www.kosmospublishers.com
contact@kosmospublishers.com
DOI: 10.37722/JAMBE.2025302

Research Article | Journal of Aquaculture, Marine Biology & Ecology | JAMBE-173
ISSN 2692-1529

Benchmarking and Fine-Tuning Monocular Depth Models for Fish Biomass Estimation in Aquaculture

Alexandrou A*, Christodoulou F, Komninos D, Dr. Ozdeger T, Seferis K

Blue Analytics LTD, Data Analytics and Cloud Services Provider, Nicosia, Cyprus

Received Date: December 12, 2025; **Accepted Date:** January 13, 2026; **Published Date:** January 21, 2026.

***Corresponding author**: Alexandrou A, *Blue Analytics LTD, Data Analytics and Cloud Services Provider, Nicosia, Cyprus;* Email: aalexandrou@blueanalytix.com

Abstract: Accurate and non-invasive fish biomass estimation is essential for efficient feeding, environmental sustainability, and productivity in modern aquaculture. While monocular vision systems offer a low-cost and scalable alternative to stereo cameras, reliable metric depth estimation from a single RGB image remains a significant challenge, particularly in complex underwater environments. This study investigates the feasibility of using state-of-the-art monocular depth foundation models for fish biomass estimation in real aquaculture cages and evaluates the impact of domain-specific fine-tuning. Four advanced monocular depth models, Depth Pro, UniDepth v2, Depth Anything v2, and the recently released Depth Anything v3, were evaluated in a zero-shot setting using real underwater footage collected from commercial sea cages. The initial evaluation showed strong scale ambiguity and systematic overestimation across all models, highlighting the need for explicit depth calibration when applying monocular approaches in underwater environments. A finetuned version of Depth Anything v2 trained on stereo-aligned underwater data delivered markedly better results. In controlled experiments using a reference object of known length, the average length estimation error was reduced from more than 7 cm to 3.84 cm. Across all test videos, the error was further reduced from 3.76 cm to 2.42 cm, demonstrating the clear benefit of underwater-specific fine-tuning. These findings demonstrate that, although off-the-shelf monocular models are limited by scale ambiguity in underwater conditions, domain-adapted monocular depth estimation can achieve reliable performance for practical biomass estimation. This work highlights the strong potential of fine-tuned monocular systems, as scalable and cost-effective alternatives for real-time aquaculture monitoring and precision feeding applications.

Introduction

Accurate estimation of fish biomass is essential for modern aquaculture, underpinning key operational decisions such as stock assessment, feeding optimization, harvesting strategies, and welfare monitoring **[1].** Traditionally, biomass estimation relies on manual sampling, which is invasive, labor-intensive, and often unreliable due to stress-induced behavioral changes and limited sampling frequency. To address these limitations, computer-vision-based approaches have gained increasing attention, offering the potential for continuous, non-invasive, and scalable monitoring directly within aquaculture cages **[1, 3].**

Stereo vision systems have demonstrated strong performance for biomass estimation due to their ability to generate metric depth maps, facilitating accurate measurement of fish length (L) and body geometry **[2,3].** However, stereo cameras remain costly, complex to deploy and maintain, and require precise calibration that makes them difficult for widespread adoption, particularly across large farms or low budget operations. In contrast, monocular RGB cameras are cheap, easy to deploy, and already common across aquaculture infrastructures, yet they inherently lack metric depth information **[4, 6]**. Their depth predictions are relative rather than absolute, which limits their applicability for tasks requiring

geometrically accurate measurements such as L estimation and biomass computation **[4, 6].**

Recent advances in monocular depth foundation models **[7, 11]** have considerably improved the accuracy and generalization capabilities of monocular depth estimation. These models leverage large-scale pre-training, transformer architectures, and metric-aware depth supervision to provide high quality depth maps across diverse scenes. Although these models demonstrate state-of-the-art performance on standard benchmarks, their suitability for underwater aquaculture environments remains largely unexplored **[12, 13].** Underwater scenes present unique challenges, including turbidity, variations in lighting, color distortion, specular reflections, and dynamic fish movement. Additionally, unstructured underwater scenes include floating or mid-water objects that increase the prediction uncertainty **[14, 15].** As a result, the performance of monocular depth foundation models in these conditions cannot be assumed and requires systematic evaluation **[12, 13]**.

This study addresses this gap by providing the first comprehensive benchmark of off-the-shelf monocular depth estimation models for fish biomass prediction in aquaculture. These aquaculture environments feature free-swimming fish suspended in open water without any ground plane or stable geometric reference, making monocular metric depth estimation significantly more challenging than in structured terrestrial scenes. We evaluate four leading depth foundation models, UniDepthv2, Depth Pro, Depth Anything v2, and Depth Anything v3, across multiple aquaculture cages with varying environmental conditions **[7, 9, 11].** Their depth predictions are compared against stereo-derived ground truth, and their effectiveness for downstream tasks such as L estimation and biomass computation is rigorously assessed. Beyond off-the-shelf performance, we further investigate the impact of domain-specific fine-tuning and simple scaling, demonstrating how adapting a depth model (Depth Anything v2) to the underwater setting improves prediction accuracy and stability. The contributions of this paper are threefold:

- We present the first benchmark evaluating state-of-the-art monocular depth models in realistic aquaculture environments for 3D fish measurement tasks.
- We quantitatively compare model performance in terms of depth error, length error, and biomass prediction accuracy across multiple unstructured underwater scenes.
- We demonstrate that fine-tuning significantly enhances model robustness and precision, highlighting the potential for monocular systems as a low-cost alternative to stereo cameras.

By providing a rigorous evaluation framework and actionable insights into the strengths and limitations of modern monocular depth models, this work contributes toward scalable, cost-efficient, and automated biomass monitoring solutions for the aquaculture industry.

The remainder of the paper is organized as follows. Section I introduces the problem setting, motivation, and key challenges associated with monocular depth estimation in underwater aquaculture environments. Section II reviews prior work on fish biomass estimation and monocular and underwater depth estimation. Section III details the datasets, benchmarked models, experimental setup, and evaluation metrics used in this study. Section IV presents quantitative and qualitative results from the zero-shot and fine-tuned evaluations. Section V discusses the implications of these findings, practical considerations for deployment, and current limitations. Finally, Section VI summarizes the main contributions and outlines directions for future research.

Related Work

A. Fish Biomass Estimation Methods

A wide range of methods have been developed for fish biomass estimation in aquaculture, each with varying levels of accuracy, cost, and practicality **[1].** Traditional approaches rely on manual sampling, in which fish are physically captured, measured, and weighed. Although this procedure provides highly reliable ground truth, it is labor-intensive, time-consuming, and disruptive to normal fish behavior. To address these limitations, vision-based systems, particularly those using stereo cameras, have gained significant traction **[1].** Stereo vision provides direct access to 3D geometry through triangulation and has demonstrated strong accuracy in both experimental and operational environments. For example, Zhang et al. **[3]** integrated stereo disparity estimation with deep learning-based fish detection to achieve precise biomass predictions, while other studies have deployed stereo pipelines in controlled settings such as aquaculture tanks or breeding boxes **[2], [16].** Despite their effectiveness, stereo systems remain expensive, mechanically complex and increase maintenance requirements.

As an alternative, monocular camera systems offer a low-cost and easily deployable solution that integrates seamlessly with existing farm infrastructure. Early monocular approaches attempted to derive biomass by incorporating reference objects of known size **[17]** or by assuming constrained fish movement patterns **[18]**. While such strategies can partially recover scale, they often rely on artificial or restrictive conditions that do not generalize to open sea cages, where fish move freely and environmental variability is substantial. For this reason, monocular methods historically suffered from unreliable metric estimation and were considered less suitable for large-scale biomass monitoring **[1]**. However, with the advancement of deep learning and the emergence of foundation models for depth estimation **[10]**, there is growing momentum toward purely vision-driven, monocular biomass assessment. These modern methods aim to infer fish geometry directly from RGB video without requiring specialized hardware, leveraging large-scale pretraining to infer metric depth more reliably than earlier monocular techniques.

B. Monocular Metric Depth Estimation

Monocular depth estimation has advanced rapidly over the past decade, driven by the emergence of large-scale RGBD datasets such as NYU Depth v2 **[19]** and KITTI **[20],** together with powerful transformer-based architectures **[21], [22].** Recent advances have enabled models not only to estimate relative depth maps, but also to evolve into general purpose metric-depth foundation models that can predict absolute depth directly from a single RGB image **[10].** These models learn strong geometric priors through a combination of synthetic supervision, large-scale real-world pre-training, and self-supervised knowledge distillation.

State-of-the-art models such as Depth Anything v2 **[7],** Depth Pro **[9],** UniDepth v2 **[8],** Metric3D v2 **[23],** and the recently introduced Depth Anything v3 **[11]** typically adopt an encoder–decoder architecture, where high-level scene features are extracted either via Vision Transformers (ViT) or convolutional backbones, followed by a dense prediction module that outputs a per-pixel depth map **[22].** Many of these models rely on synthetic datasets such as Hypersim **[24]** or SceneNet **[25]** for ground-truth metric supervision, combined with billions of unlabeled real-world images for pseudo-label pretraining through teacher–student distillation. As a result, modern monocular depth foundation models achieve strong performance across a wide range of terrestrial environments, including indoor and outdoor scenes.

C. Unstructured Underwater Depth Estimation

Despite deep learning model's success in terrestrial scenes, these models face significant challenges when deployed underwater. Light absorption, wavelength attenuation, turbidity, scattering, refraction, and color distortion alter visual cues that monocular depth models depend on, leading to deviations in metric scale when used in aquaculture cages **[26], [27].** Prior studies **[12], [28]** have shown that even high-performing terrestrial models experience performance degradation underwater unless domain adaptation or explicit scaling is applied. Vision-based methods often incorporate enhancement steps, color correction, refraction compensation, to restore image fidelity before depth estimation **[27].** Additionally, dedicated underwater monocular depth systems, such as UW-Depth **[29]**, rely on supervised training using underwater RGB–depth pairs. However, the scarcity of high-quality underwater ground-truth limits model scale and generalization.

Our work focuses on evaluating several state-of-the-art monocular depth models, such as UniDepth v2 **[8],** Depth Anything v2 **[7],** Depth Anything v3 **[11],** and Depth Pro **[9],** and addresses the existing gap in unstructured underwater settings for the purpose of fish biomass estimation. In addition to zero-shot benchmarking, we also examine simple scaling strategies to correct raw metric predictions and fine-tuning the general-purpose Depth Anything v2 model on our collected underwater stereo-aligned dataset.

Materials and methods

A. Overview

The methodology of this study follows a unified processing pipeline designed to evaluate monocular depth estimation models for fish biomass prediction in real aquaculture environments. The workflow begins with monocular RGB video captured inside commercial sea cages. These frames are processed by a monocular depth estimation model to generate a depth map, which is then transformed into metric scale through model-specific or regression-based scaling methods. Length (L) is subsequently estimated from detected keypoints on the fish's body, and biomass is computed using a species specific length–weight relationship.

For this study, all models are benchmarked using the same quantitative metrics for depth, L, and biomass accuracy, enabling a consistent comparison of scaling performance and in-domain adaptation.

An overview of the processing pipeline used for length (L) estimation and average biomass prediction is shown in Figure 1.

B. Experimental Materials: Hardware and Software Components

All experiments were carried out in commercial aquaculture facilities located in Greece, where each open-sea cage contains approximately 400,000 fish. Recordings were collected from multiple cages that naturally varied in turbidity, lighting conditions, fish density, and background structure. To ensure consistency during data collection, footage was recorded during periods of relatively stable ambient lighting and reduced surface reflection. Frames exhibiting excessive turbidity, bubbles, occlusion, or low visibility were discarded prior to analysis.

a) Camera System: A commercial stereo vision camera was used for all recordings, providing synchronized RGB frames and stereo-derived depth maps. The device includes built-in functionality for dense depth estimation and 3D point cloud generation, recording at HD resolution and 30 FPS. For underwater operation, the camera was mounted inside a waterproof housing equipped with an anti-fog treated front window. In the context of this study, the left RGB frame was treated as a monocular input, while the stereo depth map served as the ground-truth reference for evaluating monocular model performance and supervising fine-tuning.

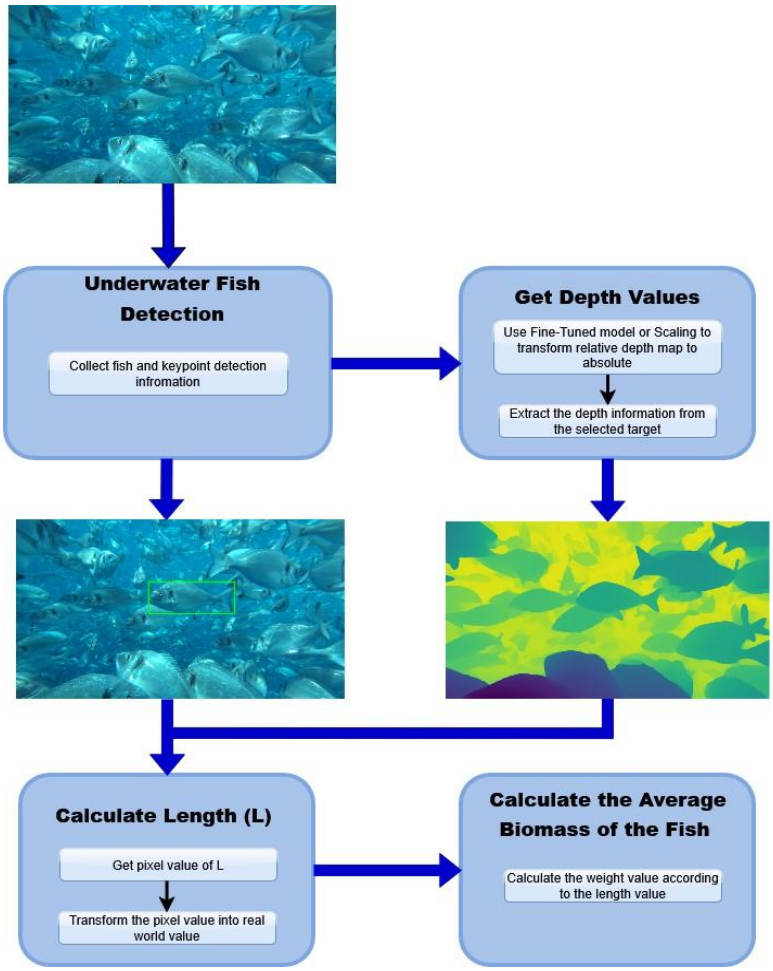

Figure 01: Schematic overview of the fish biomass estimation process [Alexandrou et al 2025, under publication]

b) Reference Objects: Reference Object played a key role in testing the models. A waterproof plastic fish of known length (L = 33 cm) was used for the underwater experiments for L and depth estimation validation.

c) Computing Hardware: All computational experiments were performed on a workstation equipped with an NVIDIA RTX 3090 GPU, Intel Core i7 processor, 32 GB RAM, and a 1 TB SSD, enabling both real-time evaluation of monocular models and efficient fine-tuning of transformer-based architectures.

C. Benchmark Monocular Depth Models

This study evaluates four state-of-the-art monocular depth estimation models, selected based on their reported performance and relevance to real-world depth estimation tasks. All models were used as provided in their original public repositories, without any architectural modifications, ensuring a fair and reproducible comparison. Each model was first assessed in a zero-shot setting using raw RGB frames as input, without retraining or adaptation, allowing an unbiased evaluation of their baseline generalization to underwater imagery. Among the evaluated models, Depth Anything v2 was additionally fine-tuned on underwater samples to investigate the impact of domain-specific adaptation.

a) UniDepth v2: UniDepth v2 **[8]** is a universal, transformer-based framework designed for zero-shot monocular metric depth estimation. The model predicts dense 3D point clouds from a single RGB image by leveraging a camera-prompting module that implicitly encodes geometric priors. UniDepth v2 uses geometric invariance losses and a learned camera representation to support robust generalization across unseen environments without requiring explicit camera intrinsics. Its architecture is particularly suitable for domains with varying imaging conditions, making it a strong candidate for underwater evaluation **[8].**

b) Depth Pro: Depth Pro **[9]** is a real-time monocular metric depth estimation model optimized for sharp, high resolution depth maps and accurate focal-length estimation. The model is specifically engineered for efficient deployment on resource-constrained devices, enabling fast inference while maintaining strong absolute depth consistency. Depth Pro's design prioritizes structural detail and high-frequency depth

information, characteristics that are valuable for detailed absolute depth estimation **[9].**

c) Depth Anything v2: Depth Anything v2 **[7]** is a largescale depth foundation model trained on billions of pseudo labeled images using self-supervised learning. It employs a vision transformer (ViT) encoder **[30]** and a DPT-based decoder **[22]** to produce accurate absolute depth maps in a wide variety of environments. The model is designed to generalize well without requiring paired depth data, making it a robust baseline for zero-shot evaluation. In this study, Depth Anything v2 also serves as the primary model for supervised fine-tuning on underwater footage, allowing a direct comparison between zero-shot and in-domain adapted performance **[7].**

d) Depth Anything v3: Depth Anything v3 **[11]** represents a significant advancement over its predecessor by extending monocular depth estimation toward full 3D scene understanding. In addition to producing dense metric depth maps from a single RGB image, the model incorporates a unified depth–ray representation that enables consistent geometry prediction and supports multi-view reasoning, camera pose estimation, and coarse 3D reconstruction. This design improves geometric stability and reduces scale drift in challenging scenes. Compared to Depth Anything v2, the new version demonstrates stronger spatial consistency and robustness to viewpoint changes, making it particularly promising for complex underwater environments where depth ambiguity, refraction, and limited visual cues are common **[11].**

D. Data Description

a) Data Sources: Data for this study were collected from operational aquaculture cages in Greece, each containing approximately 400,000 fish. This dataset supported both the zero-shot benchmarking of monocular depth models and the supervised fine-tuning of Depth Anything v2 on domains specific underwater imagery.

b) Benchmarking Dataset: All benchmark models were evaluated on a real-world dataset consisting of 20 full length underwater videos, each approximately two hours long, recorded across multiple aquaculture sites. These recordings span a wide range of environmental conditions, including varying turbidity levels, fish densities, lighting conditions, and background structures. To standardize the evaluation procedure, each video was sampled at a rate of one frame every seven frames, resulting in approximately 10,000 RGB frames per video and an estimated 2,000 visible fish instances per recording. This ensured a diverse dataset representative of real operational variability without overwhelming redundancy.

c) Fine-Tuning Dataset: To train and evaluate the finetuned monocular depth model, approximately 45,000 paired RGB frames and stereo-derived depth maps were collected from 10 different aquaculture environments. These samples were split into an 80% training set and a 20% validation set. Additionally, 15,000 previously unseen frames from different cages were reserved as a held-out test set to assess generalization performance after fine-tuning. This separation ensured that evaluation was performed on distinct temporal segments of the environment and was not influenced by data leakage.

d) Environmental Control During Data Collection: To minimize variability introduced by underwater environmental conditions, all recordings were conducted under stable daylight and adequate water clarity. The camera was installed at a fixed depth and orientation across recording sessions to maintain consistent viewing geometry and illumination conditions. Data collection was temporarily suspended during periods of elevated turbidity or particle concentration and resumed once visibility improved. No artificial illumination was employed; instead, the stereo camera's automatic exposure and white balance settings were used to compensate for brightness and color fluctuations. These measures were implemented to reduce environmental bias and ensure consistency in the depth estimation results.

e) Data Preprocessing: For the zero-shot evaluation of benchmark models, no preprocessing was applied beyond standard resolution handling, preserving the raw underwater characteristics of the input frames. For the fine-tuning experiments, RGB images and depth maps were resized to 518 × 518 pixels. RGB inputs were normalized using ImageNet statistics (mean: [0.485, 0.456, 0.406], std: [0.229, 0.224, 0.225]). To increase robustness to small viewpoint variations and horizontal flipping were applied. To further improve model performance, we incorporated online color-based data augmentation techniques, including random adjustments to image contrast and brightness. These augmentations were applied on the fly during training, without generating or storing additional training data.

E. Fine-Tuning the Monocular Depth Model

To enhance the monocular depth estimation results in underwater aquaculture environments, we fine-tuned the Depth Anything v2 model **[7]** using stereo-derived depth maps as supervisory signals. The objective was to adapt the model to domain-specific challenges such as turbidity, variable lighting, occlusion, and underwater color distortions, while preserving the strong visual priors learned from large-scale terrestrial pretraining.

Fine-tuning was performed using paired RGB–depth samples acquired inside operational aquaculture cages. The encoder weights were initialized from the pre-trained Depth Anything v2 small checkpoint, while the decoder was randomly initialized to allow for flexible adaptation to underwater geometry and texture statistics. All experiments used a maximum depth threshold of 2 meters to emphasize the foreground region where fish typically appear and where precise metric depth estimation is most critical. The same maximum depth constraint was also applied to the stereo camera depth estimates, as it was empirically found to improve depth quality and reliability within the 2-meter range, leading to better overall measurement performance in the evaluated scenarios.

Training was conducted for 50 epochs with a batch size of 8. We used the AdamW optimizer with a weight decay of 10^{-2} **[31],** an initial learning rate of 5×10^{-6}, and a linear warm-up phase over the first 3 epochs. The loss function was the scale-invariant logarithmic loss (SiLogLoss) **[32],** widely adopted for metric depth regression due to its robustness to global scale variations.

After fine-tuning, the adapted model was tested on previously unseen sequences from different cages to assess its generalization across environments. Predictions were compared directly against stereo-derived depth maps and stereo-based L and biomass estimates, which were treated as ground-truth references.

F. Length and Biomass Estimation

Length (L) was estimated from two anatomically consistent keypoints, the mouth and tail, either annotated manually or detected automatically using a keypoint-aware object detection model (e.g., Faster R-CNN **[33]**). For the fake-fish experiments, the keypoint model was not used; instead, the mouth and tail positions were manually selected using a custom annotation script. The pixel coordinates of each keypoint were projected into 3D space using the scaled depth map and the camera's intrinsic parameters:

$$X = \frac{(x - C_x)Z}{f} \quad (1)$$

$$Y = \frac{(x - C_y)Z}{f} \quad (2)$$

where (x,y) represent pixel coordinates, Z is the depth value at that pixel, C_x, C_y are the principal point offsets, and f is the focal length.

The 3D length was then computed as the Euclidean distance between the reconstructed head and tail points:

$$L = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2 + (Z_2 - Z_1)^2} \quad (3)$$

Biomass (W) was estimated using the species-specific length–weight relationship (LWR) **[34]**:

$$W = a \cdot L^b \quad (4)$$

The parameters a and b were empirically estimated using our dataset and were found to align with typical published ranges for *Sparus aurata* (gilthead seabream, locally "tsipoura").

G. Evaluation Metrics

Model performance on fish length estimation was evaluated by comparing monocular-derived length (L) predictions against stereo-derived L measurements, which served as ground-truth references. Ground truth depth, fish length, and biomass measurements were obtained using a calibrated stereoscopic camera system, where stereo-derived depth maps were used to reconstruct fish geometry and compute biomass estimates for all evaluations. Although the ultimate objective is to estimate cage-level biomass, accurate biomass prediction fundamentally depends on reliable fish length estimation, which in turn requires accurate depth reconstruction. For this reason, we focus our evaluation on L as the primary geometric measurement underpinning biomass estimation and assess all models based on their ability to recover accurate fish lengths.

We report the following quantitative metrics.

a) Per-Fish Mean Absolute Length Error (L-MAE):

$$LMAE = \frac{1}{N}\sum_{i=1}^{N}\left|\widehat{L} - L_i\right| \quad (5)$$

where $L_b i$ is the predicted standard length for fish i and L_i is the corresponding ground-truth measurement.

b) Standard Deviation of Length Errors (L-STD):

$$L - STD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left|\widehat{L} - L_i\right| - LMAE\right)^2} \quad (6)$$

This metric quantifies the variability of per-fish absolute length errors.

c) Mean Absolute Error of Video-Level Mean Lengths (Video L-MAE): For each video v, let $L_b v$ denote the mean predicted L and L_v the mean ground-truth L. The per-video mean length error is:

$$e_v = \left|\overline{\widehat{L_v}} - \overline{L_v}\right| \quad (7)$$

Averaging across all V videos yields:

$$Video\ L - MAE = \frac{1}{V}\sum_{v=1}^{V} e_v \quad (8)$$

This metric measures the absolute deviation between predicted and ground-truth average standard lengths at the video level, capturing systematic errors in aggregated length estimates rather than per-fish prediction accuracy.

d) Biomass Accuracy:

$$\Delta W = \frac{1}{N}\sum_{i=1}^{N}\widehat{W_i} - \frac{1}{N}\sum_{i=1}^{N} W_i \quad (9)$$

This metric quantifies the deviation of the average predicted biomass from the stereo-derived average biomass.

Results

A. Zero-Shot Benchmarking Results

In the zero-shot setting, all monocular depth models were evaluated on real aquaculture video without any retraining or domain adaptation. The results revealed significant overestimation and highlighted substantial scale-related errors when applied to underwater imagery.

Depth Pro [9], UniDepth v2 **[8],** and Depth Anything v3 **[11]** showed strong overestimation trends, with predicted L values significantly exceeding the stereo-derived ground truth. Depth Anything v3 **[11]** did not outperform the other models in the raw zero-shot setting, despite being the most recent architecture.

Depth Anything v2 **[7]** exhibited behavior that was strongly dependent on the maximum-depth parameter. When using the default recommended settings (20 m for indoor scenes and 80 m for outdoor scenes), the model significantly overpredicted depth, similar to the other zero-shot models. However, restricting the maximum depth to 7.5 m, based on empirical underwater visibility, resulted in improved L and depth predictions, although a residual bias remained.

The benchmark confirms that the dominant error in the zero-shot setting is linked to global scale mismatch rather than structural misinterpretation of the scene. Quantitative results for all zero-shot configurations are summarized in Table I.

B. Scaled Model Performance

To further evaluate the impact of scale mismatch, post-hoc scaling was applied to the zero-shot depth predictions by dividing the estimated depths by a constant factor of 2. This factor was derived from the average regression-based scale correction observed across all evaluated models and was conservatively rounded upward to avoid underestimation. Applying this scaling resulted in a notable reduction in both length (L) and depth errors across all models.

For Depth Pro **[9],** UniDepth v2 [**8],** and Depth Anything v3 **[11],** this simple scaling significantly reduced overestimation and brought predicted values closer to the stereo-derived ground truth. While the improvement confirms that depth structure was reasonably preserved, the dependency on a scaling factor highlights a limitation in generalizability.

The quantitative performance of the scaled models is reported in Table I.

C. Fine-Tuned Model Performance

The fine-tuned version of Depth Anything v2 **[7]** achieved the best overall performance across all experiments. After supervised training using stereo-aligned underwater data, both L and depth errors were significantly reduced compared to all zero-shot and scaled configurations.

In addition to lower mean error, the fine-tuned model also exhibited reduced variability in predictions, providing more consistent depth and L estimates across frames and scenes. These results are included in Table I, where the fine-tuned model consistently outperforms all other configurations.

D. Fake Fish Benchmarking Results

A controlled experiment using a fake fish was conducted to evaluate absolute scale accuracy. All zero-shot models substantially overestimated the length of the object, with Depth Anything v3 **[11]** and UniDepth v2 **[8]** showing the largest deviations.

Constraining Depth Anything v2 **[7]** to a maximum depth of 7.5 m reduced the error but did not fully eliminate bias. The fine-tuned version of Depth Anything v2 **[7]** produced the most accurate and stable predictions in this scenario. Scaled variants of the other models also showed improved results, coming close to the performance of the fine-tuned model.

A full summary of the fake fish benchmarking results is provided in Table II.

Discussion

A. Interpretation of Results

a) Zero-Shot Model Performance: The zero-shot evaluation clearly demonstrates that, although recent foundation models are capable of producing visually coherent and structurally consistent depth maps, their ability to preserve correct metric scale in unstructured underwater environments remains highly limited. Depth Pro **[9]**, UniDepth v2 **[8],** and Depth Anything v3 **[11]** all exhibited severe scale drift when applied directly to aquaculture footage, consistently overestimating both absolute depth and length. These errors were not random but systematic, indicating that the models' learned geometric priors do not transfer effectively to underwater optical conditions.

Among the zero-shot models, Depth Anything v2 **[7]** demonstrated a unique sensitivity to its maximum depth setting. By default, the model is configured for indoor (20 m) or outdoor (80 m) scenes. As mentioned before, when these recommended settings were applied, it produced substantial overestimation in the aquaculture environment, comparable to the other models. However, constraining the maximum depth to 7.5 m—based on empirical observations of underwater visibility in the cages—led to a marked improvement in L and depth estimation accuracy. This result highlights that environmental prior knowledge can partially mitigate scale ambiguity in foundation models. At the same time, it exposes a critical weakness: the model becomes strongly dependent on scene-specific

hyperparameters, which reduces autonomy under dynamic water conditions.

Similarly, Depth Anything v3 **[11],** despite its architectural improvements and enhanced 3D representation capacity, exhibited overestimation comparable to its predecessor. This reinforces the conclusion that architectural sophistication alone is insufficient to resolve scale ambiguity when a pronounced domain gap is present between training data and the target underwater environment.

b) Effectiveness of Scaling and Fine-Tuning.: The results clearly demonstrate that both post-hoc scaling and supervised fine-tuning can substantially reduce depth and L estimation errors in monocular systems, though with important differences in robustness and generalizability. Simple linear scaling,

Table i: Overall individual fish length (L) and depth estimation errors for monocular depth models across different configurations.

Model	L MAE (cm)	L STD (cm)	Depth MAE (cm)	Depth STD (cm)	Video-level L MAE (cm)
Benchmark (zero-shot / constrained) models					
Depth Pro	24.63	6.45	73.13	19.56	24.62
UniDepth v2	29.86	7.14	92.90	14.65	29.86
Depth Anything v3	29.13	7.01	91.98	14.15	29.76
Depth Anything v2 (max distance = 7.5 m)	3.76	3.13	15.28	14.35	2.95
Scaled models (depth / 2)					
UniDepth v2 (depth / 2)	4.22	2.63	15.72	11.06	3.59
Depth Pro (depth / 2)	4.05	3.23	12.99	11.01	2.94
Depth Anything v3 (depth / 2)	4.15	2.94	13.59	11.52	3.21
Fine-tuned model					
Depth Anything v2 (fine-tuned)	2.42	1.40	9.37	6.21	2.17

Table ii: Fake fish L prediction results for monocular depth models.

Model	Mean L (cm)	L STD (cm)	Mean L Error (cm)
Benchmark (zero-shot) models			
Depth Anything v2 (7.5 m)	26.07	2.88	7.00
Depth Pro	51.01	7.03	18.22
Depth Anything v3	65.91	7.64	33.11
UniDepth v2	59.77	8.56	28.73
Scaled models (depth / 2)			
Depth Pro (depth / 2)	23.80	3.16	7.29
UniDepth v2 (depth / 2)	25.48	2.06	7.48
Depth Anything v3 (depth / 2)	28.43	5.78	5.21
Fine-tuned model			
Depth Anything v2 (FT)	29.16	4.81	3.84

such as dividing predicted depth values by a constant factor, significantly reduced overestimation for all models, including Depth Pro **[9],** UniDepth v2 **[8],** and Depth Anything v3 **[11].** This confirms that a large portion of the zero-shot error lies in a systematic scale mismatch rather than in structural misinterpretation of the scene. In controlled settings, this approach can provide a fast and computationally negligible improvement in prediction accuracy.

However, the reliance on a scaling factor introduces an inherent limitation. The optimal factor is highly dependent on camera placement, water clarity, and scene geometry, which can change significantly across cages or even within the same cage over time. As a result, such scaling cannot be safely generalized across farms or deployment conditions without recalibration, limiting its reliability for large-scale or long-term monitoring systems. As we observed in our results, some cages required much lower scaling factors (e.g., 1.5), while others needed substantially higher values (e.g., 2.5 or even 3), further illustrating the variability and lack of generalizability.

In contrast, fine-tuning Depth Anything v2 **[7]** on stereo-aligned underwater data led to the most substantial and consistent performance improvements. This approach enabled the model to internalize underwater-specific visual cues, such as color attenuation patterns, reduced contrast, and scattering artifacts, rather than relying on external correction factors. The fine-tuned model achieved the lowest errors in both depth and L estimation across all experiments, indicating that forward domain adaptation is not only beneficial but essential for achieving stable, physically meaningful monocular depth predictions in real aquaculture environments.

c) Fake Fish Benchmark Analysis: The fake fish experiment provided a controlled geometric reference to further analyze how each model handled absolute scale under identical conditions. As shown in Table II, all zero-shot models significantly overestimated the length of the 33 cm reference object, with Depth Anything v3 [11] and UniDepth v2 [8] producing the largest deviations. These results confirm that the models were not simply misinterpreting fish shape, but were systematically mis-scaling the entire scene. In contrast, the fine-tuned Depth Anything v2 **[7]** model demonstrated a substantial reduction in error, producing the closest agreement with stereo-derived L values among all evaluated approaches.

The scaled variants further highlight the partial effectiveness of simple correction strategies. Applying a constant scaling factor reduced the magnitude of the errors across all models. However, the resulting L estimates exhibited slightly greater errors compared to the fine-tuned model. This reinforces the conclusion that manual scaling can correct global drift but cannot fully capture the complex, non-linear distortions introduced by underwater imaging.

Overall, while this experiment may not fully represent the fine-tuned model's true performance, given that it was not trained on scenarios involving a fake fish attached to a stick, the fake-fish benchmark still reinforces a key insight: although heuristic scaling can offer short-term improvements, fine-tuning remains the most reliable and principled approach for achieving accurate and consistent metric measurements in underwater monocular depth estimation.

B. Computational Requirements and Deployment Considerations

The proposed biomass estimation pipeline is designed for continuous on-farm monitoring rather than frame-synchronous real-time video processing. While input video streams are captured at 30 FPS, depth inference is performed on a subsampled set of frames to reduce computational load. On the reference hardware (NVIDIA RTX 3090), the end-to-end processing time per selected frame ranges between approximately 0.5 and 1 second, depending on scene content, particularly the number of detected fish instances.

This level of latency is sufficient for biomass estimation, as fish length distributions and aggregate biomass statistics evolve on significantly slower timescales than individual video frames. The system therefore supports near real-time or periodic processing suitable for operational monitoring and decision support in commercial aquaculture environments. This enables deployment on on-farm servers or GPU-accelerated edge systems where delayed or periodic inference is acceptable.

C. Implications for Aquaculture

The results of this study demonstrate that reliable biomass estimation can be achieved using a single monocular camera when paired with robust depth estimation and appropriate domain adaptation. This has important implications for commercial aquaculture, where cost, scalability, ease of installation, and minimal disruption to fish welfare are critical considerations.

First, the ability to estimate length and biomass from monocular imagery significantly reduces system complexity and hardware requirements. Unlike stereo vision systems, which require precise baseline calibration, synchronized sensors, and higher deployment costs, a monocular setup can be installed more easily on existing cage infrastructure. This makes large-scale, multi-cage monitoring more economically feasible, enabling farmers to deploy multiple units across sites instead of relying on a limited number of high-cost stereo systems.

Second, the fine-tuned model demonstrates strong robustness to underwater visual challenges such as turbidity, scattering, and variable lighting. This is particularly important in real-world aquaculture environments where conditions change daily and cannot be tightly controlled. The improved stability and reduced variance in predictions suggest that the system is suitable for continuous monitoring, rather than sporadic measurements.

From an operational perspective, this method supports noninvasive, near real-time biomass estimation. Since it requires only passive imaging and does not involve handling, crowding, or restraining fish, it aligns with modern animal

welfare standards. Continuous and automated biomass monitoring can provide more accurate growth curves, allowing farmers to optimize feeding strategies, reduce waste, and prevent overfeeding, which in turn helps lower operational costs and minimize environmental impact.

In addition, monocular-based biomass estimation opens the door to edge-based, real-time deployments. The relatively low computational requirements compared to dense stereo reconstruction allow the system to run on edge devices. This enables on-site processing with minimal bandwidth requirements, making it suitable for offshore farms and remote locations where constant data transmission to cloud servers may not be practical.

D. Limitations of the Current Study

Despite the encouraging results achieved in this study, several limitations must be acknowledged.

The most significant limitation is the reliance on stereo-derived depth maps as ground truth for both evaluation and fine-tuning. Although stereo vision provides reasonably accurate depth estimates, it is not immune to noise, particularly in underwater environments where turbidity, light scattering, and reflective particles can degrade depth quality. As a result, any inaccuracies present in the stereo depth maps may propagate into the monocular model during training and evaluation.

Another important limitation relates to environmental variability. Even though recordings were conducted under relatively stable conditions, underwater environments remain inherently dynamic.

The dataset used in this study, while extensive, is still limited in scope. Expanding the dataset to include a wider range of use cases, such as different aquaculture systems, fish sizes, growth stages, and species, would be beneficial in improving the generality and robustness of the model.

Furthermore, only a limited number of monocular depth models were benchmarked in this work. Although the selected models represent state-of-the-art approaches, broader evaluation across a wider range of architectures and training paradigms could provide deeper insight into the strengths and limitations of monocular depth estimation in underwater environments.

In addition, only one model (Depth Anything V2) was finetuned for underwater conditions. While this allowed a focused evaluation of domain adaptation, it does not fully represent the fine-tuning potential of other promising foundation models.

Finally, while this study demonstrates strong results using the left camera of a stereo system, the model has not yet been fully evaluated on independent monocular cameras with different sensors, lenses, and resolutions. Variations in camera quality and optical characteristics could influence depth prediction accuracy, potentially requiring additional calibration or adaptation steps.

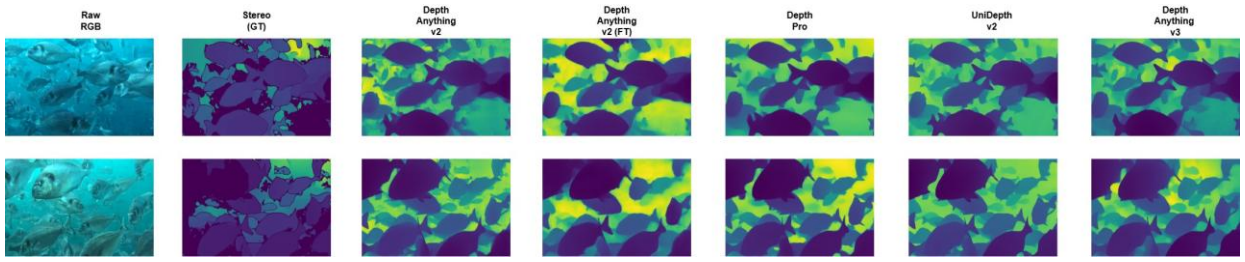

Figure 02: Visualization of depth predictions from all the models

E. Future Work

Several directions can be explored to extend and strengthen the findings of this study.

First, expanding the training and evaluation dataset to include a broader range of aquaculture environments, different fish species, and varying water conditions would significantly improve the generalization capabilities of the model. Incorporating greater diversity in fish size, morphology, and environmental characteristics will enable the development of a more robust and widely applicable system.

Second, future work should include full deployment and testing using true standalone monocular cameras, rather than relying on the left frame of a stereo system. Evaluating performance across different camera sensors, resolutions, and

lens configurations will provide a more realistic and practical assessment for real-world applications.

Another important direction is the exploration and finetuning of additional foundation depth models. While this work focused on Depth Anything V2 **[7]**, future studies could include large-scale benchmarking and fine-tuning of other state-of-the-art models.

Conclusion

This study investigated the feasibility of monocular depth estimation for fish biomass prediction in real aquaculture environments. Through a comprehensive evaluation of four state-of-the-art foundation models, Depth Anything V2 **[7],** Depth Anything v3 **[11],** Depth Pro **[9],** and UniDepth V2 **[8],** we demonstrated that monocular vision could provide a viable and cost-effective alternative to traditional stereo-based measurement systems.

The zero-shot benchmarking revealed substantial variation in model behavior when deployed in underwater conditions. Certain models exhibited strong overestimation or instability, highlighting the importance of scaling and domain awareness. These results emphasized that, although foundation models trained on terrestrial imagery provide powerful representations, underwater environments introduce unique challenges such as turbidity, color attenuation, and refraction that significantly impact depth prediction accuracy.

To address this domain gap, the Depth Anything V2 **[7]** model was further fine-tuned using paired underwater RGB–depth data. The fine-tuning process resulted in a clear and consistent reduction in both depth and length (L) errors across all datasets and evaluation scenarios. Within the scope of this study, the fine-tuned model consistently outperformed all zero-shot and scaled baselines, achieving the lowest L and depth error metrics and emerging as the most accurate and reliable approach for monocular depth estimation in underwater aquaculture environments.

Evaluation on both real fish and a controlled reference object (fake fish) further confirmed the robustness and stability of the fine-tuned approach. These experiments demonstrated the model's ability to produce realistic and consistent length estimates, which directly enable more accurate biomass estimation and bring monocular performance closer to stereo-derived ground truth values.

Despite the encouraging results, several limitations remain. Model performance can be affected by challenging underwater conditions such as turbidity, lighting variability, and noise in stereo-derived ground truth, while the current evaluation focuses on a limited set of environments and species. Future work will address these limitations by expanding the dataset to more diverse aquaculture settings, evaluating additional monocular depth models, and validating performance using standalone monocular camera systems.

Overall, the findings of this work indicate that targeted fine-tuning is the most effective strategy for addressing scale ambiguity and domain shift in monocular depth estimation for aquaculture. This approach significantly reduces hardware costs, simplifies deployment, and enables continuous, noninvasive monitoring, making it a practical and scalable solution for commercial aquaculture operations.

Data Availability: The dataset was collected in commercial aquaculture cages and is proprietary under data ownership and confidentiality agreements with the farm operator, and is therefore not publicly available.

Declaration: All experimental procedures employed non-invasive imaging techniques without physical handling of fish. Data collection was authorized by the farm operators and adhered to established aquaculture welfare protocols, with careful measures taken to minimize disturbance and ensure compliance with animal welfare standards.

Acknowledgements: Funding for this work was provided by the Next Generation EU *Restart 2016–2020* Programme (project ID ENTERPRISES/0223/Sub-Call1/0157).

References

1. D. Li, Y. Hao, and Y. Duan, "Nonintrusive methods for biomass estimation in aquaculture with emphasis on fish: A review," *Reviews in Aquaculture*, vol. 12, no. 3, pp. 1390–1411, 2020.
2. G. Feng, B. Pan, and M. Chen, "Non-Contact Tilapia Mass Estimation Method Based on Underwater Binocular Vision," *Applied Sciences*, vol. 14, no. 10, p. 4009, Jan. 2024.
3. T. Zhang, Y. Yang, Y. Liu, C. Liu, R. Zhao, D. Li, and C. Shi, "Fully automatic system for fish biomass estimation based on deep neural network," *Ecological Informatics*, vol. 79, p. 102399, 2024.
4. C. Zhang, X. Weng, Y. Cao, and M. Ding, "Monocular Absolute Depth Estimation from Motion for Small Unmanned Aerial Vehicles by Geometry-Based Scale Recovery," *Sensors*, vol. 24, no. 14, p. 4541, Jan. 2024.
5. F. Xue, G. Zhuo, Z. Huang, W. Fu, Z. Wu, and M. H. A. Jr, "Toward Hierarchical Self-Supervised Monocular Absolute Depth Estimation for Autonomous Driving Applications," Sep. 2020.
6. R. Fan, T. Ma, Z. Li, N. An, and J. Cheng, "Region-aware Depth Scale Adaptation with Sparse Measurements," Jul. 2025.
7. L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth Anything V2," Oct. 2024.
8. L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. V. Gool, and F. Yu, "UniDepth: Universal Monocular Metric Depth Estimation," Mar. 2024.
9. A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun, "Depth Pro: Sharp

Monocular Metric Depth in Less Than a Second," Apr. 2025.

10. S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Muller, "ZoeDepth:¨ Zero-shot Transfer by Combining Relative and Metric Depth," Feb. 2023.
11. H. Lin, S. Chen, J. Liew, D. Y. Chen, Z. Li, G. Shi, J. Feng, and B. Kang, "Depth Anything 3: Recovering the Visual Space from Any Views," https://arxiv.org/abs/2511.10647v1, Nov. 2025.
12. Z. Cai and C. Metzler, "Underwater Monocular Metric Depth Estimation: Real-World Benchmarks and Synthetic Fine-Tuning with Vision Foundation Models," Jul. 2025.
13. X. Yang, X. Zhang, N. Wang, G. Xin, and W. Hu, "Underwater selfsupervised depth estimation," *Neurocomputing*, vol. 514, pp. 362–373, Dec. 2022.
14. P. Agand, M. Chang, and M. Chen, "DMODE: Differential Monocular Object Distance Estimation Module without Class Specific Information," May 2024.
15. H. Hu, Y. Feng, D. Li, S. Zhang, and H. Zhao, "Monocular Depth Estimation via Self-Supervised Self-Distillation," *Sensors (Basel, Switzerland)*, vol. 24, no. 13, p. 4090, Jun. 2024.
16. N. Tonachella, A. Martini, M. Martinoli, D. Pulcini, A. Romano, and F. Capoccioni, "An affordable and easy-to-use tool for automatic fish length and weight estimation in mariculture," *Scientific Reports*, vol. 12, no. 1, p. 15642, Sep. 2022.
17. A. Masoumian, D. G. F. Marei, S. Abdulwahab, J. Cristiano, D. Puig, and H. A. Rashwan, "Absolute distance prediction based on deep learning object detection and monocular depth estimation models," Oct. 2021.
18. J. Mei, J.-N. Hwang, S. Romain, C. Rose, B. Moore, and K. Magrane, "Absolute 3d pose estimation and length measurement of severely deformed fish from monocular videos in longline fishing," 2021.
19. [Online]. Available: https://arxiv.org/abs/2102.04639
20. D. Ignatov, A. Ignatov, and R. Timofte, "Virtually Enriched NYU Depth V2 Dataset for Monocular Depth Estimation: Do We Need Artificial Augmentation?" Apr. 2024.
21. A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.
22. I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-mixer: An all-MLP architecture for vision," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 24261–24272.
23. R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision Transformers for Dense Prediction," Mar. 2021.
24. M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, K. Wang, H. Chen, G. Yu, C. Shen, and S. Shen, "Metric3Dv2: A Versatile Monocular Geometric
25. Foundation Model for Zero-shot Metric Depth and Surface Normal Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10579–10596, Dec. 2024.
26. M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind, "Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding," Aug. 2021.
27. J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, "SceneNet RGB-D: 5M Photorealistic Images of Synthetic Indoor Trajectories with Ground Truth," Jan. 2017.
28. S. P. Gonzalez-Sabbagh and A. Robles-Kelly, "A Survey on Underwater´ Computer Vision," *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 268:1– 268:39, Jul. 2023.
29. D. Akkaynak and T. Treibitz, "Sea-Thru: A Method for Removing Water From Underwater Images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 1682–1691.
30. H. Liu, M. Roznere, and A. Q. Li, "Deep Underwater Monocular Depth Estimation with Single-Beam Echosounder," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. London, United Kingdom: IEEE, May 2023, pp. 1090–1097.
31. L. Ebner, G. Billings, and S. Williams, "Metrically Scaled Monocular Depth Estimation through Sparse Priors for Underwater Robots," Oct. 2023.
32. A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.
33. I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Jan. 2019.
34. D. Eigen, C. Puhrsch, and R. Fergus, "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network," Jun. 2014.
35. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards RealTime Object Detection with Region Proposal Networks," Jan. 2016.
36. N. Jisr, G. Younes, C. Sukhn, and M. H. El-Dakdouki, "Length-weight relationships and relative condition factor of fish inhabiting the marine area of the Eastern Mediterranean city, Tripoli-Lebanon," *Egyptian Journal of Aquatic Research*, vol. 44, no. 4, pp. 299–305, 2018.

Citation: *Alexandrou A, Christodoulou F, Komninos, D, Ozdeger, T, Seferis, K (2026) Benchmarking and Fine-Tuning Monocular Depth Models for Fish Biomass Estimation in Aquaculture. Jr Aqua Mar Bio Eco: JAMBE173.*