Dataset selection is critical for effective pre-training of fish detection models for underwater video

Ayyagari, Devi; Alavi, Talukder Wasi; Singh, Navlika; Barnes, Joshua; Morris, Corey; Whidden, Christopher

Publisher's version / Version de l'éditeur:

https://doi.org/10.1093/icesjms/fsaf039

ICES Journal of Marine Science, 82, 4, 2025-04-04

ICES Journal of Marine Science, 2025, Vol. 82, Issue 4, fsaf039
https://doi.org/10.1093/icesjms/fsaf039
Received: 20 August 2024; **revised:** 13 March 2025; **accepted:** 14 March 2025
Original Article

Dataset selection is critical for effective pre-training of fish detection models for underwater video

Devi Ayyagari [1,*], Talukder Wasi Alavi[1], Navlika Singh[2], Joshua Barnes[3], Corey Morris[4], Christopher Whidden[1]

[1]Dalhousie University, Halifax, NS B3H 4R2, Canada, Faculty of Computer Science
[2]Indian Institute of Technology, Jodhpur, 342030, India
[3]National Research Council Canada, St. John's, NL, A1B 3T5, Canada
[4]Fisheries and Oceans Canada, St. John's, NL A1C 5X1, Canada

*Corresponding author. Faculty of Computer Science, Dalhousie University, Halifax, NS B3H 4R2, Canada. E-mail: devi.ayyagari@dal.ca

Abstract

Underwater digital monitoring systems using acoustics and video have the potential to transform marine monitoring and fisheries stock assessment but generate significant amounts of data, shifting the burden from data collection to data analysis. Machine learning (ML) is a potential solution but remains underutilized for marine monitoring, partly due to the time and cost of annotating new training datasets for each marine class and habitat. This raises the pivotal question: "How can we train marine machine learning models with limited annotated data?" We catalog publicly available marine datasets annotated for detection and classification, investigating the feasibility of leveraging a fish detector trained on three existing datasets to detect fish in a new small underwater marine dataset. We compare the accuracy and training time of pre-trained models to those without pre-training. We find pre-training with OzFish yields faster convergence and comparable performance with smaller training datasets. However, pre-training with some datasets reduced performance and increased training time. We expect our catalog of publicly available marine datasets will assist in the selection of pre-training datasets. Our results underscore the need for diverse, large, publicly available marine datasets with varied habitat and class distributions to develop and integrate ML models into automated systems for monitoring marine ecosystems.

Keywords: machine learning; transfer learning; public marine image datasets; fisheries; stock assessment; autonomous monitoring; remote video monitoring

Introduction

Traditionally, marine ecosystem monitoring relies on methods such as trawling, gill nets, field surveys, visual observations, and manual data collection on-site (Gallo et al. 2022). While these methods are effective, they come with limitations in scalability, cost, time, and labor, often concentrating on specific target species and with negative environmental effects. Consequently, our monitoring of underwater life remains insufficient, leading to surprises when fish or shellfish populations suddenly decline, as seen in the recent unexpected collapse of the Bering Sea snow crab population (Szuwalski et al. 2023). The swift changes in migratory and reproductive behaviors of various marine species due to climate change and warming waters (Poloczanska et al. 2016) exacerbate this issue.

Addressing these issues necessitates the urgent adoption of scalable monitoring methods. These methods are vital not only to track these changes but also to ensure the efficacy and safety of climate-related endeavors, such as expanding tidal power, offshore wind farms, and innovative carbon dioxide removal technologies. To meet the escalating demand for monitoring marine species, alternative approaches, including remote sensing (Wang et al. 2010), environmental DNA metabarcoding (Xiong et al. 2022), acoustic sensing (Yassir et al. 2023), and video cameras (Saleh et al. 2022), have been proposed. While these methods offer cost-effective and potentially scalable monitoring alternatives, the sheer volume of data they produce makes manual analysis impractical, prompting automated techniques to analyze the massive data generated.

Deep learning presents a promising solution for scaling marine monitoring by automating the processing of large datasets. Specifically, deep learning can automate two primary tasks in underwater video monitoring: detection, which identifies and localizes marine objects, and classification, which assigns species labels to the detected objects. In recent years, deep learning has achieved notable breakthroughs across domains such as healthcare, robotics, entertainment, manufacturing, and transportation (Min et al. 2017, Pierson and Gashler 2017, Da'u and Salim 2020, Gupta et al. 2021, Yang et al. 2021), demonstrating its effectiveness in addressing complex object detection and classification challenges. In the context of detecting and classifying marine species, object detection models have been widely applied to identify and categorize marine objects in various data sources, including fishing boats (Rekha et al. 2020), fish farms (Crescitelli et al. 2021), and fish counter images (Malik et al. 2023); public datasets such as fish4K (Boom et al. 2012), fishCLEF (Joly et al. 2016), Norfisk (Crescitelli et al. 2021), and OzFish (Australian Institute of Marine Science (AIMS) et al. 2019); and in relatively fewer cases, custom underwater video datasets (Lathifah et al. 2020, Marrable et al. 2022). Deep learning models learn discriminative features from a set of examples and then apply these learned features to perform tasks such as classification and detection on another set of examples, assuming identi-

cal data distributions. These models necessitate training on labeled data specific to the environment and species of interest. However, iterating this training process for each new application is time-consuming and requires experts to manually label a substantial number of images, often in the thousands, for each species. Motivated by the human brain's adaptability in transferring knowledge across tasks, researchers are actively investigating and embracing strategies to explore the possibilities of transfer learning—leveraging knowledge gained from mastering one task to excel in another (Knausgård et al. 2022, Veiga et al. 2022).

Transfer learning is commonly used to train machine learning models on smaller datasets by first training them on a large publicly available dataset and then applying the acquired knowledge to a task with limited annotated data. Transfer learning generally comprises two main phases: pre-training and fine-tuning. In the pre-training phase, a machine learning model is trained on a large publicly available dataset to learn the discriminative features of the examples in that dataset. During the fine-tuning phase, some or all of the model's parameters are adjusted using the parameters learned during pre-training to better fit the smaller dataset. This approach contrasts with training a model without any pre-training: with randomly initialized weights, where representations are learned solely from the often limited annotated dataset available.

Various studies have used transfer learning by pre-training with public marine datasets for the tasks of marine species classification and detection. For example, Knausgård et.al., (2022) pre-train the backbone of their YOLOv3 detector using weights from a YOLOv3 model trained on ImageNet (Russakovsky et al. 2015) for object classification (not detection) and pre-train their classification model with Fish4K (Boom et al. 2012). They then fine-tune with some examples from their custom temperate fish dataset to achieve 83.68% classification accuracy on a small high-resolution dataset of 1022 images of 4 temperate fish species. Siddiqui et al. (2017) extracted features from three deep learning architectures trained on ImageNet (Russakovsky et al. 2015) to train a classifier to detect and classify 16 classes of fish in Western Australia. Veiga et al. (2022) trained using a combination of the annotated samples from a publicly available dataset OzFish (Australian Institute of Marine Science (AIMS) et al. 2019) and negative examples from the custom dataset to generate annotations of a 700 min raw video footage custom dataset with the same species distribution as the OzFish dataset. While these examples showcase the application of pre-training in marine detection and classification tasks, the role of pre-training in fine-tuning remains unclear. Specifically, it is not evident whether the choice of the pre-training dataset significantly influences the fine-tuning of a marine dataset or if utilizing the largest general-purpose vision dataset is more advantageous. If the former holds true, what defines an ideal marine dataset?

In this study, we systematically study the influence of pre-training dataset selection on the fine-tuning process for fish detection in a low-lit, low-resolution underwater marine dataset. Most state-of-the-art object recognition models optimize detection and classification jointly using a shared feature extraction backbone; however, isolating detection from classification offers significant benefits for generalizing these tasks to small marine datasets. This is because generalizing detection and classification to marine monitoring presents fundamentally different challenges: Detection requires labor-intensive manual annotation to draw bounding boxes around objects, while classification must contend with large variations in species distributions, morphological changes, and temporal dynamics that complicate transfer learning. By focusing exclusively on detection, we can tailor pre-training and fine-tuning strategies to reduce annotation burdens and enhance localization performance, ultimately creating a scalable framework that is better suited to complex and resource-constrained environments such as low-light, low-resolution underwater settings.

As an initial step, we compiled a catalog of publicly available underwater marine datasets and evaluated their suitability for training machine learning models for detection and classification based on criteria such as dataset size, number of examples per marine class, and ease of use. Next, we formulated specific hypotheses and conducted systematic experiments to evaluate how pre-training dataset choice affects convergence time, performance [measured by mean average precision (mAP]), and generalization on two subsets of our custom underwater dataset. We hypothesize that pre-training with large marine datasets improves both detection performance and convergence speed for fish detection, but it does not fully overcome the challenges of small fine-tuning datasets. Our experiments demonstrate that the choice of pre-training dataset significantly impacts fish detection performance and convergence speed on our custom underwater dataset: The specific hypotheses and experiments are detailed in the "Experiments" section, and the results of the experiments are presented in the "Results" section.

Throughout the study, we maintain a consistent approach for fine-tuning: all model parameters are initialized with those learned during the pre-training phase, without selectively initializing specific layers of the fine-tuning model. Here, "marine classes" refer to categories of marine species represented in the underwater images, such as fish, crustaceans, or other marine organisms. It is important to clarify that the term "class" here is used in the context of machine learning and does not correspond to taxonomic classifications in marine biology.

In summary, the paper presents a comprehensive catalog of publicly available marine datasets assessed for their applicability to classification and detection tasks, along with a systematic exploration of the impact of pre-training dataset choice on fine-tuning in the context of fish detection. To the best of our knowledge, this study is the first comprehensive investigation to guide the selection of pre-training image datasets for underwater fish detection.

Materials and methods

Dataset curation

The process of compiling the curated list of datasets started with an initial selection of datasets: Fish4K, OzFish, and Norfisk, which we discovered during our earlier research (citation removed for anonymity for review). To expand our collection, we conducted targeted online searches utilizing keywords: "Fish," "Fish images," "Fish species," "Fish distribution," "Fish computer vision," "Norfisk," and "datasets similar to Norfisk." Additionally, we broadened our exploration to include data hosting platforms such as Kaggle and Roboflow, aiming to identify datasets pertinent to marine species classification. This systematic search approach yielded a substantial

Table 1. Public datasets with at least 1000 photographic images of marine classes annotated for the tasks of marine class detection and classification.

Dataset	Size of dataset			M	H	T	O
	No. of images	Detection	Classification				
Fish4Knowledge (Boom et al. 2012)	27 370 images	27 370 annotations	23 species	N	N	Y	N
OzFish (Australian Institute of Marine Science (AIMS) et al. 2019)	80 000 images	45K annotations over 1800 frames	500 classes	Y	N	N	N
Fathomnet (Katija et al. 2022)	109 871 images	296 795 annotations	2398 concepts	Y	Y	N	N
Deep Vision fish dataset (Allken and Rosen 2020)	4925 images	4.925 annotations	3 fish classes one class: mesopelafic fish	Y	N	N	N
NOAA Puget Sound Nearshore Fish (Farrell et al. 2023)	77 739 images	67 990 annotations on 30 384 images	2 marine animals Fish, Custaceans	Y	Y	N	N
NorFisk (Crescitelli et al. 2021)	12 514 images	12 514 annotations	2 classes	N	N	N	N
DeepFish (Garcia-d'Urso et al. 2022)	1291 images	7339 annotations	59 species	Y	N	N	Y
Open Images Google API (Kuznetsova et al. 2020) dataset	24 386 images	24 386 annotations	16 marine classes	Y	N	N	N
FishCLEF (Joly et al. 2016)	93 videos with 20 000 images	22 655 annotations	15 species	Y	Y	N	N
google_dataset _try (roboflow.com) (Xelou1 2022)	1290 images	1290 annotations	14 classes	Y	N	N	N
Brackish dataset (Pedersen et al. 2023)	21 151 images	49 266 annotations	6 marine clases	Y	Y	Y	N
Fishnet.AI (Kay and Merrifield 2021)	143 818 images	549 209 annotations	34 object classes (fish classes and humans)	Y	N	N	N
Salmon Computer Vision (Atlas et al. 2023)	1567 videos each several minutes to an hour long	532 000 annotations	15 salmon species	Y	Y	Y	N
Visual Marine Animal Tracking (VMAT; Cai et al. 2023) dataset	33 sequences, with each seq ~75s long	74 178 annotations	17 marine classes	Y	Y	Y	N
AFFiNe (Venema and de Beer 2022)	7482 images	7482 annotations	30 species	N	N	N	Y
Kakadu FishAI dataset (Jansen et al. 2024)	44 112 images	82 904 annotations	23 tropical freshwater species	Y	Y	N	N
Labeled Fishes in the Wild (Cutter et al. 2015)	929 annotated; 3167 negative images; test set: 211 frames from single video	1005 training set 2061 test set	N/A	N	Y	N	N

M: Does the dataset include images with multiple marine classes in the same image?; H: Does the dataset include images of the habitat without any marine sightings?; T: Is the dataset annotated for tracking marine classes?; O: Are the marine classes outside a water body, like on fish boats or in fish markets?

pool of datasets. Subsequently, we filtered these datasets to identify datasets containing a minimum of 1000 photographic images with labels suitable for training supervised machine learning models for classification and detection tasks while excluding those with illustrations or stampings.

The resulting selection comprises fifteen publicly accessible datasets, listed in Table 1 for recommendation. Moreover, we evaluated each dataset for the presence of multiple objects within images, the inclusion of non-fish images (i.e. negative samples without fish and only habitat), label availability for tracking, and the depiction of fish both underwater and outside of water to ensure alignment with various practical use cases. Detailed dataset descriptions are attached in the Supplementary Material. A comprehensive list of all identified datasets is also attached as Supplementary Material.

The recommended datasets have been sourced from various locations and encompass a diverse range of data acquisition methods. The largest dataset, FathomNet (Katija et al. 2022), is a collection of approximately 100 000 high-resolution expertly annotated images, initially seeded with a subset of curated imagery and metadata from the Monterey Bay Aquarium Research Institute (MBARI), National Geographic Society (NGS), and the National Oceanic and Atmospheric Administration (NOAA) and allows for collection from different data sources. Other datasets that were collected through the deployment of underwater cameras in the ocean include Fish4Knowledge (Boom et al. 2012), OzFish (Australian Institute of Marine Science (AIMS) et al. 2019), and Deep Vision fish dataset (Allken and Rosen 2020). For instance, the Fish4Knowledge dataset was acquired by placing 10 cameras in Southern Taiwan over three years to study the detection, classification, and behavioral recognition of coral reef fish, primarily during daylight hours. Another notable dataset, OzFish, was collected from 3000 baited remote underwater videos and includes annotations not only for detection and classification but also for studying the size and shape of fish tails. Similarly, the Deep Vision fish dataset gathers data from surveys conducted in the Atlantic Ocean during 2017 and 2018, with a specific focus on economically significant pelagic species.

In contrast, some datasets were procured within fish farms, such as the NOAA Puget Sound Nearshore Fish 2017–2018 dataset (Farrell et al. 2023) and the Norfisk (Crescitelli et al. 2021) datasets, which were gathered in and around shellfish aquaculture farms within a Northeast Pacific estuary and multiple fish farms in Norway, respectively. Additionally, certain datasets have been acquired outside of water, such as DeepFish (Garcia-d'Urso et al. 2022), consisting of images captured at a local wholesale fish market in El Campello, Spain, and AFFiNe (Venema and de Beer 2022), which is a dataset of 30 freshwater species of the Netherlands, photographed by anglers. Others, like Fishnet.AI (Kay and Merrifield 2021), were obtained from on-board monitoring cameras on longline tuna fishing boats in the Western and Central Pa-

cific, containing images of both fish and humans on fishing boats.

Furthermore, there is considerable variation in species and habitat distribution granularity among these datasets. For instance, the Salmon computer vision dataset (Atlas et al. 2023), gathered by deploying cameras across two First Nation-run weirs, includes labels for 15 distinct salmon species. In contrast, the Brackish dataset (Pedersen et al. 2019), derived from annotated frames of 89 videos captured by a single camera spanning a 180 km long strait between the North Sea and Kattegat, offers labels at a finer granularity, encompassing categories such as fish, small fish, crab, shrimp, jellyfish, and starfish. Conversely, the Labeled Fishes in the Wild dataset Cutter et al. (2015) lacks species labeling but provides annotations exclusively for detection purposes. As for the VMAT (Cai et al. 2023) dataset, it covers diverse marine habitats, ranging from coral reefs to seagrass beds, shallow and deep mid-water columns, and sandy and rocky seabeds with 17 different types of marine animal, including octopuses, sharks, dolphins, and nine classes of fish. Additionally, it captures a variety of swimming behaviors, including fast, medium, and slow constant swimming, darting, crawling, and stop-and-go maneuvers.

Datasets like the Open Images Google API (Kuznetsova et al. 2020) were compiled from online sources, notably Flickr. The origin of some datasets remains unclear, such as the Google_dataset_try dataset (Xelou1 2022), a small balanced dataset hosted on the Roboflow platform, and fishCLEF, a collection of 93 annotated underwater videos released as part of the larger lifeCLEF (Joly et al. 2016) competition. These carefully selected datasets offer a versatile toolkit, serving as an excellent foundation for researchers and practitioners to adapt machine learning to marine class classification and detection.

Four supplementary datasets: WildFish++ (Zhuang et al. 2021), ONC SeaTube (Hoeberechts et al. 2015), Fishnet.AI (Kay and Merrifield 2021), and DeepFish (Saleh et al. 2020) are promising underwater video datasets with significant potential for machine learning research. However, their application to supervised machine learning is limited by challenges such as restricted access or the absence of annotations. Further details about these datasets are provided in Section 4 of the Supplementary Material.

What role does the choice of the pre-training dataset play in fine-tuning the underwater marine dataset?

Experiments

We formulate specific hypotheses and design corresponding experiments listed below to investigate the impact of the pre-training dataset on fine-tuning a custom underwater dataset (The custom underwater dataset used in this study will be made available upon reasonable request. The YOLOv7 models trained on the public marine datasets Norfisk, OzFish, and NPSNF are publicly hosted on Hugging Face.) for fish detection. Figure 1 describes the workflow of the conducted experiments.

- **Hypothesis 1:** Pre-training with large datasets—datasets with at least 10 000 annotated objects in the dataset—will result in higher mAP, a metric used to assess the performance of the task of fish detection, on the fine-tuning dataset, compared to training with random initialization.
  Experiment: Conduct statistical t-tests to compare the mAP of models fine-tuned on a custom underwater video dataset pre-trained with three public marine underwater datasets, a generic computer vision dataset MS COCO ((Lin et al., 2014), with the model trained solely on the custom dataset without pre-training, on an independent test set of the custom underwater dataset used for fine-tuning.
- **Hypothesis 2:** Pre-training with large datasets results in faster convergence on the fine-tuning dataset compared to random initialization. Convergence is when the model's performance stabilizes, meaning further training will not significantly improve accuracy or reduce the error rate. It is measured in epochs, indicating the computational power required to train the model on that dataset.
  Experiment: Compare the number of epochs required for the custom underwater dataset with the same number of samples to converge, both with and without pre-training. An epoch is defined as the duration taken by the model to iterate through all samples in the training set precisely once.
- **Hypothesis 3:** Pre-training with diverse datasets enhances the model's ability to generalize effectively (higher mAP) when performing the same task on a completely unseen dataset.
  Experiment: Conduct statistical t-tests to compare the mAP of models fine-tuned on a custom underwater video dataset, pre-trained with three public marine underwater datasets, a generic computer vision dataset MS COCO, with the model trained solely on the custom dataset without pre-training, on a distinct underwater marine dataset: the Brackish dataset (Pedersen et al. 2019). This dataset has a different data distribution from the pre-training and fine-tuning datasets used in this study and is referred to as the "OOD" dataset.
- **Hypothesis 4:** Pre-training alone cannot fully address the disparity in annotated dataset sizes, i.e. models fine-tuned on larger datasets will perform (mAP) better than models fine-tuned on smaller datasets, irrespective of the dataset used for pre-training.
  Experiment: Conduct statistical t-tests to compare the fine-tuning performance (mAP) of models trained on two subsets of the custom underwater marine dataset—specifically, one subset comprising 500 samples and another containing 1000 samples—pre-trained with three public marine underwater datasets, and a general computer vision dataset, MS COCO on a test set of the custom underwater dataset used for fine-tuning.

Model

The YOLO family of models is a single-stage object detector that frames object detection as a regression problem, applying this approach to spatially separated bounding boxes and their class probabilities to identify various classes of objects in an image. The initial version, proposed by Redmon et al. (2016) in 2015, utilized a single convolutional neural network to predict bounding boxes and class probabilities directly from complete images during a single evaluation. Over the years, YOLO models have improved in accuracy and efficiency. This study uses YOLOv7 (Wang et al. 2023), which was released in 2022, for all experiments and evaluates performance using the following metrics:

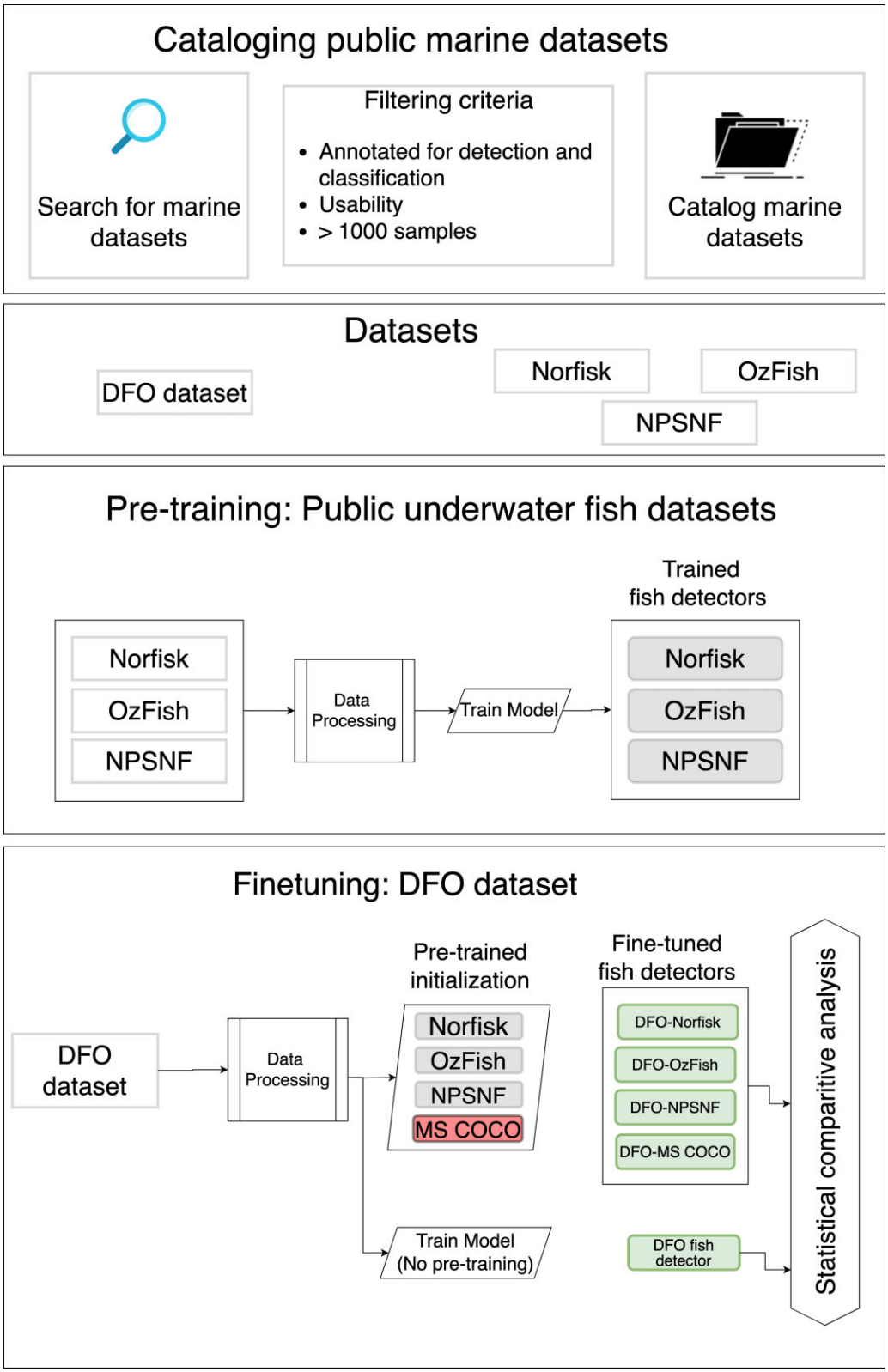

Figure 1. Study workflow: publicly available fish datasets are cataloged, and three underwater marine datasets—Norfisk, OzFish, and the NOAA Puget Sound Nearshore Fisheries dataset—are selected based on their similarity to the custom dataset. A YOLOv7 detector is trained on each of these datasets. Two training sets are created from the annotated custom underwater dataset, containing 500 and 1000 samples. For each subset, a YOLOv7 detector is trained using three approaches: (a) pre-training on the marine datasets, (b) pre-training on MS COCO, and (c) training from random initialization.

- **Intersection over union (IoU):** IoU quantifies the overlap between the predicted bounding box and the annotated fish object. Higher IoU values indicate more accurate alignment. The IoU threshold is a predefined value (e.g. 0.5), specifying the minimum overlap required for a detection to be considered valid. For instance, with a 0.5 IoU threshold, detections overlapping with the annotated fish object by at least 50% are considered valid detections.

$$\text{IoU} = \frac{\text{Area of overlap}}{\text{Area of Union}}. \quad (1)$$

- **Precision:** Precision is characterized as the proportion of correct positive fish detections among all annotated fish objects, serving as an indicator of the accuracy of positive predictions made by the object detector.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}. \quad (2)$$

- **Recall:** Recall measures how well the detector finds all the annotated fish objects. A higher recall indicates that the model can effectively recognize and recall a greater proportion of relevant fish objects present in the dataset.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \quad (3)$$

- **mAP:** Average precision (AP) serves as a singular metric capturing the trade-off between precision and recall in a model, calculated as the area under the interpolated precision-recall curve. For the task of classification with multiple classes, AP is computed individually for each class and then averaged to obtain the mAP. However, for the specific task of detection with a single class: fish, the mAP is equivalent to the AP.

$$\text{mean Average Precision (mAP)} = \frac{1}{j}\sum_{i=1}^{j} \text{AP}_i \quad (4)$$

with AP_i being the AP of the ith class and j being the total number of classes.

$$\text{Average Precision(AP)} = \sum_{k=0}^{n-1}[(\text{recall}(k) - \text{recall}(k+1)) \times \text{precision}(k)], \quad (5)$$

where n is the number of ground truth objects and k is an index representing each object in the ground truth, ranging from 0 to $n - 1$.

Datasets

To facilitate effective transfer learning from pre-training to fine-tuning, we selected pre-training datasets that closely resembled our custom dataset. The custom dataset consists of underwater images with a resolution of 1080 × 1920, captured at depths of 200–500 m, with class distribution described in Table 2. From the marine catalog presented in Table 1, we identified datasets that met the following criteria: (i) an image resolution of 1080 × 1920, (ii) images capturing fish within the water column, and (iii) at least 1000 examples of the most abundant class in the dataset. Originally, we were targeting datasets with a similar class distribution as our dataset, but we did not find any dataset with a similar distribution to the custom dataset in this study.

Table 2. Class distributions in the OzFish and custom datasets.

Class/species	Count
OzFish dataset	
Fish	30 635
Lethrinus punctulatus	2162
Lethrinus atkinsoni	929
Lutjanus sebae	815
Acanthurus triostegus	689
Lutjanus vitta	497
Thalassoma lunare	449
Pomacentrus coelestis	443
Abudefduf sexfasciatus	317
Chromis atripectoralis	230
Lutjanus bohar	215
Thalassoma lutescens	209
Epinephelus areolatus	185
Pentapodus porosus	152
Alepes vari	143
Lethrinus rubrioperculatus	143
Lethrinus ravus	143
Lethrinus nebulosus	134
Epinephelus multinotatus	131
Abalistes stellatus	122
Dascyllus aruanus	119
Acanthurus sp.	104
Caesio teres	98
Naso fageni	95
Carangoides gymnostethus	83
Siganus fuscescens	71
Epinephelus fasciatus	68
Odonus niger	62
Stegastes nigricans	53
Custom dataset	
Gadus morhua	66 741
Macrourus berglax	24 639
Skate	10 361
Sebastes mentella	8490
Anarhichas lupus	6248
Hippoglossus hippoglossus	3478

Only species with at least 50 examples are included for OzFish for brevity.

We identified the OzFish (Australian Institute of Marine Science (AIMS) et al. 2019), NPSNF (Farrell et al. 2023), and Brackish (Pedersen et al. 2019) datasets as those that satisfied all the criteria outlined above. Ozfish dataset has several subsets annotated for separate tasks such as detection, classification, fish size computation, and fish tail detection. For this study, we selected the "Bounding Box Species Annotations" subset, as it is the only subset where every frame is exhaustively annotated with bounding boxes and corresponding species labels for all marine organisms in the frame. The OzFish dataset contains 142 species of reef fish, with multiple fish per image. Similarly, the NPSNF and Brackish datasets also feature multiple fish per image but differ in that they were acquired in brackish waters and categorize marine life into broader groups, such as fish and crab, rather than more specific species specification.

Furthermore, we included the Norfisk (Crescitelli et al. 2021) dataset, which, despite its lower resolution of 253 × 337, met other criteria and contains underwater images captured in fish farms with examples of two classes: saithe (a member of the Cod family, similar to *Gadus morhua* in the custom dataset) and salmonoids. Unlike the other datasets, Norfisk has only one fish per image.

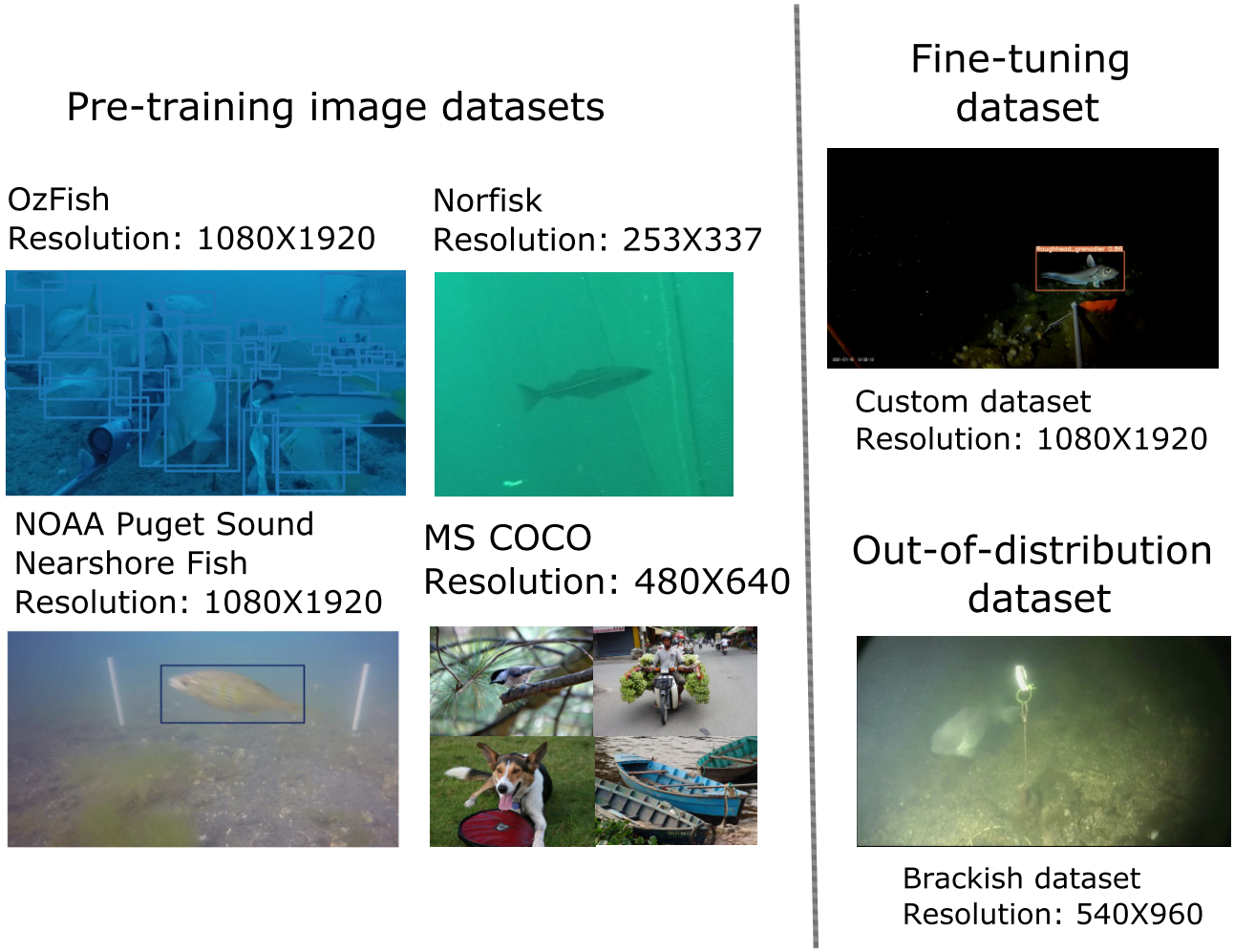

Figure 2. Sample images from the datasets used in this study. All images were obtained from publicly available image datasets. In the pre-training datasets (arranged from top left to bottom right), sample images were extracted from the OzFish dataset (Australian Institute of Marine Science et al.,2019), the Norfisk dataset (Crescitelli et al.,2021), the NPSNF dataset (Farrell et al., 2023), and the MS COCO dataset (Lin et al.,2014). For the out-of-distribution dataset, the sample image was extracted from the Brackish dataset (Pedersen et al.,2019).

The OzFish, Norfisk, and NPSNF datasets were used for pre-training, while the Brackish dataset was designated for out-of-distribution (OOD) testing. The Brackish dataset is considered "OOD" because the model was not fine-tuned on this dataset but was instead tested on it to evaluate generalization performance. We aimed to select an OOD dataset that shares similarities with one of the pre-training datasets, and the Brackish dataset met this criterion. It was acquired in environments similar to those of the NPSNF dataset, and both datasets include crab as a shared class. Although both the NPSNF and Brackish datasets use unconventional labeling, categorizing marine animals broadly as "fish," "small_fish," and "crab" without finer taxonomic resolution, they were chosen for this study due to the environmental and class overlap between them. Examples of images from all datasets used in this study can be seen in Fig. 2. The class distributions of the OzFish dataset and the custom dataset (the only datasets in this study with corresponding scientific names) are presented in Table 2. The species distribution of Skate in the custom dataset, along with the class distributions of the Norfisk, NPSNF, and Brackish datasets, are included in the Supplementary Material. For the pre-training experiments, we used the Brackish dataset (Pedersen et al. 2019), which is extended by the Brackish MOT dataset (Pedersen et al. 2023) listed in Table 1.

In machine learning, it is common practice to fine-tune models using pre-trained weights from large, publicly available datasets rather than starting with randomly initialized weights. YOLOv7 was pre-trained on the MS COCO dataset (Lin et al.,2014), and the resulting model parameters were released by its authors. For completeness, we compare models fine-tuned on marine datasets with models fine-tuned on MS COCO dataset. MS COCO is a general-purpose computer vision dataset containing 200 000 labeled images with 1.5 million object instances across 80 classes, mostly non-marine categories such as "boat," "cow," and "backpack." We did not retrain YOLOv7 on MS COCO; instead, we used the YOLOv7 model pre-trained on MS COCO and published by the original authors (Wong 2022).

Data processing

Pre-training datasets

Typically, for training a machine learning model, a dataset is partitioned into training, validation, and test sets. The model

Table 3. Training, validation, and test set sizes for the Norfisk, OzFish, and NPSNF marine underwater datasets, along with the mAP of YOLOv7 models trained on each dataset.

Dataset	Size of training set	Size of validation set	Size of test set	mAP on test set
NorFisk	8758	2502	1254	0.999
OzFish	1230	351	177	0.764
NOAA Puget Sound Nearshore Fish Dataset (NPSNF)	18 041	6171	6172	0.717

The mAP values indicate the detection performance of the models on their respective test sets.

learns features from the training set and is monitored for overfitting using the validation set during training. Performance is evaluated on the independent test set, not exposed to the model during training. In this study, training and testing of public marine datasets are restricted to images containing fish. Images with at least one annotated fish object from Norfisk, OzFish, and NPSNF datasets were randomly distributed into training, validation, and test sets with ratios of 70/20/10, 70/20/10, and 60/20/20, respectively, accounting for dataset size. The split sizes for the three datasets are detailed in Table 3, and details regarding label preparation for all datasets are provided in the Supplementary Section.

Fine-tuning dataset

The custom dataset used in this study consists of 221 videos, exhaustively annotated by an ecologist at the DFO using the VIAME (Dawkins et al. 2017) platform. These 221 videos are segmented into train, validation, and test splits in a 70-15-15 ratio based on the frequency of fish in a frame within each video. The models are trained on the data in the training split, the data in the validation split are then used to track the model parameters and avoid overfitting, and the performance of the trained models is evaluated on the data in the test split. Frames are extracted from these videos at a rate of 30 frames per second, and frames lacking at least one annotated fish object are excluded. The resulting train, validation, and test splits comprise 71 736; 19 211; and 19 210 frames with fish, respectively.

Labels generated by VIAME are transformed to adhere to the format expected by YOLOv7, and text files containing information about the fish bounding boxes are created, maintaining the same directory structure as the extracted frames. Within the training subset, frames with fish are further divided into two subsets: one with 500 samples and another with 1000 samples, both containing frames with at least one annotated fish object. These subsets ensure equal representation of examples from each class of fish, with frames possibly featuring examples from multiple fish classes or multiple instances of the same fish class.

OOD dataset

The Brackish dataset, selected to evaluate the fine-tuned model's generalization beyond its initial training distribution, comprises six labeled categories: "fish," "small_fish," "crab," "shrimp," "jellyfish," and "starfish." For the test set, we choose all images featuring at least one annotated sample of the class "fish." It is important to note that these frames may also include instances from other categories such as shrimp and crab, acknowledging and accommodating the diversity within these frames. The annotations, presented in a format similar to OzFish and NPSNF, are formatted as expected by YOLOv7, and corresponding text files are generated to facilitate the computation of mAP for the comparative analysis.

Training

Pre-training

The YOLOv7 detector is trained on the training split of each of the four public datasets, with default hyperparameters and augmentations, with weights initialized randomly. We use the default IoU and confidence thresholds of 0.65 and 0.001, respectively, to train and evaluate the detectors. Norfisk and OzFish are trained at an image resolution of 640, the recommended image resolution for training a YOLOv7 object detector, with a batch size of 32. The model is trained on one NVIDIA A100 SXM4 GPU. NPSNF, on the other hand, is trained at an image resolution of 1280 × 1280 on NVIDIA A100 SXM4 GPUs connected via NVLink with a batch size of 4, following the hardware limitations and chosen parameters on the researcher's original work on YOLOv5, as can be found here (Morris 2023). It is important to note that we do not perform any hyperparameter tuning. All models are trained till convergence, and the models with the best mAP, at the IoU and confidence thresholds of 0.65 and 0.001, respectively, are used for fine-tuning the custom dataset. The performance of the trained models on the test sets of these datasets can be found in Table 3.

Fine-tuning

Initially, we train a YOLOv7 detector on each of the two training subsets (one with 500 samples and another with 1000 samples) of the fine-tuning dataset, using the validation set to prevent overfitting. All the layers for training are randomly initialized without using weights from previous tasks to establish the baseline. These two models are referred to as the models trained without any pre-training, with random initialization. Subsequently, we fine-tune the YOLOv7 detector using the parameters learned using the four different publicly available datasets: (i) Norfisk, (ii) OzFish, (iii) NPSNF, and (iv) MS COCO on both the training subsets. To ensure statistical significance, each of these five experiments is conducted five times due to the randomness inherent in training a YOLOv7 detector. All experiments are conducted using the default hyperparameters and augmentations as provided in the default YOLOv7 repository. Training is conducted on a single NVIDIA A100 SXM4 GPU, with images trained at a resolution of 640 until convergence. The models with the best mAP on the validation splits (the data samples used to avoid overfitting and determine when the model has converged) of each of the fine-tuning datasets are used for evaluating the fine-tuned models.

Results

Detection performance on the custom dataset

- **Hypothesis 1:** In contrast to our hypothesis, pre-training with marine datasets did not consistently result in higher mAP as indicated in Table 4 and Fig. 3. Pre-training with some marine datasets—NPSNF and Norfisk—performed statistically lower than models trained with no pre-training, while pre-training with OzFish performed statistically significantly better. Models pre-

Table 4. Performance comparison of models fine-tuned on the custom underwater image dataset with subsets of 500 and 1000 frames, pre-trained with publicly available marine datasets (Norfisk, OzFish, and NPSNF), pre-trained with the general-purpose computer vision dataset MS COCO, and randomly initialized models.

Models fine-tuned with	No pre-training	Performance on test set for models pre-trained with mAP (mean ± stddev)			
		MS COCO	Norfisk	OzFish	NPSNF
500 samples	0.854 ± 0.004	0.857 ± 0.003^{+}	0.794 ± 0.003^{**+}	**0.896 ± 0.01****	0.655 ± 0.02^{**+}
1000 samples	0.879 ± 0.007	0.879 ± 0.011	0.858 ± 0.014	**0.898 ± 0.005***	0.824 ± 0.002**

Models marked with $*$ and $**$ indicate a statistically significant difference with P-values $<.05$ and .01, respectively, in the t-test compared to training with random initialization. Models marked with $+$ indicated statistical significance with a P-value $< .05$ between the models fine-tuned with 500 and 1000 samples. Bold values denote the models with the highest mAP.

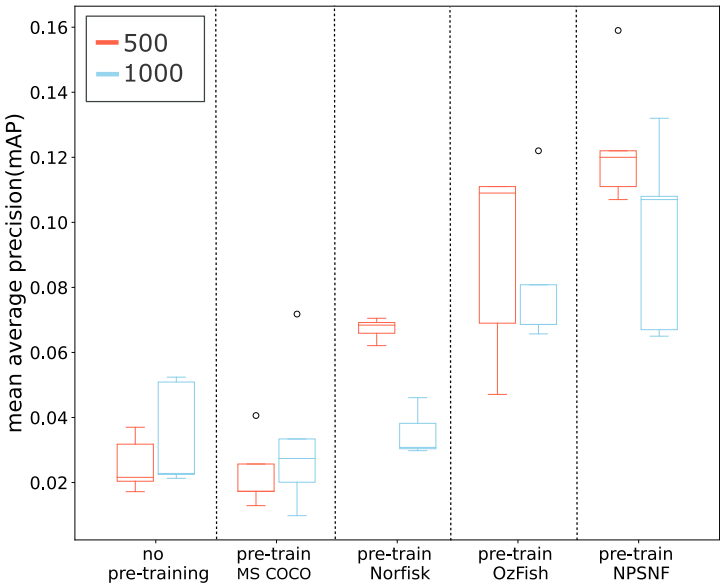

Figure 3. Performance (mAP) comparison of the models trained with and without pre-training. Models are pre-trained with the marine datasets: Norfisk, NPSNF, and OzFish, and a general-purpose image dataset—MS COCO on two annotated subsets of the custom dataset—one with 500 samples and another with 1000 samples of the custom dataset. All the trained models are tested on the test set of the custom dataset. Each fine-tuned model is trained five times to test for statistical significance.

trained with MS COCO showed no statistical difference in terms of mAP.

- **Hypothesis 2:** In contrast to our hypothesis, pre-training did not always accelerate the convergence time as indicated in Table 5. Pre-training with the marine datasets Norfisk and NPSNF took longer to converge, while pre-training with OzFish and MS COCO took less time than training without any pre-training.
- **Hypothesis 3:** In contrast to our hypothesis, pre-training with large datasets did not consistently result in better generalization on the Brackish dataset as indicated in Table 6 and Fig. 4. Pre-training with the marine datasets NPSNF and OzFish showed statistically significantly better results than the models trained without any pre-training. Models pre-trained with MS COCO showed no statistical difference in terms of mAP to the models trained without any pre-training. All the trained models exhibited overall poor performance on the Brackish dataset.
- **Hypothesis 4:** In contrast to our initial hypothesis that pre-training alone cannot fully address the disparity in annotated dataset size, pre-training with the marine dataset OzFish yielded statistically similar results with the smaller annotated dataset as observed in Table 4. However, pre-training with the datasets MS COCO, Norfisk, and NPSNF; and the models trained without any pre-training showed statistically significantly higher performance with the larger annotated dataset.

We observe that the choice of pre-training dataset significantly influences both the performance (mAP) and convergence rate of the fine-tuning task. Models pre-trained with the OzFish marine dataset outperformed those trained without pre-training and those pre-trained with other marine datasets, and they also converged the fastest. Remarkably, the models fine-tuned on a smaller subset of the custom underwater dataset with 500 samples, using the OzFish dataset, showed comparable performance to those fine-tuned on a larger subset with 1000 samples. This is a notable result, as it demonstrates that with the right pre-training datasets, similar performance can be achieved with smaller datasets. In contrast, pre-training with the Norfisk and NPSNF marine datasets resulted in slower convergence and lower mAP compared to random initialization or pre-training with MS COCO. This underscores the importance of carefully selecting the pre-training dataset, as not all marine datasets necessarily outperform MS COCO or random initialization.

Table 5. Epochs required for model convergence when fine-tuning on 500 and 1000 frames.

Models fine-tuned with	No pre-training	Number of epochs to converge (mean ± stddev)			
		MS COCO	Norfisk	OzFish	NPSNF
500 samples	362 ± 50	244 ± 40*	2732 ± 37**	**27 ± 9****	2146 ± 156**
1000 samples	256 ± 68	190 ± 22	2275 ± 32**	**55 ± 37****	1565 ± 19**

Lower values indicate faster convergence. Asterisks (∗) and (∗∗) denote statistically significant differences compared to random initialization, with *P*-values < .05 and < .01, respectively. Models marked with ∗ and ∗∗ indicate a statistically significant difference with *P*-values < .05 and < .01, respectively, in the t-test compared to training with random initialization. Bold values denote the models with the fastest convergence.

Table 6. Performance comparison of models fine-tuned on a custom underwater image dataset with subsets of 500 and 1000 frames, pre-trained with publicly available marine datasets (Norfisk, OzFish, and NPSNF), pre-trained with the general-purpose computer vision dataset MS COCO, and randomly initialized models on Brackish dataset.

Models fine-tuned with	No pretraining	Performance on Brackish dataset for models pre-trained with mAP (mean ± stddev)			
		MS COCO	Norfisk	OzFish	NPSNF
500 samples	0.026 ± 0.007	0.023 ± 0.001	0.067 ± 0.003**+	0.089 ± 0.026*	**0.124 ± 0.018****
1000 samples	0.034 ± 0.014	0.032 ± 0.021	0.035 ± 0.006	0.081 ± 0.021*	**0.096 ± 0.026***

Models marked with ∗ and ∗∗ indicate a statistically significant difference with *P*-values < .05 and .01, respectively, in the t-test compared to training with random initialization. Models marked with + indicated statistical significance with a *P*-value < .05 between the models fine-tuned with 500 and 1000 samples.Bold values denote the models with the highest mAP.

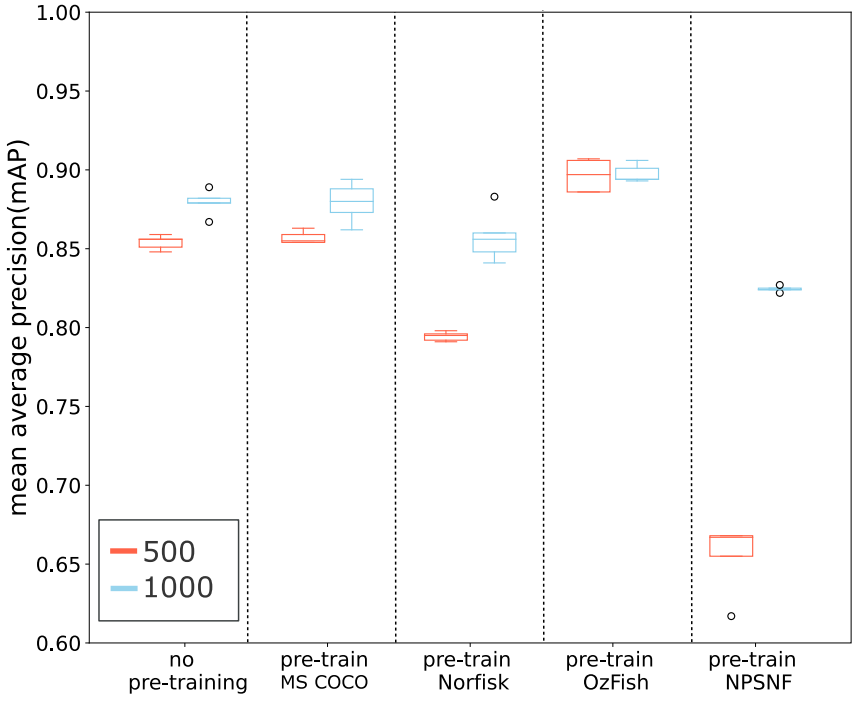

Figure 4. Performance (mAP) comparison of the models trained with and without pre-training. Models are pre-trained with the marine datasets: Norfisk, NPSNF, and OzFish, and a general-purpose image dataset—MS COCO on two annotated subsets of the custom dataset—one with 500 samples and another with 1000 samples of the custom dataset. All the trained models are tested on the Brackish dataset, the dataset not exposed to the model during training. Each fine-tuned model is trained five times to test for statistical significance.

Recall across the different classes in the custom dataset

Although the models in this study were trained exclusively for detection, we investigated whether pre-training with specific datasets influenced performance across different classes. Some pre-training datasets, such as Norfisk, include examples from a similar family—e.g. Saithe in Norfisk and *Gadus morhua* in the custom dataset, both belonging to the Cod family. To understand how the class distribution of the pre-training datasets impacted performance across different classes in the custom dataset, we evaluated the models' detection performance across these classes. Since the models were trained solely for detection, metrics such as true negatives and false negatives could not be computed. Instead, we focused on true positives and used recall as the primary evaluation metric. Table 7 shows the mean and standard deviation of recall percentages for models fine-tuned on the subset of the custom dataset with 1000 samples, as evaluated on the test set. The results for models trained on the 500-sample subset are provided in the Supplementary Material.

Table 7. Recall values for models trained on different pre-training datasets, including marine datasets (Norfisk, NPSNF, and OzFish) and the general-purpose MS COCO dataset.

Classes	Mean and stddev for recall across datasets				
	No pre-training	MS COCO	Norfisk	OzFish	NPSNF
Gadus morhua	83.02 ± 1.5	83.15 ± 1.1	81.32 ± 0.2	86.42 ± 1.5	69.96 ± 0.4
Macrourus berglax	87.92 ± 1.5	89.68 ± 0.5	83.71 ± 1.0	92.72 ± 0.7	79.50 ± 0.6
Skate	68.34 ± 5.9	64.13 ± 2.5	65.25 ± 2.0	79.50 ± 5.7	49.30 ± 0.9
Sebastes mentella	89.59 ± 0.9	90.03 ± 1.2	81.93 ± 0.9	92.23 ± 2.0	72.09 ± 1.1
Anarhichas lupus	83.03 ± 2.1	80.81 ± 4.3	76.27 ± 4.4	85.01 ± 5.5	67.61 ± 2.8
Hippoglossus hippoglossus	60.75 ± 4.3	56.54 ± 3.1	56.54 ± 3.5	72.02 ± 4.3	53.82 ± 1.1

All models were fine-tuned on 1000 samples from the custom dataset. Higher recall indicates better ability to correctly identify relevant instances.

We observed that the models had the best recall on *Sebastes mentella*, followed closely by *Macrourus berglax*, likely due to distinguishing features such as the reddish coloration of *Sebastes mentella* and the prominent eyes of Grenadier. Conversely, the models consistently struggled with Skate and *Hippoglossus hippoglossus* across all pre-training experiments. These groundfish classes remain close to the seabed, making them inherently difficult to detect and classify.

Discussion

The importance of the choice of the pre-training dataset on both the training time and the performance (mAP) of machine learning models for fish detection and classification is evident from our study. Specifically, for our dataset with deep underwater images of fish under battery-powered lighting, pre-training with OzFish demonstrated outstanding performance compared to models without pre-training. Model pre-trained with OzFish achieved an average mAP of 0.89 in 27 epochs compared to 0.85 in 362 epochs for the models trained with 500 custom underwater samples and an average mAP of 0.89 in 55 epochs compared to 0.87 in 256 epochs for models trained with 1000 underwater samples. Both results were statistically significant, with a P-value $< .01$, on the t-test conducted to account for training randomness on mAP distributions from five models, one distribution from models pre-trained with OzFish and the other from models with random initialization. Additionally, models pre-trained on the OzFish dataset consistently exhibited higher recall across all classes in the fine-tuned dataset compared to models pre-trained on other datasets. Furthermore, OzFish pre-trained models consistently exhibited higher recall across all classes and lower standard deviations in recall compared to randomly initialized models, demonstrating their robustness and reliability.

Contrary to our initial hypothesis that pre-training on larger marine datasets would enhance fine-tuning performance, pre-training with Norfisk and NPSNF yielded poorer results compared to OzFish, a smaller dataset, and random initialization with MS COCO, a large image dataset without any marine-specific categories. We attribute this to limitations in the Norfisk and NPSNF datasets. Norfisk comprises only two categories (saithe and salmonoids), is acquired on a fish farm, features one fish per image, and includes lower-resolution images, reducing the diversity and informativeness of its data. Similarly, NPSNF primarily contains single-fish-per-frame images captured in murky waters, where even human identification is challenging. These factors likely constrained the ability of models pre-trained on these datasets to transfer features effectively during fine-tuning, highlighting the importance of data diversity and quality.

We make an interesting observation in the recall performance of models fine-tuned with NorFisk. The models pre-trained on NorFisk—a dataset containing examples of "saithe," a member of the cod family—and fine-tuned with 500 examples from the custom dataset exhibited a recall for *Gadus morhua* that surpassed that of *Macrourus berglax*, the second-best performing class following *Sebastes mentella*. This represented the only occurrence where *Gadus morhua* detection showed a higher recall than *Macrourus berglax*. However, when fine-tuned with 1000 examples, the recall for *Macrourus berglax* surpassed that of *Gadus morhua*, indicating that NorFisk may harbor specific features advantageous for cod recognition, although these benefits may have been mitigated by catastrophic forgetting (Kirkpatrick et al. 2017), a prevalent phenomenon encountered when fine-tuning all model parameters during the task.

For OOD testing, all models exhibited low mAP values, reflecting the expected performance drop when machine learning models are tested on data distributions they were not trained on. While we hypothesized that exposure to more diverse pre-training datasets would enhance robustness, our results did not support this assumption, likely due to catastrophic forgetting during fine-tuning. Models pre-trained on NPSNF generalized best to the Brackish dataset, possibly due to shared environmental characteristics, such as murky waters, and the presence of crabs in both datasets. However, the significant class variance between the fine-tuned dataset, which lacks crab, and the Brackish dataset, where crab predominates, likely diminished the model's ability to retain relevant information and limited its generalization capability.

OzFish comprises images of fish captured in reef environments, characterized by matching resolution, multiple classes in a single frame, fish in the water column, and instances of occlusion. These features closely align with our custom dataset, likely facilitating better feature transfer during fine-tuning and resulting in superior performance. Despite having only 1800 images compared to the 30 384 examples in NPSNF, the largest marine pre-training dataset used, we hypothesize that OzFish's diversity and similarity to our custom dataset were critical factors in its success.

These findings highlight the importance of dataset characteristics, such as resolution, class diversity, habitat complexity, and the presence of multiple classes per image, in determining the effectiveness of pre-training. They emphasize the need for selecting datasets that closely match the target domain to

enhance model performance. However, the absence of universal standards for resolution, clarity, format, and comprehensive annotation (i.e. ensuring that all fish in the images are labeled) complicates direct comparisons of datasets' suitability for detection tasks. Establishing such standards would likely improve pre-training outcomes and facilitate the integration of multiple datasets.

Our experiments highlight the critical impact of selecting an appropriate pre-training dataset on model performance, even if ideal characteristics are not always obvious. When experimenting with various datasets is impractical, we recommend pre-training with a large image database such as MS COCO or ImageNet (Russakovsky et al. 2015). Although pre-training with MS COCO may result in similar performance (an average mAP of 0.85 and 0.87 for models trained with 500 and 1000 samples, respectively) to random initialization, it encourages faster and more stable convergence. For models trained with 500 examples from each marine class in the custom underwater dataset, those pre-trained with MS COCO converged in 244 epochs, compared to 362 epochs for models with random initialization, showing significantly quicker convergence times with a P-value $< .05$. While models trained with 1000 samples did not show statistically significant faster convergence, the standard deviation for random initialization was 62 epochs compared to 22 epochs for MS COCO pre-training. Additionally, the average epochs required for convergence were reduced to 190 for MS COCO pre-training, compared to 256 for random initialization. This smaller standard deviation, combined with stable average mAP, makes pre-training with MS COCO a preferred and stable alternative. Most open-sourced computer vision architectures publish their parameters learned on MS COCO, making pre-training with MS COCO highly accessible and ready to use.

In this study, we focus exclusively on the impact of pre-training on fine-tuning for the task of detection, specifically detecting general "fish" objects in a custom dataset. Detecting generic fish objects is not only an intermediate step toward species classification but also directly useful in marine monitoring, where species-level identification is not always required. Similar to echosounder studies that estimate fish biomass through acoustic backscatter without species-specific resolution (Peña et al. 2021), generalized fish detection can provide a general overview of biodiversity. The detected fish objects can then support downstream tasks such as species classification, population estimation, behavioral analysis, and identifying previously unseen species. Decoupling detection from classification enables us to address the unique challenges of each task independently, which can be difficult if both tasks are optimized simultaneously. For example, detection can utilize large public image datasets to learn shared anatomical features across various marine species, even when the species distributions in specific studies do not align with those in public datasets, thus enhancing the generalization of detection. Meanwhile, classification can benefit from unsupervised techniques like clustering, which are impractical when applied to entire images containing multiple objects of different classes, but become feasible when focused on detected marine objects. We believe this decoupling is a critical intermediate step toward automating the application of machine learning to marine monitoring.

We also note that the performance of the trained models could potentially be improved by further refining the hyperparameters specific to the dataset in question. Hyperparameter tuning is a time-consuming process, and achieving state-of-the-art results in terms of mAP was not the primary focus of this study. Furthermore, our study primarily focused on fine-tuning by initializing all model parameters. More nuanced experiments, such as selectively fine-tuning only the last few layers or training models to perform simultaneous detection tasks on both large annotated public datasets and small custom datasets, could reveal additional insights. Moreover, this study did not specifically analyze which dataset characteristics exerted the greatest influence. Future research could systematically conduct experiments to assess the relative importance of different attributes such as class diversity, number of examples per class, and image resolution, thereby improving the selection of pre-training datasets. Additionally, combining datasets with diverse characteristics—such as taxonomy, habitats, and imaging conditions—for pre-training could enhance model generalization and potentially outperform models pre-trained on single datasets, particularly when fine-tuned on small labeled datasets.

Efficient application of machine learning to monitor marine ecosystems hinges on three key elements: data representative of the test set distribution, sufficient computational resources, and publicly available, adaptable machine learning methodologies. Representative datasets enable accurate modeling, sufficient computational resources handle large-scale data processing, and adaptable methodologies aid in tailoring models to specific challenges in marine monitoring. For practitioners new to machine learning, we recommend following best practices such as pre-training with comprehensive datasets like MS COCO, ensuring sufficient class-specific samples (Ayyagari et al. 2023), using representative datasets for training and testing, applying data augmentation techniques, and addressing class imbalances with subsampling, oversampling, or adjusted loss functions. Additionally, practitioners should thoroughly understand their datasets and conduct basic sanity checks, such as verifying the data, ensuring initial overfitting to a single example, and establishing baseline metrics before performing more complex experiments. Comprehensive experimentation is essential to validate machine learning approaches and ensure robust and reproducible results.

In parallel with advancements in machine learning methodologies, the availability of diverse, extensive, and publicly accessible annotated marine datasets is crucial for progress in underwater data monitoring. The primary challenges in creating and publishing such datasets include data collection, the cost and effort of manual annotation, and hosting requirements. To address these challenges, we advocate for innovative approaches such as Cornell's Merlin app for bird identification and leveraging platforms like Amazon Mechanical Turk for crowdsourcing annotations. These methods could be applied to datasets such as ONC SeaTube (Hoeberechts et al. 2015) and DeepFish (Saleh et al. 2020) to automate the annotation process. Engaging the broader public through initiatives like MBARI's recently launched game, Fathomverse (Simpson 2024), can further accelerate the development of annotated marine datasets.

A significant limitation of many existing datasets is the prevalence of ambiguous labeling practices. Generic terms such as "crab," "cod," or "starfish" are often used, which can refer to multiple species across different families, reducing the utility of these datasets for precise machine learning applications. Only a few datasets, such as OzFish, DeepFish,

FishForKnowledge, and FishCLEF, provide full Latin species names. When species-level identification is not possible, researchers should adopt the lowest possible taxonomic specification (e.g. "*Sebastes* sp." instead of "Redfish") to enhance accuracy and usability. Proper taxonomic labeling is critical for improving dataset quality, interpretability, and reproducibility. It also enables researchers to identify the most suitable datasets for tasks like transfer learning, ultimately enhancing the scalability and effectiveness of machine learning models.

Publicly accessible datasets that capture diverse habitats across various marine environments, combined with rigorous taxonomic species labeling, offer a strong foundation for advancing the scalability of machine learning models in marine ecosystem monitoring. Coupled with a robust understanding of applying machine learning to small underwater datasets, these efforts can significantly improve our ability to monitor marine species, better understand the impacts of climate change, and develop effective climate-driven mitigation strategies in oceanic ecosystems.

Acknowledgments

We wish to thank Daniel Porter and Khanh Nguyen at DFO for procuring and annotating the custom dataset used in this study. We also appreciate Thomas Trappenberg for his valuable discussions on best practices for training and comparing machine learning methods. Our thanks extend to Dan Morris at the Google AI for Nature and Society program for providing the website with a list of some marine datasets, which helped us ensure that no relevant marine dataset was overlooked. Finally, we are grateful to the peer reviewers for their constructive and thoughtful comments and suggestions, which significantly improved the manuscript. This research was enabled in part by computing resources provided by ACENET and the Digital Research Alliance of Canada.

Author contributions

Conceptualization: C.M., J.B., and C.W.; Data curation: T.A., N.S., C.M., and J.B.; Formal Analysis: D.A.; Funding acquisition: C.M., J.B., and C.W.; Methodology: D.A., C.M., J.B., and C.W.; Project administration: C.W.; Resources: C.W., C.M., and J.B.; Software: D.A. and N.S.; Supervision: C.W.; Validation: D.A.; Visualization: D.A.; Writing—original draft: D.A.; Writing—review & editing: C.W. All the authors have approved the final submitted draft.

Supplementary data

Supplementary data is available at *ICES Journal of Marine Science* online.

Conflict of interest: The authors declare that the research was carried out in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This research was supported by grants from National Research Council of Canada (NRC) (OCN-110-2), Ocean Frontier Institute (OFI) (OG-202110), and also supported by the Dalhousie Faculty of Computer Science. Data were provided by the Department of Fisheries and Oceans (DFO). C.W. is supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN-2021-02988).

Data availability

The Norfisk dataset is available from Dataverse at https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/H5G3K5; the OzFish dataset can be accessed via the Pawsey Data Portal at https://storage.pawsey.org.au/public/m/FDFML/labelled/frames; the NPSNF dataset is available on GCP at https://lila.science/datasets/noaa-puget-sound-nearshore-fish; and the Brackish dataset is hosted on Roboflow at https://public.roboflow.com/object-detection/brackish-underwater. The custom dataset used in this article will be shared upon reasonable request to the corresponding author.

The YOLOv7 models trained on the the publicly available marine datasets Norfisk, OzFish and NOSNF are hosted publicly on Hugging Face. Norfisk: https://huggingface.co/Devi-Ayyagari/yolov7_norfisk; OzFish: https://huggingface.co/Devi-Ayyagari/yolov7_OzFish; NPSNF: https://huggingface.co/Devi-Ayyagari/yolov7_NPSNF

References

Allken V, Rosen S. Deep vision fish dataset. 2020. https://doi.org/10.21335/NMDC-551736490(28 February 2025, date last accessed).

Atlas WI, Ma S, Chou YC *et al.* Wild salmon enumeration and monitoring using deep learning empowered detection and tracking. *Front Mar Sci* 2023;**10**:1200408. https://doi.org/10.3389/fmars.2023.1200408.

Australian Institute of Marine Science (AIMS), University of Western Australia (UWA), Curtin University. Ozfish dataset—machine learning dataset for baited remote underwater video stations. 2019. https://doi.org/10.25845/5e28f062c5097(28 February 2025, date last accessed).

Ayyagari D, Morris C, Barnes J *et al.* Toward low-cost automated monitoring of life below water with deep learning. *Environ Data Sci* 2023;**2**:e13. https://doi.org/10.1017/eds.2023.8.

Boom BJ, Huang PX, He J *et al.* Supporting ground-truth annotation of image datasets using clustering. In: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. 2012, 1542–5.

Cai L, McGuire NE, Hanlon R *et al.* Semi-supervised visual tracking of marine animals using autonomous underwater vehicles. *Int J Comput Vision* 2023;**131**:1406–27. https://doi.org/10.1007/s11263-023-01762-5.

Crescitelli AM, Gansel LC, Zhang H. Norfisk: fish image dataset from norwegian fish farms for species recognition using deep neural networks. *Model Identif Control* 2021;**42**:1–16. https://doi.org/10.4173/mic.2021.1.1.

Cutter G, Stierhoff K, Zeng J. Automated detection of rockfish in unconstrained underwater videos using haar cascades and a new image dataset: labeled fishes in the wild. In: *2015 IEEE Winter Applications and Computer Vision Workshops*, 2015, 57–62. https://doi.org/10.1109/WACVW.2015.11.

Da'u A, Salim N. Recommendation system based on deep learning methods: a systematic review and new directions. *Artif Intell Rev* 2020;**53**:2709–48. https://doi.org/10.1007/s10462-019-09744-1.

Dawkins M, Sherrill L, Fieldhouse K *et al.* An open-source platform for underwater image and video analytics. In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2017, 898–06. https://doi.org/10.1109/WACV.2017.105.

Farrell D, Ferriss B, Sanderson B *et al.* NOAA Puget Sound Nearshore Fish 2017–2018. *Mendeley Data, V1*. 2023. https://doi.org/10.17632/n73g6ysv8c.1.

Gallo ND, Bowlin NM, Thompson AR *et al.* Fisheries surveys are essential ocean observing programs in a time of global change: a synthesis of oceanographic and ecological data from US West Coast Fisheries Surveys. *Front Mar Sci* 2022;**9**:757124. https://doi.org/10.3389/fmars.2022.757124.

Garcia-d'Urso N, Galan-Cuenca A, Pérez-Sánchez P *et al.* The DeepFish computer vision dataset for fish instance segmentation, classification, and size estimation. *Scientific Data* 2022;**9**:287. https://doi.org/10.1038/s41597-022-01416-0.

Gupta A, Anpalagan A, Guan L *et al.* Deep learning for object detection and scene perception in self-driving cars: survey, challenges, and open issues. *Array* 2021;**10**:100057. https://doi.org/10.1016/j.array.2021.100057.

Hoeberechts M, Owens D, Riddell DJ *et al.* The power of seeing: experiences using video as a deep-sea engagement and education tool. In: *OCEANS 2015—MTS/IEEE Washington*. 2015, 1–9. https://doi.org/10.23919/OCEANS.2015.7404592.

Jansen A, van Bodegraven S, Esparon A *et al.* Monitoring tropical freshwater fish with underwater videography and deep learning. *Marine and Freshwater Research* 2024:**75**. https://doi.org/10.1071/MF23166.

Joly A, Goëau H, Glotin H *et al.* Lifeclef 2016: multimedia life species identification challenges. In: N Fuhr, P Quaresma, T Gonçalveset. al. (eds), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Cham: Springer, 2016. 286–310. https://doi.org/10.1007/978-3-319-44564-9_26.

Katija K, Orenstein E, Schlining B *et al.* Fathomnet: a global image database for enabling artificial intelligence in the ocean. *Sci Rep* 2022;**12**:15914. https://doi.org/10.1038/s41598-022-19939-2.

Kay J, Merrifield M. The fishnet open images database: a dataset for fish detection and fine-grained categorization in fisheries. 2021, https://arxiv.org/abs/2106.09178. Preprint: not peer reviewed.

Kirkpatrick J, Pascanu R, Rabinowitz N *et al.* Overcoming catastrophic forgetting in neural networks. *Proc Natl Acad Sci* 2017;**114**:3521–6. https://doi.org/10.1073/pnas.1611835114.

Knausgård KM, Wiklund A, Sørdalen TK *et al.* Temperate fish detection and classification: a deep learning based approach. *Appl Intell* 2022;**52**:6988–7001. https://doi.org/10.1007/s10489-020-02154-9.

Kuznetsova A, Rom H, Alldrin N *et al.* The open images dataset v4. *Int J Comput Vision* 2020;**128**:1956–81. https://doi.org/10.1007/s11263-020-01316-z.

Lathifah HM, Novamizanti L, Rizal S. Fast and accurate fish classification from underwater video using you only look once. *IOP Conf Ser Mater Sci Eng* 2020;**982**:012003. https://doi.org/10.1088/1757-899X/982/1/012003.

Lin T.-Y., Maire M., Belongie S *et al. Microsoft coco: Common objects in context*, Vol. 13, Springer: 2014, 740–55.

Malik H, Naeem A, Hassan S *et al.* Multi-classification deep neural networks for identification of fish species using camera captured images. *PLoS One* 2023;**18**:e0284992. https://doi.org/10.1371/journal.pone.0284992.

Marrable D, Barker K, Tippaya S *et al.* Accelerating species recognition and labelling of fish from underwater video with machine-assisted deep learning. *Front Mar Sci* 2022;**9**:944582. https://doi.org/10.3389/fmars.2022.944582.

Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform* 2017;**18**:851–69. https://doi.org/10.1093/bib/bbw068.

Morris D. Github—agentmorris/noaa-fish: data preparation and model training for the NOAA Puget Sound Nearshore Fish dataset. 2023. https://github.com/agentmorris/noaa-fish. (27 June 2024, date last accessed).

Pedersen M, Bruslund Haurum J, Gade R *et al.* Detection of marine animals in a new underwater dataset with varying visibility. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019, 18–26.

Pedersen M, Lehotský D, Nikolov I *et al.* Brackishmot: the brackish multi-object tracking dataset. In: R Gade, M Felsberg, J-K Kämäräinen (eds), *Image Analysis*. Cham. Springer Nature, 2023. 17–33. https://doi.org/10.1007/978-3-031-31435-3_2.

Peña H, Macaulay GJ, Ona E *et al.* Estimating individual fish school biomass using digital omnidirectional sonars, applied to mackerel and herring. *ICES J Mar Sci* 2021;**78**:940–51. https://doi.org/10.1093/icesjms/fsaa237.

Pierson HA, Gashler MS. Deep learning in robotics: a review of recent research. *Adv Robot* 2017;**31**:821–35. https://doi.org/10.1080/01691864.2017.1365009.

Poloczanska ES, Burrows MT, Brown CJ *et al.* Responses of marine organisms to climate change across oceans. *Front Mar Sci* 2016;**3**:62. https://doi.org/10.3389/fmars.2016.00062.

Redmon J, Divvala S, Girshick R *et al.* You only look once: unified, real-time object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, 779–88.

Rekha BS, Srinivasan GN, Reddy SK *et al.* Fish detection and classification using convolutional neural networks. In: S Smys, JMRS Tavares, VE Balas, et al. (eds), *Computational Vision and Bio-Inspired Computing*. Cham: Springer, 2020, 1221–31. https://doi.org/10.1007/978-3-030-37218-7_128.

Russakovsky O, Deng J, Su H *et al.* ImageNet large scale visual recognition challenge. *Int J Comput Vision* 2015;**115**:211–52. https://doi.org/10.1007/s11263-015-0816-y.

Saleh A, Laradji IH, Konovalov DA *et al.* A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Sci Rep* 2020;**10**:14671. https://doi.org/10.1038/s41598-020-71639-x.

Saleh A, Sheaves M, Rahimi Azghadi M. Computer vision and deep learning for fish classification in underwater habitats: a survey. *Fish Fish* 2022;**23**:977–99. https://doi.org/10.1111/faf.12666.

Siddiqui SA, Salman A, Malik MI *et al.* Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data. *ICES J Mar Sci* 2017;**75**:374–89. https://doi.org/10.1093/icesjms/fsx109.

Simpson E. Fathomverse. 2024. https://www.fathomverse.game/(7 March 2025, date last accessed).

Szuwalski CS, Aydin K, Fedewa EJ *et al.* The collapse of eastern bering sea snow crab. *Science* 2023;**382**:306–10. https://doi.org/10.1126/science.adf6035.

Veiga RJM, Ochoa IE, Belackova A *et al.* Autonomous temporal pseudo-labeling for fish detection. *Appl Sci* 2022;**12**:5910. https://www.mdpi.com/2076-3417/12/12/5910.

Venema J, de Beer J. Affine—angling freshwater fish netherlands. 2022. https://www.kaggle.com/datasets/jorritvenema/affine (26 February 2025, date last accessed).

Wang CY, Bochkovskiy A, Liao HYM. Yolov7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, 7464–75.

Wang K, Franklin SE, Guo X *et al.* Remote sensing of ecology, biodiversity and conservation: a review from the perspective of remote sensing specialists. *Sensors* 2010;**10**:9647–67. https://www.mdpi.com/1424-8220/10/11/9647.

Wong KY. Github—wongkinyiu/yolov7: implementation of paper—yolov7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. 2022. https://github.com/WongKinYiu/yolov7 (27 June 2024, date last accessed).

Xelou. google_dataset_try dataset. 2022. https://universe.roboflow.com/xelou1-hotmail-fr/google_dataset_try (6 February 2025, date last accessed).

Xiong F, Shu L, Zeng H *et al.* Methodology for fish biodiversity monitoring with environmental dna metabarcoding: the primers, databases and bioinformatic pipelines. *Water Biol Secur* 2022;**1**:100007. https://doi.org/10.1016/j.watbs.2022.100007.

Yang S, Zhu F, Ling X *et al.* Intelligent health care: applications of deep learning in computational medicine. *Front Genet* 2021;**12**:607471. https://doi.org/10.3389/fgene.2021.607471.

Yassir A, Jai Andaloussi S, Ouchetto O *et al.* Acoustic fish species identification using deep learning and machine learning algorithms: a systematic review. *Fish Res* 2023;**266**:106790. https://www.sciencedirect.com/science/article/pii/S0165783623001832.

Zhuang P, Wang Y, Qiao Y. Wildfish++: a comprehensive fish benchmark for multimedia research. *IEEE T Multimedia* 2021;**23**:3603–17. https://doi.org/10.1109/TMM.2020.3028482.

Handling Editor: Howard Browman