, fish body curvature, and poor image quality remain persistent issues in free-swimming fish studies.

Table 5 summarizes the results of studies that utilized various prediction models, factors, and formulas. It also summarizes the species and conditions of the fish and the facilities used in each study.

In aquaculture practice, there is a greater emphasis on estimating the total weight of fish in a fish pond. This can be achieved by calculating the average weight of the cultured fish and combining it with the quantity of fish to make an estimation. In the process of automatic fish counting using computer vision techniques, the depth information provided by stereo vision can assist in resolving the issue of fish occlusion in high-density scenarios. The primary methods for fish counting include detecting or segmenting the fish target, or counting the number of trajectories tracked in consecutive frames. Wu et al. (2019b) utilized the SSD detection method in combination with the SORT algorithm to detect fish every ten frames in a sequence of left and right images captured by a stereo camera. The frame with the highest confidence level is then selected as the final result for fish classification and counting.

Fish biomass can also be determined directly by using laser scanning methods, which express it as the product of fish density and volume. The volume of fish is converted to total biomass by assuming that fish density equals water density. Almansa et al. (2015) used a laser scanning system to measure the ratio of fish biomass (FB) to fish layer volume (FLV) and estimated the proportion of fish biomass in the measured fish layer volume by comparing it to 1. The coefficient of variation was below 7.2%. This method can detect fish biomass in high stocking density aquaculture facilities, particularly for bottom-dwelling fish. However, it does not provide sorting information on the fish. In addition, the availability of real biomass density value of different fish species and the complexity of laser scanning equipment require further consideration.

5.3 Behavior analysis

The welfare conditions of fish are closely related to their individual or group behavior, such as health, appetite and stress levels. Traditionally, fish diseases and deaths are detected through manual observation, such as by their turning over, jumping, or splashing. However, this method is time-consuming and labor-intensive. Current research primarily focuses on automated monitoring and quantification of fish behavior using stereo vision technologies, providing farmers with more realistic, 3D information over extended periods without requiring additional labor.

The research on fish behavior analysis in aquaculture serves two main purposes. The first is to identify stress factors that cause abnormal behavioral responses in fish, such as poor water quality, overcrowding, and disease. This information is then used to improve the welfare of fish by adjusting management strategies. The second is to analyze the appetite and feeding status of fish using various behavioral and locomotor parameters, which can assist in making intelligent feeding decisions. The analysis of fish behavior encompasses individual and group behavior, which can be categorized into qualitative and quantitative indicators. Qualitative indicators typically include swimming trajectory, distribution location, behavioral changes, swimming direction etc. Quantitative indicators include fish swimming speed, acceleration, Tail-beat frequency (TBF), turning angle, distance between each fish,

and Wall-hitting rate (WHR)(An et al. 2021). There are also human defined quantitative indexes, such as flocking index of fish feeding behavior (FIFFB) (Zhou et al. 2017b), and computer vision-based feeding activity index (CVFAI)(Liu et al. 2014).

Currently, there has been progress in the research of fish behavior analysis based on 2D video. Xiao (2015) located the head and tail points of fish using an edge detection algorithm. They calculated the tail-beat frequency (TBF) and wall-hitting rate (WHR) of fish under the effect of NaOH and glyphosate and concluded that a combination of these two indexes can provide good feedback for monitoring water quality. However the authors discussed that a multi-camera view could provide richer information. Måløy et al. (2019) used a 2D-convolutional spatial network, a 3D-convolutional motion network, and LSTM to extract spatial-temporal features of salmon feeding behaviors with a prediction accuracy of 80%. However, mapping the 3D motion trajectories of fish to a 2D plane results in a loss of depth information. An et al. (2021) concluded that to successfully quantify fish behavior, it is important to improve multi-target tracking accuracy and utilize 3D devices. Current 3D-based researches can be divided into the following categories based on their practical applications:

5.3.1 Water quality bio-indicator

Water quality bio-indicators of fish behavior refer to the reflection of different water quality conditions through the movement characteristics of fish. Xu et al. (2024) utilized the improved YOLOv8 model as the object detection method, in combination with the Kalman filter, Kuhn Munkres (KM) algorithm, and kernelized Correlation Filters (KCF), to obtain 3D positional information of fish. The study analyzed the behavioral changes of fish under different levels of ammonia-nitrogen stress through qualitative and quantitative measures, including behavioral trajectory, exercise volumes, spatial distribution, and movement speed. The study revealed that various fish species exhibit distinct movement patterns when subjected to ammonia stress.

Some studies have used machine learning methods to train water quality judgement models. Cheng et al. (2019) investigated the relationship between water quality and fish behavior. The study utilized the KM algorithm's association strategy for tracking, with state updates through Kalman filtering and occlusion compensation via KCF. The velocity, acceleration, curvature, swimming distance of the center, and dispersion parameters of the fish were calculated from the 3D trajectories. They used integrated learning that combined with SVM, eXtreme Gradient Boosting, and PointNet-based classifiers for model training. A mapping model was established between the characteristic parameters of fish movement behavior and the water quality environment, ultimately enabling the judgment of normal and abnormal water quality with a recognition rate of over 95%. The use of 3D behavioral monitoring offers more comprehensive location data and overcomes occlusion issues, providing a novel approach for detecting of abnormalities in aquaculture water bodies.

5.3.2 Abnormal behavior recognition

Abnormal behavior of fish is closely related to their welfare status and usually includes hypoxic behavior, stress behavior and disease behavior. Hypoxic behavior is common in aquaculture. Some researchers have applied stereo vision technology to monitor hypoxic behavior of fish in the early years (Israeli and Kimmel 1996). By using multiple vertically

Table 5 Summary of applications of stereo vision in mass estimation of fish

Mode	References	Facility	Fish		Factor	Formular	Result
			Species	Condition			
Single-factor	Liang and Chiou (2009)	2D	Taiwan tilapia	Dead fish	Length, height, perimeter or area	Linear regression	Area-mass estimation was more accurate with $R^2=0.9722$
	Balaban et al. (2010a)	2D	Pollock	Dead fish	Area	Linear, power, and Polynomial	Area-mass power model was the best with $R^2=0.993$
	Balaban et al. (2010b)	2D	Four types of salmon	Dead fish	Area	Linear, Linearized, power, Polynomial	Area-mass power model was the best with $R^2=0.987$
	Hufschmied et al. (2011)	2D	Siberian sturgeon	Swimming through a sorting device	Area	Linear	Average relative error of the model was ±5.7%.
	Serna and Ollero (2001)	Stereo	Gilthead sea breams	Lively fish	Length	Power curve	Standard deviation of fish mass was 8.40 g.
	Martinez-de Dios et al. (2003)	Stereo	Gilthead sea breams	Lively fish	Length	Power curve	Errors was lower than 5%.
	Tonachella et al. (2022)	Stereo	Gilthead seabream and European sea bass	Lively fish	Length	Power curve	Mean error of 3%.
	Shi et al. (2022)	Stereo	Spotted knifejaw	Lively fish	Area or Length	Linear, logarithmic, power, exponential, quadratic	Area-mass linear model was the best fitting formular ($R^2=0.96$).
	Da Silva Vale et al. (2020)	Stereo	Nile Tilapia	Lively fish	Volume or Length	Regression	Average error from volume was 7.85%, and from length was 9.79%.
	Yu et al. (2022)	Stereo	Oplegnathus punctatus	Lively fish	Perimeter	Linear, power curve, square root	The square root model had the highest accuracy with MAPE = 6.54
	Wang et al. (2024)	Stereo	bass	Lively fish	Perimeter	Linear, logarithmic, power	The power model was more precise with $R=0.98$, MAE = 14.72 and RMSE = 18.13
	Hao et al. (2024)	Stereo	Oplegnathus punctatus	Lively fish	Length	Least square (LS)	$R^2=0.89$, RMSE = 6.58 g and MAE = 8.02

Table 5 (continued)

Mode	References	Facility	Fish		Factor	Formular	Result
			Species	Condition			
Multiple-factor	Costa et al. (2013)	2D	European seabass	Lively fish	Area, major and minor axis length, perimeter and centroids size	Partial least squares (PLS)modelling	Correlation coefficient between the observed and predicted mass was 0.9875. RMSE = 16.03.
	Viazzi et al. (2015)	2D	Jade perch Scortum barcoo	Lively fish	Length, area, height	Polynomial, Linear, power	The best single-factor model: area-mass with $R^2 = 0.98$ The model using all three factors with $R^2 = 0.99$ were more accurate.
	Fernandes et al. (2020)	2D	Nile Tilapia	Lively fish	Area, length, height and eccentricity	Multiple linear regression models	The best model had area and square of area as predictors with $R^2 = 0.95$.
	Pache et al. (2022)	2D	A variant of the catfish species	Lively fish	31 extractors such as Hu Image moments, contour properties, major axis, minor axis, area, and perimeter etc.	Linear Regressor, Random Forest Regressor, Multi-layer Perceptron and Support Vector Regression	The best result was the Linear Regressor obtained a $R^2 = 0.76$ and MAE = 0.83 g.
	Beddow et al. (1996)	Stereo	Atlantic salmon	Lively fish	Body depth and truss dimensions	Multifactor regressions (Logarithmic)	Prediction accuracy within ±5%
	Mathiassen et al. (2011)	3D camera and diode laser module	Herring	Frozen	View area, length, width, middle cross-sectional area, maximum cross-sectional area, volume, thickness	Multiple linear regression	Using several 2D and 3D features enabled more accurate mass estimation than using 3D volume only with RMSE of 5.6 g.
	Saberioon and Císař (2018)	RGBD camera	Seabass	Lively fish	Total dorsal area, total dorsal length, maximum width, five equidistant fish widths.	SVM and RF	SVM algorithm with R^2 of 0.872 and RMSE of 0.13 gave a better prediction of mass compared to RF.
	Hsieh and Lee (2023)	Stereo	Oplegnathus punctatus	Lively fish	Length and height	Polynomial regression	Error of 5.035%

Table 5 (continued)

Mode	References	Facility	Fish		Factor	Formular	Result
			Species	Condition			
	Zhang et al. (2024)	Stereo	oplegnathus punctatus	Lively fish	Length or height	Regression	R^2 of 0.98 and a MRE of 2.87%

placed cameras, the 3D coordinate values under hypoxic conditions were obtained. The results show that the spatial distribution of fish tends to move upward under hypoxia. Bao et al. (2018) obtained the 3D motion trajectories of fish under different dissolved oxygen concentrations by converting the 2D center point coordinates of fish to 3D coordinates based on the imaging relationship between mirrors and a single camera. The results showed that a decrease in dissolved oxygen content led to an increase in the average height of swimming fish's position in the tank. Additionally, a large number of fish began to float or even appeared to be moribund when the dissolved oxygen content was lower than 2 mg/L.

To investigate how fish exhibit behavior in response to stimuli, Butail and Paley (2012) developed a high frame-rate tracking framework and reconstructed a 3D model of the fish body using generative modeling techniques. Instantaneous states were optimized using simulated annealing (SA) and used as a search strategy in state space, which allowed for larger variation in pose and shape of fish. The study found that the curvature profile of the fish differs between gliding turns without spooking and fast-starting states with spooking. The system can reconstruct the position, orientation, and shape of individual fish in a dense school and can be used to recognize abnormal behavior of fish group by investigating their fast-starting time characteristics. To investigate the effects of benthic trawl aquaculture on benthic fish, Williams et al. (2013) studied the avoidance behavior of benthic vermilion snapper in response to a camera truck using multiple benthic stereo cameras. The tracked parameters included individual fish mean distance off the bottom, individual velocity, and group locomotion parameters such as group velocity, speed, and swimming. The results indicate that fish respond to external stimuli in three stages: conscious non-avoidance during the initiation stage, avoidance without speed increase during the intermediate stage, and changing swimming direction and accelerating away from the threatening behavior during the third stage.

AlZu'bi et al. (2015) utilized behavior analysis software to extract motion features of zebrafish such as speed, deceleration/acceleration, activity time, sharp movements, occupancy, and spatial activity distribution. The feature vectors extracted from the stationary state were used as a baseline to normalize the remaining states. quadratic discriminant analysis (QDA), kNN, and SVM classifiers were utilized to distinguish between normal and abnormal fish activity. The study showed that the fish exhibit distinct abnormal behavior patterns in response to stress, which can also aid in identifying sick fish in aquaculture.

5.3.3 Feeding behavior

Currently, research on fish feeding behavior is focused more on the feeding intensity identification. Ye et al. (2016) utilized traditional segmentation and optical flow methods to extract motion characteristics such as speed and turning rate from Nile tilapia shoals. They employed composite entropy to assess the shoal's feeding intensity and developed corresponding feeding strategies. In recent years, more studies have trained deep learning networks using manually annotated datasets of fish feeding intensity to achieve automatic recognition. Such research typically categorizes fish feeding intensity into 'none,' 'weak,' 'medium,' and 'strong,' achieving accuracy rates of around 92–97% (Feng et al. 2022; Zheng et al. 2023; Zhang et al. 2024). Atoum et al. (2014) trained three machine learning classifiers, Maximum Likelihood Estimation (MLE), Adaboost, and SVM, using labeled videos data to classify fish activity and estimate active feeding duration. Additionally, a

feed detector based on a correlation filter was developed to detect overfeeding and prevent it. However, the accuracy of the single-camera based system in acquiring fish movement parameters needs to be further improved. Stereo vision technology has been proven to be more accurate in acquiring a variety of fish morphology and movement parameters, as it better reflects the 3D distribution and movement trajectories of fish in real situations. Therefore, it has the potential to assess the willingness of fish to feed more accurately. However, the application of stereo vision technology is still in its early stages, limited to tasks such as fish body tracking and motion parameter acquisition.

6 Challenges and future prospect

The pursuit of large-scale, high-efficiency, and sustainable production in the aquaculture industry requires more precise and intelligent technologies for production management. The application of stereo vision in fish phenotypes and behavioral analysis has shown positive increase, but their applications are still limited now. Several factors may contribute to this phenomenon. Firstly, the complex underwater environment and the characteristics of freely swimming fish pose challenges that have slowed the initial adoption of stereo vision technology in research. Additionally, due to the need for prior calibration and subsequent stereo matching, many aquaculture enterprises or farmers tend to favor simpler operation with monocular cameras. Thirdly, applications such as fish abnormal behavior recognition and feeding behavior analysis rely not only on stereo vision-based 3D tracking methods but also require additional aquaculture expertise. This necessitates researchers possessing interdisciplinary knowledge and building multimodal annotated datasets.

Based on the critical analysis of current stereo vision technology applications and the actual needs of aquaculture, this section discusses the current challenges and prospects for the future development from three main aspects: data acquisition, modeling techniques, and applications.

6.1 Establishment of a large-scale public multimodal stereo vision database

The availability of large and diverse datasets is crucial for the successful application of stereo vision combined with deep learning techniques to solve aquaculture problems. There are few public stereo vision datasets available for underwater fish, and benchmark datasets are virtually non-existent. Most current studies collect data in laboratory environments with simple water quality, shallow depths, small numbers of fish, and single species. It is hard to guarantee the performance of models trained on this basis in practical aquaculture applications. Some recent 2D image-based studies have proposed automatic fish classification using transfer learning or unsupervised learning strategies to compensate for limited labeled data. However, real-time processing is limited due to the significant amount of computation required (Maccarone et al. 2023).

The findings of this review call for the creation of publicly accessible large-scale stereo datasets for underwater fish. Taking reference from well-established 3D datasets for human behavior, future establishment of large-scale fish datasets should consider different aquaculture scenarios, diverse fish species coverage, and comprehensive annotations, including water quality parameters, behavioral annotations, pose descriptions, semantic labeling and

more. In specific construction, a multi-source annotated dataset for fish farming processes can be built from four modules: farming scenarios, farming environments, farming subjects, and farming equipment. Farming scenarios should include land-based factory farming, offshore aquaculture, ponds, and net cages. Farming environments should encompass meteorological and water quality data relevant to fish growth. Within farming subjects, besides basic fish information and aquaculture knowledge, core content crucial for future integration and construction should include original fish image and video data acquired from various cameras and sensors. Establishing large-scale multi-source annotated databases is crucial for advancing intelligent aquaculture, and facilitating research on multi-source data mapping models, which will support precise control and decision-making in future intelligent farming practices.

Meanwhile, a specific platform for sharing underwater fish stereo data can be established. It is necessary to develop uniform data quality standards and labeling guidelines. Additionally, further attention should be given to the development of self-labeling techniques for fish stereo image data.

6.2 Developing and applying innovative cutting-edge stereo vision technologies

Further research and applications for end-to-end more accurate estimation from underwater multi-view images that strike a balance between accuracy and efficiency - Underwater images often lack matching and texture features due to the complex underwater scene, light refraction, attenuation, and increased turbidity. This can cause difficulties and errors in stereo matching. CNN-based end-to-end depth estimation models have become a research focus. These models can generate dense depth maps directly from stereo image pairs and have made progress in reducing memory usage (Hamid et al. 2022). Considering the expensive cost of acquiring ground truth for underwater depth maps, future research should focus on the application of unsupervised techniques in depth estimation networks. Future research should focus on the challenges of occlusion, duplication, and weak texture of underwater fish images while considering the need for low-cost and high-accuracy in production applications.

Advanced research on utilizing simultaneous multi-target 3D localization technique based on the comprehensive information acquired from stereo vision - The applications of current stereo vision models based on deep learning technology in aquaculture mainly focus on obtaining the depth information for individual fish respectively. The 3D coordinates of key points on the fish body can be calculated, providing real-time spatial distribution information and other related downstream tasks like 3D tracking. Compared to 2D computer vision, stereo vision technology has not been fully utilized in such applications. In the future, it is recommended to conduct in-depth study of stereo vision-based multi-target synchronous 3D localization technology for underwater fish. This will provide valuable information, such as aggregation index and vitality of fish groups, which can be used to establish a more accurate mapping model of fish feeding status and behavioral characteristics.

Establishing a multi-task learning framework based on stereo vision for fish phenotype extraction and behavioral analysis - Most current research on 3D object detection, key-point detection, and the tracking of fish adopts a sequential processing mode, employing separate models for each task. This approach not only wastes computational resources

but also overlooks the correlations between depth features extracted from different tasks. Multi-task learning frameworks can enhance the performance and reduce inference time of adjacent tasks by sharing information. These frameworks have made significant progress in domains such as autonomous driving. Chen et al. (2022) proposed a multi-task learning framework called FishNet for fish visual recognition based on 2D images. It integrates fish detection, instance segmentation, and pose estimation into an end-to-end network, achieving a balance between accuracy and speed. Stereo vision technology can provide more comprehensive and accurate information on fish phenotypes and behaviors in aquaculture. Therefore, further research is needed to establish a multi-task learning network based on stereo vision for 3D detection, segmentation, and tracking of fish.

Advancing the high-quality 3D reconstruction and 3D panoramic stitching methods for dynamic fish targets based on multi-view depth feature fusion - The non-rigid deformation of the free-swimming fish body, irregular fast movement, self-obscuration and mutual occlusion can cause appearance changes, making it challenging to rely solely on traditional image feature matching for 3D reconstruction of the fish body. Current research hotspots involve extracting deep features of multi-view images for 3D reconstruction based on neural networks. One of the hotspots in this field is the use of radiance field-based 3D reconstruction methods that infer the position and intensity of light sources. Waternerf, which was proposed by Sethuraman et al. (2023), has demonstrated the ability of NeRF to render the underwater scenes. The recent development of D-NeRF has demonstrated the ability of NeRF to handle non-rigid object motion in reconstructing dynamic scenes (Pumarola et al. 2021). The latest technique of 3D Gaussian splatting proposed a more efficient reconstruction approach, where the scene point cloud obtained through Structure from Motion (SfM) is modeled as a 3D Gaussian image and rendered in real-time through splatting to generate the 3D model (Kerbl et al. 2023). Another possible solution is the Skinned Multi-Person Linear (SMPL) model which is popular in human 3D pose estimation. Wu et al. (2022a) created a 3D template mesh model by pre-scanning a real fish body. Accurate 4D (3D space+time) shape of swimming fish can be reconstructed by optimizing the fish body model for each frame. The mesh vertices can be tuned using Linear Blend Skinning (LBS), which is potentially advantageous when dealing with fish bodies with non-rigid deformations. In addition, the trajectories of fish in large-scale aquaculture settings may extend beyond the field of view of a single stereo vision system. Further research and application should be conducted on finding correspondence relationships among images through alignment (registration) algorithms and integrating fish trajectories from multiple viewpoints seamlessly using 3D panoramic stitching techniques. It is important to focus on solving potential problems such as scene movement, blurring, or ghosting caused by different image exposures.

Developing end-to-end multimodal data processing models for richer information understanding - In practical aquaculture applications, processing data from multiple modalities is a developing trend. Decisions in actual aquaculture management rely not only on fish phenotypic characteristics and movement trajectories, but also on information such as water quality parameters, temperature and feeding strategies, etc. In addition, the integration of stereo vision technology with acoustic techniques can be employed to overcome the challenges posed by turbid and dim underwater environments. In the future, when constructing the model, it is important to consider the input, fusion, and alignment of multimodal data to achieve a more comprehensive understanding of information and integrated processing.

6.3 Integrating management platform with edging computing technologies and embedded smart device

Applying a management model for integrating platforms with edge computing for costs reduction and increase efficiency - Deep learning-based stereo data processing models for aquaculture applications rely heavily on computational and memory resources, which can be challenging, time-consuming, and costly. Therefore, an integrated management center and edge computing collaborative management model is the future development direction. The integration platform's uniformly configured algorithmic model can maximize the use of computing power, reducing the cost of terminal equipment. Additionally, the deployment of edge computing allows for real-time data storage and response, improving efficiency and speed. Future attention should be given to the use of edge acceleration technologies, which increase available memory and enable more efficient workflows.

Developing underwater intelligent unmanned equipment combined with stereo vision technology to improve dynamic monitoring capabilities - Remote-controlled underwater unmanned intelligent devices, such as autonomous underwater vehicles (AUVs) and underwater remotely operated vehicles (ROVs), have been be used for water quality testing, fish behavior monitoring, and automatic feeding (Wu et al. 2022b). By combining with the advantages of stereo vision in robot navigation and path planning, the development of underwater intelligent unmanned equipment that integrates data acquisition, processing, and remote feeding functions can help solve the challenges of dynamic tracking and accurate feeding of fish in large-scale aquaculture facilities, such as intelligent marine ranches and pond-type intelligent fish farms.

Integrating of stereo vision models in cutting-edge applications in the aquaculture industry, **such as aquaponics and digital twin factories**.

Currently, in the intelligent aquaculture industry, typical models include factory-scale recirculating aquaculture systems (RAS) and innovative aquaponics smart farming modes. Factory-scale RAS modes utilize automation control, intensive farming practices, and water recirculation to enhance efficiency and yield while reducing environmental impact. Aquaponics smart factory farming combines fish farming with hydroponic cultivation, maximizing water resource reuse and land productivity, thereby promoting environmental sustainability. In practical applications, the implementation of smart factories heavily relies on comprehensive and precise data perception. Integrating stereo vision technology into different production stages of smart factories, such as aquaculture and environmental monitoring, enhances the perception accuracy of physical entities within aquaculture facilities, leading to greater economic benefits. Furthermore, digital twin technology facilitates intelligent decision-making and precise control of aquaculture systems through perception, recognition, simulation, and optimization. It synergizes effectively with stereo vision technology to achieve efficient, energy-saving, and environmentally friendly production goals.

7 Conclusions

This paper presents a comprehensive and critical review of the use of stereo vision technology in acquiring fish phenotype and analyzing their behavior in aquaculture. Firstly, the stereo data acquisition of underwater fish is analyzed, including the preparation and calibration

of hardware equipment. This is followed by an introduction of the commonly used image enhancement and augmentation methods, as well as the current status of underwater fish stereo datasets. Additionally, this paper focuses on the processing of underwater fish stereo data using different deep learning techniques, including detection, segmentation, tracking and 3D reconstruction. In particular, various strategies proposed by current research to address the specific challenges of aquaculture applications are discussed in detail. Finally, the paper discusses the applications of stereo vision techniques for fish size measurement, biomass estimation, and behavior analysis. The review demonstrates that stereo vision has significant potential in providing more accurate information than monocular vision. However, much of the current research is still based on 2D image-based data processing methods, leaving significant room for future development. The primary challenges identified by the review include insufficient datasets, non-rigid fish body deformation and frequent occlusion, data collection limitations due to underwater environments, and computational burden concerns. In the future, it is recommended to focus on the application of cutting-edge deep learning techniques to develop a multimodal data processing model as well as 3D reconstruction model that can balance the speed and accuracy. This is the research direction for the application of stereo vision in aquaculture for fish body phenotype acquisition and behavior analysis.

Acknowledgements This work was supported in part by the 'National Key Research and Development Program of China' (Grant Numbers: 2023YFD2400401, 2023YFD2400400), the National Natural Science Foundation of China 'Intelligent Identification Method of Underwater Fish Morphological Characteristics Based on Binocular Vision' (Grant Number: 62206021), and the National Natural Science Foundation of China 'Analysis and feature recognition on feeding behavior of fish school in facility farming based on machine vision' (Grant Number: 62076244). The authors thank Professor Yanqing Duan (University of Bedfordshire) for English editing and review.

Author contributions The article's concept and framework were designed by Yaxuan Zhao and Huihui Yu, who also draft the manuscript. Yingyi Chen determined the core content of the article. Ling Xu and Hanxiang Qin conducted the literature search and collection and assisted in revising the article. The manuscript was revised and supervised by Huihui Yu and Yingyi Chen, who also obtained funding.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no known conflict of interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aguiar J, Pinto AM, Cruz NA, Matos AC (2016) The Impact of Convergence Cameras in a stereoscopic system for AUVs. In: Campilho A, Karray F (eds) Image analysis and recognition. Springer International Publishing, Cham, pp 521–529. https://doi.org/10.1007/978-3-319-41501-7_58

Ahmed MS, Aurpa TT, Azad MAK (2022) Fish Disease Detection using image based machine learning technique in aquaculture. J King Saud Univ - Comput Inform Sci 34:5170–5182. https://doi.org/10.1016/j.jksuci.2021.05.003

Akkaynak D, Treibitz T (2019) Sea-Thru: A Method for Removing Water From Underwater Images. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Long Beach, CA, USA, 1682–1691. https://doi.org/10.1109/CVPR.2019.00178

Almansa C, Reig L, Oca J (2015) The laser scanner is a reliable method to estimate the biomass of a Senegalese sole (Solea senegalensis) population in a tank. Aquacult Eng 69:78–83. https://doi.org/10.1016/j.aquaeng.2015.10.003

AlZu'bi H, Al-Nuaimy W, Buckley J, Sneddon L, Iain Y (2015) Real-time 3D fish tracking and behaviour analysis. In: 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT). IEEE, Amman, Jordan, 1–5. https://doi.org/10.1109/AEECT.2015.7360567

An D, Huang J, Wei Y (2021) A survey of fish behaviour quantification indexes and methods in aquaculture. Reviews Aquaculture 13:2169–2189. https://doi.org/10.1111/raq.12564

Atoum Y, Srivastava S, Liu X (2014) Automatic feeding control for dense aquaculture fish tanks. IEEE Signal Proc Let 22:1089–1093. https://doi.org/10.1109/LSP.2014.2385794

Balaban MO, Chombeau M, Cırban D, Gümüş B (2010a) Prediction of the weight of alaskan Pollock using image analysis. J Food Sci 75:E552–E556. https://doi.org/10.1111/j.1750-3841.2010.01813.x

Balaban MO, Ünal Şengör GF, Soriano MG, Ruiz EG (2010b) Using image analysis to predict the weight of alaskan Salmon of different species. J Food Sci 75:E157–E162. https://doi.org/10.1111/j.1750-3841.2010.01522.x

Bao YJ, Ji CY, Zhang B, Gu JL (2018) Representation of freshwater aquaculture fish behavior in low dissolved oxygen condition based on 3D computer vision. Mod Phys Lett B 32:1840090. https://doi.org/10.1142/S0217984918400900

Beddow TA, Ross LG, Marchant JA (1996) Predicting salmon biomass remotely using a digital stereo-imaging technique. Aquaculture 146:189–203. https://doi.org/10.1016/S0044-8486(96)01384-1

Ben Tamou A, Benzinou A, Nasreddine K (2022) Targeted data Augmentation and hierarchical classification with Deep Learning for Fish species Identification in underwater images. J Imaging 8:214. https://doi.org/10.3390/jimaging8080214

Boldt JL, Williams K, Rooper CN, Towler RH, Gauthier S (2018) Development of stereo camera methodologies to improve pelagic fish biomass estimates and inform ecosystem management in marine waters. Fish Res 198:66–77. https://doi.org/10.1016/j.fishres.2017.10.013

Boulch A, Guerry J, Le Saux B, Audebert N (2018) SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. Computers Graphics 71:189–198. https://doi.org/10.1016/j.cag.2017.11.010

Boutros N, Shortis MR, Harvey ES (2015) A comparison of calibration methods and system configurations of underwater stereo-video systems for applications in marine ecology: stereo-video calibration and configuration. Limnol Oceanogr Methods 13:224–236. https://doi.org/10.1002/lom3.10020

Brown MZ, Burschka D, Hager GD (2003) Advances in computational stereo. IEEE Trans Pattern Anal Mach Intell 25:993–1008. https://doi.org/10.1109/TPAMI.2003.1217603

Butail S, Paley DA (2012) Three-dimensional reconstruction of the fast-start swimming kinematics of densely schooling fish. J R Soc Interface 9:77–88. https://doi.org/10.1098/rsif.2011.0113

Cai L, He L, Xu Y, Zhao Y, Yang X (2010) Multi-object detection and tracking by stereo vision. Pattern Recogn 43:4028–4041. https://doi.org/10.1016/j.patcog.2010.06.012

Chang J-R, Chen Y-S (2018) Pyramid stereo matching network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 5410–5418. https://doi.org/10.48550/arXiv.1803.08669

Chen Z, Cao L, Wang Q, Cai Y (2022) FishNet: Fish visual recognition with one stage multi-task learning. IET Image Proc 16:3237–3246. https://doi.org/10.1049/ipr2.12556

Cheng XE, Du SS, Li HY, Hu JF, Chen ML (2018) Obtaining three-dimensional trajectory of multiple fish in water tank via video tracking. Multimed Tools Appl 77:24499–24519. https://doi.org/10.1007/s11042-018-5755-5

Cheng S, Zhao K, Zhang D (2019) Abnormal Water Quality Monitoring based on visual sensing of three-Dimensional Motion Behavior of Fish. Symmetry 11:1179. https://doi.org/10.3390/sym11091179

Cheng R, Zhang C, Xu Q, Liu G, Song Y, Yuan X, Sun J (2020) Underwater fish body length estimation based on binocular image Processing. Information 11:476. https://doi.org/10.3390/info11100476

Chuang M-C, Williams K, Towler R (2015) Tracking live Fish from Low-contrast and low-frame-rate stereo videos. IEEE T Circ Syst Vid 25:167–179. https://doi.org/10.1109/TCSVT.2014.2357093
Costa C, Loy A, Cataudella S, Davis D, Scardi M (2006) Extracting fish size using dual underwater cameras. Aquacult Eng 35:218–227. https://doi.org/10.1016/j.aquaeng.2006.02.003
Costa C, Antonucci F, Boglione C, Menesatti P, Vandeputte M, Chatain B (2013) Automated sorting for size, sex and skeletal anomalies of cultured seabass using external shape analysis. Aquacult Eng 52:58–64. https://doi.org/10.1016/j.aquaeng.2012.09.001
Cren EDL (1951) The length-weight relationship and Seasonal cycle in gonad weight and Condition in the Perch (Perca fluviatilis). J Anim Ecol 20:201. https://doi.org/10.2307/1540
Da Silva Vale RT, Ueda EK, Takimoto RY, Castro Martins TD (2020) Fish volume monitoring using Stereo Vision for Fish farms. IFAC-PapersOnLine 53:15824–15828. https://doi.org/10.1016/j.ifacol.2020.12.232
De Verdal H, Vandeputte M, Pepey E, Vidal M-O, Chatain B (2014) Individual growth monitoring of European sea bass larvae by image analysis and microsatellite genotyping. Aquaculture 434:470–475. https://doi.org/10.1016/j.aquaculture.2014.09.012
Deng Y, Tan H, Tong M, Zhou D, Li Y, Zhu M (2022) An Automatic Recognition Method for Fish Species and length using an underwater Stereo Vision System. Fishes-Basel 7:326. https://doi.org/10.3390/fishes7060326
Deng Y, Tan H, Zhou D, Li Y, Zhu M (2023) An automatic body length estimating method for Micropterus salmoides using local water surface stereo vision. Biosyst Eng 235:166–179. https://doi.org/10.1016/j.biosystemseng.2023.09.013
Dubrovinskaya E, Dalgleish F, Ouyang B, Casari P, Kobe (2018) 1–8. https://doi.org/10.1109/OCEANSKOBE.2018.8559113
Dunbrack RL (2006) In situ measurement of fish body length using perspective-based remote stereo-video. Fish Res 82:327–331. https://doi.org/10.1016/j.fishres.2006.08.017
FAO (2024) The state of World fisheries and Aquaculture 2024 – Blue Transformation in action. Rome. https://doi.org/10.4060/cd0683en
Feng S, Yang X, Liu Y, Zhao Z, Liu J, Yan Y, Zhou C (2022) Fish feeding intensity quantification using machine vision and a lightweight 3D ResNet-GloRe network. Aquacult Eng 98:102244. https://doi.org/10.1016/j.aquaeng.2022.102244
Fernandes AFA, Turra EM, De Alvarenga ÉR, Passafaro TL, Lopes FB, Alves GFO, Singh V, Rosa GJM (2020) Deep Learning image segmentation for extraction of fish body measurements and prediction of body weight and carcass traits in Nile tilapia. Comput Electron Agr 170:105274. https://doi.org/10.1016/j.compag.2020.105274
Fu X, Cao X (2020) Underwater image enhancement with global–local networks and compressed-histogram equalization. Sig Process Image Commun 86:115892. https://doi.org/10.1016/j.image.2020.115892
Fusiello A, Trucco E, Verri A (2000) A compact algorithm for rectification of stereo pairs. Mach Vis Appl 12:16–22. https://doi.org/10.1007/s001380050120
Garcia R, Prados R, Quintana J, Tempelaar A, Gracias N, Rosen S, Vågstøl H, Løvall K (2020) Automatic segmentation of fish using deep learning with application to fish size measurement. Ices J Mar Sci 77:1354–1366. https://doi.org/10.1093/icesjms/fsz186
Gümüş B, Balaban MO (2010) Prediction of the weight of Aquacultured Rainbow Trout *(Oncorhynchus mykiss)* by image analysis. J Aquat Food Prod Technol 19:227–237. https://doi.org/10.1080/10498850.2010.508869
Hamid MS, Abd Manap N, Hamzah RA, Kadmin AF (2022) Stereo matching algorithm based on deep learning: a survey. J King Saud University-Computer Inform Sci 34:1663–1673. https://doi.org/10.1016/j.jksuci.2020.08.011
Hao Y, Guo S, Zhou X, Yin H (2024) Underwater swimming fish mass estimation based on binocular vision. Aquacult Int 32:7973–7995. https://doi.org/10.1007/s10499-024-01550-z
Hao M, Yu H, Li D (2016) The measurement of fish size by Machine Vision - A review. In: Li D, Li Z (eds) Computer and Computing technologies in Agriculture IX. Springer International Publishing, Cham, pp 15–32. https://doi.org/10.1007/978-3-319-48354-2_2
Harvey E, Shortis M (1995) A system for stereo-video measurement of sub-tidal organisms. Mar Technol Soc J 29:10–22
Harvey E, Cappo M, Shortis M, Robson S, Buchanan J, Speare P (2003) The accuracy and precision of underwater measurements of length and maximum body depth of southern bluefin tuna (Thunnus maccoyii) with a stereo–video camera system. Fish Res 63:315–326. https://doi.org/10.1016/S0165-7836(03)00080-8
Hile R (1940) AGE AND GROWTH OF THE CISCO, LEUCICHTHYS ARTEDI (LE SUEUR), IN THE LAKES OF THE NORTHEASTERN. Bull Bureau Fisheries 48:211

Hsieh Y-Z, Lee P-Y (2023) Analysis of Oplegnathus Punctatus Body parameters using underwater Stereo Vision. IEEE Trans Emerg Top Comput Intell 1–13. https://doi.org/10.1109/TETCI.2023.3290022

Huang H, Zhou H, Yang X, Zhang L, Qi L, Zang A-Y (2019a) Faster R-CNN for marine organisms detection and recognition using data augmentation. Neurocomputing 337:372–384. https://doi.org/10.1016/j.neucom.2019.01.084

Huang T-W, Hwang J-N, Romain S, Wallace F (2019b) Fish Tracking and Segmentation from stereo videos on the Wild Sea Surface for Electronic monitoring of rail fishing. IEEE T Circ Syst Vid 29:3146–3158. https://doi.org/10.1109/TCSVT.2018.2872575

Huang K, Li Y, Suo F, Xiang J (2020) Stereo Vison and Mask-RCNN Segmentation Based 3D Points Cloud Matching for Fish Dimension Measurement. In: 2020 39th Chinese Control Conference (CCC). IEEE, Shenyang, China, 6345–6350. https://doi.org/10.23919/CCC50068.2020.9188604

Hufschmied P, Fankhauser T, Pugovkin D (2011) Automatic stress-free sorting of sturgeons inside culture tanks using image processing: automatic stress-free sorting of sturgeons. J Appl Ichthyol 27:622–626. https://doi.org/10.1111/j.1439-0426.2011.01704.x

Israeli D, Kimmel E (1996) Monitoring the behavior of hypoxia-stressed Carassius auratus using computer vision. Aquacult Eng 15:423–440. https://doi.org/10.1016/S0144-8609(96)01009-6

Jin L, Liang H (2017) Deep learning for underwater image recognition in small sample size situations. In: OCEANS 2017 - Aberdeen. IEEE, Aberdeen, United Kingdom, 1–4. https://doi.org/10.1109/OCEANSE.2017.8084645

Ju R, Xu X, Yang Y, Wu G (2013) Stereo GrabCut: interactive and consistent object extraction for stereo images. In: Huet B, Ngo C-W, Tang J, Zhou Z-H, Hauptmann AG, Yan S (eds) Advances in Multimedia Information Processing – PCM 2013. Springer International Publishing, Cham, pp 418–429. https://doi.org/10.1007/978-3-319-03731-8_39

Kendall A, Martirosyan H, Dasgupta S, Henry P, Kennedy R, Bachrach A, Bry A (2017) End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE international conference on computer vision. 66–75. https://doi.org/10.48550/arXiv.1703.04309

Kerbl B, Kopanas G, Leimkühler T, Drettakis G (2023) 3d gaussian splatting for real-time radiance field rendering. ACM Trans Graphics 42:1–14

Koh ME, Fong MWK, Ng EYK (2023) Aqua3DNet: real-time 3D pose estimation of livestock in aquaculture by monocular machine vision. Aquacult Eng 103:102367. https://doi.org/10.1016/j.aquaeng.2023.102367

Komeyama K, Tanaka T, Yamaguchi T, Asaumi S, Torisawa S, Takagi T (2018) Body Measurement of Reared Red Sea Bream using Stereo Vision. J Robot Mechatron 30:231–237. https://doi.org/10.20965/jrm.2018.p0231

Kumari D, Kaur K (2016) A survey on stereo matching techniques for 3D vision in image processing. Int J Eng Mater Manuf 4:40–49. https://doi.org/10.5815/ijem.2016.04.05

Li J (2013) Application of image enhancement method for digital images based on Retinex theory. Optik 124:5986–5988. https://doi.org/10.1016/j.ijleo.2013.04.115

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

Li P, Chen X, Shen S (2019) Stereo r-cnn based 3d object detection for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7644–7652. https://doi.org/10.48550/arXiv.1902.09738

Li C, Anwar S, Porikli F (2020a) Underwater scene prior inspired deep underwater image and video enhancement. Pattern Recogn 98:107038. https://doi.org/10.1016/j.patcog.2019.107038

Li D, Hao Y, Duan Y (2020b) Nonintrusive methods for biomass estimation in aquaculture with emphasis on fish: a review. Reviews Aquaculture 12:1390–1411. https://doi.org/10.1111/raq.12388

Li D, Li X, Wang Q, Hao Y (2022) Advanced techniques for the Intelligent diagnosis of Fish diseases: a review. Animals 12:2938. https://doi.org/10.3390/ani12212938

Liang Y-T, Chiou Y-C (2009) Machine vision-based automatic raw fish handling and weighing system of Taiwan Tilapia. In: Chien B-C, Hong T-P, Chen S-M, Ali M (eds) Next-generation Applied Intelligence. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp 711–720. https://doi.org/10.1007/978-3-642-02568-6_72

Liu Z, Li X, Fan L, Lu H, Liu L, Liu Y (2014) Measuring feeding activity of fish in RAS using computer vision. Aquacult Eng 60:20–27. https://doi.org/10.1016/j.aquaeng.2014.03.005

Liu X, Yue Y, Shi M, Qian Z-M (2019) 3-D Video Tracking of multiple fish in a Water Tank. IEEE Access 7:145049–145059. https://doi.org/10.1109/ACCESS.2019.2945606

Liu C, Tao L, Kim Y-T (2020) VLW-Net: a very light-weight convolutional neural network (CNN) for single image dehazing. In: Advanced Concepts for Intelligent Vision Systems: 20th International Conference, ACIVS 2020, Auckland, New Zealand, February 10–14, 2020, Proceedings 20. Springer, 433–442. https://doi.org/10.1007/978-3-030-40605-9_37

Liu H, Suo F, Li Y, Xiang J (2022) Research on A Binocular Fish Dimension Measurement Method Based on Instance Segmentation and Fish Tracking. In: 2022 34th Chinese Control and Decision Conference (CCDC). IEEE, Hefei, China, 2791–2796. https://doi.org/10.1109/CCDC55256.2022.10034386

Liu H, Ma X, Yu Y, Wang L, Hao L (2023) Application of Deep Learning-based object detection techniques in Fish Aquaculture: a review. J Mar Sci Eng 11:867. https://doi.org/10.3390/jmse11040867

Maccarone A, Drummond K, McCarthy A, Steinlehner UK, Tachella J, Garcia DA, Pawlikowska A, Lamb RA, Henderson RK, McLaughlin S, Altmann Y, Buller GS (2023) Submerged single-photon LiDAR imaging sensor used for real-time 3D scene reconstruction in scattering underwater environments. Opt Express 31:16690. https://doi.org/10.1364/OE.487129

Måløy H, Aamodt A, Misimi E (2019) A spatio-temporal recurrent network for salmon feeding action recognition from underwater videos in aquaculture. Comput Electron Agr 167:105087. https://doi.org/10.1016/j.compag.2019.105087

Mao J, Xiao G, Sheng W, Qu Z, Liu Y (2016) Research on realizing the 3D occlusion tracking location method of fish's school target. Neurocomputing 214:61–79. https://doi.org/10.1016/j.neucom.2016.05.067

Marrable D, Tippaya S, Barker K, Harvey E, Bierwagen SL, Wyatt M, Bainbridge S, Stowar M (2023) Generalised deep learning model for semi-automated length measurement of fish in stereo-BRUVS. Front Mar Sci 10:1171625. https://doi.org/10.3389/fmars.2023.1171625

Martinez-de Dios JR, Serna C, Ollero A (2003) Computer vision and robotics techniques in fish farms. Robotica 21:233–243. https://doi.org/10.1017/S0263574702004733

Mathiassen JR, Misimi E, Toldnes B, Bondø M, Østvik SO (2011) High-speed weight estimation of whole Herring (*Clupea harengus*) using 3D machine vision. J Food Sci 76. https://doi.org/10.1111/j.1750-3841.2011.02226.x

Mayer N, Ilg E, Hausser P, Fischer P, Cremers D, Dosovitskiy A, Brox T (2016) A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 4040–4048. https://doi.org/10.48550/arXiv.1512.02134

Mei J, Hwang J-N, Romain S, Rose C, Moore B, 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). IEEE, 2175–2179. https://doi.org/10.1109/ICASSP39728.2021.9414803

Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R (2021) Nerf: representing scenes as neural radiance fields for view synthesis. Commun ACM 65:99–106. https://doi.org/10.1145/3503250

Mujtaba DF, Mahapatra NR Fish species classification with data augmentation. In: 2021 International Conference on Computational Science and, Intelligence C (2021) (CSCI). IEEE, 1588–1593. https://doi.org/10.1109/CSCI54926.2021.00307

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

Odone F, Trucco E, Verri A (2001) A trainable system for grading fish from images. Appl Artif Intell 15:735–745. https://doi.org/10.1080/088395101317018573

Pache MCB, Sant'Ana DA, Rezende FPC, De Andrade Porto JV, Rozales JVA, De Moraes Weber VA, Da Silva Oliveira Junior A, Garcia V, Naka MH, Pistori H (2022) Non-intrusively estimating the live body biomass of Pintado Real® fingerlings: a feature selection approach. Ecol Inf 68:101509. https://doi.org/10.1016/j.ecoinf.2021.101509

Palconit MGB, Ii RSC, Alejandrino JD, Pareja ME, Almero VJD, Bandala AA, Vicerra RRP, Sybingco E, Dadios EP, Naguib RNG (2021) Three-Dimensional Stereo Vision Tracking of multiple free-swimming fish for low Frame Rate Video. J Adv Comput Intell Intell Inf 25:639–646. https://doi.org/10.20965/jaciii.2021.p0639

Pautsina A, Císař P, Štys D, Terjesen BF, Espmark ÅMO (2015) Infrared reflection system for indoor 3D tracking of fish. Aquacult Eng 69:7–17. https://doi.org/10.1016/j.aquaeng.2015.09.002

Pedersen M, Haurum JB, Bengtson, SH, Moeslund TB (2020) 3d-zef: a 3d zebrafish tracking benchmark dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 2426–2436. https://doi.org/10.1109/CVPR42600.2020.00250

Perez J, Attanasio AC, Nechyporenko N, Sanz PJ (2017) A deep learning approach for underwater image enhancement. In: Biomedical Applications Based on Natural and Artificial Computing: International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2017, Corunna, Spain, June 19–23, 2017, Proceedings, Part II. Springer, 183–192. https://doi.org/10.1007/978-3-319-59773-7_19

Perez D, Ferrero FJ, Alvarez I, Valledor M, Campo JC (2018) Automatic measurement of fish size using stereo vision. In: 2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC). IEEE, Houston, TX, USA, 1–6. https://doi.org/10.1109/I2MTC.2018.8409687

Prabhakar CJ, Praveen Kumar PU (2010) Underwater image denoising using adaptive wavelet subband thresholding. In: 2010 International Conference on Signal and Image Processing. IEEE, Chennai, India, 322–327. https://doi.org/10.1109/ICSIP.2010.5697491

Priyadharsini R, Sree Sharmila T, Rajendran V (2018) A wavelet transform based contrast enhancement method for underwater acoustic images. Multidim Syst Sign Process 29:1845–1859. https://doi.org/10.1007/s11045-017-0533-5

Pumarola A, Corona E, Pons-Moll G, Moreno-Noguer F (2021) D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10318–10327. https://doi.org/10.48550/arXiv.2011.13961

Qi CR, Su H, Mo K, Guibas LJ (2017a) Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 652–660. https://doi.org/10.48550/arXiv.1612.00593

Qi X, Liao R, Jia J, Fidler S, Urtasun R (2017b) 3d graph neural networks for rgbd semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. 5199–5208. https://doi.org/10.1109/ICCV.2017.556

Qian ZM, Chen YQ (2017) Feature point based 3D tracking of multiple fish from multi-view images. PLoS ONE 12:e0180254. https://doi.org/10.1371/journal.pone.0180254

Qian Z, Shi M, Wang M, Cun T (2017) Skeleton-based 3D Tracking of multiple fish from two orthogonal views. In: Yang J, Hu Q, Cheng M-M, Wang L, Liu Q, Bai X, Meng D (eds) Computer vision. Springer Singapore, Singapore, pp 25–36. https://doi.org/10.1007/978-981-10-7299-4_3

Ravanbakhsh M, Shortis MR, Shafait F, Mian A, Harvey ES, Seager JW (2015) Automated fish detection in underwater images using shape-based level sets. Photogram Rec 30:46–62. https://doi.org/10.1111/phor.12091

Reza AM (2004) Realization of the contrast Limited Adaptive Histogram Equalization (CLAHE) for real-time image enhancement. The Journal of VLSI Signal Processing-Systems for Signal. Image Video Technol 38:35–44. https://doi.org/10.1023/B:VLSI.0000028532.53893.82

Risholm P, Mohammed A, Kirkhus T, Clausen S, Vasilyev L, Folkedal O, Johnsen Ø, Haugholt KH, Thielemann J (2022) Automatic length estimation of free-swimming fish using an underwater 3D range-gated camera. Aquacult Eng 97:102227. https://doi.org/10.1016/j.aquaeng.2022.102227

Rosen S, Jörgensen T, Hammersland-White D, Holst JC (2013) DeepVision: a stereo camera system provides highly accurate counts and lengths of fish passing inside a trawl. Can J Fish Aquat Sci 70:1456–1467. https://doi.org/10.1139/cjfas-2013-0124

Saad A, Jakobsen S, Bondø M, Mulelid M, Kelasidi E (2024) StereoYolo+DeepSORT: a framework to track fish from underwater stereo camera in situ. In: Sixteenth International Conference on Machine Vision (ICMV 2023). SPIE, 321–329. https://doi.org/10.1117/12.3023414

Saberioon MM, Cisar P (2016) Automated multiple fish tracking in three-dimension using a structured light sensor. Comput Electron Agr 121:215–221. https://doi.org/10.1016/j.compag.2015.12.014

Saberioon M, Císař P (2018) Automated within tank fish mass estimation using infrared reflection system. Comput Electron Agr 150:484–492. https://doi.org/10.1016/j.compag.2018.05.025

Salman A, Jalal A, Shafait F, Mian A, Shortis M, Seager J, Harvey E (2016) Fish species classification in unconstrained underwater environments based on deep learning: fish classification based on deep learning. Limnol Oceanogr Methods 14:570–585. https://doi.org/10.1002/lom3.10113

Serna C, Ollero A (2001) A Stereo Vision System for the Estimation of Biomass in Fish farms. IFAC Proc Volumes 34:185–191. https://doi.org/10.1016/S1474-6670(17)32814-8

Sethuraman AV, Ramanagopal MS, Skinner KA (2023) Waternerf: neural radiance fields for underwater scenes. In: OCEANS 2023-MTS/IEEE US Gulf Coast. IEEE, pp 1–7. https://doi.org/10.23919/OCEANS52994.2023.10336972

Shafait F, Harvey ES, Shortis MR, Mian A, Ravanbakhsh M, Seager JW, Culverhouse PF, Cline DE, Edgington DR (2017) Towards automating underwater measurement of fish length: a comparison of semi-automatic and manual stereo–video measurements. Ices J Mar Sci 74:1690–1701. https://doi.org/10.1093/icesjms/fsx007

Shen J, Xu W, Luo Y, Su P-C, Cheung SS (2014) Extrinsic calibration for wide-baseline RGB-D camera network. In: 2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP). IEEE, 1–6. https://doi.org/10.1109/MMSP.2014.6958798

Shi S, Wang X, Li H (2019) Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 770–779. https://doi.org/10.1109/CVPR.2019.00086

Shi C, Wang Q, He X, Zhang X, Li D (2020) An automatic method of fish length estimation using underwater stereo system based on LabVIEW. Comput Electron Agr 173:105419. https://doi.org/10.1016/j.compag.2020.105419

Shi C, Zhao R, Liu C, Li D (2022) Underwater fish mass estimation using pattern matching based on binocular system. Aquacult Eng 99:102285. https://doi.org/10.1016/j.aquaeng.2022.102285
Shorten C, Khoshgoftaar TM (2019) A survey on Image Data Augmentation for Deep Learning. J Big Data 6:60. https://doi.org/10.1186/s40537-019-0197-0
Shortis M (2019) Camera calibration techniques for Accurate Measurement Underwater. In: McCarthy JK, Benjamin J, Winton T, Van Duivenvoorde W (eds) 3D Recording and Interpretation for Maritime Archaeology. Springer International Publishing, Cham, pp 11–27. https://doi.org/10.1007/978-3-030-03635-5_2
Silva C, Aires R, Rodrigues F (2023) A compact underwater stereo vision system for measuring fish. Aquaculture Fisheries. https://doi.org/10.1016/j.aaf.2023.03.006. S2468550X23000539
Somerton DA, Williams K, Campbell MD (2017) Quantifying the behavior of fish in response to a moving camera vehicle by using benthic stereo cameras and target tracking. Fish B-Noaa 115:343–354. https://doi.org/10.7755/FB.115.3.5
Strachan NJC (1993) 2-Length measurement of fish by computer vision. Comput Electron Agr 8:93–104. https://doi.org/10.1016/0168-1699(93)90009-P
Sun J, Zheng N-N, Shum H-Y (2003) Stereo matching using belief propagation. IEEE Trans Pattern Anal Mach Intell 25:787–800. https://doi.org/10.1109/TPAMI.2003.1206509
Suo F, Huang K, Ling G, Li Y, Xiang J (2020) Fish Keypoints Detection for Ecology Monitoring Based on Underwater Visual Intelligence. In: 2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV). IEEE, Shenzhen, China, 542–547. https://doi.org/10.1109/ICARCV50220.2020.9305424
Tanaka T, Ikeda R, Yuta Y, Tsurukawa K, Nakamura S, Yamaguchi T, Komeyama K (2019) Annual monitoring of growth of red sea bream by multi-stereo-image measurement. Fisheries Sci 85:1037–1043. https://doi.org/10.1007/s12562-019-01347-7
Tillett R, McFarlane N, Lines J (2000) Estimating dimensions of Free-Swimming Fish using 3D point distribution models. Comput Vis Image Und 79:123–141. https://doi.org/10.1006/cviu.2000.0847
Tonachella N, Martini A, Martinoli M, Pulcini D, Romano A, Capoccioni F (2022) An affordable and easy-to-use tool for automatic fish length and weight estimation in mariculture. Sci Rep 12:15642. https://doi.org/10.1038/s41598-022-19932-9
Torisawa S, Kadota M, Komeyama K, Suzuki K, Takagi T (2011) A digital stereo-video camera system for three-dimensional monitoring of free-swimming Pacific bluefin tuna, *Thunnus orientalis*, cultured in a net cage. Aquat Living Resour 24:107–112. https://doi.org/10.1051/alr/2011133
Tran MT, Kim DH, Kim CK, Kim HK, Kim SB (2018) Determination of Injury Rate on Fish Surface Based on Fuzzy C-means Clustering Algorithm and L*a*b* Color Space Using ZED Stereo Camera. In: 2018 15th International Conference on Ubiquitous Robots (UR). IEEE, Honolulu, HI, USA, 466–471. https://doi.org/10.1109/URAI.2018.8441790
TW F (1904) The rate of growth of fishes. Twenty-second annual report 141–241
Ubina NA, Cheng SC, Chang CC, Cai SY, Lan HY, Lu HY (2022) Intelligent Underwater Stereo Camera Design for Fish Metric Estimation using Reliable object matching. IEEE Access 10:74605–74619. https://doi.org/10.1109/ACCESS.2022.3185753
Viazzi S, Van Hoestenberghe S, Goddeeris BM, Berckmans D (2015) Automatic mass estimation of Jade perch Scortum barcoo by computer vision. Aquacult Eng 64:42–48. https://doi.org/10.1016/j.aquaeng.2014.11.003
Voskakis D, Makris A, Papandroulakis N (2021) Deep learning based fish length estimation. An application for the Mediterranean aquaculture. OCEANS 2021: San Diego – Porto. IEEE, San Diego, CA, USA, pp 1–5. https://doi.org/10.23919/OCEANS44145.2021.9705813
Wang D, Lim KB (2011) Obtaining depth map from segment-based stereo matching using graph cuts. J Vis Commun Image Represent 22:325–331. https://doi.org/10.1016/j.jvcir.2011.02.001
Wang SH, Zhao J, Liu X, Qian Z-M, Liu Y, Chen YQ (2017) 3D tracking swimming fish school with learned kinematic model using LSTM network. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, New Orleans, LA, 1068–1072. https://doi.org/10.1109/ICASSP.2017.7952320
Wang H, Ji X, Zhao H, Yue Y (2020a) Semantic Segmentation of Freshwater Fish Body Based on Generative Adversarial Network. In: 2020 IEEE International Conference on Mechatronics and Automation (ICMA). IEEE, Beijing, China, 249–254. https://doi.org/10.1109/ICMA49215.2020.9233767
Wang Z, Zheng L, Liu Y, Li Y, Wang S (2020b) Towards real-time multi-object tracking. In: European conference on computer vision. Springer, 107–122. https://doi.org/10.1007/978-3-030-58621-8_7
Wang C, Li Z, Wang T, Xu X, Zhang X, Li D (2021) Intelligent fish farm—the future of aquaculture. Aquacult Int 29:2681–2711. https://doi.org/10.1007/s10499-021-00773-8
Wang G, Li X, Yu J, Xu W, Akhter M, Ji S, Hao Y, Li D (2024) Stereo matching and 3D reconstruction with NeRF supervision for accurate weight estimation in free-swimming fish. Comput Electron Agr 225:109255. https://doi.org/10.1016/j.compag.2024.109255

Wei G, Wei Z, Huang L, Nie J, Chang H (2018) Robust underwater fish classification based on Data Augmentation by adding noises in Random Local regions. In: Hong R, Cheng W-H, Yamasaki T, Wang M, Ngo C-W (eds) Advances in Multimedia Information Processing – PCM 2018. Springer International Publishing, Cham, pp 509–518. https://doi.org/10.1007/978-3-030-00767-6_47

Weng X, Wang J, Held D, Kitani K (2020) 3d multi-object tracking: A baseline and new evaluation metrics. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 10359–10366. https://doi.org/10.1109/IROS45743.2020.9341164

Williams K, Wilson CD, Horne JK (2013) Walleye pollock (Theragra chalcogramma) behavior in midwater trawls. Fish Res 143:109–118. https://doi.org/10.1016/j.fishres.2013.01.016

Williams K, Lauffenburger N, Chuang M-C, Hwang J-N, Towler R (2016) Automated measurements of fish within a trawl using stereo images from a camera-trawl device (CamTrawl). Methods Oceanogr 17:138–152. https://doi.org/10.1016/j.mio.2016.09.008

Wu H, He S, Deng Z et al (2019a) Fishery monitoring system with AUV based on YOLO and SGBM. In: 2019 Chinese Control Conference (CCC). IEEE,: 4726–4731. https://doi.org/10.23919/ChiCC.2019.8866087

Wu Z-Y, Tseng S-L, Lin H-Y, Chen H-Y, Luan TV (2019b) Incorporating Stereo with Convolutional Neural Networks for Real-Time Fish Detection and Classification. In: 2019 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM). IEEE, Bangkok, Thailand, 83–88. https://doi.org/10.1109/CIS-RAM47153.2019.9095805

Wu R, Deussen O, Li L (2022a) DeepShapeKit: accurate 4D shape reconstruction of swimming fish. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, Kyoto, Japan, 526–531. https://doi.org/10.1109/IROS47612.2022.9982097

Wu Y, Duan Y, Wei Y, An D, Liu J (2022b) Application of intelligent and unmanned equipment in aquaculture: a review. Comput Electron Agr 199:107201. https://doi.org/10.1016/j.compag.2022.107201

Wu Z, Zhou Z, Allibert G, Stolz C, Demonceaux C, Ma C (2022c) Transformer fusion for indoor rgb-d semantic segmentation. Available SSRN 4251286. https://doi.org/10.2139/ssrn.4251286

Xiao G (2015) Water quality monitoring using abnormal tail-beat frequency of crucian carp. Ecotoxicol Environ Saf 111:185–191. https://doi.org/10.1016/j.ecoenv.2014.09.028

Xiao G, Fan WK, Mao JF, Cheng ZB, Zhong D-H, Li Y (2016) Research of the Fish Tracking Method with Occlusion Based on Monocular Stereo Vision. In: 2016 International Conference on Information System and Artificial Intelligence (ISAI). IEEE, Hong Kong, China, 581–589. https://doi.org/10.1109/ISAI.2016.0129

Xu H, Zhang J (2020) Aanet: Adaptive aggregation network for efficient stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 1959–1968. https://doi.org/10.48550/arXiv.2004.09548

Xu W, Liu C, Wang G, Zhao Y, Yu J, Muhammad A, Li D (2024) Behavioral response of fish under ammonia nitrogen stress based on machine vision. Eng Appl Artif Intel 128:107442. https://doi.org/10.1016/j.engappai.2023.107442

Yang L, Liu Y, Yu H, Fang X, Song L, Li D, Chen Y (2021a) Computer Vision Models in Intelligent aquaculture with emphasis on Fish Detection and Behavior Analysis: a review. Arch Computat Methods Eng 28:2785–2816. https://doi.org/10.1007/s11831-020-09486-2

Yang X, Zhang S, Liu J, Gao Q, Dong S, Zhou C (2021b) Deep learning for smart fish farming: applications, opportunities and challenges. Reviews Aquaculture 13:66–90. https://doi.org/10.1111/raq.12464

Yang Y, Xu Y, Zhang C, Xu Z, Huang J (2022) Hierarchical Vision Transformer with Channel Attention for RGB-D Image Segmentation. In: Proceedings of the 4th International Symposium on Signal Processing Systems. ACM, Xi'an China, 68–73. https://doi.org/10.1145/3532342.3532352

Yao Y, Luo Z, Li S, Fang T, Quan L (2018) Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European conference on computer vision (ECCV). 767–783. https://doi.org/10.48550/arXiv.1804.02505

Ye X, Xu H, Ji X, Xu R (2018) Underwater image Enhancement using stacked generative adversarial networks. In: Hong R, Cheng W-H, Yamasaki T, Wang M, Ngo C-W (eds) Advances in Multimedia Information Processing – PCM 2018. Springer International Publishing, Cham, pp 514–524. https://doi.org/10.1007/978-3-030-00764-5_47

Ye Z, Zhao J, Han Z, Zhu S, Li J, Lu H, Ruan Y (2016) Behavioral characteristics and statistics-based imaging techniques in the assessment and optimization of tilapia feeding in a recirculating aquaculture system. Trans ASABE 59(1):345–355. https://doi.org/10.13031/trans.59.11406

Yu X, Wang Y, Liu J, Wang J, An D, Wei Y (2022) Non-contact weight estimation system for fish based on instance segmentation. Expert Syst Appl 210:118403. https://doi.org/10.1016/j.eswa.2022.118403

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

Zbontar J, LeCun Y (2015) Computing the stereo matching cost with a convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 1592–1599. https://doi.org/10.48550/arXiv.1409.4326
Zhang Z (2000) A flexible new technique for camera calibration. IEEE Trans Pattern Anal Mach Intell 22:1330–1334. https://doi.org/10.1109/34.888718
Zhang T, Yang Y, Liu Y, Liu C, Zhao R, Li D, Shi C (2024) Fully automatic system for fish biomass estimation based on deep neural network. Ecol Inform 79:102399. https://doi.org/10.1016/j.ecoinf.2023.102399
Zheng K, Yang R, Li R, Guo P, Yang L, Qin H (2023) A spatiotemporal attention network-based analysis of golden pompano school feeding behavior in an aquaculture vessel. Comput Electron Agr 205:107610. https://doi.org/10.1016/j.compag.2022.107610
Zhao S, Zhang S, Liu J, Wang H, Zhu J, Li D, Zhao R (2021) Application of machine learning in intelligent fish aquaculture: a review. Aquaculture 540:736724. https://doi.org/10.1016/j.aquaculture.2021.736724
Zhou C, Yang X, Zhang B, Lin K, Xu D, Guo Q, Sun C (2017a) An adaptive image enhancement method for a recirculating aquaculture system. Sci Rep 7:6243. https://doi.org/10.1038/s41598-017-06538-9
Zhou C, Zhang B, Lin K, Xu D, Chen C, Yang X, Sun C (2017b) Near-infrared imaging to quantify the feeding behavior of fish in aquaculture. Comput Electron Agr 135:233–241. https://doi.org/10.1016/j.compag.2017.02.013
Zhou C, Lin K, Xu D, Chen L, Guo Q, Sun C, Yang X (2018a) Near infrared computer vision and neuro-fuzzy model-based feeding decision system for fish in aquaculture. Comput Electron Agr 146:114–124. https://doi.org/10.1016/j.compag.2018.02.006
Zhou C, Xu D, Lin K, Sun C, Yang X (2018b) Intelligent feeding control methods in aquaculture with an emphasis on fish: a review. Reviews Aquaculture 10:975–993. https://doi.org/10.1111/raq.12218
Zhou K, Meng X, Cheng B (2020a) Review of stereo matching algorithms based on deep learning. Comput Intell Neurosci 2020(1):8562323. https://doi.org/10.1155/2020/8562323
Zhou X, Koltun V, Krähenbühl P (2020b) Tracking objects as points. In: European conference on computer vision. Springer, 474–490. https://doi.org/10.1007/978-3-030-58548-8_28
Zhou M, Shen P, Zhu H, Shen Y (2023) In-Water fish body-length measurement system based on Stereo Vision. Sensors 23:6325. https://doi.org/10.3390/s23146325
Zion B (2012) The use of computer vision technologies in aquaculture – a review. Comput Electron Agr 88:125–132. https://doi.org/10.1016/j.compag.2012.07.010

Authors and Affiliations

Yaxuan Zhao[1,2] · Hanxiang Qin[2,3,4,5] · Ling Xu[2,3,4,5] · Huihui Yu[1,2] · Yingyi Chen[2,3,4,5]

✉ Huihui Yu
yuhh1990@126.com

✉ Yingyi Chen
chenyingyi@cau.edu.cn

[1] School of Information Science & Technology, Beijing Forestry University, Beijing 100083, People's Republic of China

[2] National Innovation Center for Digital Fishery, Beijing 100083, People's Republic of China

[3] Key Laboratory of Smart Farming Technologies for Aquatic Animal and Livestock, Ministry of Agriculture and Rural Affairs, Beijing 100083, People's Republic of China

[4] Beijing Engineering and Technology Research Center for Internet of Things in Agriculture, Beijing 100083, People's Republic of China

[5] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, People's Republic of China