MANUSCRIPT UNDER REVIEW (April 2026, minor revisions)

Title: Advancing automated fish size estimation from images: Applications, challenges, and a case study for images without a specified reference object

Running title: Automated Fish Size Estimation

Authors:

Catarina NS Silva[1,2], Ricardo Cardoso Pereira[3], Freddie Heather[4], Sean Simmons[5], Asta Audzijonyte[4,6]

Affiliations:

[1]Centre for Functional Ecology - Science for People & the Planet (CFE), Associate Laboratory TERRA, Department of Life Sciences, University of Coimbra, 3030-790 Coimbra, Portugal – catarina.s.silva@uc.pt
[2]Centre for Functional Ecology - Science for People & the Planet (CFE), Associate Laboratory TERRA, University of Coimbra Campus at Figueira da Foz, Quinta das Olaias 3080-183 Figueira da Foz, Portugal – catarina.s.silva@uc.pt
[3] University of Coimbra, CISUC/LASI - Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering, Coimbra, Portugal
[4] Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania, Australia
[5] Angler's Atlas, Goldstream Publishing, Prince George, British Columbia, Canada
[6] Centre for Marine Socioecology, Tasmania, Australia

Keywords

Fish size, Computer vision, Machine learning, Artificial intelligence, Fisheries management, Monitoring, Recreational fisheries

Abstract

Automatic and accurate estimation of fish sizes from images and videos is essential for many monitoring, fisheries management, stock assessment and conservation efforts. However, current methods often rely on physical reference objects or stereo-camera systems that are not always available. This paper explores the advancements, applications, and challenges of automated fish body-size estimation from images, using artificial intelligence (AI) and machine learning (ML) methods. We first introduce key concepts in AI and ML for a non-specialised audience and review existing literature on models used for fish size estimation. We identify key barriers such as a lack of high quality and publicly available datasets, image variability, scattered efforts and the challenges of model generalisation across diverse species. Then we present a novel framework for size estimation from monocular (non-stereo) images without a specified reference object, using a fishing tournament dataset from an angling app. Our approach utilises an efficient, pretrained deep learning-based feature extraction tool integrated with an automated regression pipeline and can be run on a single computer with a GPU. Our findings demonstrate a promising pathway for size estimation in images without a reference object, where estimated fish lengths were mostly within 10% of their true length. Future research and collaborative efforts should focus on improving the diversity and public availability of training data and integrating image metadata to enhance the accuracy of size estimation. Additionally, it is essential to rigorously test and refine the robustness of current models in real-world fisheries applications and to adopt standardised, comparable metrics for evaluating and benchmarking fish size estimation models across studies.

Table of Contents

1. Introduction

In fish population and fisheries research and management, data on fish body sizes, or length frequencies, can provide important insights into population status and species biology. Size data underpins many ecological and management frameworks, including size-spectrum theory, biomass estimation, and length-based stock assessments (Froese et al., 2018; Hordyk et al., 2014; Pauly, 1987). Body size is correlated with key biological rates such as growth, metabolism and reproduction (Brown et al., 2004; Peters, 1983), it affects species trophic position and predation risk (Jennings et al., 2001) and is therefore a valuable metric for understanding individual fitness and broader ecosystem dynamics. Importantly, unlike ageing, fish length can be estimated non-invasively from visual observations, making it particularly useful for underwater monitoring or aquaculture settings where destructive sampling is not feasible or required.

The increasing availability of digital imagery has created new opportunities to automate the collection of fish and aquatic invertebrate body size data, for example through underwater cameras, onboard catch monitoring systems, angler or small-scale fisheries contributed photos, citizen science, and cameras installed on fish cleaning stations. Machine learning (ML), which is a branch of artificial intelligence (AI) focused on algorithms that improve their performance through data-driven optimisation, and computer vision, a subfield of AI focused on image-based tasks, now provide powerful tools to automate the extraction of fish length information from such datasets. Computer vision has already been applied successfully for automated species identification in fisheries catch and bycatch monitoring, behavioural analysis such as feeding or interactions with fishing gears, and estimation or monitoring of weights or lengths in fisheries (onboard, recreational fishing) and aquaculture (both underwater and out-of-water) settings (Abangan et al., 2023; Barbedo, 2022; Lonati et al., 2024; Sheaves et al., 2020). In both aquaculture and fisheries, AI tools are now increasingly used for both real-time or post hoc processing of visual data. Yet, while numerous studies have focused on developing ML models for fish detection and species identification, fewer have addressed the challenge of estimating fish size. To date, there is still no generic, widely accepted, and publicly available framework for extracting size information from the wide variety of currently available image sources. As a result, the large volumes of fish imagery generated across fisheries, ecological monitoring, and citizen science efforts remain underutilised for fish population assessment.

Several key factors continue to limit AI applications in real life fish or catch monitoring situations, including the high variability in lighting, water clarity and background conditions across images, as well as diverse fish morphologies (Huang et al., 2025). Further, the lack of coordinated efforts to build large, diverse, and annotated datasets for training size estimation models and the commercial nature of many existing tools restrict open access, reproducibility, and community-driven development. This is unfortunate, both because there is an urgent need for better monitoring tools to address growing pressures on fish populations, and because the widespread adoption of digital technology could provide unprecedented amount of fish size data for research. Bridging this gap between the technological and data availability potential and practical applications requires funding and closer integration between model developers and end users in fisheries science, aquaculture and ecology. For non-specialist users such as fisheries managers, conservation practitioners, and field ecologists, understanding the capabilities and limitations of different model types is crucial for informed adoption (Wing & Woodward, 2024). Equally important is the development of open-source user-friendly tools, standardised datasets, and collaborative initiatives that foster transparency and accelerate progress across disciplines.

In this study we aim to: 1) introduce key machine learning (ML) concepts relevant to fish body size estimation for a non-specialised audience, 2) summarise current main approaches and studies employing ML for fish size estimation from images, 3) present a case study and an open source model on estimating fish sizes in monocular images (i.e. images captured with a single camera, without stereo or depth information) without a specified reference object and 4) provide recommendations for improving the transparency and uptake of ML tools for automated fish size estimation.

2. Automated approaches to measure fish

The process of automating fish size measurements in images has traditionally been divided into two main tasks. The first task is to accurately detect fish of interest in the image and determine the two points between which the length should be estimated. This task is done by object detection models, which either place a bounding box around the detected fish individual or segment it from the image, i.e. trace its overall shape. The models then need to determine which two points to use for fish size estimation, typically the tip of the snout and the end of the tail (e.g. Abinaya et al., 2022). Alternatively, the model might first estimate the area of the traced fish shape and derive the fish length from that (Climent-Perez et al., 2024). Moreover, if a fish is in a distorted position, individual length must be reconstructed by measuring fish along a set of defined and consistent body segments (e.g. Yu et al., 2023). This task of finding a fish may seem straightforward, given the rapid advances of object detection models, however the challenge is still far from resolved in real world applications, where light and background conditions, as well as fish positions and distance to the camera are highly variable (reviewed in section 2.1).

Assuming the first task of identifying the points for size measurement is completed with sufficient accuracy, the second task is to relate the pixel distance or area to a real size measure. Converting pixel distance can be done using a reference object of known length (e.g., a ruler) present in the image, using stereo images where the distance to the camera and 3D shape of the fish can be reconstructed from two images taken from slightly different angles, or estimating distance of the fish to the camera (image depth) in cases where camera and lens parameters are known. Alternatively, deep learning models such as visual transformers (presented in section 3) can be trained to directly predict fish length from image features, bypassing explicit object detection and measurement steps. In such cases, the exact methodology of measurement becomes a “black box”, as the internal reasoning used by the model to estimate size is often not interpretable.

While this study focuses on AI based fish size estimation, accurate detection of fish and its shape has been an inextricably linked part of the process and is often the main determinant of the model performance (see also review by Barbedo, 2022). Therefore, before progressing to size estimation models, we briefly review the progress of object detection models in fish image analysis.

2.1. Object detection models and their applications for fish studies

Object detection models are now widespread, from smartphones, to cameras on streets, self-service supermarket checkouts and self-driving cars. The evolution of object detection models has been reviewed by Zou et al. (2019), but here we provide a brief overview that focuses on fisheries and aquaculture studies (see also (Barbedo, 2022). The literature on detection models can be highly technical and inaccessible for an ecologically oriented audience, so we also summarise relevant terms and methods (Table S1). Simply speaking, object detection models aim to find the location of an object (also known as localisation) and to do it as fast as possible. There is usually a trade-off between accuracy and speed, with different models developed for specific purposes, e.g. high-speed models

for real-time applications, such as cameras on fishing gears that might determine fishing decisions, versus slower but higher accuracy models for post-sampling or catch image analyses.

In the context of image analysis, it is important to distinguish between **classification** and **object detection** models. Classification models assign an entire image to a category, such as fish species, but are not concerned with where in the image the fish is located. In contrast, object detection models both locate and identify individual objects within an image. **Segmentation** models are closely related to object detection but go a step further by classifying each pixel in the image as belonging to the object of interest or not (e.g., answering the question: "Is this pixel part of the fish?"). While classification models do not provide spatial information, they are often simpler, faster, and require less annotated data, as they are trained using a dataset containing only one label per image. Classification models are useful when the goal is, for example, to confirm species presence or when there is only one primary subject in the image. In contrast, detection or segmentation models require more complex annotations such as bounding boxes or outlines and class names for every object of interest. However, the combined localisation and identification capability makes object detection and segmentation essential for more complex scenarios such as images from fishing vessels or underwater surveys, where multiple fish of different species may appear in the same frame and need to be individually recognised (Zaidi et al., 2022).

Both classification and detection models start with the same 'backbone', an algorithm that transforms the original image pixels into feature maps — internal representations that capture patterns such as lines, edges, shapes, and textures in different parts of the image (see below and Table S1 for glossary). In convolutional neural networks (CNNs, Li et al., 2022), these maps are built through a mathematical operation called convolution, which scans the image with filters of increasing size. In Vision Transformers or ViTs (Dosovitskiy et al., 2021; Khan et al., 2022), the image is instead split into small patches, which are then processed in sequence to learn relationships between features across the entire image (see Table S1 for a glossary). The exact details of feature extraction are technical, but it is important to note that in deep learning models, the relevant features to be identified are not defined by the user but are learned by the model during the training stage. It is tempting to think that those features represent increasingly abstract and recognisable patterns in the image, moving from identification of lines and curves to more specific shapes (e.g. circles or tail-like shapes), to anatomical features as eye, head or fins. However, neural networks are largely black boxes, and increasingly complex features may not at all correspond to traits or patterns that humans consciously recognise as important. The extracted features are then passed onto a subsequent neural network or, in the case of the study presented here, a regression model, which finds optimal weights (parameters) for each specific feature to reduce the cost function (e.g. difference between the known fish species or size and model estimation). Some models may pass the same feature maps to several subsequent models (or 'heads'), where one model finds optimal weights for feature maps to produce accurate bounding boxes, while another may find different weighting to yield most accurate species classification.

Development of object classification models started in 1990s, initially focusing on recognition of handwritten digits and human faces (Lecun et al., 1998; Turk & Pentland, 1991). These models already used simple CNNs but used specific handwritten algorithms to produce classification results. A major breakthrough in computer vision came with the release of advanced image classification AlexNet model in 2012 (Krizhevsky et al., 2012). AlexNet combined the concepts of convolution and neural networks into a large-scale deep learning model trained on millions of images and introduced additional operations to improve the training process. AlexNet's ability to learn high-level features from millions of diverse images also led to the expansion of transfer learning applications. Transfer learning

is an approach where a model trained to recognise diverse classes (such as “car,” “fish,” or “tree”) in a large, generic dataset (e.g., ImageNet for object classification or COCO for object detection) is adapted to a more specific task, such as fish species detection or segmentation. By leveraging features learned from millions of images, transfer learning reduces the need for large, annotated datasets in the target domain (e.g. fish species) and often improves model performance and generalisation.

AlexNet’s success triggered a rapid switch to deep CNN for feature extraction in most vision tasks. Many subsequent object detection frameworks used AlexNet’s convolutional layers to extract and process feature maps and only added additional ‘heads’ to e.g. produce bounding-boxes and classification of objects within those bounding boxes. This was the start of modern **two-stage object detection** models, initiated by R-CNN in 2014 (Ren et al., 2015). Two-stage models first generate candidate regions where the object is likely to occur (proposals made from the entire image feature maps) and then analyse each of these regions separately, honing on feature maps of specific areas, to classify objects and refine bounding box coordinates. These models include R-CNN, Fast R-CNN and Faster R-CNN (Girshick, 2015; Girshick et al., 2014; Ren et al., 2015), each with increasing improvements in accuracy and processing speed. Two-stage models are generally slower, but have higher accuracy and performance, especially in detecting objects of difference sizes and distances to the camera. In contrast, **one-stage** models were developed in 2016 with the goal of improving computational speed. One-stage models were first developed by YOLO or ‘you only look once’ (Redmon et al., 2016) and later with single shot detector SSD (Liu et al., 2016) or Retina-Net models (T. Y. Lin et al., 2017). One-stage models process the entire image in one pass, setting bounding boxes and probabilities of object classes all in one head from full image feature maps. Progressive versions of YOLO (v2 to v12, released in February 2025) further improved model performance, speed and accuracy, although these models still struggle with small objects in noisy images (Murat & Kiran, 2025).

A major advancement in object classification and detection, and in deep learning applications in general, was achieved through the development of **transformers**, introduced by Google in 2017. Unlike CNNs that look for relationships in adjacent parts of an image (through sliding matrix operations), transformers analyse all input data at once, by finding most important relationships among input tokens (local visual features). This means that transformers employ a self-attention mechanism to evaluate the global context of an input, calculating dynamic attention scores that estimate the relevance of specific visual features or words based on their context in the entire image or text (Vaswani et al., 2017). Transformers were instrumental in the advancement of large language models (LLMs), where interpreting the meaning of a word requires the full context of a sentence. **Visual transformers (ViTs)** also skip the image convolution step (local feature extraction) and instead analyse features of the entire image at once (Dosovitskiy et al., 2021). Similarly to large language models, ViTs split the image into patches (like words in a sentence) and processes these patches like a sequence of tokens, looking for general relationships among them. While ViTs require larger training datasets, they have outperformed CNNs in all benchmarking image datasets (Dosovitskiy et al., 2021).

Finally, latest computer vision applications use **hybrid models**, combining both CNNs and ViTs (Wu et al., 2021). Here CNNs first process the image to produce localised feature maps, which are then passed onto ViTs to assess image-wide relationships. These models improve accuracy, but such improvements generally require increased computation cost and larger training datasets compared to CNN-only approaches (unless pretrained feature extractions can be used, as in the case study presented here).

While object detection algorithms have made great progress, their performance in real world fish detection is still relatively poor (see also section 2.3). Major challenges remain in correctly identifying fish in challenging underwater environments, with scattered or low light levels and complex habitats (in kelp forests, mangroves or reefs), where fish body parts might be obscured. Tracking fish individuals, or any other objects in videos also remains a challenge and a key area of ongoing research (Abangan et al., 2023; Zou et al., 2019), as traditional object detectors are generally focused on single images. Yet, fish tracking is essential for size estimation in underwater video-based monitoring (BRUVs, ROVs), where a fish is tracked until its orientation enables sufficiently accurate size measurement. To date, most BRUV analyses still use manual labour for fish tracking, which is time consuming and introduces a major bottleneck in image processing. Another challenge for fish specific applications relate to the low performance of models in detecting small objects, especially in crowded conditions.

2.2 Models for fish size estimation from images

Fish size estimation from images can be approached through different methods, which differ fundamentally in their use of training data, reliance on reference objects, and complexity of modelling. **Reference object-based methods** do not necessarily require training of models for size estimation. Instead, they use physical objects of known dimensions to calibrate the scale of the image. These objects can, for example, be rulers, checkerboards, conveyor belts (of known width), ArUco markers (square pieces of paper or other material with a black border and an internal pattern of white and black squares), colour plates or coins. By measuring the number of pixels corresponding to the reference object, the pixel length of the fish can be converted into real-world units. For accuracy, the fish and the reference object must be at approximately the same distance from the camera, and some calibrations may need to be done to account for the curvature of fish, camera distortions and light conditions.

Reference object-based approaches are conceptually simple and can provide reliable size estimates when objects of known dimensions are consistently present in the image. However, their application is constrained by several practical and methodological limitations. First, placing or ensuring the presence of a reference object in the field can be logistically challenging, particularly in natural underwater environments where conditions are dynamic, and access is limited. Second, the accuracy of size estimation depends heavily on the relative position and distance of the fish to the reference object, which is not always controllable in natural conditions. Third, reference objects may not always remain visible or within the same focal plane as the target organism, leading to errors in scaling and measurement. Finally, this approach is less suited for large-scale monitoring, autonomous data collection, or opportunistic imagery (e.g., from citizen science or archival footage), where reference objects are typically absent. These constraints limit the generalisability of reference object-based methods and underscore the need for more automated, reference-free approaches that can operate effectively across diverse sampling contexts

Similarly, methods using a **fixed and known distance between the camera and the object or 3D reconstruction from stereo images** (Marrable et al., 2022) also do not require ML for fish size estimation. They rely on explicit geometric relationships rather than learned features and in theory can provide accurate size estimates for all detected objects. Fixed distance approaches are often used in aquaculture or commercial fisheries settings, and they require fixed camera setups, which restrict their scalability in natural or underwater environments. In contrast, stereo images are more relevant in observations of natural ecosystems but require calibrations to account for specific camera parameters and are sensitive to lighting conditions (Garner et al., 2021; Tonachella et al., 2022; Voskakis et al.,

2021). In both cases, accurate fish detection is still needed to identify the individuals to be measured, and in videos, robust tracking algorithms are required to follow individuals until their orientation is suitable for measurement (Monkman et al., 2019).

A new emerging approach is **monocular depth estimation** ("monovision"), where deep learning models attempt to reconstruct the distance to objects using a single camera image. This technique has been explored in fields such as robotics and autonomous driving (Afshar et al., 2023) and in forest-specific applications (Jia et al., 2025) but, to our knowledge, it has not yet been applied to fish size estimation. Its accuracy depends on having detailed knowledge of camera parameters (e.g., focal length, sensor size, lens distortions) and is further complicated by underwater imaging conditions such as light scattering, turbidity, and refraction. While promising, monovision remains limited and largely unexplored in the context of fisheries and ecological monitoring.

Species-specific models can estimate size from measurable body proportions or landmarks, using known relationships between morphology and total length derived from sample data. Typically, they rely on regression or other classical inferential models (e.g. Álvarez-Ellacuría et al., 2020). Some are purely statistical, while others incorporate machine learning algorithms such as random forests or shallow neural networks trained on shape descriptors. These models can operate without reference objects, but usually require species-specific training data and calibration. Their main limitation lies in generalisability, as performance may decline when fish morphology varies significantly across age, sex, or environmental conditions.

Finally, it is important to note that the AI field is developing rapidly. While most applications to date use separate object detection or segmentation and size measurement methods, recent deep learning approaches, including CNNs and transformer-based architectures, can now **jointly detect fish and predict size** (i.e. "end-to-end AI" approach) directly from image features such as shape, texture, and proportions (Jareño et al., 2024). These models learn complex, non-linear relationships between image data and size measurements by training on large, annotated datasets. Unlike detection and classification models, "end-to-end AI" approaches do not require explicit equations (as in stereo images or statistical model) or predefined features (reference objects); instead, they automatically extract and encode relevant visual patterns for size estimation. This makes them especially useful when reference objects are unavailable or impractical, such as in recreational angler photos or underwater footage from ecological surveys (such as the case study presented here). However, their accuracy depends heavily on the quality, size and representativeness of training, and the models have not been yet widely deployed or tested in real life situations.

2.3 Overview of published studies using automated and ML tools for fish size estimation

This section provides a synthesis of the current literature using automated and ML methods for fish size estimation, highlighting the diversity of methods, common limitations, and gaps in consistency across studies. To compile relevant literature on automated and ML applications for fish size estimation, we conducted a systematic search using two major academic databases: Web of Science and Google Scholar. The search queries used were *((ALL=("machine learning") AND ALL=("fish size")) OR (ALL=("fish length") AND ALL=("machine learning")))* for Web of Science and *"machine learning" AND "fish size" OR "machine learning" AND "fish length"* for Google Scholar. The search was performed in April 2025. The screening and data extraction were conducted by CNS Silva and F Heather. While our search focused on the specific application of "machine learning" to ensure a review of contemporary data-driven architectures, we acknowledge that the broader field of digital

photogrammetry has a longer history of fish size estimation using classical image processing techniques.

The Web of Science search yielded 19 results, while Google Scholar returned 1050 results. Due to the broad scope of Google Scholar, many of the retrieved publications were unrelated, with results screened and selected to only include studies that explicitly applied automated techniques to estimate fish sizes from images. A total of 39 relevant publications, including scientific papers, books, reports, a thesis and conference proceedings, were identified as relevant and included in this review (Figure S1, Table S2).

There has been a clear and consistent rise in the number of studies applying automated and ML techniques for fish size estimation over the past several years (Figure 1), with most research concentrated in fisheries and aquaculture and relatively limited application in broader ecological contexts, such as monitoring and conservation. Out of 39 reviewed studies, 24 were done for above-water, 14 for underwater environments, and one study was conducted in both above- and underwater environments (Yu et al., 2023). The number of images used for training varied from a few hundred to tens of thousands, with the median value of approximately 1100 images, with studies from aquaculture generally using smaller training datasets (for the full distribution of range of species and images used see Figure S2). As highlighted above, ML based studies for fish size identification generally first employed an object detection model and then used different methods to estimate length of these objects. In many cases, published studies focused on the performance of the object detection model, with less emphasis on the accuracy of the size estimation method itself. Below we briefly review the main object detection techniques used and then concentrate on the size estimation part.

Object detection

Deep learning-based object detection has become the most widely used strategy for automating fish recognition in images, though the choice of architecture has evolved over time. Among the reviewed studies, earlier ones (e.g. Álvarez-Ellacuría et al., 2020; Monkman et al., 2019) predominantly used two-stage models (e.g. R-CNN, Faster R-CNN, Mask R-CNN), while one-stage models (e.g. YOLO) gained more popularity in recent years (e.g. Jansi Rani et al., 2024; Karoline & Nogueira, 2024). Studies employed a variety of deep learning architectures such as Mask R-CNN, YOLO, ResNet, and EfficientNet models, with many leveraging transfer learning (e.g. (Marrable et al., 2023; Monkman et al., 2019).

Many reviewed studies used data augmentation to train the detection models, where images were rotated, flipped or transformed in other ways to create ‘new’ images, increasing the size of the training data to enhance model performance (e.g. Fernandes et al., 2020; Shibata et al., 2024). Notably only 17 out of 39 (or 44%) of studies explicitly reported the number of images used in different stages of model development. Performance of object detection models was reported using a range of metrics, including accuracy, Intersection over Union (IoU, percent overlap between predicted and true bounding box, where 1.0 = perfect overlap and 0.0 = no overlap), precision (percentage of correctly identified objects among all identified objects), recall (percentage of all real objects identified by the model), and F1 score (combination of precision (P) and recall (R); 2*P*R/(P+R)). Top models achieved accuracies above 90% and IoU scores often exceeding 80% (e.g. Garcia et al., 2020; Rocha et al., 2024; Yu et al., 2021), which highlights the effectiveness of deep learning for automatic fish detection in images, while also reflecting a potential publication bias, as studies reporting higher-performing models are more likely to be published. However, as accuracy can be misleading in imbalanced datasets, some studies also reported precision and recall (n=8 and n=9, respectively), which provide

a more informative assessment of detection performance. Where available, these metrics were generally high (>90%), supporting the conclusion that deep learning approaches are effective for fish detection, although inconsistencies in reporting across studies limit direct comparison.

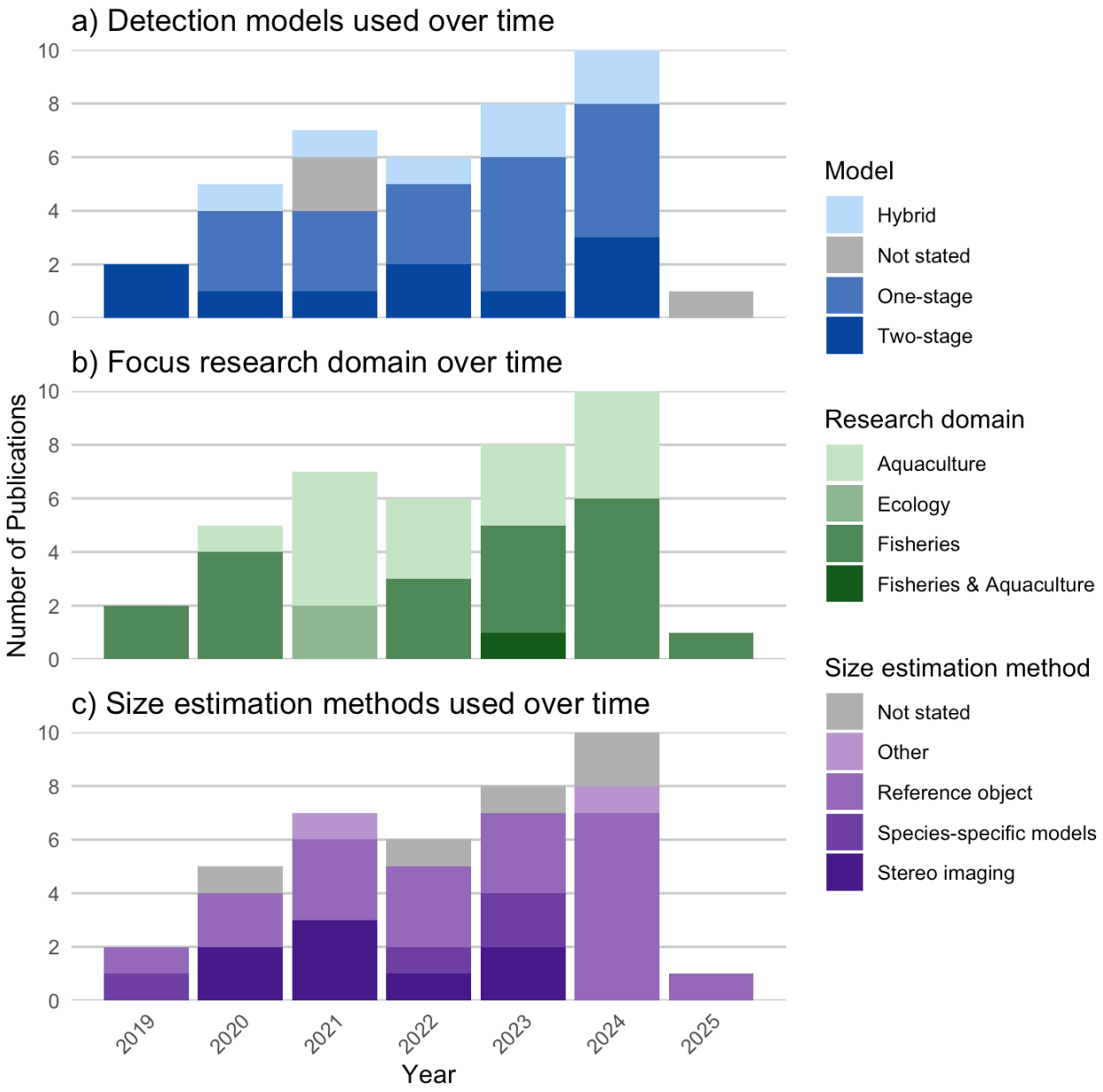

Figure 1: Temporal trends in publications by (a) detection model architectures used, (b) research disciplines of application and (c) size estimation methods used as mentioned in each publication. *Computer vision (non-ML)*: Use traditional, non-learning-based image processing techniques (e.g. Grab Cut, morphological operations); *Hybrid models*: Combine different architectures or stages (e.g., two-stage detector + one-stage detector); *One-stage models*: Perform object localisation and classification in a single step (e.g. YOLOv3–v9, CenterNet); *Two-stage models*: Detect object regions first, then classify them (e.g. R-CNN, Faster R-CNN, Mask R-CNN, Haar classifier). *Other*: Includes approaches without the need for reference objects such as end-to-end automated systems (e.g. ViTs). *Reference object*: use objects of known size placed in the scene to calibrate image scale. *Species-specific models*: Applies empirical equations linking measurable features (e.g., body depth) to total length for a given species (e.g. length estimation from body ratio equations). *Stereo imaging:* Uses

dual-camera setups to estimate 3D structure and fish size from disparity. Note that literature search was conducted in April 2025 so this year is not complete.

Size estimation approaches and performance

Among reviewed studies, reference object-based methods were the most used for fish size estimation (51.3%), followed by stereo imaging (20.5%), while species-specific models (10.3%) and other methods such as end-to-end automated systems using ViTs (5.1%) were rarer. Notably, 12.8% of studies did not explicitly mention the methods used for size estimation. Some studies were difficult to classify within our defined categories. For example, Álvarez-Ellacuría et al. (2020) developed an Unsupervised Deep Morphological Open-source System (UDMOS) for estimating fish length in images without the need for reference objects. The pipeline combined an initial object detection stage with post-processing steps based on brightness, pixel ratios, and histogram-based distance estimation to calculate fish size. This approach integrated detection and measurement in a single workflow but remained a modular pipeline rather than an end-to-end learning model, as the size estimation relied on rule-based heuristics rather than direct prediction from image features.

Reviewed studies reported several metrics for the performance of fish size estimation including absolute errors, relative errors, a correlation coefficient and standard deviation expressed as mean or median standard deviation (Table 1). The most reported error metrics were mean percent error or MPE (n=9), with values ranging from 0.37% to 6.69%, and mean absolute error or MAE (n=9), with values ranging from 0.22 cm to 5.36 cm and R^2 (n=5), with values ranging from 0.998 to 0.7. Surprisingly, around 20% of studies (n=8) did not report any error metric for size estimation. On average, studies reported only one error metric and only one study reported three different metrics (Nguyen et al., 2024). This lack of standardised performance metrics limited our ability to compare fish size estimation accuracy across studies.

Table 1: Summary of commonly reported performance metrics for fish size estimation in the reviewed studies.

Name	Acronym	Description
Mean absolute error	MAE	Average magnitude of errors, regardless of direction
Mean absolute error deviation	MAED	Variability of absolute errors across samples; if absolute errors for all fish size estimates were identical, MAED would be 0
Mean absolute percentage error	MAPE	Average of absolute relative errors expressed as a percentage; scale-independent accuracy metric (MAPE = MARE × 100)
Mean absolute relative error	MARE	Average absolute prediction error relative to the true value; scale-independent metric of accuracy.
Mean percent error	MPE	Average error relative to true value expressed in percentage
Mean bias error	MBE	Average of errors; indicates whether predictions are systematically over- or under-estimated (positive = overestimation, negative = underestimation)
Mean relative error	MRE	Average error relative to true value

Mean squared error	MSE	Average of squared errors; penalizes larger errors more strongly
Pearson's correlation coefficient	*r*	Measures linear correlation between predictions and true values
R-squared (coefficient of determination)	R^2	Proportion of variance in true values explained by the model
Root mean squared error	RMSE	Square root of MSE
Square root of the mean squared deviation	RMSD	Average magnitude of prediction errors, calculated as the square root of the mean of squared differences between predicted and true values
Standard deviation of errors	SDE	Spread (variability) of prediction errors around the mean

Although no statistically significant difference was detected in mean percent error (MPE) between studies using reference objects and those using stereo imaging (Wilcoxon rank-sum test, $p = 0.23$), studies employing reference objects tended to show lower error values ($\bar{x} = 2.5\%$, SD = 2.3%) than those using stereo imaging ($\bar{x} = 5.3\%$, SD = 1.6%); however only 9 studies reported values of mean percent error.

The number of species considered across the reviewed studies varied widely, ranging from a single species to as many as 319 (see Figure S2 for more details on distribution of species numbers across studies). Marrable et al. (2023) trained a one-stage model (YOLOv5) on 319 species of fish and used underwater stereo images to automate length measurements. This generalised semi-automated model performed similarly to the accuracy of measurements taken by humans (Pearsons's correlation coefficient of 0.99) and was able to estimate sizes of numerous species with varying colour, texture and morphometrics. However, as we already noted in the overview of object detection models for underwater image analyses, the authors of this study also could not fully automate the matching of individual fish across the two stereo images, and to identify the optimal frame for fish measurement, i.e., where the fish is most perpendicular to the cameras.

Approximately half of reviewed studies (16 out of 39) focused on developing size estimation models for only one species. While species-specific models may achieve high accuracy for the target organism, their broader applicability is limited, as they often fail to generalise to species with different body shapes, morphologies, or colour patterns.

3. Case Study - A machine learning based image classification method to estimate fish sizes from monocular images without a specified reference object

In this study, we present an end-to-end automated pipeline designed to predict fish lengths from image-based visual embeddings (compact numerical representations of images that capture their essential visual features) using a visual transformer (ViT) and an automated machine learning (AutoML) regression model. We use a ViT DINOv2 (Oquab et al., 2023), which does not fit neatly into the conventional categories of one-stage, two-stage, or hybrid detection models mentioned above, since it is primarily a self-supervised vision transformer designed for general-purpose image feature extraction rather than object detection. On its own, DINOv2 generates embeddings that can be directly

used for classification tasks, which conceptually aligns more closely with a one-stage approach. However, when integrated into more complex architectures, DINOv2 can form a part of a two-stage or hybrid detection pipeline. Our pipeline, described below, consists of a structured sequence of tasks, beginning with feature extraction, followed by data postprocessing, model training, and evaluation. The following sections describe the methodology and the experimental results.

3.1. Methods

Dataset and pre-processing. Data consisted of 2865 images of fish (Table S3) collected from the Angler´s Atlas angler app. For all Angler´s Atlas data, fish images and size were collected through app-based fishing tournaments, conducted in Canada and USA during 2022-2023. For these tournaments anglers must provide the size of a fish and upload an image of that fish next to a ruler for fish length verification by the Angler's Atlas team. Anglers can also upload a "hero shot" which consist of anglers holding the fish, without having a ruler or another reference object in the background. This reporting design means that for each "hero shot" image used to train the model, there is a verified and accurate fish length information available. Not all tournaments incentivised catching the largest fish; depending on the rules some tournaments collected photos across a range of fish sizes. In this study we only used "hero shots" for ML based fish length estimation, whereas fish length information data was checked and verified by the Angler's Atlas team. As a result, all images used for the model training include a human holding a fish out of water, at various angles and distance to the camera. We selected images that had only one fish per image. The images were initially screened manually to remove low quality images (e.g. no fish in the image, obscured fish, image too blurred or with very poor lighting). The final dataset included 2865 images from 90 fish species, with fish body sizes ranging from 10 to 79.6 cm (Table S3).

The images were automatically pre-processed as a part of the pipeline through a series of transformations. As any deep-learning model, DINOv2 requires input images with specific characteristics. To meet these requirements, our images were first resized to 244 × 244 pixels, then centre-cropped to 224 × 224 pixels (keeping the central region and cropping the edges), and their pixel values were normalised with a mean and standard deviation of 0.5. The 224 × 224 input size is a widely adopted standard in computer vision, as it was originally established in the ImageNet benchmark and has since became the default resolution for many pre-trained models, including in DINOv2.

Visual feature extraction model. In this study we use DINOv2 Vision Transformer Giant, developed by Facebook Research (Oquab et al., 2023) and pre-trained on large-scale datasets to extract visual embeddings from the images. DINOv2 is comprehensively described by Oquab et al. (2023) with full details on the model design, training data, and optimisation procedure. The model was downloaded and applied to extract visual embeddings from fish images, and we did not apply any fine-tuning or architectural modifications to the model.

After obtaining the visual image embeddings (where each image was now represented by a vector of 1536 numbers), the pipeline proceeded to post-process the data (Figure 2). This step involved standardising both the visual features (embeddings) and the target values (fish lengths) for model training to enhance model stability and improve learning efficiency. The values in the visual embedding vector were standardised to zero mean and unit variance using z-score standardisation ($z = (x - \mu) / \sigma$, where x is the raw embedding vector and μ, σ are the mean and standard deviation computed over the training embeddings). These standardised embedding vectors were then used as input features to the regression model. Fish length values corresponding to each image were used for the cost

function to train the model, and these values were also normalised to be between zero and one (see link to the model code for further details). Once scaled, the dataset was divided into training and test sets, with 75% of the data allocated for training (2148 images) and the remaining 25% reserved for evaluation (717 images). This last step of dividing the data was done 10 times, each with a random split between the training and test datasets and subsequent regression step, repeated for each of the random image splitting replications (Figure 2).

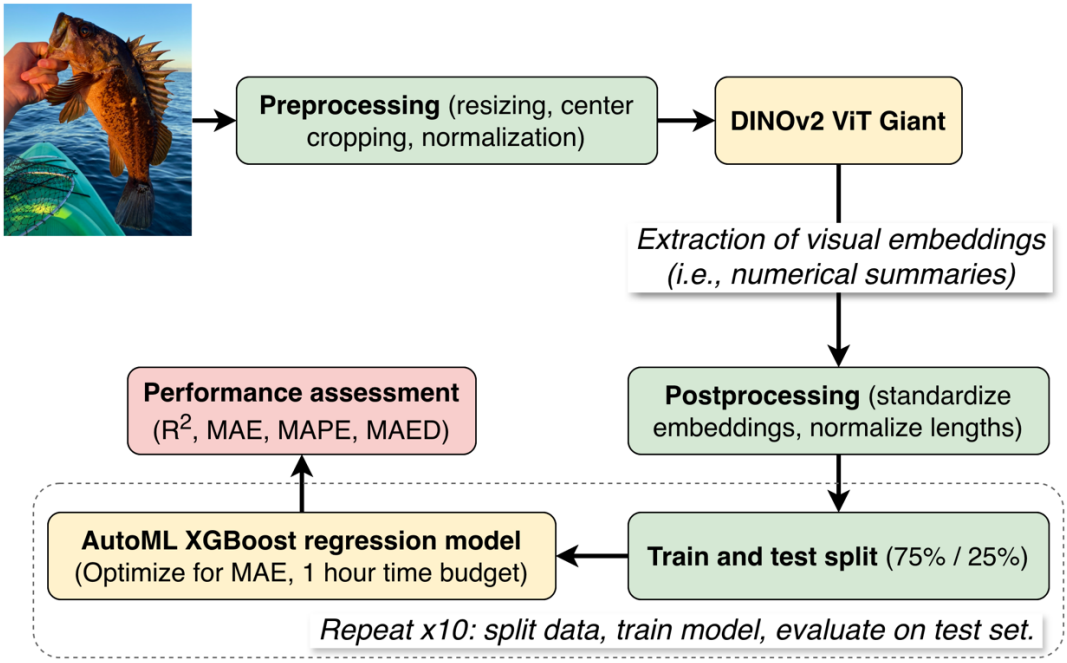

Figure 2. Workflow of the case study pipeline for fish size estimation. The process includes image preprocessing, feature extraction using DINOv2 Giant, post-processing, and dataset partitioning into training and test sets. AutoML with XGBoost was then applied to build predictive models, and performance was assessed using standard evaluation metrics.

Size regression model: Extracted embeddings were used as inputs for the downstream prediction task, focused on training a machine learning model to predict fish lengths from their corresponding embeddings. For this we employed an AutoML approach, which automates the end-to-end workflow of building predictive models. AutoML searches over a set of algorithms and/or their hyperparameters to discover the configuration that optimises a user-specified performance metric. We constrained the pipeline to use XGBoost (Extreme Gradient Boosting), a gradient-boosted decision tree algorithm, to model the relationship between image embeddings and fish size. XGBoost is an efficient implementation of gradient-boosted decision trees and has consistently shown top-performance for tabular data, which matches the nature of our dataset after feature extraction (Chen & Guestrin, 2016). Early-stopping (see definition in Table S1) and sub-sampling allow the process to abandon underperforming models quickly, squeezing more experiments into a fixed time. After that time, AutoML returns the model with the best performance. The training was completed on a single GPU, with the training process taking only 1.5 hours per run, performed on a workstation equipped with an NVIDIA GeForce RTX 4060 GPU.

In our pipeline, AutoML was configured to use mean absolute error (MAE) as the primary optimisation metric, ensuring that the model is trained to minimise absolute deviations between predicted and true fish lengths. We allowed the AutoML system to train the regression models for up to one hour. During

this time, it automatically tested different model hyperparameters (settings such as learning rate, depth of the network, or batch size, see Table S1 of terms) using an internal hold-out validation split which selected the combination that produced the lowest error on the internal validation set.

Performance evaluation metrics: Following model training, the pipeline moved to the evaluation phase, where the performance of the trained model was tested on unseen data. The model was loaded, and predictions generated for the test set. Since the target values were previously scaled during preprocessing, an inverse transformation was applied to restore them to their original range. The evaluation process measured the model's performance using four key metrics: R^2 score, mean absolute error (MAE), mean absolute percentage error (MAPE), and mean absolute error deviation (MAED). These metrics collectively provided a comprehensive assessment of the model's predictive accuracy.

Hold-out validation: To systematically assess the model's performance, the pipeline executed the training and evaluation process ten times, each as a fully independent run. In each run, the dataset was randomly divided into training and test sets using a standard hold-out split, a new model was trained from scratch using only the training set, and then evaluated on the held-out test set. This train-test step was repeated 10 times. No model weights or information were carried over between runs, ensuring no data leakage from the test set into the training process. By averaging results over these ten independent runs, we reduced the impact of randomness in both data partitioning and model training, yielding a more reliable estimate of overall performance (Figure 2).

3.2. Results

On average there were 32 images per species on the final dataset used for model training, ranging from 1 to 694 images per species (Table S3, Figure S3). Across all species, the overall mean fish length was 41.78 cm (SD = 20.57 cm), illustrating that the dataset included a broad range of body sizes (Figures 3, S3).

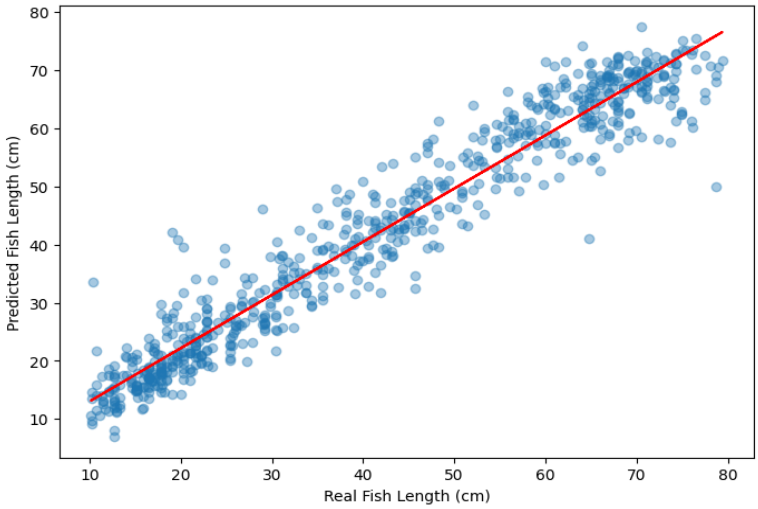

Figure 3. Linear regression (R^2 = 0.927) between real and predicted fish lengths (cm) by the fish size regression model.

The results of the fish length prediction pipeline demonstrated a strong overall performance, with an R^2 score of 0.927, indicating that the model successfully explained 92.7% of the variance in fish length across fish sizes ranging from 10 to 79.6 cm (Figure 3). MAE was at 3.98 cm, indicating that, on

average, the model's estimates deviated by about 3.98 cm from the actual fish lengths. This corresponded to an overall MAPE of 11.84% and this error metric tended to decrease with fish size, where smallest fish (10-20 cm long) had a mean MAPE of 20% whereas length of fishes above 40 cm was estimated with MAPE of 6-9% (Figure 4). Finally, the MAED was around 3.86 cm, which suggested that the errors of length estimates were not highly variable across the dataset. Overall, these findings indicated that the deep learning-derived features captured essential patterns that correlate with fish size, allowing for accurate regression predictions, as seen in Figures 3 and Figure S5.

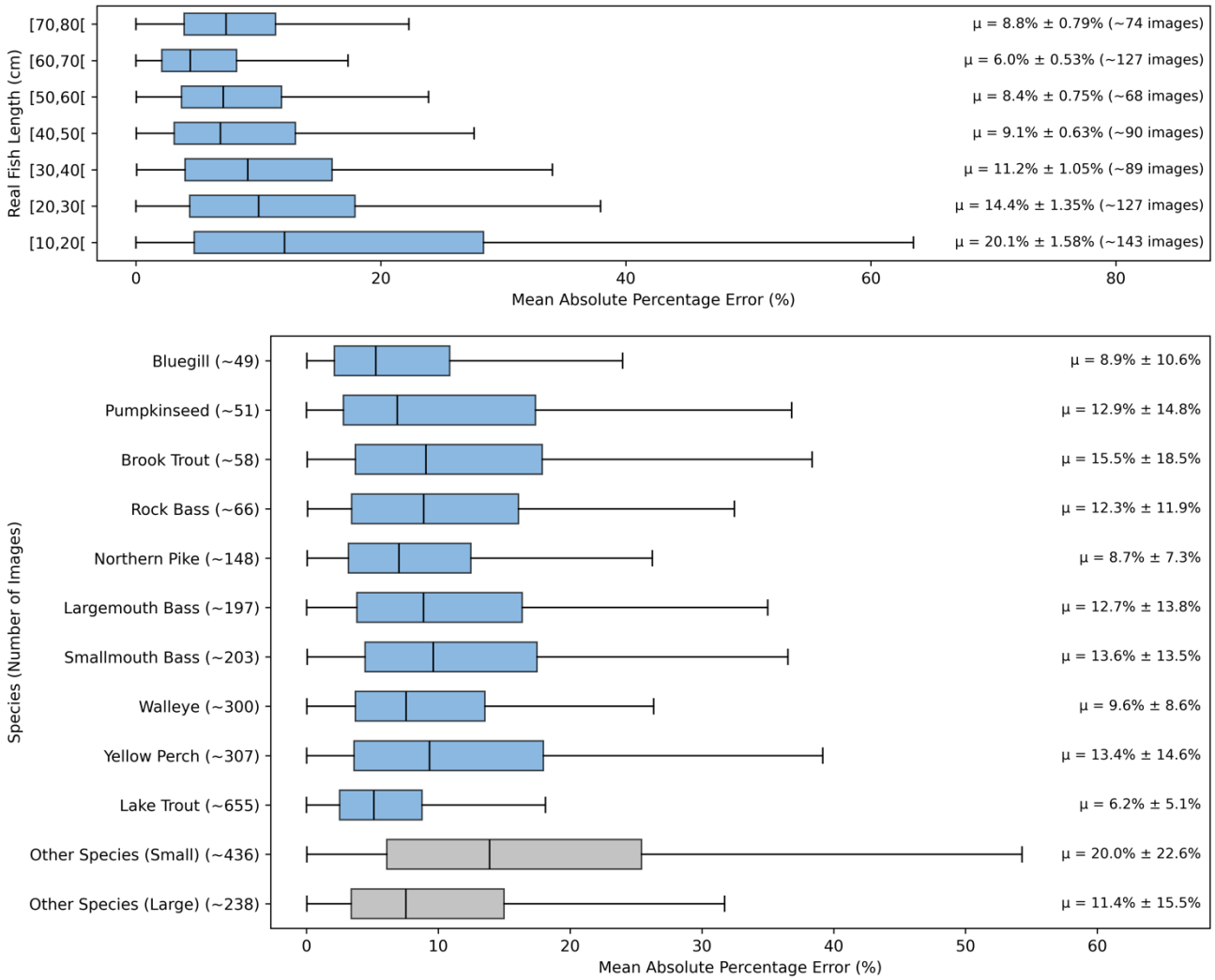

Figure 4. Mean absolute percentage error (MAPE) for each of the 10 cm fish length groups, are shown in the top plot. The bottom plot shows MAPE for the most abundant species and for the rare species classified into small and large species (based on their median length). The boxplots show median estimated across 10 runs from the test dataset, while the μ values shown on the right are the means with full variability across the 10 estimates shown with +/- values. The edges of the boxes are the 25th and 75th percentiles (i.e., the interquartile range), and the whiskers extend to the minimum and maximum observed values.

Images with the smallest MAPE generally corresponded to species with a high number of training images, such as lake trout (n = 694) and included images with well framed and cropped fish individuals (Figures 4 and 5). By contrast, images with the largest MAPE often contained substantial visual noise, such as the full bodies of anglers or distracting backgrounds. These patterns suggest that both training sample size and image composition played roles in determining model accuracy for size estimation.

Figure 5. Examples of five images with the lowest (top row, most accurate predictions) and highest (bottom row, least accurate predictions) mean absolute percentage error (MAPE) in size estimation.

3.3. Discussion on the case study findings

Estimating fish length from monocular images without a reference object remains a long-standing challenge in fisheries monitoring because most established approaches rely on rulers or stereo-camera systems to convert pixel dimensions into real-world sizes (Monkman et al., 2019; Shibata et al., 2024). Our study demonstrates that accurate size estimation is possible even in the absence of explicit references by leveraging visual transformers (ViTs) for feature extraction and a relatively simple, non-neural network model (AutoML) for regression analysis. The pipeline presented here achieved strong predictive performance across a wide fish size range (10-80 cm), despite a relatively small dataset of 2865 images and the complexity of images ("hero shots") that often included diverse backgrounds, angler poses, and variable fish orientations. Estimates were particularly reliable for individuals larger than 40 cm, with errors in the range of 6-9%, consistent across different fish sizes (Figures 3, 4). Compared to traditional morphometric methods or earlier CNN-based pipelines, this end-to-end approach reduces the need for manual calibration or controlled imaging conditions, highlighting the utility of ViTs and automated modelling frameworks for scalable, real-world ecological applications. Notably, the performance of our model was achieved using very limited resources. The dataset used for training was relatively small and the architecture of the model was compact, making use of a pretrained ViT to extract digital features from images and a widely used regression modelling tool to convert extracted features into size estimation. This model architecture meant that the training could be completed using low computing resources. Assuming the model performance can be maintained with new datasets (see below), the pipeline presented here could therefore be used even by small research or management teams across the world to develop more specific fish size estimation applications.

Comparing our results with those from recent studies, our model performance appears competitive, especially given the absence of explicit reference objects (Table 2). The choice of error metric should be context-dependent (Chai & Draxler, 2014), as different metrics capture different aspects of model

performance, and their suitability varies based on factors such as the size distribution of the target species and the intended application (e.g., regulatory enforcement, stock assessment, or citizen science data collection). For instance, MAE is commonly used in ML for overall absolute model performance but may be less meaningful when comparing across fishes of varying sizes, given that 1cm error in 10cm fish and 100cm fish is quite different. Percentage or relative error metrics such as MAPE are more suitable for these purposes, because they scale errors relative to true fish size, but they can become unreliable when the true values are very small and the division by small denominators disproportionately inflates the error. The most encouraging aspect is that in our case study relative errors were rather consistent (ca 10%) across size groups, suggesting this limitation did not strongly affect our results.

Table 2: Comparison of datasets (number of images used for training, number of species included), approach used (model type, size estimation method) and reported performance metrics in fish length estimation across reviewed studies that reported mean absolute errors (MAE) and our case study.

Study	**Dataset (n, species)**	**Approach**	**Performance (MAE)**
This study	2865 images, 90 spp.	ViT (DINOv2) embeddings + XGBoost regression; no reference object	3.98 cm
Monkman et al., 2019	734 images, 1 spp	Two-stage; reference object	2.20 cm
Bravata et al., 2020	623 images, 22 spp	One-stage; stereo imaging	3.22 cm
Tseng et al., 2020	5000 images, 8 spp	One-stage; reference object	5.36 cm
Ovalle et al., 2022	4782 images, 14 spp	Two-stage; reference object	0.92 cm
Mots'oehli et al., 2024	40000 images, 163 spp	One-stage; reference object	2.30 cm
Jansi Rani et al., 2024	3726 images, 1 spp	One-stage; not stated	0.22 cm
Gao et al., 2024	1000 images, 2 spp	One-stage; not stated	0.80 cm
Climent-Perez et al., 2024	1185 images, 59 spp	Hybrid; reference object	1.27 cm
Nogueira, 2024	362 images, 4 spp	One-stage; reference object	1.44 cm
Cao et al., 2024	3080 images, 1 spp	One-stage; reference object	0.33 cm

To our knowledge this is the first publicly available model aiming to estimate fish sizes from monocular (non-stereo) images without a specified reference object across a wide range of fish species. One somewhat similar application is presented by Jareño et al., (2024), who also used an automated approach to classify fish sizes from images. However their study was based on a limited number of species and images were collected in a controlled fish market setting, where fish were photographed in standardised boxes (one box per image). While they used 8505 images to train a species classification model for 19 different species, the size classification model was trained and applied only for one species *Pagrus pagrus* and size determination only focused on classifying fish into five distinct size groups, rather than an actual size estimation in centimetres. Other models have or are being developed in commercial aquaculture settings and in the commercial or recreational fisheries space, but with a proprietary restriction. However such models have limited application scope, and their performance cannot be properly compared.

Lowest model performance was observed in smallest fish individuals and there are several potential reasons for this. First, some photos with small fish included mostly background information with fish only taking a small fraction of the image, potentially providing insufficient information for the model. Similar challenges have been documented in broader computer vision literature, where small objects

are harder to detect as they provide less contextual information and are more easily overwhelmed by background noise (Cheng et al., 2023; Hua & Chen, 2025; Nikouei et al., 2025, Zou et al., 2019). The challenge of extracting relevant information from small objects might have been amplified by our choice of a pretrained vision transformer DINOv2 to extract image features (embeddings). While ViT include a powerful attention mechanism to find relationships across different parts of the image, they encode images in fixed-resolution patches and if small fish occupy only a few patches, this limits the morphological detail that is captured. Vision Transformers have been shown to struggle with encoding fine-grained local textures and edges due to their patch-based representation (Azad et al., 2023). Although this limitation has primarily been documented in medical imaging, the same challenge may also affect fish size estimation. This limitation is not unique to ViTs however, as CNNs also face difficulties detecting and accurately representing small objects in cluttered or noisy images, because local features can be lost or suppressed during successive pooling and convolution operations (Muksit et al., 2022).

A second reason could be the fact that images of small fish were mostly dominated by diverse species, while larger fish sizes included popular angling species with few species and more images per species (Figures S3, S4). This combination of high species diversity but limited per-species representation may have also limited the model's ability to learn consistent features for small fish, thereby reducing prediction accuracy. The uneven representation of images across species is both a strength and limitation for training our model. On one hand, a diverse range of species likely have increased a probability that the model is indeed learning to recognise individual fish length, rather than other features of the image (e.g. learn species identity and estimate the size based on that). On the other hand, species with limited image data may have yielded less reliable size predictions. In addition, a certain degree of measurement error is inherent to all size data, as it depends on the smallest increment of the measuring instrument (e.g., 1/4 inch). The same absolute error thus represents a much larger relative error for a small fish (e.g., 5 cm length) than for a large fish (e.g., 70 cm length), which may further contribute to the higher variability observed among smaller individuals.

Although our pipeline appears to have reasonable overall performance, a central question is how does the model actually estimate fish length from monocular images without a reference object? Since DINOv2 embeddings capture holistic visual features, the model is unlikely to measure length in a geometric sense. Instead, it probably relies on combinations of cues: the relative shape and proportions of the fish body, the scaling of textures (e.g., fin-to-body ratio, eye size), and potentially even contextual information such as the angler's hands or arm span. Because hands are often in consistent positions when holding fish in our dataset, they could act as implicit size references, as "proxy landmarks," even though this was not explicitly modelled. This possibility highlights both the promise and the ambiguity of using deep learning and generic vision transformers as they may exploit contextual cues in ways not transparent to researchers. Deep learning models for wildlife detection have been shown to use background or human-related cues (such as traps, vegetation type, or camera placement) rather than the animal itself (Beery et al., 2018), and the 'black-box' aspect of deep learning models remains a challenge in our understanding. Further, we also found that image background and fish positioning influenced model performance, with uniform backgrounds and horizontally oriented fish generally producing lower errors. This finding echoes broader work in computer vision emphasising the importance of image composition and signal-to-noise ratio in model performance (Beery et al., 2018). Future work should explore the features that the model uses for training by applying tools such as attention heatmaps, Grad-CAM adaptations for ViTs, or attribution maps to highlight which regions of the image contribute most to the prediction (Playout et al., 2022). Such approaches could confirm whether the model is primarily "looking at" the fish body or drawing

information from surrounding human features, which has important implications for model transferability to other contexts (e.g., underwater images where no humans are present).

Overall, this case study offers a promising tool for automatic fish size measurements without a reference object. Our approach was deliberately simplified to demonstrate a reproducible and computationally accessible workflow, with the aim of lowering barriers to adoption while highlighting opportunities for further methodological development. However, before the model can be applied in real world applications it is essential to further test its performance. A critical test will be to assess whether fish sizes can be estimated from images without humans in the background. Given the challenge of the problem, this study originally aimed to ease the task by specifically using only “hero shot” photos, i.e. photos of humans holding a fish. The next stage is to see whether the model can be adapted to other types of images. Yet, even if the model only works well in these specific settings (humans holding a fish) it would still be useful for automatic analyses of many photos shared and contributed by anglers. Should the model’s performance be maintained across more diverse types of photos, it would provide an exciting opportunity to develop a more generic model for fish size estimation. Either way, future improvements should be made by training the model further with more examples of small-bodied species images in varied environmental conditions. Refining and automating preprocessing steps to reduce background noise will also help the model focus on relevant features. From a modelling perspective, integrating species-specific priors (e.g. expected body shapes or growth trajectories) could guide predictions, particularly in ambiguous cases. Such auxiliary metadata incorporated into model training, like fish species, habitat, or measurement conditions could provide additional contextual cues to enhance prediction accuracy. Finally, approaches that combine multiple architectural paradigms - for example, integrating CNNs with ViTs - may yield richer feature representations, enhancing robustness across variable image conditions (Haruna et al., 2025). With continued improvements in training data, model architecture, and real-world validation, publicly available and relatively generic fish size estimation models could significantly enhance data collection in fisheries science, aquaculture and management.

4. Conclusions and future directions

The application of machine learning to automated fish length estimation has expanded rapidly, but the field remains fragmented and lacks consistent methodological and reporting standards. Our literature analyses showed that studies vary widely in how they describe model architectures, training sample sizes, and performance metrics, which complicates cross-study comparisons and hinders reproducibility, potentially because this is a relatively new and rapidly evolving field and methodology is not yet rigorously developed. We recommend that it should be mandatory to provide clear descriptions of model types and architectures, training sample sizes, and preprocessing procedure. In addition, studies should transparently describe the size estimation strategy used (e.g., reference object, stereo vision or monocular estimation) and the environmental or contextual conditions under which images were collected. Without such transparency, it is difficult to evaluate the strengths and limitations of different approaches. Finally, the choice of fish size performance metric should be context-dependent and studies should report a comprehensive set of evaluation metrics, as different metrics capture different aspects of performance.

A second key recommendation is the adoption of benchmark datasets that would allow fairer model comparison and reproducibility. Progress in computer vision has been fuelled by the availability of large-scale, well-curated, open datasets such as ImageNet (Deng et al., 2009) and COCO (Lin et al., 2014). By contrast, in fisheries science, datasets remain relatively scarce, fragmented, domain-specific and often not publicly shared. Existing resources such as Fish4Knowledge (Boom, 2016),

DeepFish (Saleh et al., 2020) and OzFish (AIMS et al, 2019) represent valuable progress, but most are restricted to either underwater imagery or fish in trays, and only a subset provide annotated fish body size information. Developing an open, standardised dataset of fish images paired with accurate size data, covering a wide range of species, habitats, and imaging conditions, is crucial for building models with broad generalisability, cross-study evaluation, community involvement and faster innovation.

Our case study illustrates the potential for image-based estimation of fish lengths without the use of explicit reference objects. Traditional methods typically rely on rulers, stereo cameras, or laser scales to provide geometric reference (e.g. Bravata et al., 2020; Fernandes et al., 2020). While these approaches are often highly accurate, they can be impractical in citizen science or recreational fisheries, where standardised calibration tools are absent. By contrast, our end-to-end approach achieved strong predictive performance across a broad fish size range, despite noisy, heterogeneous "hero shot" imagery. These results suggest that transformer-based and hybrid (CNNs and transformers) methods may offer a complementary route for fisheries monitoring, particularly in contexts where manual calibration is not feasible. However, further work and broader collaboration is required to ensure the models are advanced to a high standard and are available to the community. As we already mentioned, several commercial systems for automated fish size estimation already exist (such as e.g. UMITRON LENS for aquaculture and I-Ocean's AI Fish Size Estimation Camera), but these models remain proprietary and largely inaccessible for research purposes. Open-source alternatives, such as NOAA's VIAME toolkit or the model presented here are urgently needed to enable collaboration, wide adoption and further development, standardisation, and open benchmarking efforts. Greater emphasis on model interpretability will be important for adoption in management contexts. Tools such as attention maps and gradient-based attribution methods can help reveal which parts of the image the model uses for prediction, clarifying whether features such as hands or fish outlines serve as implicit reference cues (Cao et al., 2024; Playout et al., 2022).

Finally, models must be scalable and user-friendly. Real-time applications for citizen science, aquaculture monitoring, or stock assessments will require lightweight architectures and straightforward deployment pipelines. Collaboration with fishers, managers, and regulators will also be critical for ensuring ethical and effective implementation (Probst, 2020). AI tools now make it relatively easy to convert advanced code into user-friendly applications, and we encourage all model developers to take this step. Even if some models cannot be publicly hosted due to computing resource limitations, user-friendly applications can be hosted on public repositories and used locally by research and management teams.

In conclusion, automated fish length estimation using machine learning represents a rapidly maturing but still fragmented field. Our study demonstrates that end-to-end pipelines using transformer-based embeddings and AutoML regression can achieve strong predictive performance even under noisy, reference-free conditions. To build on these advances, the field must prioritise standardised reporting, multi-metric evaluation, open benchmark datasets, publicly available and user-friendly models and explainable approaches. With these foundations in place, automated fish length estimation has the potential to become a scalable, reliable tool for fisheries monitoring, aquaculture, and citizen science, supporting more sustainable management of aquatic resources.

Acknowledgements

This work was supported by national funds through FCT - Fundação para a Ciência e Tecnologia, I.P., in the framework of the Project UID/04004/2025 - Centre for Functional Ecology - Science for the

People & the Planet with DOI identifier 10.54499/UID/04004/2025 and within the scope of the research unit UID/00326 - Centre for Informatics and Systems of the University of Coimbra. It was also supported by the National Network for Advanced Computing FCCN - FCT – RNCA (project number 2022.38089.CPCA.A0). CNSS was supported by FCT - Fundação para a Ciência e Tecnologia, I.P., in the framework of the project 2022.01002.CEECIND/CP1714/CT0017. RCP was partially supported by a contract from the Horizon Europe project REDUCE (Grant Agreement number 101135583). AA and FH were funded by the Australian Research Council Discovery Project DP220102446. Additional funding was provided through the grant to AA from European Regional Development Fund (project No 01.2.2-LMT-K-718-02-0006) under grant agreement with the Research Council of Lithuania (LMTLT) and Pew Fellowship in Marine Conservation to AA. This work was also funded by European funds through the European Regional Development Fund (FEDER), under the Centro 2030 Programme, project “MARCentro+ Inovação e Sustentabilidade na Gestão dos Recursos Marinhos da Região Centro” (CENTRO2030-FEDER- 02614400).

Data Availability Statement

The model’s source code is available on https://github.com/fishsizeproject/fish-length-predictor. Due to privacy reason for angler generated data, images used for model training could not be made publicly available and should requested from Angler’s Atlas by contacting SS.

References

Abangan, A. S., Kopp, D., & Faillettaz, R. (2023). Artificial intelligence for fish behavior recognition may unlock fishing gear selectivity. In *Frontiers in Marine Science* (Vol. 10). Frontiers Media S.A. https://doi.org/10.3389/fmars.2023.1010761

Abinaya, N. S., Susan, D., & Sidharthan, R. K. (2022). Deep learning-based segmental analysis of fish for biomass estimation in an occulted environment. *Computers and Electronics in Agriculture*, *197*(November 2021), 106985. https://doi.org/10.1016/j.compag.2022.106985

Afshar, M. F., Shirmohammadi, Z., Ghahramani, S. A. A. G., Noorparvar, A., & Hemmatyar, A. M. A. (2023). An Efficient Approach to Monocular Depth Estimation for Autonomous Vehicle Perception Systems. *Sustainability (Switzerland)*, *15*(11). https://doi.org/10.3390/su15118897

Álvarez-Ellacuría, A., Palmer, M., Catalán, I. A., & Lisani, J. L. (2020). Image-based, unsupervised estimation of fish size from commercial landings using deep learning. *ICES Journal of Marine Science*, *77*(4), 1330–1339. https://doi.org/10.1093/icesjms/fsz216

Australian Institute of Marine Science (AIMS), U. of W. A. (UWA), C. U. (2019). *OzFish Dataset - Machine learning dataset for Baited Remote Underwater Video Stations*.

Azad, R., Kazerouni, A., Azad, B., Khodapanah Aghdam, E., Velichko, Y., Bagci, U., & Merhof, D. (2023). Laplacian-Former: Overcoming the Limitations of Vision Transformers in Local Texture Detection. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *14222 LNCS*, 736–746. https://doi.org/10.1007/978-3-031-43898-1_70

Barbedo, J. G. A. (2022). A Review on the Use of Computer Vision and Artificial Intelligence for Fish Recognition, Monitoring, and Management. In *Fishes* (Vol. 7, Issue 6). MDPI. https://doi.org/10.3390/fishes7060335

Beery, S., Horn, G. Van, & Caltech, P. P. (2018). Recognition in Terra Incognita. *Proceedings of the European Conference on Computer Vision (ECCV)*, 456–473. https://beerys.github.io/CaltechCameraTraps/

902 Boom, B. J. (2016). Fish4Knowledge Database Structure, Creating and Sharing Scientific Data.
903 In R. B. Fisher, Y.-H. Chen-Burger, D. Giordano, L. Hardman, & F.-P. Lin (Eds.),
904 *Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data* (pp. 73–
905 82). Springer International Publishing. https://doi.org/10.1007/978-3-319-30208-9_7
906 Bravata, N., Kelly, D., Eickholt, J., Bryan, J., Miehls, S., & Zielinski, D. (2020). Applications of
907 deep convolutional neural networks to predict length, circumference, and weight from
908 mostly dewatered images of fish. *Ecology and Evolution*, *10*(17), 9313–9325.
909 https://doi.org/10.1002/ece3.6618
910 Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M., & West, G. B. (2004). Toward a metabolic
911 theory of ecology. *Ecology*, *85*(7), 1771–1789. https://doi.org/https://doi.org/10.1890/03-
912 9000
913 Cao, D., Guo, C., Shi, M., Liu, Y., Fang, Y., Yang, H., Cheng, Y., Zhang, W., Wang, Y., Li, Y., & Xia,
914 X. Q. (2024). A method for custom measurement of fish dimensions using the improved
915 YOLOv5-keypoint framework with multi-attention mechanisms. *Water Biology and*
916 *Security*, *3*(4). https://doi.org/10.1016/j.watbs.2024.100293
917 Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?
918 -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*,
919 *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014
920 Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the*
921 *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-*
922 *August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785
923 Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., & Han, J. (2023). Towards Large-Scale
924 Small Object Detection: Survey and Benchmarks. *IEEE Transactions on Pattern Analysis*
925 *and Machine Intelligence*, *45*(11), 13467–13488.
926 https://doi.org/10.1109/TPAMI.2023.3290594
927 Climent-Perez, P., Galán-Cuenca, A., Garcia-d'Urso, N. E., Saval-Calvo, M., Azorin-Lopez, J., &
928 Fuster-Guillo, A. (2024). Simultaneous, vision-based fish instance segmentation, species
929 classification and size regression. *PeerJ Computer Science*, *10*.
930 https://doi.org/10.7717/peerj-cs.1770
931 Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale
932 Hierarchical Image Database. *IEEE Computer Vision and Pattern Recognition*.
933 Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani,
934 M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is
935 Worth 16x16 Words: Transformers for Image Recognition at Scale. *International*
936 *Conference on Learning Representations*. https://github.com/
937 Fernandes, A. F. A., Turra, E. M., de Alvarenga, É. R., Passafaro, T. L., Lopes, F. B., Alves, G. F.
938 O., Singh, V., & Rosa, G. J. M. (2020). Deep Learning image segmentation for extraction of
939 fish body measurements and prediction of body weight and carcass traits in Nile tilapia.
940 *Computers and Electronics in Agriculture*, *170*.
941 https://doi.org/10.1016/j.compag.2020.105274
942 Froese, R., Winker, H., Coro, G., Demirel, N., Tsikliras, A. C., Dimarchopoulou, D., Scarcella,
943 G., Probst, W. N., Dureuil, M., Pauly, D., & Anderson, E. (2018). A new approach for
944 estimating stock status from length frequency data. *ICES Journal of Marine Science*, *75*(6),
945 2004–2015. https://doi.org/10.1093/icesjms/fsy078
946 Gao, T., Xiong, Z., Li, Z., Huang, X., Liu, Y., & Cai, K. (2024). Precise underwater fish
947 measurement: A geometric approach leveraging medium regression. *Computers and*
948 *Electronics in Agriculture*, *221*. https://doi.org/10.1016/j.compag.2024.108932
949 Garcia, R., Prados, R., Quintana, J., Tempelaar, A., Gracias, N., Rosen, S., Vågstøl, H., & Løvall,
950 K. (2020). Automatic segmentation of fish using deep learning with application to fish size

measurement. *ICES Journal of Marine Science*, *77*(4), 1354–1366. https://doi.org/10.1093/icesjms/fsz186

Garner, S. B., Olsen, A. M., Caillouet, R., Campbell, M. D., & Patterson, W. F. (2021). Estimating reef fish size distributions with a mini remotely operated vehicle-integrated stereo camera system. *PLoS ONE*, *16*(3 March). https://doi.org/10.1371/journal.pone.0247985

Girshick, R. (2015). Fast R-CNN. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1440–1448. https://doi.org/10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014, October 22). Rich feature hierarchies for accurate object detection and semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2014.81

Haruna, Y., Qin, S., Adama Chukkol, A. H., Yusuf, A. A., Bello, I., & Lawan, A. (2025). Exploring the synergies of hybrid convolutional neural network and Vision Transformer architectures for computer vision: A survey. In *Engineering Applications of Artificial Intelligence* (Vol. 144). Elsevier Ltd. https://doi.org/10.1016/j.engappai.2025.110057

Hordyk, A., Ono, K., Valencia, S., Loneragan, N., & Prince, J. (2014). A novel length-based empirical estimation method of spawning potential ratio (SPR), and tests of its performance, for small-scale, data-poor fisheries. *ICES Journal of Marine Science*, *72*(1), 217–231. https://doi.org/10.1093/icesjms/fsu004

Hua, W., & Chen, Q. (2025). A survey of small object detection based on deep learning in aerial images. *Artificial Intelligence Review*, *58*(6), 162. https://doi.org/10.1007/s10462-025-11150-9

Huang, T., Zang, X., Kondyukov, G., Hou, Z., Peng, G., Pander, J., Knott, J., Geist, J., Melesse, M. B., Jacobson, P., & Deng, Z. D. (2025). Towards automated and real-time multi-object detection of anguilliform fishes from sonar data using YOLOv8 deep learning algorithm. *Ecological Informatics*, *91*. https://doi.org/10.1016/j.ecoinf.2025.103381

Jansi Rani, S. V., Ioannou, I., Swetha, R., Dhivya Lakshmi, R. M., & Vassiliou, V. (2024). A novel automated approach for fish biomass estimation in turbid environments through deep learning, object detection, and regression. *Ecological Informatics*, *81*. https://doi.org/10.1016/j.ecoinf.2024.102663

Jareño, J., Bárcena-González, G., Castro-Gutiérrez, J., Cabrera-Castro, R., & Galindo, P. L. (2024). Enhancing Fish Auction with Deep Learning and Computer Vision: Automated Caliber and Species Classification. *Fishes*, *9*(4). https://doi.org/10.3390/fishes9040133

Jennings, S., Pinnegar, J. K., Polunin, N. V. C., & Boon, T. W. (2001). Weak cross-species relationships between body size and trophic level belie powerful size-based trophic structuring in fish communities. *Journal of Animal Ecology*, *70*(6), 934–944. https://doi.org/10.1046/j.0021-8790.2001.00552.x

Jia, J., Kang, J., Chen, L., Gao, X., Zhang, B., & Yang, G. (2025). A Comprehensive Evaluation of Monocular Depth Estimation Methods in Low-Altitude Forest Environment. *Remote Sensing*, *17*(4). https://doi.org/10.3390/rs17040717

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. *ACM Computing Surveys*, *54*(10). https://doi.org/10.1145/3505244

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25*. http://code.google.com/p/cuda-convnet/

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal Loss for Dense Object Detection. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-October*, 2999–3007. https://doi.org/10.1109/ICCV.2017.324

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. *European Conference on Computer Vision*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 21–37). Springer International Publishing.

Lonati, M., Jahanbakht, M., Atkins, D., Bierwagen, S. L., Chin, A., Barnett, A., & Rummer, J. L. (2024). Novel use of deep neural networks on photographic identification of epaulette sharks (Hemiscyllium ocellatum) across life stages. *Journal of Fish Biology*. https://doi.org/10.1111/jfb.15887

Marrable, D., Barker, K., Tippaya, S., Wyatt, M., Bainbridge, S., Stowar, M., & Larke, J. (2022). Accelerating Species Recognition and Labelling of Fish From Underwater Video With Machine-Assisted Deep Learning. *Frontiers in Marine Science*, *9*(August), 1–11. https://doi.org/10.3389/fmars.2022.944582

Marrable, D., Tippaya, S., Barker, K., Harvey, E., Bierwagen, S. L., Wyatt, M., Bainbridge, S., & Stowar, M. (2023). Generalised deep learning model for semi-automated length measurement of fish in stereo-BRUVS. *Frontiers in Marine Science*, *10*. https://doi.org/10.3389/fmars.2023.1171625

Monkman, G. G., Hyder, K., Kaiser, M. J., & Vidal, F. P. (2019). Using machine vision to estimate fish length from images using regional convolutional neural networks. *Methods in Ecology and Evolution*, *10*(12), 2045–2056. https://doi.org/10.1111/2041-210X.13282

Mots'oehli, M., Nikolaev, A., Igede, W. B., Lynham, J., Mous, P. J., & Sadowski, P. (2024). FishNet: Deep Neural Networks for Low-Cost Fish Stock Estimation. *2024 IEEE International Conference on Omni-Layer Intelligent Systems, COINS 2024*. https://doi.org/10.1109/COINS61597.2024.10622134

Muksit, A. Al, Hasan, F., Hasan Bhuiyan Emon, M. F., Haque, M. R., Anwary, A. R., & Shatabda, S. (2022). YOLO-Fish: A robust fish detection model to detect fish in realistic underwater environment. *Ecological Informatics*, *72*. https://doi.org/10.1016/j.ecoinf.2022.101847

Murat, A. A., & Kiran, M. S. (2025). A comprehensive review on YOLO versions for object detection. In *Engineering Science and Technology, an International Journal* (Vol. 70). Elsevier B.V. https://doi.org/10.1016/j.jestch.2025.102161

Nguyen, H. T. P., Jun, M., & Jeong, H. (2024). Real-time estimation of olive flounder growth in indoor aquaculture using cameras combined with a grid. *Journal of the World Aquaculture Society*. https://doi.org/10.1111/jwas.13108

Nikouei, M., Baroutian, B., Nabavi, S., Taraghi, F., Aghaei, A., Sajedi, A., & Moghaddam, M. E. (2025). Small Object Detection: A Comprehensive Survey on Challenges, Techniques and Real-World Applications. *Intelligent Systems with Applications*, 200561. https://doi.org/10.1016/j.iswa.2025.200561

Nogueira, B. K. C. (2024). *Automated Fish Size Measurement System for Long-Term Growth Studies in the Azores*.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang,

P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., … Bojanowski, P. (2023). *DINOv2: Learning Robust Visual Features without Supervision*. http://arxiv.org/abs/2304.07193

Ovalle, J. C., Vilas, C., & Antelo, L. T. (2022). On the use of deep learning for fish species recognition and quantification on board fishing vessels. *Marine Policy*, *139*. https://doi.org/10.1016/j.marpol.2022.105015

Pauly, D. , & M. G. R. (Eds.). (1987). *Length-based methods in fisheries research. ICLARM Conference Proceedings, 13. International Center for Living Aquatic Resources Management*.

Peters, R. H. (1983). The Ecological Implications of Body Size. In *Cambridge Studies in Ecology*. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511608551

Playout, C., Duval, R., Boucher, M. C., & Cheriet, F. (2022). Focused Attention in Transformers for interpretable classification of retinal images. *Medical Image Analysis*, *82*. https://doi.org/10.1016/j.media.2022.102608

Probst, W. N. (2020). How emerging data technologies can increase trust and transparency in fisheries. *ICES Journal of Marine Science*, *77*(4), 1286–1294. https://doi.org/10.1093/icesjms/fsz036

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/https://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, *28*, 91–99. https://doi.org/10.4324/9780080519340-12

Rocha, W. S., da Fonseca, T. F. C., Watanabe, C. Y. V., da Costa Dória, C. R., & Sant'Anna, I. R. A. (2024). Automatic measurement of fish from images using convolutional neural networks. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-19180-1

Saleh, A., Laradji, I. H., Konovalov, D. A., Bradley, M., Vazquez, D., & Sheaves, M. (2020). A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-71639-x

Sheaves, M., Bradley, M., Herrera, C., Mattone, C., Lennard, C., Sheaves, J., & Konovalov, D. A. (2020). Optimizing video sampling for juvenile fish surveys: Using deep learning and evaluation of assumptions to produce critical fisheries parameters. *Fish and Fisheries*, *21*(6), 1259–1276. https://doi.org/10.1111/faf.12501

Shibata, Y., Iwahara, Y., Manano, M., Kanaya, A., Sone, R., Tamura, S., Kakuta, N., Nishino, T., Ishihara, A., & Kugai, S. (2024). Length estimation of fish detected as non-occluded using a smartphone application and deep learning method. *Fisheries Research*, *273*, 106970. https://doi.org/https://doi.org/10.1016/j.fishres.2024.106970

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. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-19932-9

Tseng, C. H., Hsieh, C. L., & Kuo, Y. F. (2020). Automatic measurement of the body length of harvested fish using convolutional neural networks. *Biosystems Engineering*, *189*, 36–47. https://doi.org/10.1016/j.biosystemseng.2019.11.002

Turk, M., & Pentland, A. (1991). Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, *3*(1). http://direct.mit.edu/jocn/article-pdf/3/1/71/1932018/jocn.1991.3.1.71.pdf

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *31st Conference on Neural Information Processing Systems*.

Voskakis, D., Makris, A., & Papandroulakis, N. (2021). Deep learning based fish length estimation. An application for the Mediterranean aquaculture. *OCEANS 2021*.

Wing, K., & Woodward, B. (2024). Advancing artificial intelligence in fisheries requires novel cross-sector collaborations. *ICES Journal of Marine Science*. https://doi.org/10.1093/icesjms/fsae118

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). CvT: Introducing Convolutions to Vision Transformers. *Proceedings of the IEEE International Conference on Computer Vision*, 22–31. https://doi.org/10.1109/ICCV48922.2021.00009

Yu, C., Hu, Z., Han, B., Wang, P., Zhao, Y., & Wu, H. (2021). Intelligent measurement of morphological characteristics of fish using improved u-net. *Electronics (Switzerland)*, *10*(12). https://doi.org/10.3390/electronics10121426

Yu, Y., Zhang, H., & Yuan, F. (2023). Key point detection method for fish size measurement based on deep learning. *IET Image Processing*, *17*(14), 4142–4158. https://doi.org/10.1049/ipr2.12924

Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M., & Lee, B. (2022). A survey of modern deep learning based object detection models. In *Digital Signal Processing: A Review Journal* (Vol. 126). Elsevier Inc. https://doi.org/10.1016/j.dsp.2022.103514

Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2019). *Object Detection in 20 Years: A Survey*. http://arxiv.org/abs/1905.05055