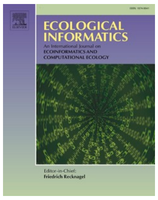
Ecological Informatics 78 (2023) 102303

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Semi-supervised and weakly-supervised deep neural networks and dataset for fish detection in turbid underwater videos

Mohammad Jahanbakht[a], Mostafa Rahimi Azghadi[a,b,*], Nathan J. Waltham[a,b]

[a] Centre for Tropical Water and Aquatic Ecosystem Research (TropWATER), James Cook University, Douglas 4811, QLD, Australia
[b] College of Science and Engineering, James Cook University, 1 James Cook Dr, Douglas 4811, QLD, Australia

ARTICLE INFO

Keywords:
Weakly-supervised classifier
Self-supervised learning
Deep neural networks
Contrastive learning
Transfer learning
XGBoost ensemble
Fish detection
Highly turbid waters

ABSTRACT

Fish are key members of marine ecosystems, and they have a significant share in the healthy human diet. Besides, fish abundance is an excellent indicator of water quality, as they have adapted to various levels of oxygen, turbidity, nutrients, and pH. To detect various fish in underwater videos, Deep Neural Networks (DNNs) can be of great assistance. However, training DNNs is highly dependent on large, labeled datasets, while labeling fish in turbid underwater video frames is a laborious and time-consuming task, hindering the development of accurate and efficient models for fish detection. To address this problem, firstly, we have collected a dataset called FishInTurbidWater, which consists of a collection of video footage gathered from turbid waters, and quickly and weakly (i.e., giving higher priority to speed over accuracy) labeled them in a 4-times fast-forwarding software. Next, we designed and implemented a semi-supervised contrastive learning fish detection model that is self-supervised using unlabeled data, and then fine-tuned with a small fraction (20%) of our weakly labeled FishInTurbidWater data. At the next step, we trained, using our weakly labeled data, a novel weakly-supervised ensemble DNN with transfer learning from ImageNet. The results show that our semi-supervised contrastive model leads to more than 20 times faster turnaround time between dataset collection and result generation, with reasonably high accuracy (89%). At the same time, the proposed weakly-supervised ensemble model can detect fish in turbid waters with high (94%) accuracy, while still cutting the development time by a factor of four, compared to fully-supervised models trained on carefully labeled datasets. Our dataset and code are publicly available at the hyperlink FishInTurbidWater.

1. Introduction

Managing coastal ecosystems for species protection has traditionally relied on species abundance and richness data, which is usually collected following accepted protocols and rigor. In estuaries and nearshore coastal waters fish are a popular species surveyed by researchers, given their distribution and abundance can generally be associated with some environmental condition or habitat association (Aguzzi et al., 2019; Whitfield, 2017).

Managing these coastal ecosystems requires data and information derived from field campaigns using labor-intensive and expensive tools like nets, traps, and pots. However, these methods can have inherent confounding problems. Overcoming these challenges is now possible with advancements in technology (Dutta and Arhonditsis, 2023). In aquatic science, scientists have teamed up with artificial intelligence programmers to develop and maximize the use of underwater video cameras because it offers a more affordable and rapid approach that presents reduced risk to operators (Ditria et al., 2021; Heggie and Ogburn, 2021; Jahanbakht et al., 2022b). Using this technology also increases sampling accuracy, replicability, and reproducibility over traditional sampling methods (Harvey et al., 2012; McIvor et al., 2022), which is the basis of any sound scientific investigation (Saleh et al., 2022b). However, the use of this technology has putative challenges, too, namely relating to large amounts of video data to process, usually back in the laboratory using computer software. This is time-consuming and can be biased if using multiple operators, and video files recorded in turbid water or where the camera is moving, make the processing difficult (Donaldson et al., 2019).

One approach to facilitate this processing is to utilize the unparalleled power of Deep Neural Networks (DNNs) in image and video

[*] Corresponding author at: Centre for Tropical Water and Aquatic Ecosystem Research (TropWATER), James Cook University, Douglas 4811, QLD, Australia.
E-mail addresses: mohammad.jahanbakht@jcu.edu.au (M. Jahanbakht), mostafa.rahimiazghadi@jcu.edu.au (M. Rahimi Azghadi), nathan.waltham@jcu.edu.au (N.J. Waltham).

https://doi.org/10.1016/j.ecoinf.2023.102303
Received 26 January 2023; Received in revised form 6 September 2023; Accepted 8 September 2023
Available online 11 September 2023

understanding. Compared to traditional image processing techniques, DNNs require a longer training time and demand more computing resources (Lai, 2019). On the other hand, their learning capabilities result in achieving higher accuracies and better generalization, while requiring no expert preprocessing, thanks to their inherent feature extraction (Jahanbakht et al., 2021; Lai, 2019).

The state-of-the-art studies around DNN-based image classification cover fish detection in clear waters (Iqbal et al., 2021; Shammi et al., 2021), controlled aquaculture farms (Lau and Lai, 2021), and out of water (Smadi et al., 2022). Fish detection in turbid underwater situations with limited vision is extremely difficult to impossible, in underwater video senses (Sheaves et al., 2016), and has thereby been mainly conducted by sonar imaging systems (Tarling et al., 2022). These sonar systems are expensive, with invasive ultrasonic radiations that can affect marine ecosystems (Pirotta et al., 2022). Furthermore, echo-sounder data labeling requires expert engagement, which will further increase the project costs. Although fish detection in turbid waters with visible light cameras faces serious challenges including low light, color distortion, obstructive suspended sediments, and high-dynamic underwater movements, when combined with deep learning, it can provide significant benefits compared to sonar systems (Jahanbakht et al., 2021).

To unlock these benefits new methods and algorithms need to be devised to efficiently work with underwater videos, including situations with low visibility. For example, both King et al. (2018) and Donaldson et al. (2019) reported the usability of various underwater camera deployments (i.e., floating/submerged, fixed/moving, and baited/not-baited) on surveying fish assemblages in turbid waters. While the former entirely relied on humans to count fish, the latter used traditional image processing techniques to increase video quality (color enhancement), before the human agents engaged in fish detection. None of these papers used advanced DNN techniques in their projects.

To enhance marine visual monitoring in turbid and low-visibility waters for improved aquatic ecological experiments, we propose a novel approach. As part of this approach, we first collect and present a new dataset named FishInTurbidWater. This dataset includes fish video data series collected within a major Australian shipping port facility, where water quality conditions can include high turbidity due to ship movements, strong tidal range (up to 8 m in a 6-h time period), and ocean currents that together can contribute to resuspension of benthic sediments into the water column in the region (Waltham et al., 2015).

We then perform a rapid and inaccurate labeling of fish presence in our video data frames, which results in a dataset for weak supervision. We then use the weakly labeled FishInTurbidWater to develop a semi-supervised contrastive learning model (CNT), as well as a new weakly-supervised ensemble DNN architecture with transfer learning from ImageNet. In the next step, we analyze our two models in terms of development time and accuracy. The proposed workflow is visually illustrated in Fig. 1, which covers data gathering to two independent image processing outcomes.

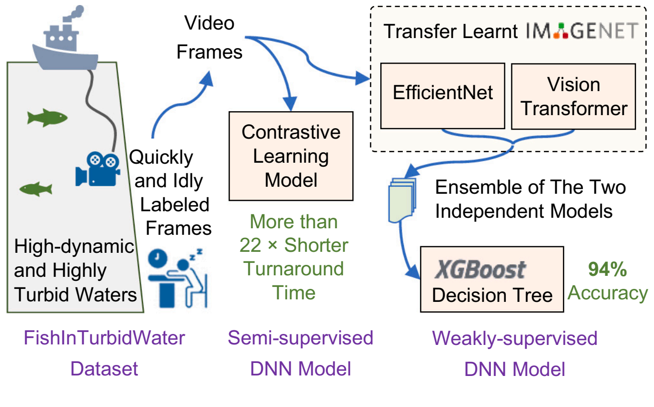

Video Frames
Transfer Learnt IMAGENET
EfficientNet
Vision Transformer
Contrastive Learning Model
Quickly and Idly Labeled Frames
Ensemble of The Two Independent Models
High-dynamic and Highly Turbid Waters
More than 22 × Shorter Turnaround Time
XGBoost Decision Tree
94% Accuracy
FishInTurbidWater Dataset
Semi-supervised DNN Model
Weakly-supervised DNN Model

Fig. 1. Graphical abstract of the proposed workflow from lazy image labeling to two state-of-the-art deep neural network designs with short turnaround time and high accuracy.

The term weakly-supervised in Fig. 1 refers to a machine learning model that is trained by a weakly labeled dataset (like our FishInTurbidWater). This contrasts with a fully-supervised model that is trained by a carefully labeled dataset. In the meantime, if a machine learning model is trained by no labeled data, it is called a self-supervised model. In state-of-the-art applications, a self-supervised model receives extra training from a fully-supervised dataset to increase its detection capacity. In this case, the resulting model is called a semi-supervised model. Finally, when we independently train multiple models and then use a dedicated model to merge their outputs into a single final output, the whole structure is called an ensemble model.

Using our proposed novel approach, we present some new results for deep learning in fish recognition. We demonstrate that our contrastive semi-supervised model is developed over five times faster than the weakly-supervised ensemble DNN, while using only a small fraction of weakly labeled FishInTurbidWater data. We also show that the ensemble model achieves 4.6% higher accuracy. In addition, we demonstrate that both our newly developed models can be developed much quicker compared to a fully-supervised approach, mainly due to the lower data annotation time they require. These results demonstrate that our proposed approaches can significantly facilitate the use of deep learning in aquatic ecology studies, e.g. those measuring fish abundance in turbid waters using underwater cameras.

To summarize, our specific contributions are as follows:

1. Collecting and presenting a new fish dataset in turbid waters.
2. Weakly labelling this dataset to provide the world's first weakly labeled fish dataset in turbid waters.
3. Developing two novel deep learning models to achieve two different goals.
   a. A semi-supervised model for accelerating model development speed from data collection to deployment.
   b. A weakly-supervised model for accelerating the development time, while also conserving the accuracy, compared to the fast semi-supervised model.

The remainder of this paper is organized as follows. Section 2 describes the weakly-supervised database and proposes two DNN structures as a novel solution to the problem of fish detection in turbid waters. We evaluate the accuracy of the proposed models, compare them with each other, and with the other fully-supervised models in Section 3. A detailed discussion of our findings is carried out in Section 4. The paper is concluded in Section 5.

2. Material and methods

This section starts by describing our labeled dataset and then proceeds by proposing two DNN models to enhance the capability of marine scientists in turbid underwater and coastal environment monitoring.

2.1. Weakly labeled public FishInTurbidWater dataset

Machine learning models learn and understand how to detect objects in a video frame, using essential training datasets that provide examples of the target in its various situations. In supervised learning, these datasets must be accurately labeled by human experts. However, it is not usually easy to obtain substantial amounts of underwater images with the object of interest being present. Besides, adding manual tags to each image is a time-consuming and tedious task that requires many hours in a laboratory, labeling images by a human. To avoid these difficulties, here, a weakly-supervised dataset, called FishInTurbidWater is created and made publicly available. To the authors' best knowledge, FishInTurbidWater is the first labeled image set of fish species in turbid underwater environments.

FishInTurbidWater includes realistic underwater videos captured by dropping a camera-array in two geolocations around the Port of Mackay, QLD, Australia. The camera-array was equipped with 15 waterproof cameras attached to a rope at 1 ~ m distance. In each drop, the cameras capture 60 min of MP4 video footage. The video frames are not only very low in visibility, but also shaky which is due to maritime waves, operator/boat movements, and occasionally dragging overwater or colliding with the seabed. For sample frames included in the dataset, see Fig. 2.

Nephelometric Turbidity Units (NTU) in our two data gathering Ports are recorded as part of a broader water quality monitoring program, which has been in place since 2014 (Waltham et al., 2021). As better illustrated in Fig. 3, the turbidity level during this period was variable between 1 and 105 NTU. More than 66% of the time, this NTU level is greater than 5.0 (local regional water quality guidelines for the protection of marine ecosystems is 1 NTU (Authority, 2010)).

To quickly label the dataset, an overall 1800 min of MP4 videos were 4× fast-forwarded and inattentively labeled by a human agent, resulting in a weakly labeled dataset with a binary label of fish present in the frame or no fish present. After this weak labeling method, we investigated the labeling accuracy of randomly sampled video frames. Some frames with fish in them were labeled as no fish, while some other frames without any fish present were labeled fish.

Our approach of labeling all the frames at nearly one-fourth of the careful labeling time resulted in addressing the putative problem of limited access to labeled data. However, the labeling is weakly done and its effect on model performance needs to be investigated. The dataset is available to the public and is shared at FishInTurbidWater for the benefit of the community.

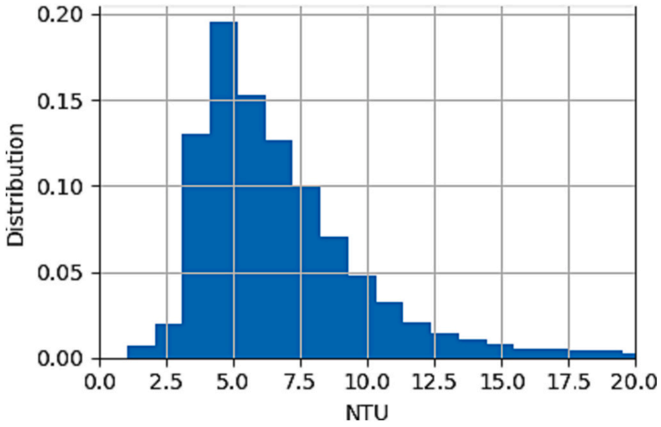


Fig. 3. Distribution plot of the measured Nephelometric Turbidity Units (NTU) in the same geolocation as our FishInTurbidWater dataset gathering.

2.2. *Semi-supervised contrastive learning*

To further reduce the impact of reliance on a high volume of accurately labeled data, we also opted to use another well-known technique in deep learning, i.e., contrastive learning. For this, we developed a two-phase semi-supervised contrastive learning approach as illustrated in Fig. 4. Our proposed model consists of a self-supervised contrastive learning phase (phase 1), followed by fully-supervised incremental fine-tuning learning (phase 2). It is worth noting that, for the full supervision training stage of our model, we use our weakly labeled fish dataset. This is in contrast to prior works that mostly use carefully labeled data for the

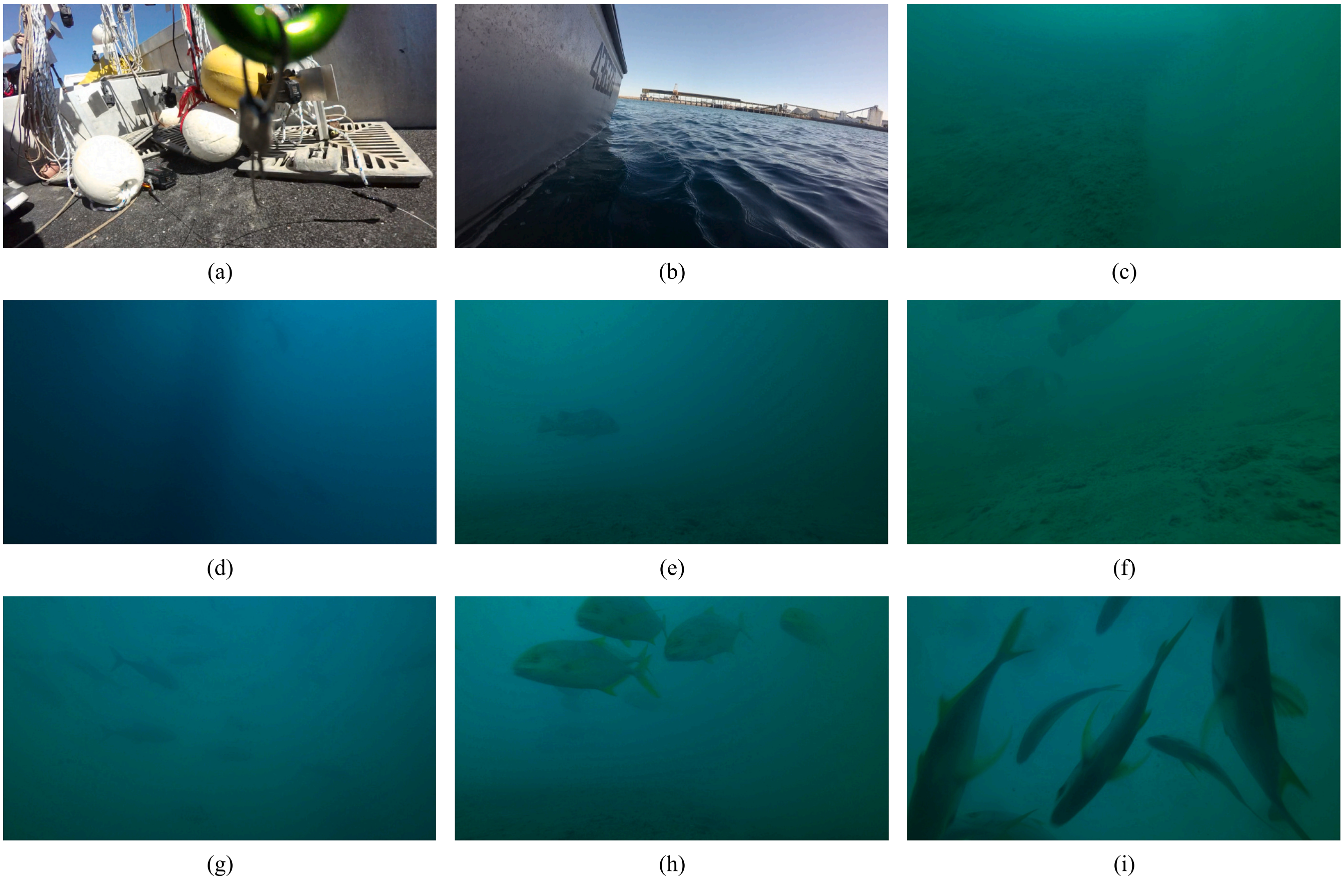


Fig. 2. FishInTurbidWater image samples, which includes (a) waiting onboard for deployment, (b) overwater port viewing, (c) seabed turbidity increasing by port activities, (d) wharf encountering, (e-f) seafloor visiting, and (g-i) lots of fish exhibiting.

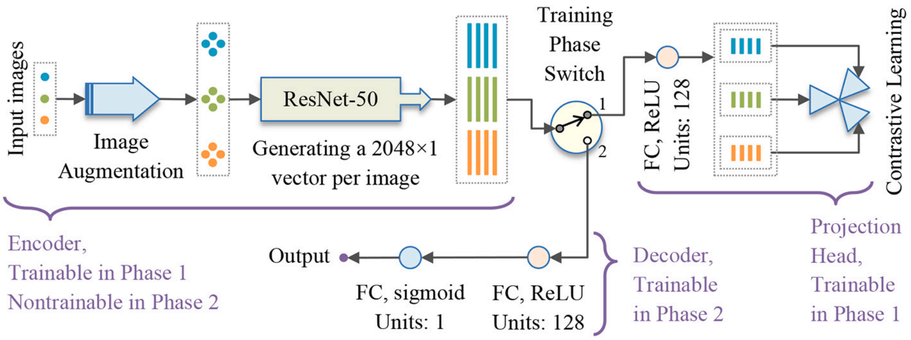

Input images
Image Augmentation
ResNet-50
Generating a 2048×1 vector per image
Training Phase Switch
1
2
FC, ReLU Units: 128
Contrastive Learning
Encoder, Trainable in Phase 1 Nontrainable in Phase 2
Output
FC, sigmoid Units: 1
FC, ReLU Units: 128
Decoder, Trainable in Phase 2
Projection Head, Trainable in Phase 1

Fig. 4. The proposed semi-supervised ResNet-based contrastive learning structure with two orderly training phases. The model trains from no labeled data during the self-supervised first phase and then fine-tunes with a small 20% subset of weakly labeled FishInTurbidWater data during the second phase.

fine-tuning stage (Saleh et al., 2022b).

Implemented using the Keras APIs under the TensorFlow framework, the first self-supervised contrastive learning phase composes of a trainable encoder, followed by a trainable projection head. The encoder takes a batch of input images and augments them with random horizontal flipping, random ±7^o rotation, and random ±40% brightness.

The augmented images are then input to a ResNet-50 model to be mapped from an RGB matrix into an encoded 2048 × 1 vector (He et al., 2016). This vector is then input to the subsequent projection head, which maps the encoded vector into its 128 × 1 representative vector. These vectors are then used in the self-supervised contrastive learning mechanism to learn features without labels (Khosla et al., 2020). In other words, the ResNet-50 will be trained to reduce the distance between in-group clusters of vector batches, while simultaneously pushing apart out-group clusters (Khosla et al., 2020).

After the successful completion of the first self-supervised phase, the model enters its second phase to undergo fully-supervised incremental learning. In this phase, the encoder's mode changes into non-trainable, the projection head switches off, and a new trainable decoder is attached to the encoder tail (see Fig. 4). The decoder's job is to map the encoded vector into a binary fish/no-fish answer. This fully-supervised setting allows very efficient use of limited label information (Khosla et al., 2020). In other words, the compact decoder with only 129 neurons can be easily trained by a fraction as small as 20% of the whole training dataset.

2.3. Weakly-supervised DNN ensemble

A second approach to counteract the problem of limited labeled data availability is to combine weak supervision (Laradji et al., 2021) and transfer learning. Transfer learning is a technique that provides an opportunity to ensure that we make the best use of available labeled data. For instance, in this work, we have used the open-source fully-labeled ImageNet dataset to pre-train two well-known DNNs, i.e., EfficientNet (Tan and Le, 2019) and ViT (Dosovitskiy et al., 2020), with a great capacity for image classification, while requiring fewer computational resources compared to other DNNs. ImageNet is a huge visual dataset with fourteen million images, designed for object recognition tasks (Lab, 2023). This free dataset has been hand-annotated with bounding boxes to indicate what objects are present and where in the images.

ImageNet consists of one thousand object classes, with tens of maritime categories, including goldfish, ray fish, jellyfish, spiny lobster, crayfish, grey whale, starfish, anemone fish, garfish, lionfish, pufferfish, sea.

snake, etc. (Lab, 2023). Pre-training our DNNs with this dataset means, they can distinguish between one thousand different objects in the first place. However, this is subject to having a clear vision of those objects, which is not the case in our highly turbid underwater environment. Therefore, after pre-training and transfer learning on ImageNet, DNNs must be retrained on realistic underwater videos, which in our case are weakly labeled.

We investigate how training on these weakly labeled data can improve the accuracy of our DNNs in detecting fish in turbid waters. Below, we discuss the two selected DNNs that were pre-trained using ImageNet to provide transfer learning for weakly-supervised training using our FishInTurbidWater dataset.

2.3.1. EfficientNet

The first model selected is EfficientNet-B7, which is a Convolutional Neural Network (CNN) with efficient scaling factors (Tan and Le, 2019). To elaborate, all CNNs are a cascade of convolutional layers that can scale the input image's resolution (width and height) and channel (number of color channels). CNN architectures can also scale by the number of consecutive convolutional layers (depth).

As illustrated in Fig. 5a, the main advantage of EfficientNet is that it can uniformly scale all depth, channel, and resolution dimensions by a simple yet effective compound rate (Tan and Le, 2019). This rate has been applied in nine stages with different repeat amounts. In other words, each stage in Fig. 5a is a Conv layer that repeats exactly for the given number of times. Changing this compound rate gives the DNN designer power to arbitrarily adjust the CNN's scaling, which in turn results in the currently existing B0 to B7 variants. Among these variants, the EfficientNet-B7 is proven to produce more accurate results (Tan and Le, 2019).

2.3.2. Vision transformer (ViT)

The second model of choice in this paper is the Vision Transformer (ViT). The original Transformer architecture has been widely used in natural language processing tasks. This was until Dosovitskiy et al. (2020) turned it into an accurate model in computer vision applications. As shown in Fig. 5b, the Transformer's attention mechanism is applied directly to the sequences of image patches in ViT (Jahanbakht et al., 2022a). This is shown to perform very well on image classification tasks (Dosovitskiy et al., 2020) and therefore has been used in our study.

2.3.3. Transfer learning and weak supervision

After pre-training of both EfficientNet and ViT using ImageNet, all their layers, except the last Fully Connected (FC) layer are frozen to transfer their learning of one thousand ImageNet classes. The last FC layer is then fine-tuned in a weakly-supervised fashion by our weakly labeled video dataset. The learning rate during the weak supervision process was kept as low as 0.0001 to minimize the undesirable effects of mislabeled frames. In other words, this lower learning rate prevents the FC layer from training instabilities, which can be imposed by bad labels (Gotmare et al., 2018).

2.3.4. XGBoost ensemble of DNNs

After successfully and independently training the two EfficientNet and ViT DNNs with ImageNet transfer learning followed by weak supervision using FishInTurbidWater, we merge their isolated single fish/

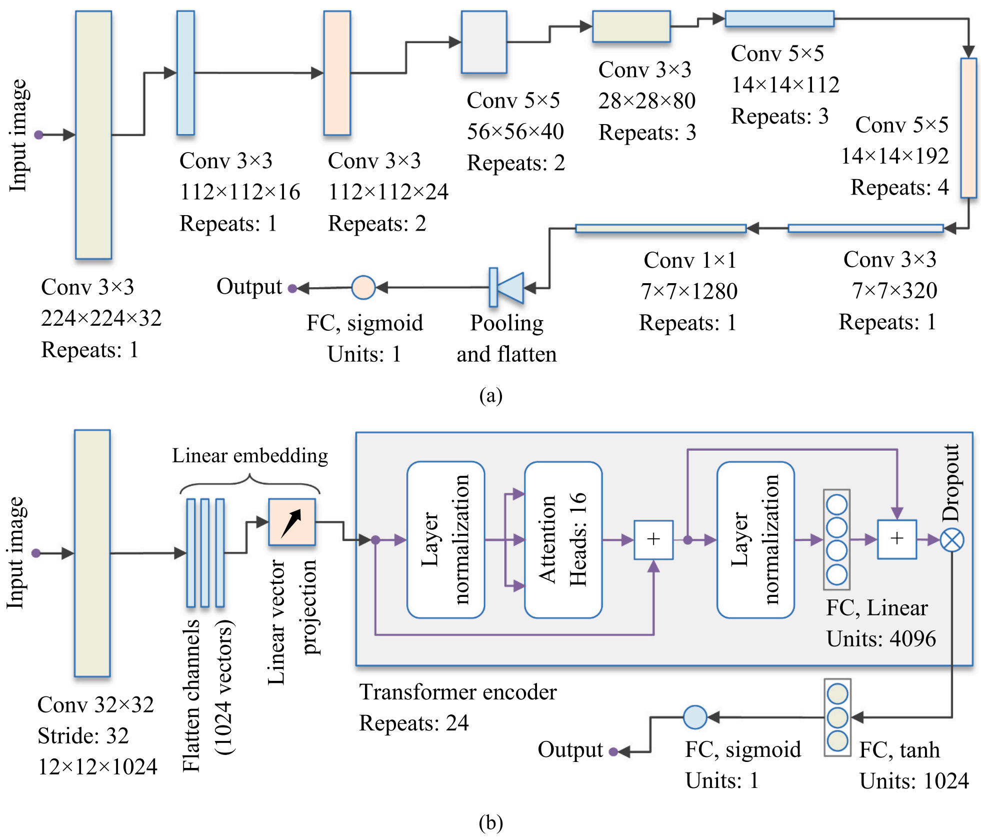


Fig. 5. The employed architecture of (a) EfficientNet and (b) ViT classifier DNNs. Each model takes a video frame at its input and returns the probability of fish existence. Conv, pooling, and FC layers are respectively referring to convolutional, max pooling, and fully connected layers. Besides, the dropout layer has used a 0.1 dropout ratio.

no-fish results into an answer. To this end, the novel XGBoost ensemble is used.

Gradient boosting refers to a class of ensemble Decision Tree (DT) algorithms that can be used in classification, regression, and ranking applications (Badirli et al., 2020). However, gradient boosting leaves the width and depth dimensions of its DTs, along with the number of DTs in the ensemble adjustable. This leads to a vast number of hyperparameters that cannot be easily optimized for a given application. In this regard, XGBoost was originally introduced by Chu (2023) as an efficient and distributed gradient boosting algorithm with rapidly optimized hyperparameters. XGBoost outperforms individual DNN models when it makes the final decision in their ensemble (Shwartz-Ziv and Armon, 2022). It is therefore beneficial to use an ensemble of DNNs with XGBoost, which can perform better than any individual model, as well as other classical ensemble techniques (Shwartz-Ziv and Armon, 2022).

As illustrated in Fig. 6, the XGBoost block consists of m distinct DT networks (i.e., T_1 to T_m). Each DT T_i receives three inputs, including the probability of fish existence by EfficientNet and ViT, along with the residual error r_{i-1} of the previous DT T_{i-1} in the row. The final output is calculated by the weighted sum DT outputs, as

$$F_{out} = \sum_{i=1}^{m} a_i T_i(X, r_{i-1}), \tag{1}$$

where a_i coefficients are the regularization parameters found by the XGBoost optimization algorithm.

3. Results

To evaluate the performance of all three DNNs explained in the previous Section, 30% of the whole FishInTurbidWater was carefully labeled manually to make a solid validation dataset. This validation dataset has never been exposed to these models during their training phase, and it has been used only for calculating the performance results presented in this Section.

As described in the previous Section, our weakly-supervised DNN ensemble structure consists of two weakly-supervised DNN models, i.e., EfficientNet and ViT, whose outputs are merged via an XGBoost

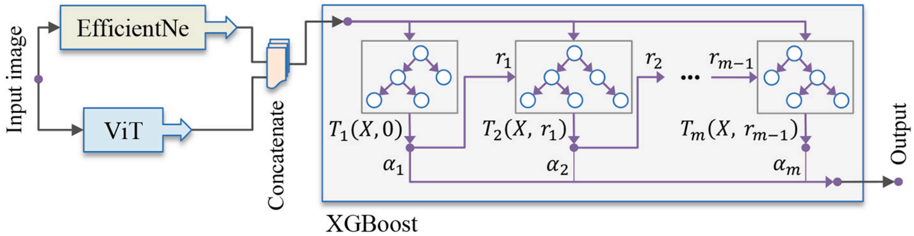


Fig. 6. The XGBoost ensemble structure, which concatenates the probability outputs of EfficientNet and ViT from Fig. 5a and Fig. 5b into an optimized network of m decision tree classifiers. The final binary output is the weighted sum of all decision trees' outputs.

ensemble. To demonstrate the impact of our proposed ensemble model, we investigated the individual performance of each model and compared it to the ensemble. The results are shown in Table 1.

Accuracy in Table 1 simply measures how often the classifier correctly predicts fish presence. Precision explains how many of the predicted fish frames turned out to contain fish. F1-score gives an idea about precision and sensitivity combination. In other words, it measures the harmonic mean of precision and sensitivity. Besides, the Area Under the Curve of Receiver Operating Characteristics (AUC-ROC) in this table is a performance measurement of a classifier against various probability thresholds. This metric indicates how certainly a classifier can distinguish between different classes. However, AUC-ROC is not available for XGBoost, as its class-separation mechanism is not based on probabilities. According to Table 1, XGBoost outperforms both the EfficientNet and ViT networks at all performance metrics showing the impact of our proposed ensemble technique.

The confusion matrix of our proposed DNNs, i.e., the first semi-supervised contrastive learning and the second weakly-supervised XGBoost ensemble are compared together in Fig. 7. The zero and one values in this matrix respectively represent fish absence and presence. Based on the confusion matrix, both models have close to 50% true-negative detections (i.e., 49.6% and 49.7%). However, the XGBoost Ensemble shows a better performance by halving the false-negative detections from 10.6% to 5.8%.

Additionally, the performance of both our semi-supervised and weakly-supervised models are compared in Table 2 with other recent publications in the literature. In this table, Sun et al. (2022) and Yu et al. (2023) used out-of-the-box RCNN and YOLO models, and they trained these models with clear-water datasets. On the other hand, Soom et al. (2022) designed their own DNN model, based on traditional CNNs, and they trained it in multiple water quality scenarios, including turbid waters.

According to Table 2, the XGBoost ensemble shows great performance at all three metrics, i.e., accuracy, precision, and F1-score. This better performance was achieved using a weakly labeled dataset, compared to fully labeled datasets, (Soom et al., 2022; Sun et al., 2022; Yu et al., 2023).

Finally, the time and accuracy trade-off in our proposed models is compared to that of a typical fully-supervised CNN (Deep and Dash, 2019) in Table 3. Even though we are training two independent DNNs in our proposed ensemble model, its turnaround time is more than 4× shorter than the conventional supervised networks.

With transfer learning from ImageNet and retraining on our weakly-supervised FishInTurbidWater dataset, our XGBoost ensemble model in Table 3 has comparable accuracy with a state-of-the-art fully-supervised CNN (Deep and Dash, 2019), which is trained on the clear-water DeepFish dataset. In the meantime, the proposed semi-supervised DNN has a noticeably short turnaround time, which is due to the self-supervised nature of contrastive learning.

To better illustrate the video qualities and the performance of the contrastive learning, EfficientNet, ViT, and the XGBoost ensemble in fish classification, some typical outputs of the DNN models are presented in Fig. 8. Here, a ✓ and a × respectively indicate fish or no-fish classification.

Table 1
Performance metrics of the two weakly-supervised DNN models (i.e., EfficientNet and ViT) and their XGBoost ensemble. These results are obtained using the carefully labeled FishInTurbidWater validation dataset.

Metric	EfficientNet	ViT	XGBoost
Accuracy	90.0%	93.0%	**94.0%**
Precision	90.3%	95.0%	**99.5%**
F1-score	0.893	0.925	**0.936**
AUC-ROC	0.974	0.953	N/A

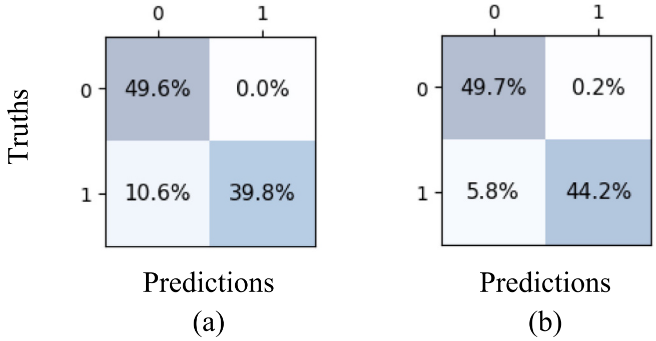


Fig. 7. Comparing the confusion matrix of our (a) Semi-supervised Contrastive Learning model with (b) the Weakly-supervised XGBoost Ensemble.

Table 2
Comparing both our semi-supervised contrastive learning DNN and the weakly-supervised XGBoost ensemble of DNNs with recent publications in the literature.

Metric	Semi-supervised Contrastive Learning	Weakly-supervised XGBoost Ensemble	(Sun et al., 2022)	(Yu et al., 2023)	(Soom et al., 2022)
Customized DNN	✓	✓	×	×	✓
Accuracy	89.4%	**94.0%**	N/A	N/A	91.6%
Precision	**99.8%**	99.5%	51%	90.2%	88.1%
F1-score	0.883	**0.936**	0.817	0.859	0.919
AUC-ROC	0.917	N/A	N/A	N/A	N/A

Table 3
Performance comparison between a fully-supervised DNN model, which is trained on a clear-water image dataset (Deep and Dash, 2019) and our two proposed XGBoost ensemble and contrastive learning DNNs, which are trained on our FishInTurbidWater dataset.

Metric	Semi-supervised CNT Learning (turbid water)	Weakly-supervised XGBoost Ensemble (turbid water)	Typical Supervised Deep Learning (clear-water)
Labeling Time	Very Short	Short	Very Long
Training Time	1.0 h	6.9 h	∼3.5 h
Turnaround Time	4 h	22 h	94 h
Accuracy	89.4%	94.0%	98%

4. Discussions

The development of deep learning methodologies continues to advance at an astonishing rate and be applied to various applications ranging from biomedical (Azghadi et al., 2020), hydrological processes in river channels (Talukdar et al., 2023) and agricultural (Olsen et al., 2019) systems, to marine (Laradji et al., 2021; Saleh et al., 2022b), and environmental (Jahanbakht et al., 2022a) sciences. The application of deep learning technologies has been also used in profiling the ecosystem services of estuarine habitats by community members (Yee et al., 2023). In this paper, we extend the application of deep learning methodologies to advance state-of-the-art underwater fish video processing techniques applied to turbid waters.

Processing fish images in turbid water has been addressed in previous literature. One study included the use of bait positioned close to the camera lens to attract fish to the camera for identification (Donaldson et al., 2019). Another study used underwater cameras equipped with clear liquid optical chambers to reduce light scatter that occurs when passing through turbid water (Jones et al., 2021). Further, Xu and Matzner (2018) examined the effects of water turbines on local fish species with low accuracy owing to local vagaries in conditions (bubbles

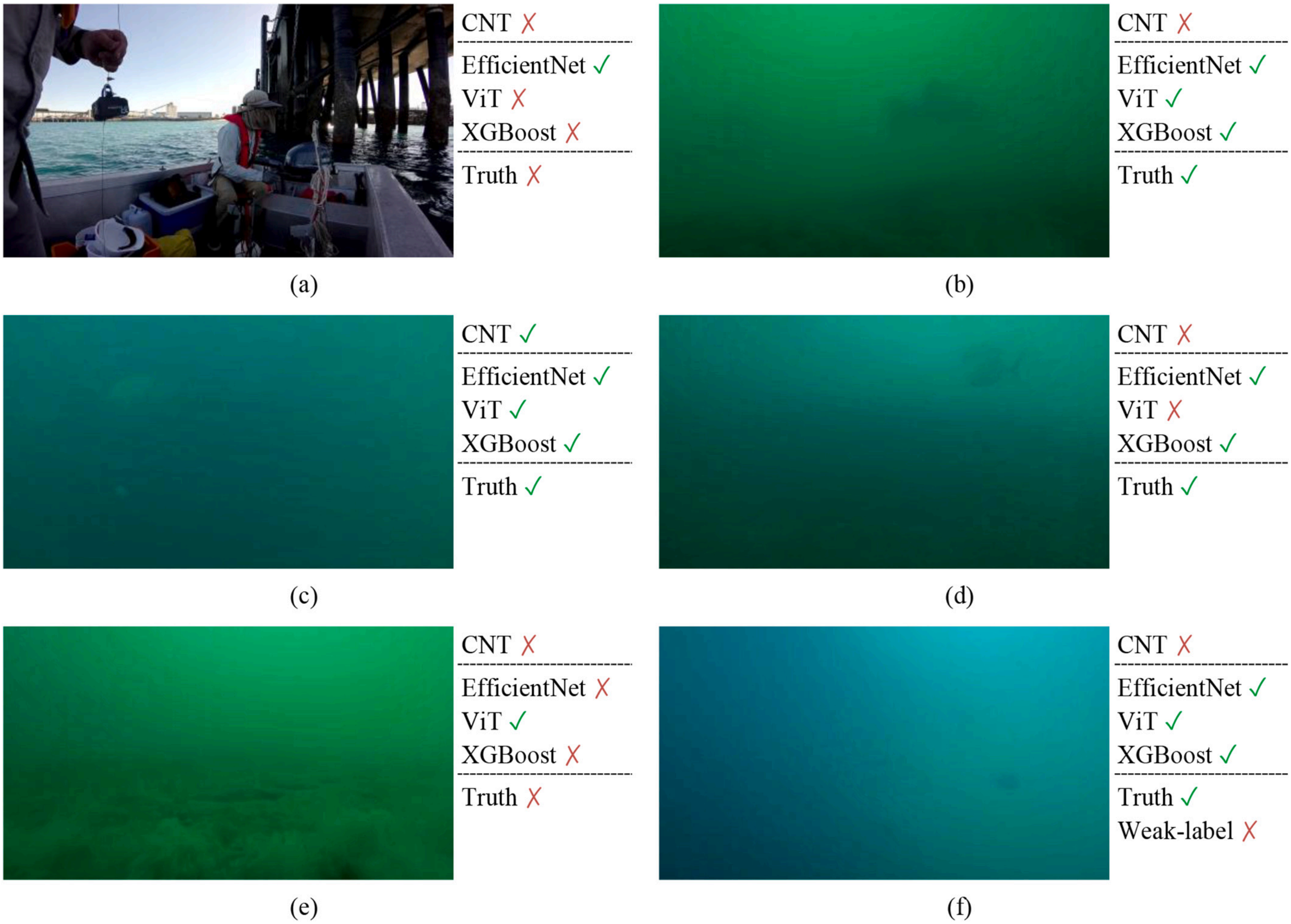


Fig. 8. The output of contrastive learning model (CNT), EfficientNet, ViT, and XGBoost ensemble are compared with the ground-truth for fish detection in some typical frames, taken from the entire dataset. In this example, XGBoost makes its own independent decision based on the best of EfficientNet and ViT outputs. It is worth noting that the video frame (f) was incorrectly labeled as no-fish in the initial weakly-supervised dataset.

from fast-flowing water and debris which interfered in fish observations). These studies mostly conclude that harsh environmental challenges such as water turbidity hinder the wide adoption of computer vision and deep learning technologies for fish habitat monitoring.

The need for improved ways to process video under conditions where the data has been compromised is necessary (Morán-López and Tolosa, 2023; Soom et al., 2022). For example, in situations where video images are affected by poor water conditions including high turbidity, the data has been disregarded with only video that occurs during times with more improved water clarity retained for statistical analysis. This not only increases costs for data collection but could also bias data collection to only periods when environmental conditions are more ideal (Donaldson et al., 2019). Our proposed methodologies enhance the capability of marine scientists who use underwater computer vision technologies in their turbid environment monitoring worldwide, which has previously been avoided due to impracticality. To address this challenge, we made the following contributions.

We collected the FishInTurbidWater dataset and quickly and weakly labeled it to contribute to the first weakly-supervised fish dataset in turbid waters. We, then used this dataset to develop two novel deep learning networks, one using semi-supervised contrastive learning for significantly accelerated model deployment time, and one weakly-supervised model to shorten deployment time, while providing high accuracy.

For semi-supervised contrastive learning, we first trained a self-supervised contrastive learning model and then fine-tuned it on our weakly labeled dataset. This semi-, weakly-supervised approach requires only a small portion of our weakly labeled dataset, which makes the development cycle from data labeling to final model inferencing very fast (i.e., 4 h) while producing a relatively high accuracy of 89%.

For the ensembling approach, we first trained two state-of-the-art DNNs on ImageNet. We then performed transfer learning of these two DNNs on our weakly labeled dataset. This generated two weakly-supervised DNNs, which we ensembled using the XGBoost technique. This novel ensemble technique significantly improves the overall accuracy compared to the two weakly-supervised individual DNNs.

Despite not having access to a control clear water dataset to run our two proposed semi- and weakly-supervised methods on, we have compared them to a state-of-the-art fully-supervised method, which is for clear water. Compared to a fully-supervised underwater fish classification model that needs many hours of turnaround time from dataset collection to model deployment (as shown in Table 3**Table 3**), both our approaches are significantly faster, while providing slightly reduced accuracies. This suggests a trade-off between development time and budget, and the accuracy required.

We show that our XGBoost ensemble model outperforms other recent publications in the literature (Soom et al., 2022; Sun et al., 2022; Yu et al., 2023). This ensemble model shows very high True Positive Rate (TPR) and True Negative Rate (TNR) of 0.88 and 0.99, while providing low False Negative Rate (FNR) and False Positive Rate (FPR) of 0.11 and 0.01. In this regard, TPR (i.e., sensitivity, recall, or hit rate) refers to the fraction of video frames with fish prediction, conditioned on fish being truly present. TNR (i.e., specificity and selectivity) is the portion of the no-fish frames that have been correctly classified. Conversely, FNR or miss rate and FPR or fall-out are respectively calculating the portion of fish and no-fish frames that were incorrectly classified. Sensitivity, specificity, miss rate, and fall-out are the most used metrics to measure the true and false classification probabilities.

We also show that our semi-weakly-supervised model can be developed roughly 23.5 times faster than a fully-supervised DNN, at the cost of nearly a 9% drop in its accuracy. This tradeoff is 4.3 times faster development for a 4% accuracy degradation. The methods presented in this paper, therefore, can assist marine scientists and environmental managers in fast and improved fish detection and monitoring in turbid water conditions.

It is worth noting that the typical operation of fully-supervised models in Table 3 is extracted from a recent survey paper by Saleh et al. (2022b), which is mainly focused on clear-water scenarios. Therefore, a more accurate comparison between our semi- and weakly-supervised models with a fully-supervised deep learning in both clear and turbid waters is required. We believe that applying the clear-water models of Saleh et al. (2022b) in turbid water situations (like our FishInTurbidWater dataset) would dramatically degrade their performance.

Besides, the approximate turnaround time in Table 3 is measured against a non-experienced human agent. Here, we added an extra 30% to the labeling time of fully-supervised networks to consider the necessary double-checks and quality controls. Overall, these numbers are only rough estimates of the required time, and they heavily depend on and vary with the human agent's circumstances. For instance, one human agent can eventually acquire experience, leading to faster labeling.

Although the proposed techniques in this paper are for binary fish video frame/image classification, future research could investigate the development of similar semi- and weakly-supervised deep learning techniques for fish species classification, and/or other applications such as fish counting, segmentation, and localization (Saleh et al., 2022a). These may need the collection and processing of new datasets in turbid waters.

In addition, while the proposed models have been successfully tested on data collected in turbid underwater conditions in two geolocations around the Port of Mackay in Australia, future studies can employ them for underwater fish video classification in other turbid waters as well.

In summary, the methodology we have introduced in this paper holds considerable promise for the ecological informatics community. It has the potential to significantly alleviate the financial and labor-intensive aspects of annotating data, a critical step in developing deep learning models. By doing so, it promises to expedite the progress of fish recognition models designed for challenging turbid water conditions. Ultimately, this could make deep learning a more viable and valuable tool for a wider range of aquatic ecology researchers.

5. Conclusion

The problem of fish detection in turbid underwater video frames was addressed by collecting a new dataset. The usually slow dataset labeling process was sped up by a factor of four, resulting in a quick compilation but also occasional incorrect labels. This rapid/weak labeling is known to influence the accuracy of traditional DNNs. To mitigate this problem, two different weak supervision approaches, one using contrastive learning and the other one using an XGBoost ensemble, were proposed. The contrastive learning model was self-supervised with no labeled data but was then fine-tuned in a semi- and weak-supervised manner, using only 20% of our weakly labeled dataset. This resulted in 23.5 times faster deployment time, compared to a fully-supervised model, while lowering the accuracy by 9%. The ensemble model, on the other hand, suffered only an accuracy loss of 4% compared to the fully-supervised model, while being 4.3 times faster in development. The result of this work can facilitate developing fast and efficient fish abundance and surveying applications in turbid underwater videos, assisting in coastal ecosystem management and decision making.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The FishInTurbidWater underwater video dataset is collected at the Port of Mackay, QLD, Australia, and it is weakly labeled in Centre for Tropical Water and Aquatic Ecosystem Research (TropWATER), James Cook University. Both the semi-supervised contrastive learning and the weakly-supervised DNN ensemble are implemented by the Keras APIs of TensorFlow framework in Python 3.9, and they are accessible through the same FishInTurbidWater link.

Acknowledgments

We thank the First Nation (Australian Aboriginal and Torres Strait Islander) people on whose land this research was completed and pay our respect to their leaders past, present, and emerging. We thank Mackay Marina for access to the port waters to deploy the underwater cameras, North Queensland Bulk Ports for funding this research as part of North Queensland Bulk Ports/James Cook University Partnership (2020-2025), and Dr. Bradly (JCU) who assisted with fieldwork. Underwater video surveys were completed in accordance with Ethics Approval (JCU A2615).

References

Aguzzi, J., Chatzievangelou, D., Marini, S., Fanelli, E., Danovaro, R., Flögel, S., Lebris, N., Juanes, F., De Leo, F.C., Del Rio, J., Thomsen, L., Costa, C., Riccobene, G., Tamburini, C., Lefevre, D., Gojak, C., Poulain, P.-M., Favali, P., Griffa, A., Purser, A., Cline, D., Edgington, D., Navarro, J., Stefanni, S., D'Hondt, S., Priede, I.G., Rountree, R., Company, J.B., 2019. New high-tech flexible networks for the monitoring of deep-sea ecosystems. Environ. Sci. Technol. 53, 6616–6631.

Authority, G.B.R.M.P., 2010. Water Quality Guidelines for the Great Barrier Reef Marine Park. Great Barrier Reef Marine Park Authority.

Azghadi, M.R., Lammie, C., Eshraghian, J.K., Payvand, M., Donati, E., Linares-Barranco, B., Indiveri, G., 2020. Hardware implementation of deep network accelerators towards healthcare and biomedical applications. IEEE Trans. Biomed. Circ. Syst. 14, 1138–1159.

Badirli, S., Liu, X., Xing, Z., Bhowmik, A., Doan, K., Keerthi, S.S., 2020. Gradient boosting neural networks: Grownet. arXiv 07971 1–6.

Chu, T., 2023. Story and Lessons Behind the Evolution of XGBoost. https://sites.google.com/site/nttrungmtwiki.

Deep, B.V., Dash, R., 2019. Underwater fish species recognition using deep learning techniques. In: Proc. 6th International Conference on Signal Processing and Integrated Networks (SPIN, Noida, India), pp. 665–669.

Ditria, E.M., Connolly, R.M., Jinks, E.L., Lopez-Marcano, S., 2021. Annotated video footage for automated identification and counting of fish in unconstrained seagrass habitats. Front. Mar. Sci. 8, 629485–629490.

Donaldson, J.A., Drews, P., Bradley, M., Morgan, D.L., Baker, R., Ebner, B.C., 2019. Countering low visibility in video survey of an estuarine fish assemblage. Pac. Conserv. Biol. 26, 190–200.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., others, 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 1–22.

Dutta, M.K., Arhonditsis, G., 2023. Empowering novel scholarship at the intersection of machine learning/deep learning and ecology. Eco. Inform. 102249.

Gotmare, A., Keskar, N.S., Xiong, C., Socher, R., 2018. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. arXiv 1–15.

Harvey, E.S., Newman, S.J., McLean, D.L., Cappo, M., Meeuwig, J.J., Skepper, C.L., 2012. Comparison of the relative efficiencies of stereo-BRUVs and traps for sampling tropical continental shelf demersal fishes. Fish. Res. 125, 108–120.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. In: Proc. 14th European Conference on Computer Vision (ECCV), Amsterdam, Netherlands, pp. 630–645.

Heggie, K., Ogburn, M.B., 2021. Rapid video assessment detects qualitative differences in oyster reef habitat. Mar. Ecol. Prog. Ser. 667, 219–224.

Iqbal, M.A., Wang, Z., Ali, Z.A., Riaz, S., 2021. Automatic fish species classification using deep convolutional neural networks. Wirel. Pers. Commun. 116, 1043–1053.

Jahanbakht, M., Xiang, W., Hanzo, L., Rahimi Azghadi, M., 2021. Internet of underwater things and big marine data analytics – a comprehensive survey. IEEE Commun. Surv. Tutor. 23, 904–956.

Jahanbakht, M., Xiang, W., Azghadi, M.R., 2022a. Sediment prediction in the great barrier reef using vision transformer with finite element analysis. Neural Netw. 152, 311–321.

Jahanbakht, M., Xiang, W., Waltham, N.J., Azghadi, M.R., 2022b. Distributed deep learning and energy-efficient real-time image processing at the ddge for fish segmentation in underwater videos. IEEE Access 10, 117796–117807.

Jones, R.E., Unsworth, R.K.F., Hawes, J., Griffin, R.A., 2021. Improving benthic biodiversity assessments in turbid aquatic environments. Aquat. Conserv. Mar. Freshwat. Ecosyst. 31, 1379–1391.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D., 2020. Supervised contrastive learning. Adv. Neural Inf. Proces. Syst. 33, 18661–18673.

King, A.J., George, A., Buckle, D.J., Novak, P.A., Fulton, C.J., 2018. Efficacy of remote underwater video cameras for monitoring tropical wetland fishes. Hydrobiologia 807, 145–164.

Lab, S.V., 2023. ImageNet of Stanford University. https://www.image-net.org/.

Lai, Y., 2019. A comparison of traditional machine learning and deep learning in image recognition. In: Proc. 3rd International Conference on Electrical, Mechanical and Computer Engineering (ICEMCE), Guizhou, China, pp. 12148–12156.

Laradji, I.H., Saleh, A., Rodriguez, P., Nowrouzezahrai, D., Azghadi, M.R., Vazquez, D., 2021. Weakly supervised underwater fish segmentation using affinity LCFCN. Sci. Rep. 11.

Lau, P.Y., Lai, S.C., 2021. Localizing fish in highly turbid underwater images. In: Proc. International Workshop on Advanced Imaging Technology (IWAIT), Kagoshima, Japan, pp. 61–69.

McIvor, A.J., Spaet, J.L.Y., Williams, C.T., Berumen, M.L., 2022. Unoccupied aerial video (UAV) surveys as alternatives to BRUV surveys for monitoring elasmobranch species in coastal waters. ICES J. Mar. Sci. 79, 1604–1613.

Morán-López, R., Tolosa, O.U., 2023. Modelling dynamic fish-waterfall interactions with digital image field data: from rescaling to weir removal for migratory freshwater fish. Eco. Inform. 77, 102183.

Olsen, A., Konovalov, D.A., Philippa, B., Ridd, P., Wood, J.C., Johns, J., Banks, W., Girgenti, B., Kenny, O., Whinney, J., others, 2019. DeepWeeds: a multiclass weed species image dataset for deep learning. Sci. Rep. 9, 1–12.

Pirotta, E., Booth, C.G., Calambokidis, J., Costa, D.P., Fahlbusch, J.A., Friedlaender, A.S., Goldbogen, J.A., Harwood, J., Hazen, E.L., New, L., Santora, J.A., Watwood, S.L., Wertman, C., Southall, B.L., 2022. From individual responses to population effects: Integrating a decade of multidisciplinary research on blue whales and sonar. Anim. Conserv. 25, 796–810.

Saleh, A., Sheaves, M., Jerry, D., Azghadi, M.R., 2022a. Applications of Deep learning in fish habitat monitoring: a tutorial and survey, pp. 1–26 arXiv:2206.05394v1.

Saleh, A., Sheaves, M., Rahimi Azghadi, M., 2022b. Computer vision and deep learning for fish classification in underwater habitats: a survey. Fish Fish. 23, 977–999.

Shammi, S.A., Das, S., Hasan, M., Noori, S.R.H., 2021. FishNet: Fish classification using convolutional neural network. In: Proc. 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, pp. 1–5.

Sheaves, M., Johnston, R., Baker, R., 2016. Use of mangroves by fish: new insights from in-forest videos. Mar. Ecol. Prog. Ser. 549, 167–182.

Shwartz-Ziv, R., Armon, A., 2022. Tabular data: deep learning is not all you need. Inf. Fusion 81, 84–90.

Smadi, A.A.L., Mehmood, A., Abugabah, A., Almekhlafi, E., Al-smadi, A.M., 2022. Deep convolutional neural network-based system for fish classification. International. J. Electr. Comput. Eng. 12.

Soom, J., Pattanaik, V., Leier, M., Tuhtan, J.A., 2022. Environmentally adaptive fish or no-fish classification for river video fish counters using high-performance desktop and embedded hardware. Eco. Inform. 72, 101817.

Sun, H., Yue, J., Li, H., 2022. An image enhancement approach for coral reef fish detection in underwater videos. Eco. Inform. 72, 101862.

Talukdar, G., Bhattacharjya, R.K., Sarma, A.K., 2023. Understanding the effect of long term and short term hydrological components on landscape ecosystem. Eco. Inform. 77, 102267.

Tan, M., Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proc. 36th International Conference on Machine Learning, Long Beach, USA, pp. 6105–6114.

Tarling, P., Cantor, M., Clapes, A., Escalera, S., 2022. Deep learning with self-supervision and uncertainty regularization to count fish in underwater images. PLoS One 17, 759–765.

Waltham, N., McKenna, S., York, P., Devlin, M., Campbell, S., Rasheed, M., Da Silva, E., Petus, C., Ridd, P., 2015. Port of Mackay and Hay point ambient marine water quality monitoring program (July 2014 to July 2015). In: Centre for Tropical Water and Aquatic Ecosystem Research (TropWATER), pp. 1–96.

Waltham, N.J., Iles, J.A., Johns, J., 2021. Port of Mackay and Hay point ambient marine water quality monitoring program: annual report 2020-2021. Centre for Tropical Water and Aquatic Ecosystem Research (TropWATER) 1–81.

Whitfield, A.K., 2017. The role of seagrass meadows, mangrove forests, salt marshes and reed beds as nursery areas and food sources for fishes in estuaries. Rev. Fish Biol. Fish. 27, 75–110.

Xu, W., Matzner, S., 2018. Underwater fish detection using deep learning for water power applications. In: Proc. International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, USA, pp. 313–318.

Yee, S.H., Sharpe, L.M., Branoff, B.L., Jackson, C.A., Cicchetti, G., Jackson, S., Pryor, M., Shumchenia, E., 2023. Ecosystem services profiles for communities benefitting from estuarine habitats along the Massachusetts coast, USA. Eco. Inform. 77, 102182.

Yu, G., Cai, R., Su, J., Hou, M., Deng, R., 2023. U-YOLOv7: a network for underwater organism detection. Eco. Inform. 75, 102108.