



Fully Documented Fisheries

Final report

Author(s): A.T.M. van Helmond (Ed.)

Wageningen University &
Research report: C076/23

Fully Documented Fisheries

Final report

A.T.M. van Helmond (Ed.)

Wageningen Marine Research
IJmuiden, November 2023

CONFIDENTIAL no

Wageningen Marine Research report C076/23

Keywords: **fisheries, monitoring, machine learning**

Client: Ministry of Agriculture, Nature and Food Quality
Attn.: Mw. E. Smith
Postbus 20401
2500 EK, Den Haag

This report can be downloaded for free from <https://doi.org/10.18174/642652>
Wageningen Marine Research provides no printed copies of reports

Wageningen Marine Research is ISO 9001:2015 certified.

© Wageningen Marine Research

Wageningen Marine Research, an institute within the legal entity Stichting Wageningen Research (a foundation under Dutch private law) represented by
Drs.ir. M.T. van Manen, Director Operations

KvK nr. 09098104,
WMR BTW nr. NL 8113.83.696.B16.
Code BIC/SWIFT address: RABONL2U
IBAN code: NL 73 RABO 0373599285

Wageningen Marine Research accepts no liability for consequential damage, nor for damage resulting from applications of the results of work or other data obtained from Wageningen Marine Research. Client indemnifies Wageningen Marine Research from claims of third parties in connection with this application.
All rights reserved. No part of this publication may be reproduced and / or published, photocopied or used in any other way without the written permission of the publisher or author.

A_4_3_2 V32 (2021)

Contents

Preface	4
Summary	5
1 General Introduction	7
2 Automatic discard registration in cluttered environments using deep learning and object tracking: class imbalance, occlusion, and a comparison to human review	10
3 An integrated end-to-end deep neural network for automated detection of discarded fish species and their weight estimation	11
4 Technical report	12
4.1 Overview	12
4.2 Image-Acquisition System	13
4.2.1 General	13
4.2.2 Mechanical Design	14
4.2.3 Line Scan Design	16
4.2.4 Parallel Framework	17
4.3 Image-Processing System	18
4.4 Communication System	19
5 Multi-stage image-based approach for fish detection and weight estimation	20
6 Improving automated discards registration using active learning technique for object detection	21
7 General discussion	22
References	25
Justification	28

Preface

This report combines different outputs from the Fully Documented Fisheries (FDF) project during the period 2019 and 2023, including a project background (chapter 1, General introduction), two peer reviewed scientific publications on the main project results (chapters 2 and 3), a technical report of the developed camera-system (chapter 4), two scientific draft manuscripts (work in progress) on the latest project results (chapter 5 and 6) and a general discussion (chapter 7) on the completion of the project.

Chapter 1, the General introduction describes the project background, providing context to the research questions, and further explanation for the need of the FDF-project from a fisheries management point of view.

Chapter 2 describes a method for detecting and counting demersal fish species in complex, cluttered, and occluded environments that can be installed on the conveyor belts of fishing vessels. Fishes on the conveyor belt were recorded using a colour camera (RGB) and were detected using a deep neural network. To improve the detection, synthetic data was generated for rare fish species. The fishes were tracked over the consecutive images using a multi-object tracking algorithm, and based on multiple observations, the fish species was determined. The results were compared with human EM review, a weighted counting error of 20% was achieved, compared to a counting error of 7% for human EM review on the same recordings.

Chapter 3 explains an integrated end-to-end simultaneous detection and weight prediction method for fishes as they appear on the sorting belt of fishing vessels based on the state-of-the-art deep convolutional neural network. The performance of the network was evaluated per species and under different occlusion levels. Additionally, a new dataset is presented containing images of fish, which, unlike common object detection datasets, also contains weight measurements and occlusion level per individual fish. For the used dataset, the model was able to detect discards with a macro F1 -score of 94.1 % and a weighted F1 -score of 93.9 %. On average the weight could be estimated with a mean absolute error of 29.7 grams.

Chapter 4 provides a technical report of the camera-system developed within the project.

Chapter 5 describes a novel multi-stage machine learning approach for fish detection and weight estimation in catches in Dutch demersal trawlers. Experimental results demonstrated a F1 score of more than 95% for species identification along with a mean absolute error of 27.183 grams in weight estimation.

Chapter 6 presents the results of an active machine learning technique. Based on certainty estimations the most informative training images can be selected and used to fine-tune data sets for model training. First results demonstrated, that our approach allows reaching the same object detector performance as with the random sampling with 400 fewer labelled images.

Chapter 7 is a general discussion of the study as a whole, providing the context of the completion of the project.

Summary

This report provides a background, overview and general discussion of the project 'Fully Documented Fisheries' (FDF) funded by the European Maritime and Fisheries Fund (EMFF).

Over the last decade Electronic Monitoring (EM) emerged as a promising monitoring tool for commercial fisheries. In general, EM systems consist of various activity sensors, a GPS, computer hardware and multiple cameras, which allow for a detailed and complete catch registration, including the discarded part of the catch. The constant video surveillance of the autonomously operating EM-system increases monitoring coverage in space and time, resulting in significantly improved quality of catch and effort information, e.g. increased representativeness and reduced bias, without requiring additional on-board personnel. However, species identification and catch estimation through manual review of video data can be very labour intensive and time consuming and withholds a wider implementation of EM in fisheries management. Automating the video review process is a possible solution to reduce the workload of manual review and facilitate processing large amounts of video data, and, eventually, the implementation of EM on a larger scale. Within this study we investigated the possibilities to automate the process of video review in the Dutch beam trawl fishery. For the purpose of data collection a camera-system was developed. The first prototype consisted of a metal box fixed on a conveyor belt. The sealed box, protects the camera from salt, humidity, and allows for optimal illumination and avoiding the presence of the personnel inside the field of view of the camera itself. The Algorithms performing data analysis, i.e. species detection and weight estimation, are deployed on a separate computer. The results indicate that there is a large potential for the application of computer vision technology to automatically register catch on board commercial fishing vessels. A key element in successfully implement image detection algorithms to automatically register weight and species was to control the external factors that have a negative effect on image quality, e.g. dirty camera lenses, crew blocking the view, reflection by illumination. Eventually, a camera-system based on innovative line scan technology above the catch sorting conveyor belt seems to be the best possible solution to automatically register discards on the Dutch beam trawlers. Recognizing that a line scan camera is not part of a standard onboard camera system, integration with commercial EM technology is an essential next step to enable a wider use of the developed technology. Future research should also focus on a wider implementation in the European fleet, possibly include demersal trawlers from other EU members states, i.e. Denmark and Belgium.

Key findings:

- To successfully monitor catches under the landing obligation there is a need for better methods and technologies to improve the application of EM. The implementation of computer vision technology is the logical next step in the innovation process of EM. Automizing the video review process is a possible solution to enable processing large amounts of video data (chapter 1).
- Deep neural network approaches (machine learning applications) can successfully detect, fish species and predict weights of discarded catch under challenging situations, e.g. cluttered and occluded catch on conveyor belts. However, detection rates and accuracy in weight prediction decreases with an increased level of occlusion (chapter 2 and 3).
- A key element in successfully implementing image detection algorithms to automatically register weight and species was to control the external factors that have a negative effect on image quality, e.g. dirty camera lenses, crew blocking the view, reflection by illumination (chapter 4).

-
- There is a large variation in the catch composition and appearance of the fishes over the season and at different fishing grounds, challenging the detection performance. It is important to gather a large and varied data set to properly train the deep neural network. Active learning allows to select training samples more effectively than by random sampling (chapter 6). When new species can occur, or fish can appear in completely new configurations, a generic-fish detector is more effective than a method that detects and classifies the fish instances at once (chapter 5).
 - To deal with the large amount of variation in practical settings, future work, needs to further develop methods for continuous and active learning. Furthermore, new AI methods for data augmentation, domain adaptation, self-supervised learning and the use of foundation models need to be explored to further improve generalization with limited availability of annotated data (chapter 7).
 - To deal with the challenge of occlusion (chapter 2 and 3), it is advisable to not only look at advanced AI methods, but to also explore options to mechanically spread out the catch to lower the level of occlusion (chapter 7).
 - The FDF-technology should be prepared for a wider implementation in the European fleet, possibly include demersal trawlers from other EU members states, i.e. Denmark and Belgium, in future research (chapter 7).

1 General Introduction

Over the last decades Electronic Monitoring (EM) has emerged as a successful and cost efficient technology to improve catch monitoring programmes of fisheries around the world (van Helmond et al., 2020). Through enhanced registration of fishing effort and location and the ability of a 100 % observing coverage, EM has the potential to provide improved, and more representative, coverage of fishing activities compared to any conventional monitoring methods. In general, an EM system consists of various activity sensors, a GPS, computer hardware and multiple cameras, which allow for detailed recording of fishing information, i.e. catch and effort data (McElderry et al., 2003; Ames et al 2005). The ability of EM systems to operate remotely creates a substantial advantage of EM over the traditional at-sea observer programmes. The constant surveillance of an EM-system increases monitoring coverage in space and time, resulting in significantly improved quality of catch and effort information, e.g. increased representativeness and reduced bias, without requiring additional on-board personnel (Kindt-Larsen et al., 2011; Needle et al., 2015; Ulrich et al., 2015; Mortensen et al., 2017). Already in 1999, after a first trial with camera systems on board fishing vessels, to cope with management reforms and gear theft in British Columbia, Canada, it was quickly recognized that cameras on board fishing vessels could be used for monitoring catches in fisheries. Since then, a long series of studies around the world proved the potential of EM to improve data collection for fisheries management (van Helmond et al., 2020). However, in spite of the obvious advantages of EM, the uptake of EM in Europe so far has remained low. The fishers consider EM an intrusion in their private workspace (Baker et al., 2013; Plet-Hansen et al., 2017) and argue that camera surveillance reflects a governmental mistrust against them (Mangi et al., 2015). Nevertheless, the European Commission and the European Fisheries Control Agency (EFCA) put EM forward as the best possible candidate to control the landing obligation of the European Union (EU), making the implementation of EM in European fisheries a controversial subject (the European Commission's proposal to revise the fisheries control system, centred on the amendment of the Control Regulation 1224/2009).

Besides the strong resistance to use cameras on board for surveillance from the fisher's point of view, there are also several practical challenges for the implementation of EM in the context of the European landing obligation. Continuous EM monitoring on board a fishing fleet results in large quantities of video data. To implement an EM programme on this scale considerable costs are needed for video processing, transfer and review. Processing and storing large amounts of video data requires large IT infrastructures, and transmitting significant volumes of video data from vessels at-sea through satellite connection is constrained by high costs. Currently, video footage is transferred by physically mailing or collecting hard drives or, in the best case scenario, through wireless transfer over mobile networks from the harbour (Michelin et al., 2018; van Helmond et al., 2020). Species identification and catch estimation through manual review of video data can be very labour intensive and time consuming. Particularly in fisheries with mixed catches of similar looking species, the capacity of human resources needed to review footage of the catch is substantial (Needle et al., 2015; van Helmond et al., 2015; Plet-Hansen et al., 2019).

To reduce the effort needed on video review in EM programmes, the so called 'audit-approach' is often implemented. With this method only a small sample of the collected footage is reviewed. A random check, often between 10-20% of the total amount of video footage collected is validated against self-recorded catch data by fishers (van Helmond et al., 2020). Even though only a minority of these reports are audited with video, the fishers do not know which hauls will be audited and when, which still creates an incentive to report all catches accurately (James et al., 2019). Another challenge for EM implementation under the EU landing obligation is the ability to quantify the catch, in particular the discarded part of the catch. Different procedures have been used for estimating catch quantities from EM footage in different studies. In general, video reviewers attempt to estimate discards from footage that is collected during the sorting process on board, when the catch is displayed on the sorting belt. However, this method can be challenging when estimating large volumes of catch. Estimating the quantity of a species in highly mixed catches is often less accurate compared

to quantity estimations in clean catches, i.e. not mixed with other species (van Helmond et al., 2015). This study suggested that distinguishing a specimen in large volumes of bycatch, particularly when similar-looking species are targeted in mixed fisheries, could be difficult. Also, other studies found a tendency of EM in underestimating discards and smaller fish (e.g. sole below 24 cm) compared with on-board observations (van Helmond et al., 2017; Mortensen et al., 2017). The accuracy of video observation should be monitored and improved where needed (Needle et al., 2015; Ulrich et al., 2015; van Helmond et al., 2020).

In an attempt to improve catch quantification of EM the implementation of protocols to improve the display of catch in front of the cameras were investigated. One approach, used in an EM trial in Denmark, required the crew to sort discards into baskets and show the baskets to the cameras before discarding, making it possible for video reviewers to estimate discard quantities by counting the number of baskets collected by the fishers (Ulrich et al. 2015). In another study, the crew was requested to display undersized sole on the sorting belt after the complete catch was processed (van Helmond et al., 2017). The implementation of these simple protocols significantly improved the accuracy of EM in estimating the quantities of discarded catch. However, a crucial element in both cases is the involvement and cooperation of the fishing crew. In the latter study it was calculated that the protocol requires an additional three minutes of processing time per haul for a single species. Given the large number of species that are regulated under the landing obligation for this fishery, implementing the protocol comes with a significant cost for the fishing industry; the extra time needed to conduct a simple protocol probably would exceed 12 h per fishing trip (van Helmond et al., 2017). These methods used to increase visibility, including the audit approach to reduce review time, as described above, shift the burden of catch monitoring to the fishing fleet, i.e. the work of data collection is mostly internalized by fishers. Therefore, sampling and recording discards under the EU landing obligation will result in a substantial increase in workload for the crew on fishing vessels. Of course, the fishers need to comply with regulations, but the success of monitoring with EM likely depends on the burden that it imposes on fishers and their crews.

So far EM seems to be a promising option in monitoring catches under the forthcoming EU landing obligation (Kindt-Larsen et al., 2011; Mangi et al., 2013). But, this is mostly the case for fisheries where it is easy to detect individual fish, e.g. hook and line (McElderry et al., 2003; Ames et al., 2007; Stanley et al., 2011) or where EM focusses on a single species that is easy to detect with video review (Kindt-Larsen et al., 2011; Ulrich et al., 2015). However, the majority of the European fishing fleet consist of mixed fisheries with substantial discard volumes (Uhlmann et al., 2014). To successfully monitor catches under the landing obligation there is a need for better methods and technologies to improve the application of EM. The implementation of computer vision technology is the logical next step in the innovation process of EM. Automizing the video review process is a possible solution to facilitate processing large amounts of video data, and, eventually, bring down the costs and burden of data collection at sea (i.e. Allken et al., 2019, 2021; Tseng & Kuo, 2020). Even more, fast data processing allows for direct automated image reviewing on board. In other words, with this technology in place real-time catch recording could be achieved by a computer directly counting the fish passing the cameras and only generating a list of species in the catch as output. This would mean that the transmission of large amounts of video footage from a vessel at sea to servers on land, to allow for further data analysis will not be necessary anymore (Michelin et al., 2018; van Helmond, 2021).

The immediate benefit from a more efficient automated catch recording system is a more realistic approach, i.e. increased feasibility, of the implementation of the landing obligation for European fisheries. Full automatization and digitalisation reduces the effort, and costs, of data collection, but also results in increased catch data quantity and quality, easier data access and, as a result, better options for data analysis, and better traceability of catches. Increased traceability and transparency are the two key features of the EU fisheries control system. Validation of the electronic catch recording systems on board, i.e. certification of algorithm standards for species identification and catch volume quantification, could eventually result in compliance by design and increase the efficiency of monitoring the landing obligation. This could also initiate a shift from a landing obligation to a 'registration obligation', making the fishing industry more efficient in complying with the EU Technical Regulations. Not having the additional workload and costs of landing the unretained and unmarketable

(and unprofitable) part of the catch possibly will increase the level of compliance and supports the implementation of the reformed EU management regulations (Msomphora and Aanesen, 2015). Ultimately, automated catch registration also provides indirect incentives and more intrinsic motivation for the fishing industry to get involved in EM. Automatization of the catch recording process supports a complete “net-to-plate” overview which provides an increased market access and potential economic benefits for fishers (Michelin and Zimring, 2020). Being transparent is an essential element in providing the proof of fishing sustainably, customers should be able to check the origin of wild caught fish, linking automated catch registration with blockchain technology makes this possible. To service this growing market an increasing number of seafood retailers are supporting sustainability labels such as MSC. The ability to collect and ‘submit’ their own catch data, potentially creates a sense of ownership and engagement in fisheries management. This can be fuelled by advantages in terms of better fishing opportunities or healthy fisheries, because, better data provides improved fisheries management.

So far, the development of computer vision technology, in fisheries monitoring, is still at an initial stage. One of the two main challenges is the lack of high quality labelled data necessary for model development. A more common choice of the deep learning based analysis of the image/visual data is a use of supervised learning approach, which requires extensive amount of training data, i.e. annotated images (Mohri et al., 2012). However, the amount of data can be reduced by making use of active learning techniques and self-supervised learning (Gal & Ghahramani, 2016; Wang et al., 2023). Additionally, automated annotating tools are becoming available aiming at reducing the annotation time and improving accuracy of annotation (Kirillov et al., 2023). Recognising species in digital images of fish catches is often a specialised task, which is especially challenged by the onboard conditions where fish visibility is often poor. This task takes time due to large number and variation of fish in images (French et al., 2015, 2020; Allken et al., 2019, 2021). Another challenge is the lack resources, expertise, or risk tolerance of EM providers to succeed in technological development and integrate it into their product workflow (Michelin and Zimring, 2020). Expertise from computer sciences, i.e. technical universities, is needed to speed up the process (van Helmond, 2021).

In an attempt to automate the process of video review for catch recording on mixed demersal fisheries, the Dutch Ministry of Agriculture, Nature and Food Quality commissioned the Fully Documented Fisheries (FDF) project to a consortium of Wageningen University and Research (WUR) and partners from the fishing industry started. Automated catch recording requires innovation of EM review strategies, involving computer vision and machine learning technology. It is the aim of the project to make a complete record of the discarded part of the catch, at least for the quota restricted species under the landing obligation, without interference of the normal catch handling processes on board.

2 Automatic discard registration in cluttered environments using deep learning and object tracking: class imbalance, occlusion, and a comparison to human review

Rick van Essen¹, Angelo Mencarelli², Aloysius van Helmond³, Linh Nguyen¹, Jurgen Batsleer³, Jan-Jaap Poos^{3,4} and Gert Kootstra¹.

¹ Farm Technology Group, Wageningen University and Research, 6700 AA Wageningen, The Netherlands

² Greenhouse Horticulture Unit, Wageningen University and Research, 6700 AP Wageningen, The Netherlands




³ Wageningen Marine Research, Wageningen University and Research, PO Box 68, 1970 AB IJmuiden, The Netherlands

⁴ Aquaculture and Fisheries Group, Wageningen University and Research, 6700 AA, Wageningen, The Netherlands



Original Article

Automatic discard registration in cluttered environments using deep learning and object tracking: class imbalance, occlusion, and a comparison to human review

Rick van Essen ¹, Angelo Mencarelli², Aloysius van Helmond³, Linh Nguyen ¹, Jurgen Batsleer³, Jan-Jaap Poos ^{3,4}, and Gert Kootstra^{1,*}

¹Farm Technology Group, Wageningen University and Research, 6700 AA Wageningen, The Netherlands

²Greenhouse Horticulture Unit, Wageningen University and Research, 6700 AP Wageningen, The Netherlands

³Wageningen Marine Research, Wageningen University and Research, PO Box 68, 1970 AB IJmuiden, The Netherlands

⁴Aquaculture and Fisheries Group, Wageningen University and Research, 6700 AA, Wageningen, The Netherlands

*Corresponding author: tel: +31 317 480 302; e-mail: gert.kootstra@wur.nl

van Essen, R., Mencarelli, A., van Helmond, A., Nguyen, L., Batsleer, J., Poos, J.-J., and Kootstra, G. Automatic discard registration in cluttered environments using deep learning and object tracking: class imbalance, occlusion, and a comparison to human review. – ICES Journal of Marine Science, 78: 3834–3846.

Received 1 June 2021; revised 27 October 2021; accepted 28 October 2021; advance access publication 27 November 2021.

This paper presents and evaluates a method for detecting and counting demersal fish species in complex, cluttered, and occluded environments that can be installed on the conveyor belts of fishing vessels. Fishes on the conveyor belt were recorded using a colour camera and were detected using a deep neural network. To improve the detection, synthetic data were generated for rare fish species. The fishes were tracked over the consecutive images using a multi-object tracking algorithm, and based on multiple observations, the fish species was determined. The effect of the synthetic data, the amount of occlusion, and the observed dorsal or ventral fish side were investigated and a comparison with human electronic monitoring (EM) review was made. Using the presented method, a weighted counting error of 20% was achieved, compared to a counting error of 7% for human EM review on the same recordings.

Keywords: by-catch registration, computer vision, deep learning, electronic monitoring, object detection, object tracking.

Introduction

Fisheries management often relies on population models that integrate fisheries data, including catch estimates (Beverton and Holt, 1957; Rijnsdorp *et al.*, 2007; Bradshaw *et al.*, 2018). Accurate catch estimates are, therefore, important for sustainable fisheries management. However, in many fisheries, only part of the catch is landed and sold, the rest may be thrown overboard [“discarded”; (Kelleher, 2005)]. Discarding of fish occurs because of market conditions or fishery management regulations, such as minimum landing sizes or quotas (Catchpole *et al.*, 2005; Rochet and Trenkel, 2005; Poos

et al., 2009). Discarded fish can make up large part of the catch and, because the process of discarding happens at sea, it often goes unrecorded.

Attempts to collect information on discarded catch are generally done at sea via on-board observer programmes (Fernandes *et al.*, 2011; Snyder and Erbaugh, 2020). In these programmes, trained personnel collect numbers, weights, length, age, and species compositions of the discarded part of the catch (Uhlmann *et al.*, 2013). However, at-sea observer programmes are expensive and time-consuming, thus often cover only a small fraction of the overall fishing effort of fishing fleets (Benoit and Allard,

2009; Stock *et al.*, 2019). Increasing the sampling coverage of an observer programme requires a substantial amount of, often unavailable, financial and labour resources. As a consequence, fisheries management is compromised, because substantial parts of the catch remain unregistered (Crowder and Murawski, 1998; Punt *et al.*, 2006). To improve sustainable management of fish populations, better data collection of catch quantities per species is required.

Video-based monitoring on-board fishing vessels, commonly described as Electronic Monitoring (EM), also described as Remote Electronic Monitoring (REM), allows catches to be observed remotely by human experts without requiring additional on-board personnel (McElderry *et al.*, 2003; Kindt-Larsen *et al.*, 2011; Stanley *et al.*, 2014; Hold *et al.*, 2015; van Helmond *et al.*, 2017). EM systems enable continuous catch monitoring over long periods, making them more suitable for monitoring discards from commercial fishing vessels than human on-board observers. Hence, they have the ability to provide more representative coverage of the fleet than any other observer programme (van Helmond *et al.*, 2020). However, because analysis of EM video data requires human observations, costs are still relatively high, which together with the amount of human resources needed, is a limiting factor in the uptake of EM (Needle *et al.*, 2014; Mortensen *et al.*, 2017). To reduce the workload and improve the sampling frequency, a reform of EM data processing is necessary.

Automated on-board image registration is the logical next step in achieving complete registration of discarded catch (French *et al.*, 2020). Training a computer to recognize fish during the sorting process on board a fishing vessel is challenging, due to variability in fish appearance. For example, fish from the same species are not identical in size, colour, and patterning, while different fish species also share similarities, e.g. all flatfish species have a white ventral side. Also, catch is often loaded in bulk on a sorting belt for processing on board, and consequently, fishes are randomly positioned on the belt, overlapping each other, resulting in severe occlusions. Therefore, automated image recognition approaches for monitoring systems need to be able to deal with complex images where multiple individuals and species are partly visible.

Related work in automatic registration of discards

Just training a computer to recognize fish during the sorting process on a fishing vessel is not sufficient for complete registration of discards. For registration of all discards on a conveyor belt, a counting procedure also needs to be included. Automatic registration of discards consists of three subtasks: (1) object localization, (2) object classification, and (3) object counting. Object localization involves locating fishes in an image. Object classification assigns a fish species to that located object, which then allows counting per species. We refer to steps (1) and (2) jointly as object detection. The biggest challenge in object counting is to track the fish over multiple video frames to prevent double counting.

Several machine-vision approaches have been proposed to perform these tasks. Traditional approaches for object detection rely on tailor-made image-processing algorithms, including dedicated handcrafted algorithms to extract image features. Zion *et al.* (1999) and Storbeck and Daan (2001), for instance, constructed a system that performs image classification based on shape descriptors of the fish in the images.

Deep-learning techniques such as a convolutional neural networks (CNN) offer a better approach for object detection (Le-

Cun *et al.*, 2015). These CNNs consist of a large number of interconnected artificial neurons in different layers. The connection strengths (or weights) between the neurons are optimized based on a training set. By feeding a large number of training images, accompanied by the required output, CNNs can be trained to perform specific tasks, such as object localization, classification, and detection. These CNN-based approaches are a subset of the deep learning techniques that are increasingly being used to solve complex problems in marine science and marine resource management (Malde *et al.*, 2019; Beyan and Browman, 2020).

Several CNN-based approaches have been suggested for detection and classification of fish in images. Shafait *et al.* (2016) and Siddiqui *et al.* (2017) developed an approach that classified images taken underwater containing single fish, while Lu *et al.* (2019) designed a CNN that classified images of fish catch landed on the deck of fishing vessels. Eickholt *et al.* (2020) implemented a CNN that can classify living fishes as they pass through a tunnel under barriers, in order to detect invasive species. Allken *et al.* (2019) implemented a CNN that classified images with multiple fish in a controlled environment.

The most recent approaches in the literature are from French *et al.* (2020) and Tseng and Kuo (2020). Both use Masked Region-based CNN (Mask R-CNN) to classify fish using on-board CCTV videos. Tseng and Kuo (2020) used videos acquired on the deck of a longliner and was able to detect *tuna*, *marlin*, *shark*, *buoys*, and 'others' with an accuracy of 83% (F1-score) and an error in counting of 21%. French *et al.* (2020) detected *cod*, *haddock*, *whiting*, *saithe*, *monk*, *Norway pout*, *plaice*, *dab*, and *grey gurnard* on a conveyor belt on-board a fishing trawler. They reported a mean class accuracy of 63%, in a research vessel experiment with known quantities of fish, and a mean class accuracy of 57–59% on a commercial vessel. Human reviewers are reported to have a mean class accuracy ranging between 74 and 86% on a sample of the dataset (French *et al.*, 2020).

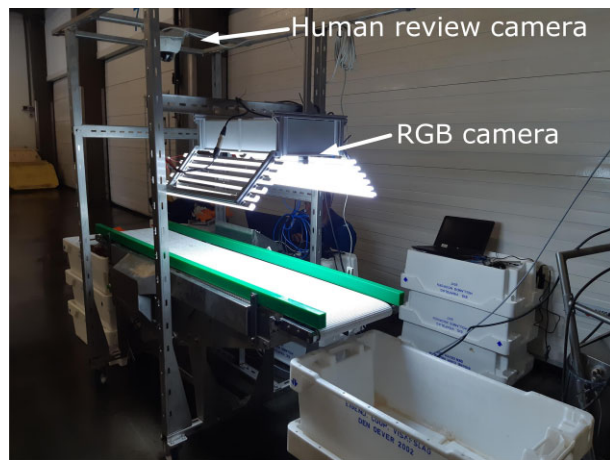
Contributions of this paper

Despite tremendous progress in the field, there are no automatic catch registration systems available yet that can deal well with a continuous stream of video data from the complex on-board situation. Moreover, the detection of fish that are partially or fully occluded by fish or other objects is a challenge that has not been systematically studied in the literature. No related work targets demersal flatfish species, most of which are characterized by a coloured dorsal and white ventral side, which is challenging for image detection. Some of these species can be more frequently observed than others, which can cause challenges in training CNN's. Finally, with the exception of French *et al.* (2020) and Qiao *et al.* (2020), the machine-vision methods have not been compared to the accuracy of a human EM review.

The objective of our work was to develop and evaluate a method for detecting and counting multiple demersal species on conveyor belts on board fishing vessels. An acquisition system was developed to get high-quality video frames of fish on the conveyor belt. Detected fish on the conveyor belt were tracked in video frames to (i) automate the counting process while avoiding double counting and (ii) to improve the detection performance. We study (1) the effect of adding synthetic data to the recorded images to improve performance for infrequently observed species; (2) the performance of the system as a function of the amount of occlusion and whether the ventral or dorsal side of the fish is observed; and (3) a comparison of our method with human EM review.

Table 1. Number of annotations per species in each of the three hauls.

Species	Haul 1	Haul 2	Haul 3	Total
Common sole	192	486	34	712
Dab	0	412	9	421
Gurnard	0	38	0	38
Lemon sole	0	196	0	196
Lesser spotted dogfish	9	27	6	42
Plaice	199	502	213	914
Pouting	8	3	2	13
Ray	20	29	162	211
Turbot	6	16	0	22
Whiting	53	518	171	742

**Figure 1.** Overview of the camera setup indicating the position of the RGB and the EM camera for human review.

Material and methods

Image data

Discarded catch

The discarded catches of beam trawlers in the North Sea were collected from one haul per week during the last 3 weeks of October 2019. These catches included the following fish species: *common sole*, *dab*, *gurnard*, *lemon sole*, *lesser spotted dogfish*, *plaice*, *pouting*, *ray*, *turbot*, and *whiting*. Debris present in the catch (like sea stars, stones, and wood) was also collected by the trawlers. It should be noted that *gurnard* and *ray* are not single species, but consist of two or more similar looking species. *Lesser spotted dogfish*, *pouting*, and *turbot* were among the species that were consistently caught in low numbers, often present at less than ten individuals per haul (Table 1). *Plaice* and *sole* were consistently caught in high numbers, generally present at more than 100 individuals per haul. Mixed discarded fish and debris were put in 8–10 boxes per haul. The total number of fish per box for each species was counted as ground-truth against which the results of the automated method and the EM review could be compared.

Data acquisition

The boxes were then emptied on a conveyor belt similar to the on-board situation. Above the conveyor belt, a RGB camera (IDS Imaging Development Systems GmbH, Germany) and an EM camera system (VIVOTEK Inc., Taiwan) were mounted (Figure 1). The

speed of the conveyor belt was 30 cm/s. The RGB camera recorded images with a resolution of 1600×1200 pixels (Figure 2a) at a time interval of 0.5 s. This resulted in approximately seven consecutive images in which each fish was visible. The EM camera system recorded videos at a resolution of 1920×1200 pixels (Figure 2b), at a time interval of 0.5 s, and was used to make comparisons with human EM observers. To estimate the count error of the method and to compare this to those from human EM observers, one box of fish for each haul was recorded four times.

For each haul, the recorded images were split into three batches. Within these batches, the hauls were combined to form one batch. The first batch of 3744 images was annotated and used for training the network. The second batch of 877 images was annotated and used for validating the network. These validation images were used to optimize some of the hyperparameters, see the "Automatic fish counting" section. The last batch of 610 images was used to evaluate the overall performance of the neural network. Care was taken that images in the different batches came from different boxes.

Adding synthetic data

To avoid problems with the detection of infrequently observed fish species in the recorded dataset, synthetic images of those species were generated (Figure 2c) and added to the training dataset. For each species, 6–12 samples were manually segmented from images in the dataset. A total of five different images without fish were used as background images. Based on a normal distribution (mean 6, SD 2), a number of fishes were randomly positioned on a background image. Each fish was randomly rotated and scaled ($\pm 15\%$ independently in x - and y -direction) and recoloured (hue $\pm 20\%$, saturation $\pm 10\%$, and brightness $\pm 10\%$), using a random uniform distribution. Finally, segmented debris was added to the images.

Data annotation

Recorded images from the RGB camera were annotated manually using bounding boxes, with a custom-made annotation tool that facilitates annotation by tracking bounding boxes over consecutive images. Synthetic images in the training set were automatically annotated and were manually refined. Each fish was assigned a unique id and orientation (dorsal or ventral for the flatfish species and rays). An estimation for the percentage of occlusion (0–20%, 20–40%, 40–60%, 60–80%, or 80–90%) was annotated for each fish in every image. Fishes in images occluded by more than 90% were not annotated.

Automatic fish counting

Figure 3 presents the method used for counting fish that are moving on the conveyor belt. From the images, fishes were detected using an object detector (see the "Fish detection" section). Since the fishes were visible in multiple consecutive images, a method was developed to track the fish in the video (see the "Tracking multiple fish over consecutive images" section). Finally, the number of tracked fishes was counted, resulting in an estimation of the number of fish for each species.

Fish detection

A YOLOv3 deep neural network (Redmon and Farhadi, 2018) was used for joint object localization and object classification. YOLO

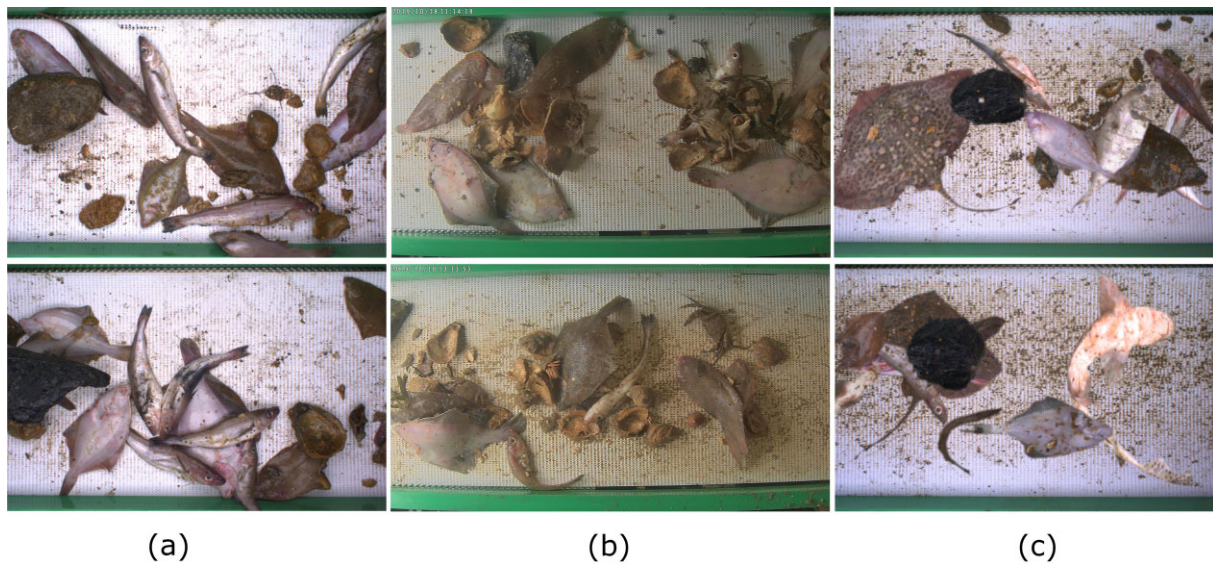


Figure 2. Examples of recorded RGB camera (a), EM camera (b), and synthetic (c) images.

outputs a vector Y for each detected object with the coordinates that represent the centre of the bounding box around the detection (b_x , b_y), the dimensions of the bounding box (b_w , b_h), the objectness p_c (probability of the bounding box), and the confidence scores for all possible classes ($c_1, c_2, c_3, \dots, c_n$):

$$Y = [b_x, b_y, b_h, b_w, p_c, c_1, c_2, c_3, \dots, c_n]. \quad (1)$$

The YOLO network consists of a Darknet 53 feature extractor backbone (Redmon and Farhadi, 2018) and three blocks that perform the bounding box detection at different spatial scales (Figure 4). In the backbone block, feature maps are extracted at three spatial scales. These features are connected with the detections that have the same spatial scale using skip connections (Redmon and Farhadi, 2018). In this way, meaningful information from earlier feature maps could be used for better object detection. The first detection block has the smallest detection layer and is responsible for detecting large objects in the image. The second block is responsible for medium sized objects and the third for the small scale objects.

The detections at the three scales are concatenated together. Since most detections have a low objectness, all detections with an objectness lower than $p_{c,min}$ are removed from the detections. Overlapping detections are removed using non-maximum suppression (NMS). NMS calculates the intersection over union (IoU) between all detections of the same class. If the intersection is higher than $IoU_{nms,min}$, the detection with the lowest objectness is removed. For more details on the network, we refer to Redmon *et al.* (2016) and Redmon and Farhadi (2017, 2018).

To improve fish detection, the network was pre-trained on images from the general COCO dataset (Ultralytics, 2021), a large image data set containing many different classes of objects (Lin *et al.*, 2014). To specialize the network in detecting fish, it was then fine-tuned using our collected data set.

With data augmentation, the original training data is randomly transformed to train the network more robustly to variations in the appearance of objects in the images. We applied real-time augmentation, where each training image is randomly transformed

on-the-fly at each iteration during the training procedure. This ensures that the network sees different variations of the images with each epoch. The augmentations include spectral transformations (hue, saturation, and brightness) of the image by changing the pixel values, and spatial transformations by rotating, scaling, shearing, and flipping the image (Redmon *et al.*, 2016). The default parameters were used for the augmentation.

The Ultralytics implementation of YOLOv3 in Pytorch (Ultralytics, 2021) was used. The network was trained on a NVIDIA GeForce GTX 1080 Ti using the training and validation dataset. A batch size of 12 was used, the learning rate was set to 10^{-3} , the momentum was set to 0.937, and the decay rate was set to 5×10^{-5} . The input image size, minimum objectness, $p_{c,min}$, and minimum NMS IoU, $IoU_{nms,min}$, hyper-parameters were selected on the validation dataset using grid search (Table 2). The networks were fine-tuned for 800 epochs (number of passes through the full training set), the epoch with the highest performance (F1-score, see the "image level evaluation" section) on the validation set was used.

Tracking multiple fish over consecutive images

Since each fish is visible in multiple images, the detected fishes should be tracked over consecutive images to avoid double counts. The tracking is inspired by the Sort method (Bewley *et al.*, 2016), using trackers with associated Kalman filters to predict fish position and size on the next image. Each tracker is modelled as:

$$x = [u, v, s, r, \dot{u}, \dot{v}, \dot{s}, \dot{r}]^T, \quad (2)$$

where u and v are the horizontal and vertical center coordinates of the bounding box, s the surface area, r the aspect ratio (width/height), and \dot{u} , \dot{v} , \dot{s} , and \dot{r} the corresponding first derivatives (velocities) of the tracker. The trackers are associated with the output of the fish detection using the Hungarian assignment algorithm (Kuhn, 1955), which assigns a detected fish to a tracker. This is done by maximising the sum IoU of all detection-tracker combinations in an image. In other words, the fish detections are assigned to the trackers in such a way that the total

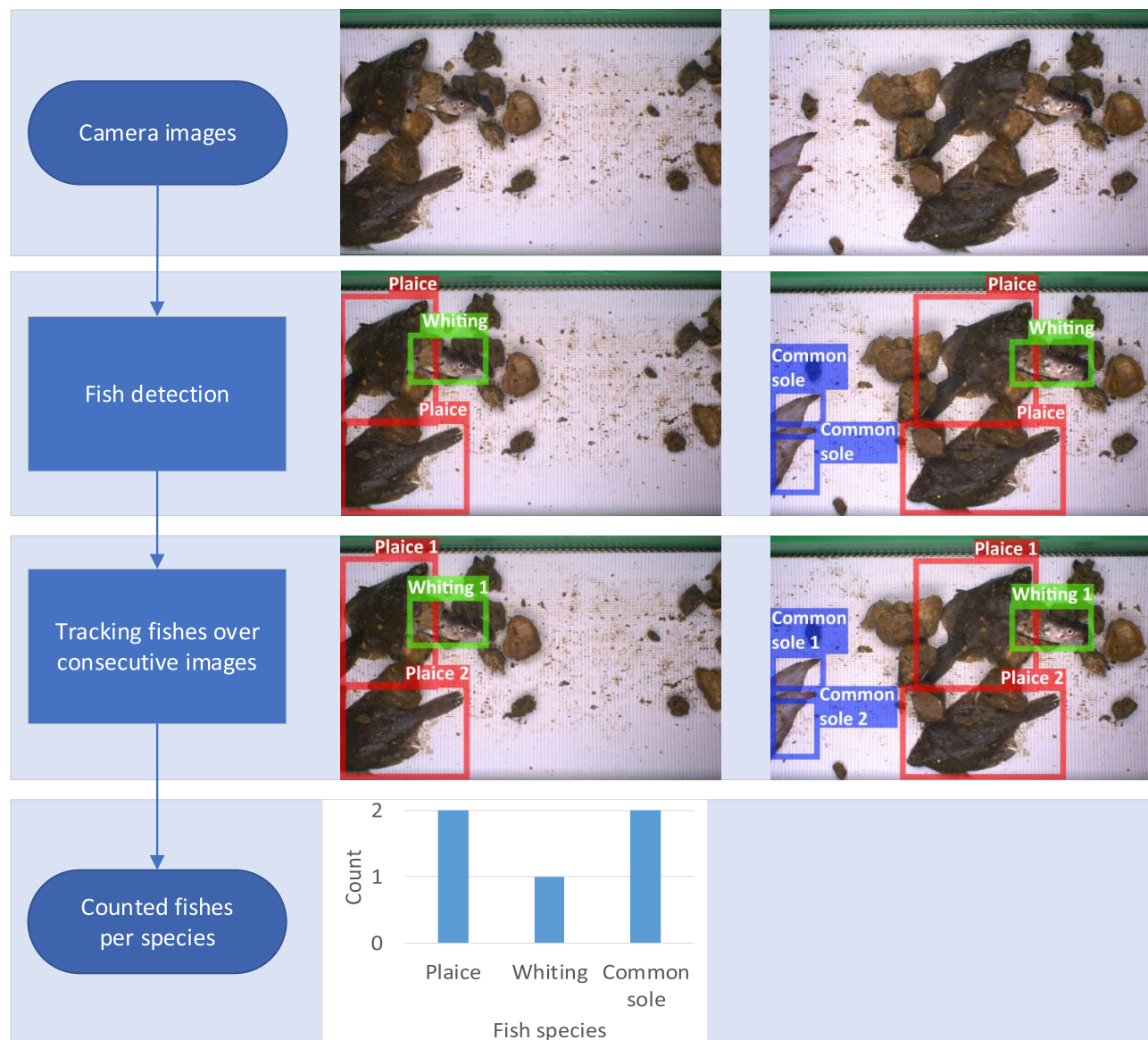


Figure 3. Overview of the fish counting method presented in this paper. The left-hand side breaks down the process used for counting fish. The images in the centre and right represent two images taken in sequence, while the conveyor belt moves under the camera from left to right. Bounding boxes indicate fish as localized and classified by YOLOv3. The bar chart at the bottom indicates the total fish counts for different species at the end of the image sequence.

overlap is maximized. When the IoU of a detection-tracker combination is higher than a given threshold, the Kalman filter is updated with the bounding box of the detection. In addition to the Sort method, the translation between the current and previous image is estimated using ORB feature matching (Rublee *et al.*, 2011) and used to update the \hat{u} and \hat{v} states. Trackers that were not updated in the last n_{max} images were automatically removed. A tracker should be updated at least n_{init} times before it is regarded as a count. The minimum IoU, n_{max} , and n_{init} parameters were optimized on the validation dataset. Their values were 0.2, 2, and 2, respectively.

The tracker is also used to predict the fish species, based on all m observations $Z = \{z_1, z_2, \dots, z_m\}$. The probability of each tracker belonging to species i , given its observations Z , $P(i|Z)$, can be calculated using Bayes' theorem. To facilitate the detection of

infrequent species, we set all priors equal. Then the equation simplifies to:

$$P(i|Z) = \frac{P(Z|i)}{P(Z)}, \quad (3)$$

$$P(Z|i) = \prod_{j=1}^m P(z_j|i), \quad (4)$$

$$P(Z) = \sum_{k=1}^n P(Z|k), \quad (5)$$

where n is the total number of species and m the number of observations for the tracker. The class with the highest probability is used as the class belonging to the tracked fish. The number of fish for each species on the conveyor belt is counted by the number of trackers for each class.

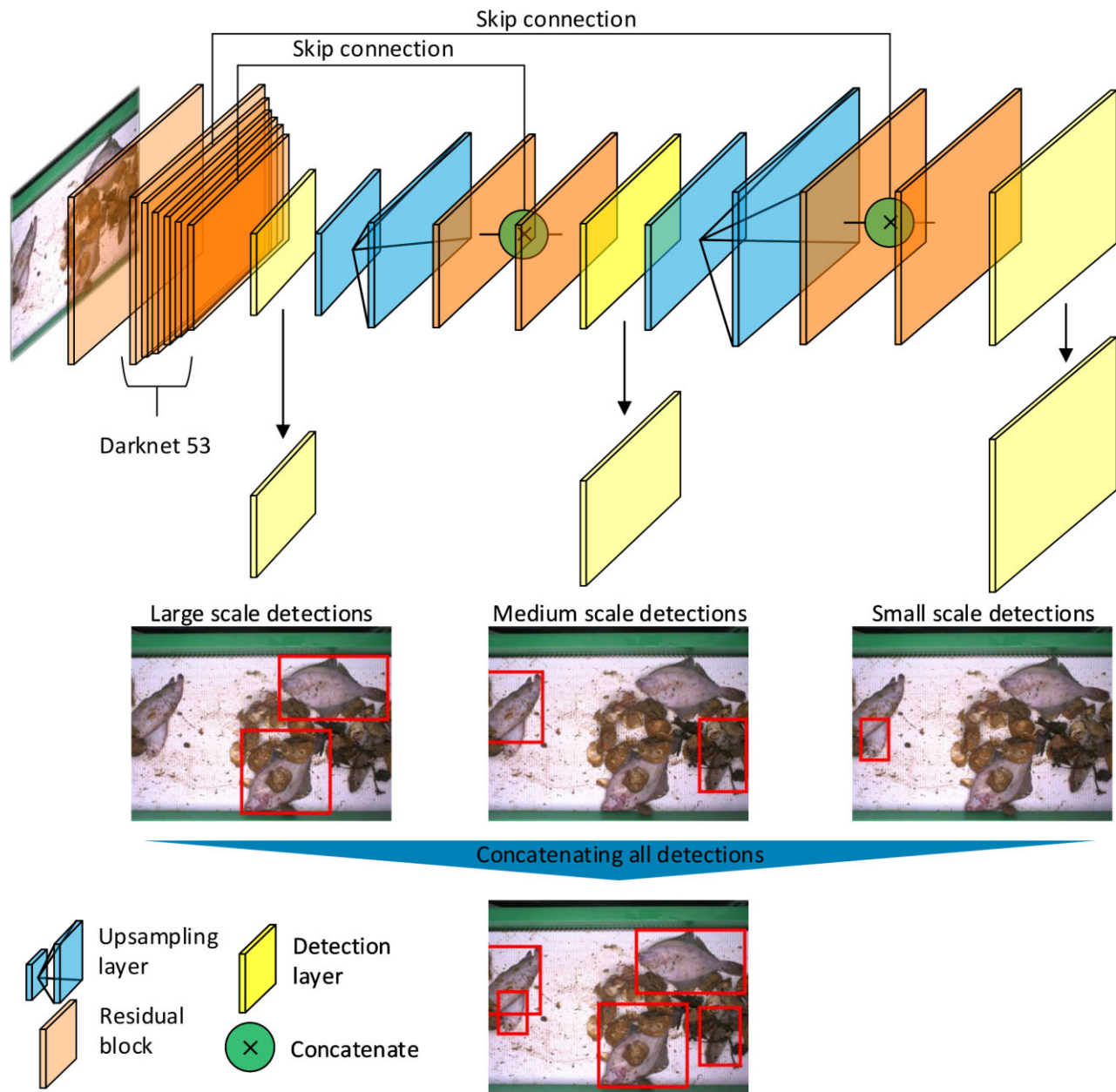


Figure 4. Simplified schematic overview of the YOLOv3 convolutional neural network. The Darknet 53 feature extractor extracts features from different resolutions, and is used to detect fishes at three different scales. For each scale, skip connections concatenate the features for the corresponding scale. The detections at the different scales are merged for the final detection. Image inspired by Kathuria (2018).

Table 2. The optimized image size, minimum objectness, and minimum NMS IoU for the models using different amounts of synthetic data.

Number of synthetic images	Detection parameters		
	Image size	Minimum objectness	Minimum NMS IoU
0	416 × 416	0.3	0.5
50	512 × 512	0.4	0.4
100	512 × 512	0.4	0.4
200	512 × 512	0.4	0.6

Evaluation

The final error in counting the fishes per species is the result of errors in both object detection and tracking. To gain an insight into these errors, the performance of the discard registration was evaluated on two levels: at image level and at batch level.

Image level evaluation

To evaluate whether a network prediction of a fish in an image is correct, the IoU between the predicted bounding box and the ground-truth bounding box was calculated. A network prediction of a fish is associated with a ground-truth bounding box if the IoU is larger than 0.5. When the predicted species matches the true

species, the prediction was marked as a true positive (TP). If the species did not match, the ground-truth bounding box was marked as a false negative (FN) and the corresponding network prediction as a false positive (FP). All the network predictions that did not have an associated ground-truth box were marked as FPs and all ground-truth bounding boxes that did not have an associated network prediction were marked as false negative.

Based on the number of TP, FP, and FN, the precision and recall were calculated as a measure of the performance of the neural network:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (7)$$

The F1-score is the harmonic mean between the precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (8)$$

The F1-score, F1_i , for each individual species i was combined using the macro and weighted average:

$$\text{F1}_{\text{macro}} = \frac{1}{n} \sum_{i=1}^n \text{F1}_i \quad (9)$$

$$\text{F1}_{\text{weighted}} = \frac{\sum_{i=1}^n \omega_i \cdot \text{F1}_i}{\sum_{i=1}^n \omega_i} \quad (10)$$

where n is the total number of species and ω_i the number of fishes of species i . Similar calculations are done for the precision and recall. Besides these performance measures, a confusion matrix, and a precision-recall curve is used. The confusion matrix summarizes the number of correctly and incorrectly classified fishes, with their count values broken down for each species. Therefore, it shows which prediction errors are made. Each row in the confusion matrix represents the fish species and each column the predicted species. Ideally, all values should be on the diagonal, meaning a perfect prediction. The precision-recall curve plots for each fish species, the trade-off between precision and recall for different threshold values of the confidence (objectness multiplied by class confidence). The area under the precision-recall curve is a robust measure of the classification performance.

Batch level evaluation

The estimated number of fishes on the conveyor belt was evaluated by the absolute percentage error (APE) for each species i :

$$\text{APE}_i = \left| \frac{\omega_i - \hat{\omega}_i}{\omega_i} \right| \quad (11)$$

with ω_i the ground-truth number of fishes (manually counted per box) and $\hat{\omega}_i$ the estimated number of fishes. Using Equations (9) and (10), the $\text{APE}_{\text{macro}}$ and $\text{APE}_{\text{weighted}}$ can be calculated as a measure for the average absolute percentage error across all species. The Pearson correlation coefficient is calculated between the estimated number of fishes and the ground-truth. Over- and under-estimation of the number of fishes is tested by the difference between the regression coefficient and 1.0, using the t -statistics ($\alpha = 0.05$).

Experiments

Effect of synthetic data

To study the effect of adding synthetic data to the detection performance, four different YOLO models were trained and evaluated at image level. The models have either 0, 50, 100, or 200 synthetic images added to the original training set. The models used the optimized hyperparameters for each model (Table 2). Each model was evaluated at image level and the model with the highest weighted F1-score on the validation dataset was used in the rest of the experiments. The confusion matrix of the best model on the test set is shown.

Effect of occlusion and orientation

To study the effect of occlusion and fish orientation, two experiments were done. First, the performance of the detection method as a function of the amount of occlusion was evaluated at image level. Second, the effect of orientation (dorsal or ventral) of fish species with a distinct ventral side (*common sole*, *dab*, *lemon sole*, *lesser spotted dogfish*, *plaice*, *ray*, and *turbot*) was evaluated.

Comparison with human EM review

To create a comparison with the current practice of on-board video review, the recordings of the EM camera in the setup were assessed using the Black Box Analyzer software (Anchorlab, 2021). A standard protocol from AnchorLab was used: viewing the video at a fast pace, pausing the video whenever a species of interest was visible, and annotating that individual fish. An individual fish was only annotated in a single image, so if that fish was also visible in subsequent frames it was not annotated again, to prevent double counting.

Both human EM review and automatic fish counting are compared with the ground-truth at batch level. This is done four times, using the same fishes with different compositions on the conveyor belt (see the "Image data" section).

Results

Effect of synthetic data

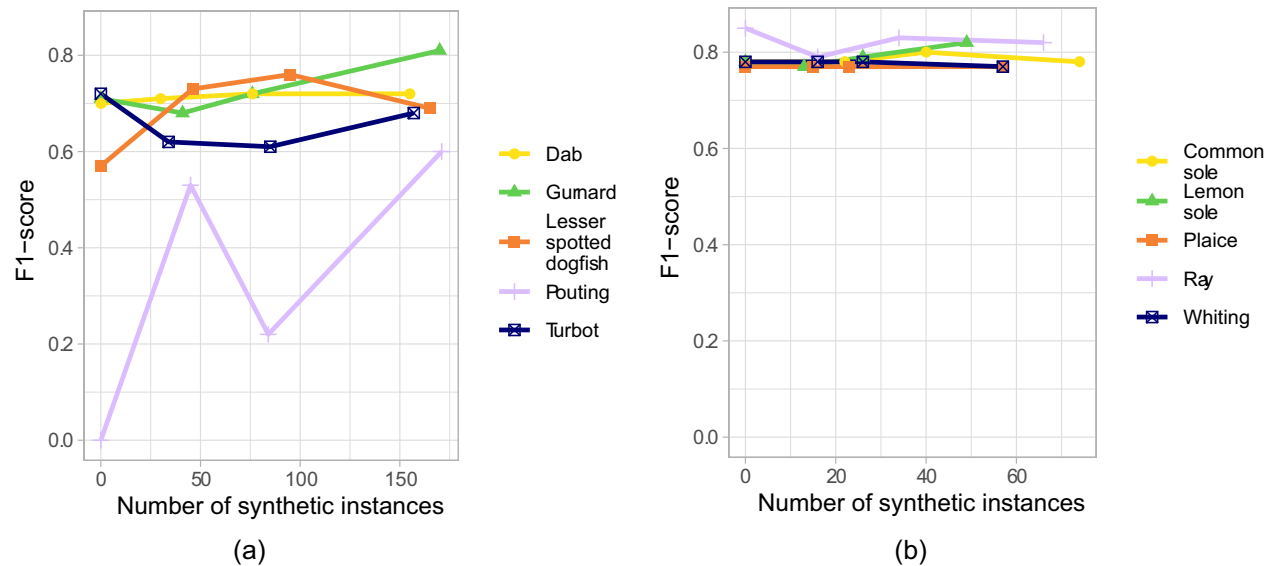
Adding synthetic images to the training data improved the weighted and macro recall for the validation dataset (Table 3). In other words, adding synthetic data lowers the chance that the network is missing fish. Adding synthetic data had a limited effect on the precision, but substantially improved the macro recall and consequently the F1-score. This makes sense since the frequency of the less frequent fish species is mainly increased in the training dataset. Since these species are also less frequent in the validation dataset, the effect of improving the detection of these species is weighted out in the weighted F1-score. The model having 200 additional synthetic images has both the highest weighted and macro F1-score and is used in the rest of the experiments described in this paper.

In general, adding synthetic observations has a positive effect on the infrequent species (Figure 5). The F1-score of *dab*, *gurnard*, *lesser spotted dogfish*, and *pouting* improved. The *turbot* shows a small decrease. The influence of adding synthetic observations to more frequent fish species is low, keeping the F1-score at a high level.

The weighted and macro F1-score on the test dataset is 0.80 and 0.70, respectively (Table 4). *Pouting* and *Gurnard* are poorly

Table 3. Validation precision, recall, and F1-scores for trained with different number of added synthetic images.

Added number of synthetic images	Macro			Weighted		
	Precision	Recall	F1	Precision	Recall	F1
0	0.77	0.63	0.69	0.84	0.71	0.77
50	0.88	0.62	0.70	0.85	0.70	0.77
100	0.81	0.67	0.73	0.84	0.73	0.78
200	0.85	0.69	0.76	0.85	0.73	0.79

**Figure 5.** Influence of adding synthetic instances on the five infrequent (a) and five more-frequent (b) fish species. Since the number and species of fishes on a synthetic image are drawn from a distribution, the number of synthetic instances can be different from the number of synthetic images.**Table 4.** Precision, recall and F1-score for the different fish species on the test dataset.

Species	Precision	Recall	F1
Common sole	0.90	0.84	0.87
Dab	0.82	0.61	0.70
Gurnard	0.90	0.29	0.44
Lemon sole	0.82	0.78	0.80
Lesser spotted dogfish	0.88	0.52	0.66
Plaice	0.82	0.83	0.82
Pouting	0.33	0.07	0.12
Ray	0.93	0.74	0.83
Turbot	0.93	0.87	0.90
Whiting	0.84	0.79	0.81
Weighted average	0.85	0.77	0.80
Macro average	0.82	0.63	0.70

detected and have low F1-scores. Most errors in the test dataset are made by missing fishes (predicting background instead of a fish species, Table 5). Only confusing *dab* for *plaice* and *pouting* for *whiting* occurs more frequently. All except one *pouting* is detected as *whiting* in the test dataset.

Figure 6 shows the precision–recall curves for the infrequent species (a) and the more-frequent species (b). Looking at the area

under the curve (AUC), it can be seen that in general the more-frequent species are better detected by the method than the infrequent species. *Turbot* and *lesser spotted dogfish* (curve hidden behind the *turbot* curve) show the optimal precision–recall curve, but it needs to be noted that this is based on only 24 and 14 test samples from these species. For *pouting*, there is no threshold that results in high precision and recall. *Whiting* shows a drop in precision when the recall increases, meaning that in some cases a higher confidence does not always lead to a better classification.

Effect of occlusion and orientation

There was a strong negative correlation between occlusion percentage and F1-score (Figure 7a), indicating that model performance decreases with increasing occlusion. The decreasing model performance can be explained by both the effect of the lower visibility of the fish at high occlusion levels and the effect of less training data for highly occluded fish.

The side of the fish facing the camera (dorsal or ventral side) for the flatfish species also influenced model performance (Figure 7b). Fishes with the dorsal side facing the camera have both a higher macro and weighted F1-score. This can be explained by both the lower number of training samples having the ventral side facing the camera and a higher difficulty classifying flat fish by their ventral side.

Table 5. Confusion matrix of the detections in the test dataset. The cell colours are the percentages (for each fish species), using the scale on the right.

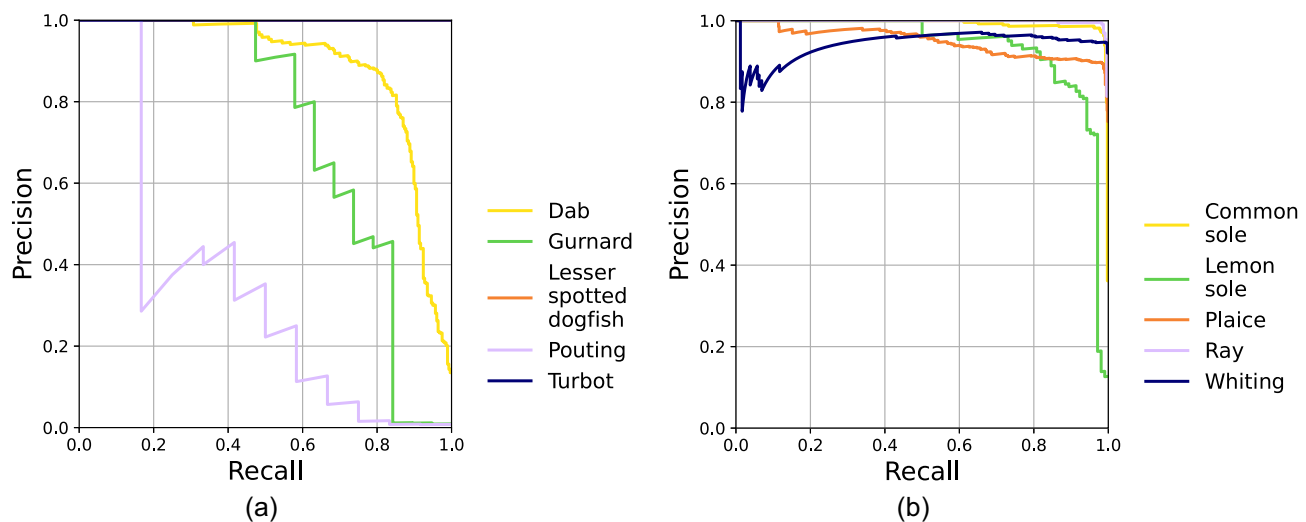
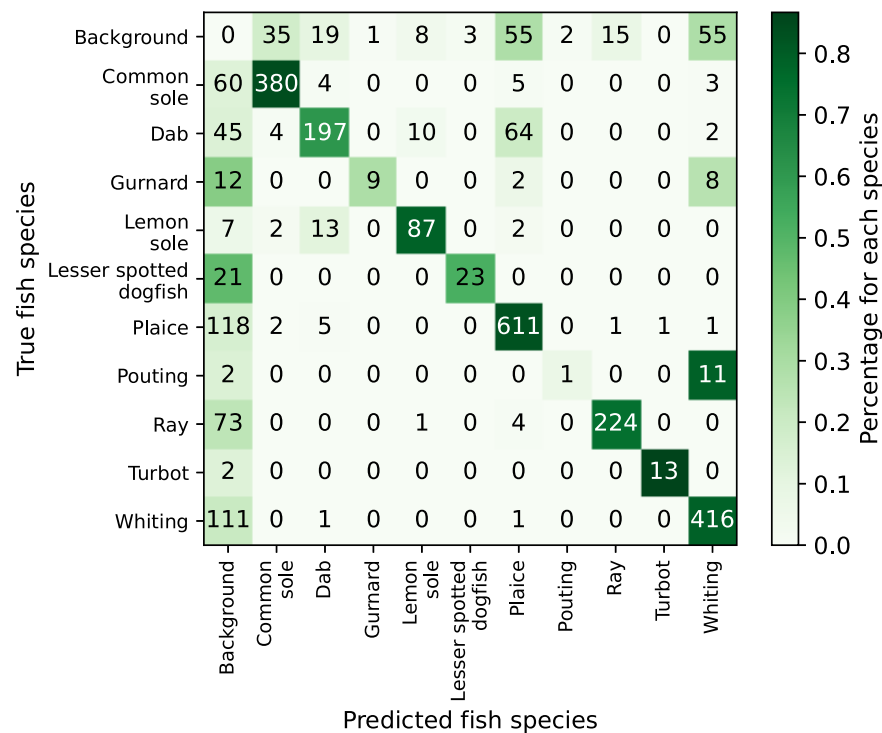


Figure 6. Precision–recall curves for the five infrequent (a) and more-frequent (b) fish species. Note that the line for *lesser spotted dogfish* is behind the line of *turbot* on the top of the graph and the line for *ray* behind the line for *common sole*.

Comparison with human EM review

Both automatic counts and human EM review counts were strongly positively correlated with the ground-truth counts with an r of 0.978 for automatic counting and 0.994 for human EM review (Figure 8). Generally, human EM review significantly underestimates ($p = 3.33 \times 10^{-11}$) the number of fishes, whereas automatic counting significantly overestimates the number of fishes ($p = 2.39 \times 10^{-8}$). Since

human EM review only annotates each fish once, it is more likely to underestimate the number of fish than to overestimate it. The overestimation of automatic counting can be caused by failed tracking of some fishes. If tracking fails, a new tracker can be created and thereby double count a fish.

Human EM review has a weighted APE of 0.07 and a macro APE of 0.29 (Table 6). Automatic counting results in a weighted APE of

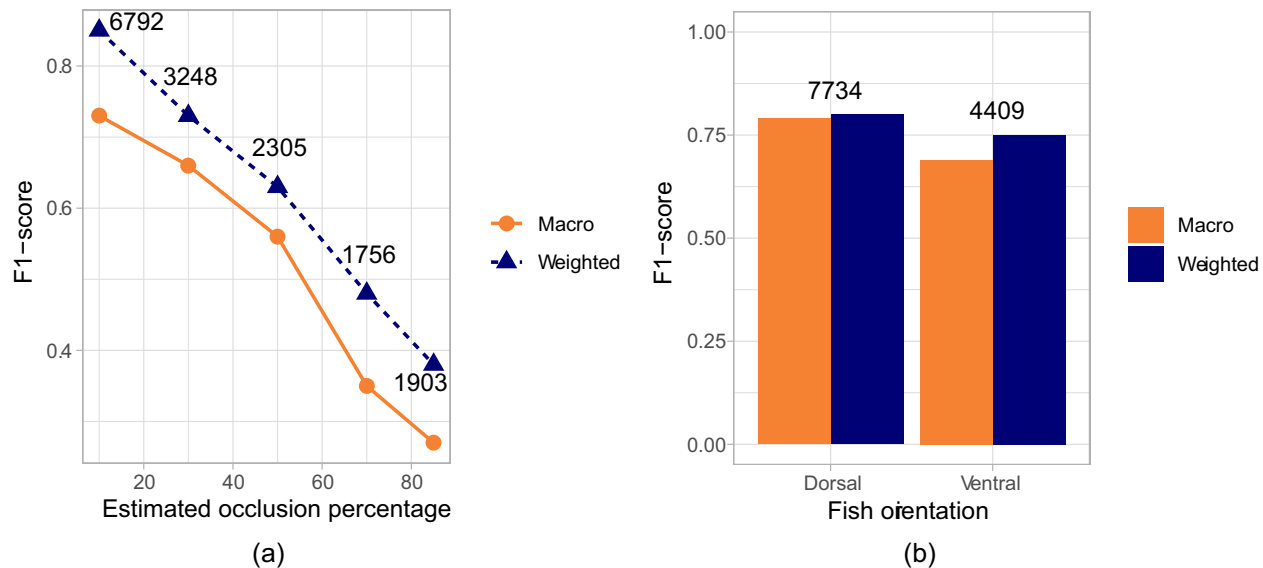


Figure 7. Influence of occlusion (a) and fish orientation (b) on the macro and weighted F1-score for the validation set. The fish orientation is only shown for the species with a distinct ventral side (*common sole*, *dab*, *lemon sole*, *lesser spotted dogfish*, *plaice*, *ray*, and *turbot*). The number of observations for each group in the training set is indicated in the text in the figure.

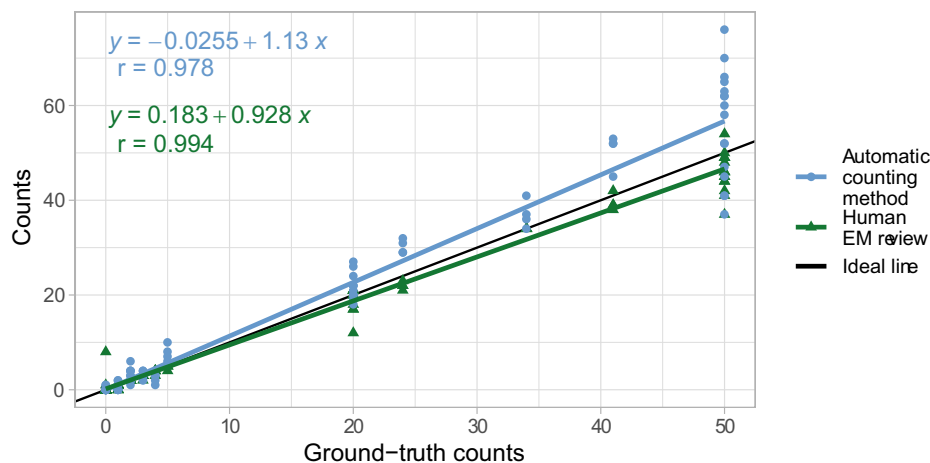


Figure 8. Regression between the estimated number of fish per species for the automatic counting method and human EM review with the ground-truth counts. Estimations are included for each haul and repetition. The Pearson correlation coefficients and equations are shown in the upper left corner.

0.20 and macro APE of 0.28. For most species, human EM review shows a lower counting error than automatic counting. Due to a high APE for *pouting*, which was observed only once in the ground-truth, the macro APE for human EM review is higher than for automatic counting. The automatic counting method has a high APE for *gurnard*, *pouting*, *turbot*, and *whiting*. The high APE for *gurnard*, *pouting*, and *turbot* is explained by the low number of observations for these species, where a small error in counting has a large influence on the APE.

Discussion

The presented method was able to detect the fishes on the conveyor belt with a macro F1-score of 70% and a weighted F1-score of 80%. The YOLOv3 neural network achieved a higher accuracy than the

mask-RCNN network used by French *et al.* (2020), with a reported mean class accuracy between 57 and 63% in a similar environment. YOLOv3 in general has a higher detection performance, compared with the Faster-RCNN (the detection part of mask-RCNN; Redmon and Farhadi, 2018, preprint: not peer reviewed.). However, it must be noted that different datasets and different fish species were used, rendering a direct comparison of our method with the work of French *et al.* (2020) impossible. The presented method shows an average and weighted absolute percentage error in counting of 28% and 20%, respectively. This error is comparable with the counting error of 21% presented in Tseng and Kuo (2020), however the environments differ.

Deviations between the number of counted fish and the ground-truth number of fish result from two sources of error: errors in fish detection and errors in tracking. Because the number of fish

Table 6. Ground truth count, count error \pm SD, and Absolute Percentage Error (APE) for the human EM review and automated count method. Ground truth and count errors are provided for all species. Macro and weighted APE is also provided for the combined species. The counts of the three hauls are combined, and mean count error and APE are based on the four repeats of these three hauls.

Fish species	Ground truth count	Human		Automatic	
		EM review		Counting method	
		Count error	APE	Count error	APE
Common sole	72	-4.5 ± 0.6	0.06	10 ± 5.8	0.14
Dab	50	-2.5 ± 4.4	0.05	-8.8 ± 2.9	0.18
Gurnard	4	0.3 ± 0.5	0.06	-1.8 ± 1.0	0.44
Lemon sole	20	-4 ± 2.7	0.20	-0.8 ± 1.0	0.04
Lesser spotted dogfish	5	-0.3 ± 0.5	0.05	0.3 ± 1.0	0.05
Plaice	111	-8.3 ± 5.4	0.07	22.3 ± 7.4	0.20
Pouting	1	2 ± 4.0	2.00	-1.0 ± 0.0	1.00
Ray	39	0.0 ± 0.0	0.00	4.0 ± 2.9	0.10
Turbot	3	-1.0 ± 0.0	0.33	-1.0 ± 0.8	0.33
Whiting	79	-3.3 ± 1.5	0.04	27.8 ± 8.0	0.35
Macro APE			0.29		0.28
Weighted APE			0.07		0.20

results from the tracking algorithm that is repeatedly informed about a single object, the method can partly recover from errors in detection. The fish detection can only cause a deviation in the number of counted fish when the detection systematically misclassifies a specific fish over consecutive images. When a tracker fails to track a fish over consecutive images, a new tracker can be created leading to a double count. Most errors made by the fish detection are detecting background instead of a fish (Table 5). Consequently, the tracking method will create too few trackers. However, the automatic count method is overestimating the number of fish (Figure 8 and Table 6). This can only be explained by imperfections in the working of the tracking method, which creates double counts for some fishes. This error compensates for the number of missed fishes caused by the detection and as a net result overestimates the number of fish. Overestimation of *plaice* and *whiting* is relatively high, compared to other species. Since the detection of these species works relatively well compared to other species (Figure 5), a higher percentage of the fish will create trackers. Consequently, more double counts are present, resulting in a higher overestimation of these species. The method worked well for *lesser spotted dogfish* and *ray*, resulting in a low APE. This can be explained by their size: the bounding boxes of large fishes overlap more between the consecutive images leading to better tracking. A way to improve the tracking is by selecting the tracking hyperparameters for each fish species.

Adding synthetic data improved detection of infrequent fish species, without affecting the detection of more-frequent species. Without adding new examples, segmenting fish from one image and adding them to a synthetic image, overlapping with other fishes, provided the neural network with new information that enhanced detection. Adding real images of these fish species, however, is expected to add more new information and will further improve detection, compared to adding synthetic images. In situations where collecting large numbers of rare fish is difficult, adding synthetic images is useful to help with counting fish on board fishing vessels.

The presence of occlusions hampered robust fish detection. The performance of the network was degraded almost linearly when

the level of occlusion increased. However, there is a correlation between the level of occlusion and the number of training examples. By adding more data (either real or synthetic) for heavily occluded fish, the performance could be improved. Although a larger training set may improve the system performance, the possibility to use some mechanical solutions, such as a precision pacing conveyor, can physically spread and separate the stacked by-catch to simplify the scene. Such a mechanical solution is expected to enable the most gain in performance, however it takes up more of the limited space available on board fishing vessels.

The tracking method could be improved by using a neural network to estimate the visual similarity between the tracker and the detection. Using this similarity, together with the position for association of the trackers and detections, the method will not match visually different fishes (Wojke *et al.*, 2017, preprint: not peer reviewed.). Another option could be to use a line scan camera, which instead of recording the whole image at once, builds it line by line. This eliminates the need for a tracking method and thereby removes one source of errors in counting.

Automated registration of catches or discards will increase the monitoring coverage of commercial fishing activities. Such improvement in the on-board monitoring process could be relevant for a wide range of specific fisheries management applications (Catchpole *et al.*, 2005; Dickey-Collas *et al.*, 2007; Uhlmann *et al.*, 2013). One of the most profound changes in European fisheries policy is the introduction of a Landing Obligation, or discard ban, for all quota-regulated species that are not covered under the prohibited species list (Article 15, Regulation (EU) 1380/2013; Borges, 2013). This discard ban encompasses a transition from a landing to a catch quota regime (i.e. landings and discards), to ensure fishing opportunities reflect the total catch of a stock. Fishers are obliged to record or land the complete catch of species under quotas, including the undersized, unmarketable part of the catch. This recording or landing of the catch is a labour intensive task, which is difficult to control by the responsible authorities (van Helmond *et al.*, 2017). The success of the landing obligation likely depends on the ability to efficiently record all catches, and, at the same time, reduce the burden that it imposes on the fishing industry. Automated registration

of catches by species, as presented here, can contribute to accurate catch estimates, especially for "data limited stocks".

Conclusion

Demersal fish can be registered with a weighted counting error of 20%, in an on-board-like conveyor belt setup using a two-stage detection and tracking system, with colour images. The neural network responsible for the fish detection yielded a weighted F1-score of 0.79, after adding 200 synthetic images. The higher the occlusion of the fishes on the conveyor belt, the lower the detection performance, with a weighted F1-score of 0.85 for fish with 0–20% occlusion and 0.38 for 80–90% occlusion. Fishes with their dorsal side facing the camera showed better detection, with a 5%-point higher weighted F1-score, compared to fishes with the ventral side facing the camera. Human EM review resulted in a weighted APE of 0.07. Without human intervention, the proposed method resulted in a weighted APE of 0.20. The method allows automatic monitoring on board of all vessels. Therefore, the use of automatic fish registration is a promising method to increase the monitoring coverage.

Funding

The study was carried out under the Fully Documented Fisheries project initiated by the Dutch Ministry of Agriculture, Nature and Food Quality and funded by the European Maritime and Fisheries Fund.

Data availability

Code used to produce the results in this paper is available at http://github.com/Rick-v-E/automatic_discard_registration. The images, annotations and network weights are available at <https://dx.doi.org/10.4121/16622566>.

Acknowledgements

The authors would like to thank the skippers and crew of participating vessels for their help in providing fish; Anchor Lab for their help by uploading the dataset in the EM system; Menko Dijkstra and Thomas Smith for their help in annotating the data; and Arjan Vroegop and Toon Tielen for building the hardware and recording software.

REFERENCES

- Allken, V., Handegard, N. O., Rosen, S., Schreyeck, T., Mahiout, T., and Malde, K. 2019. Fish species identification using a convolutional neural network trained on synthetic data. *ICES Journal of Marine Science*, 76: 342–349.
- Anchorlab 2021. Black box analyzer. <http://www.anchorlab.dk/EFM.a.spx> (last accessed 13 January 2021).
- Benoît, H., and Allard, J. 2009. Can the data from at-sea observer surveys be used to make general inferences about catch composition and discards?. *Canadian Journal of Fisheries and Aquatic Sciences*, 66: 2025–2039.
- Beverton, R. J., and Holt, S. J. 1957. On the Dynamics of Exploited Fish Populations. Ministry of Agriculture, Fisheries and Food HMSO, London.
- Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. 2016. Simple online and realtime tracking. *In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468.
- Beyan, C., and Browman, H. I. 2020. Setting the stage for the machine intelligence era in marine science. *ICES Journal of Marine Science*, 77: 1267–1273.
- Borges, L. 2013. The evolution of a discard policy in Europe. *Fish and Fisheries*, 16: 534–540.
- Bradshaw, C. J. A., Prowse, T. A. A., Drew, M., Gillanders, B. M., Donnellan, S. C., and Huveneers, C. 2018. Predicting sustainable shark harvests when stock assessments are lacking. *ICES Journal of Marine Science*, 75: 1591–1601.
- Catchpole, T., Frid, C., and Gray, T. 2005. Discarding in the English north-east coast nephrops norvegicus fishery: the role of social and environmental factors. *Fisheries Research*, 72: 45–54.
- Crowder, L. B., and Murawski, S. A. 1998. Fisheries bycatch: implications for management. *Fisheries*, 23: 8–17.
- Dickey-Collas, M., Pastoors, M., and Keeken, O. A. 2007. Precisely wrong or vaguely right: simulations of noisy discard data and trends in fishing effort being included in the stock assessment of North Sea plaice. *ICES Journal of Marine Science*, 9: 64.
- Eickholt, J., Kelly, D., Bryan, J., Miehl, S., and Zielinski, D. 2020. Advancements towards selective barrier passage by automatic species identification: applications of deep convolutional neural networks on images of dewatered fish. *ICES Journal of Marine Science*, 77: 2804–2813.
- Fernandes, P., Coull, K., Davis, C., Clark, P., Catarino, R., Bailey, N., Fryer, R. *et al.* 2011. Observations of discards in the Scottish mixed demersal trawl fishery. *ICES Journal of Marine Science*, 68: 1734–1742.
- French, G., Mackiewicz, M., Fisher, M., Holah, H., Kilburn, R., Campbell, N., and Needle, C. 2020. Deep neural networks for analysis of fisheries surveillance video and automated monitoring of fish discards. *ICES Journal of Marine Science*, 77: 1340–1353.
- Hold, N., Murray, L. G., Pantin, J. R., Haig, J. A., Hinz, H., and Kaiser, M. 2015. Video capture of crustacean fisheries data as an alternative to on-board observers. *ICES Journal of Marine Science*, 72: 1811–1821.
- Kathuria, A. 2018. What's new in yolo v3? <https://towardsdatascience.com/yolo-v3-object-detection-53fb7d3bfe6b> (last accessed 15 February 2021).
- Kelleher, K. 2005. Discards in the World's Marine Fisheries: An Update, 470. Food and Agriculture Organization of the United Nations.
- Kindt-Larsen, L., Kirkegaard, E., and Dalskov, J. 2011. Fully documented fishery: a tool to support a catch quota management system. *ICES Journal of Marine Science*, 68: 1606–1610.
- Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2: 83–97.
- LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. *Nature*, 521: 436–44.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. *et al.* 2014. Microsoft COCO: common objects in context. *In Proceedings of the European Conference on Computer Vision*, pp. 740–755. Springer.
- Lu, Y.-C., Tung, C., and Kuo, Y.-F. 2019. Identifying the species of harvested tuna and billfish using deep convolutional neural networks. *ICES Journal of Marine Science*, 77: 1318–1329.
- Malde, K., Handegard, N. O., Eikvil, L., and Salberg, A.-B. 2019. Machine intelligence and the data-driven future of marine science. *ICES Journal of Marine Science*, 77: 1274–1285.
- McElderry, H., Schrader, J., and Illingworth, J. 2003. The efficacy of video-based monitoring for the halibut fishery. *Research Document*, 2003/042. Canadian Science Advisory Secretariat.
- Mortensen, L., Ulrich, C., Olesen, H., Bergsson, H., Berg, C., Tzamouranis, N., and Dalskov, J. 2017. Effectiveness of fully documented fisheries to estimate discards in a participatory research scheme. *Fisheries Research*, 187: 150–157.
- Needle, C., Dinsdale, R., Buch, T., Catarino, R., Drewery, J., and Butler, N. 2014. Scottish science applications of remote electronic monitoring. *ICES Journal of Marine Science*, 72: 1214–1229.
- Poos, J. J., Bogaards, J. A., Quirijns, F. J., Gillis, D. M., and Rijnsdorp, A. D. 2009. Individual quotas, fishing effort allocation, and over-quota

- discarding in mixed fisheries. *ICES Journal of Marine Science*, 67: 323–333.
- Punt, A. E., Smith, D. C., Tuck, G. N., and Methot, R. D. 2006. Including discard data in fisheries stock assessments: two case studies from south-eastern australia. *Fisheries Research*, 79: 239–250.
- Qiao, M., Wang, D., Tuck, G. N., Little, L. R., Punt, A. E., and Gerner, M. 2020. Deep learning methods applied to electronic monitoring data: automated catch event detection for longline fishing. *ICES Journal of Marine Science*, 78: 25–35.
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. 2016. You only look once: unified, real-time object detection. *In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788.
- Redmon, J., and Farhadi, A. 2017. Yolo9000: better, faster, stronger. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271.
- Redmon, J., and Farhadi, A. 2018. Yolo3: an incremental improvement. *arXiv preprint arXiv : 1804.02767*. preprint: not peer reviewed.
- Rijnsdorp, A., Daan, N., Dekker, W., Poos, J., and Van Densen, W. 2007. Sustainable use of flatfish resources: addressing the credibility crisis in mixed fisheries management. *Journal of Sea Research*, 57: 114–125.
- Rochet, M.-J., and Trenkel, V. M. 2005. Factors for the variability of discards: assumptions and field evidence. *Canadian Journal of Fisheries and Aquatic Sciences*, 62: 224–235.
- Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. 2011. ORB: an efficient alternative to SIFT or SURF. *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 2564–2571.
- Shafait, F., Mian, A., Shortis, M., Ghanem, B., Culverhouse, P. F., Edgington, D., Cline, D. *et al.* 2016. Fish identification from videos captured in uncontrolled underwater environments. *ICES Journal of Marine Science*, 73: 2737–2746.
- Siddiqui, S., Malik, I., Shafait, F., Mian, A., Shortis, M., and Harvey, E. 2017. Automatic fish species classification in underwater videos: exploiting pretrained deep neural network models to compensate for limited labelled data. *ICES Journal of Marine Science*, 75: 90326628.
- Snyder, H., and Erbaugh, J. 2020. Fishery observers address arctic fishery discards. *Environmental Research Letters*, 15: 22525219.
- Stanley, R., Karim, T., Koolman, J., and McDerry, H. 2014. Design and implementation of electronic monitoring in the British Columbia groundfish hook and line fishery: a retrospective view of the ingredients of success. *ICES Journal of Marine Science*, 72: 1230–1236.
- Stock, B., Ward, E., Thorson, J., Jannot, J., and Semmens, B. 2019. The utility of spatial model-based estimators of unobserved bycatch. *ICES Journal of Marine Science*, 76: 255–267.
- Storbeck, F., and Daan, B. 2001. Fish species recognition using computer vision and a neural network. *Fisheries Research*, 51: 11–15.
- Tseng, C.-H., and Kuo, Y.-F. 2020. Detecting and counting harvested fish and identifying fish types in electronic monitoring system videos using deep convolutional neural networks. *ICES Journal of Marine Science*, 77: 1367–1378.
- Uhlmann, S. S., van Helmond, A. T. M., Stefánsdóttir, E. K., Sigurðardóttir, S., Haralabous, J., Bellido, J. M., Carbonell, A., *et al.* 2013. Discarded fish in European waters: general patterns and contrasts. *ICES Journal of Marine Science*, 71: 1235–1245.
- Ultralytics 2021. Yolo3. <https://github.com/ultralytics/yolov3> (last accessed 13 January 2021).
- van Helmond, A., Chen, C., and Poos, J.-J. 2017. Using electronic monitoring to record catches of sole (*Solea solea*) in a bottom trawl fishery. *ICES Journal of Marine Science*, 74: 1421–1427.
- van Helmond, A. T., Mortensen, L. O., Plet-Hansen, K. S., Ulrich, C., Needle, C. L., Oesterwind, D., Kindt-Larsen, L. *et al.* 2020. Electronic monitoring in fisheries: lessons from global experiences and future opportunities. *Fish and Fisheries*, 21: 162–189.
- Wojke, N., Bewley, A., and Paulus, D. 2017. Simple online and real-time tracking with a deep association metric. *arXiv preprint arXiv : 1703.07402*. preprint: not peer reviewed.
- Zion, B., Shklyar, A., and Karplus, I. 1999. Sorting fish by computer vision. *Computers and Electronics in Agriculture*, 23: 175–187.

Handling Editor: Cigdem Beyan

3 An integrated end-to-end deep neural network for automated detection of discarded fish species and their weight estimation

Maria Sokolova¹, Manuel Cordova¹, Henk Nap¹, Aloysius van Helmond², Michiel Mans¹, Arjan Vroegop³, Angelo Mencarelli³ and Gert Kootstra¹

¹ Wageningen University and Research, Farm Technology Group, Wageningen, 6700 AA, The Netherlands

² Wageningen University and Research, Wageningen Marine Research, IJmuiden, 1970 AB, The Netherlands

³ Wageningen University and Research, Greenhouse Horticulture Unit, Wageningen, 6700 AP, The Netherlands

An integrated end-to-end deep neural network for automated detection of discarded fish species and their weight estimation

Maria Sokolova^{1,*}, Manuel Cordova¹, Henk Nap¹, Aloysius van Helmond², Michiel Mans¹, Arjan Vroegop³, Angelo Mencarelli³, and Gert Kootstra¹

¹Wageningen University and Research, Farm Technology Group, Wageningen, 6700 AA, The Netherlands

²Wageningen University and Research, Wageningen Marine Research, IJmuiden, 1970 AB, The Netherlands

³Wageningen University and Research, Greenhouse Horticulture Unit, Wageningen, 6700 AP, The Netherlands

*Corresponding author: tel: +31 6 45097195; e-mail: maria.sokolova@wur.nl.

Sustainable management of aquatic resources requires efficient acquisition and processing of vast amounts of information to check the compliance of fishing activities with the regulations. Recent implementation of the European Common Fisheries Policy Landing Obligation implies the declaration of all listed species and sizes at the harbour. To comply with such regulation, fishers need to collect and store all discards onboard the vessel, which results in additional processing time, labour demands, and costs. In this study, we presented a system that allowed image-based documentation of discards on the conveyor belt. We presented a novel integrated end-to-end simultaneous detection and weight prediction pipeline based on the state-of-the-art deep convolutional neural network. The performance of the network was evaluated per species and under different occlusion levels. The resulting model was able to detect discards with a macro F1-score of 94.10% and a weighted F1-score of 93.88%. Weight of the fish could be predicted with mean absolute error, mean absolute percentage error, and root squared error of 29.74 (g), 23.78%, and 44.69 (g), respectively. Additionally, we presented a new dataset containing images of fish, which, unlike common object detection datasets, also contains weight measurements and occlusion level per individual fish.

Keywords: computer vision, fisheries, occlusion, YOLOv5.

Introduction

Sustainable management of aquatic resources requires versatile data from fishing activities, including catch information, which is needed to ensure fisheries state compliance with the regulations of fishing activities. In the European Union, implementation of the Landing Obligation, part of the reformed European Common Fisheries Policy, implies the discard ban of quota-related fish species and sizes (Article 15 of Regulation (EU) No 1380/2013). This means that instead of throwing the unwanted and non-target fraction of catch overboard—referred to as discarding—fishers are obliged to sort, store, and declare the complete catch, including discards, in their logbooks. The regulation was introduced to stipulate selectivity improvement of fishing activities, especially in demersal trawl fisheries, which are typically characterized by mixed catches and high discard rates (Kennelly and Broadhurst, 2021). Non-target species extraction has a substantial negative impact on ecosystem, and consequently on economic and social sustainability of future commercial fisheries. Discards monitoring is crucial as it contributes to a more precise estimation of the fishing activities and thus a better record of the total fishing mortality. Discard registration includes species determination and measurement of individual fish weight and length. Biomass estimation of catches is an important parameter in monitoring fish populations (Lado, 2016).

Currently, there are several incentives to monitor the number of discards on board fishing vessels, which include fish-

eries observer programmes and remote electronic monitoring (REM) systems. Observer programmes are carried out by employing specifically trained personnel who join fishing trips and record discard composition, including Protected, Endangered, and Threatened species, amount, and morphometrics, such as weight and length. Such programmes are expensive and time-consuming, and the sampling intensity, i.e. coverage of the fishing fleet, is low and potentially biased (Benoît and Allard, 2009). An alternative to observer programmes is the implementation of REM systems on board, which accumulates the information from several sensors integrated on the fishing vessel. This allows collecting diverse data about fishing activities, including location, duration, catch amount, and composition (van Helmond *et al.*, 2020). In REM, the discard quantities and composition can be registered via closed-circuit television (CCTV) cameras placed at the conveyor belt of the fishing vessel. A complete manual video analysis is infeasible due to being time-consuming and labour-intensive (Underwood *et al.*, 2014; van Helmond *et al.*, 2020). This is a bottleneck for the system implementation on a fleet-level scale. This limitation can be overcome with integration of the component, which can reliably detect species and predict the individuals' weight, providing the total number and weight of discarded fish per species.

In return, automated analysis of fish discards on board commercial demersal trawlers is challenged by the high number of species and the quality of the obtained images. Moreover, the

Received: 25 April 2023; Revised: 6 July 2023; Accepted: 7 July 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of International Council for the Exploration of the Sea. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

images can contain many overlapping fish and debris, which may cause drastic occlusions. Several studies (French *et al.*, 2015, 2020; Tseng and Kuo, 2020) reported that the quality of obtained images with REM systems is often suboptimal for efficient fish detection. This is due to changing illumination conditions above the conveyor belt and on the deck of fishing vessels. Besides, the camera lens can be blurred by the water drops presence, originating from high humidity or weather conditions. In addition to detecting fish species, several contributions demonstrated automated weight and length estimation (Balaban *et al.*, 2010; Saberioon and Císař, 2018; Konovalov *et al.*, 2019; Tseng and Kuo, 2020; Ovalle *et al.*, 2022). However, those studies applied fish-metrics estimation as an additional step after object detection. Multi-step approaches are often based on feature engineering and direct use of objects' physical properties, which are then converted into weight or length estimations. Such an approach is not always semantic and heavily relies on the objects' detected area. This complicates the real-world application on board the fishing vessels, where observation conditions are typically highly variable.

In this study, we proposed a solution to outrange the existing challenges with a dedicated image-acquisition system and image processing pipeline for efficient discard species and weight registration. The system provided high-resolution images of discards on the conveyor belt. The image processing pipeline and an integrated end-to-end approach provided simultaneous detection, classification, and weight estimation under different occlusion levels of fish. Efficiency of the discards registration system provides the ground for a win-win situation for fishers, fisheries administrators, and scientists. Specifically, discards documentation eliminates the need to land the (by regulation) unmarketable, and therefore, useless part of the catch. This gives an opportunity to alter the Landing Obligation towards a more efficient and practical discards Registration Obligation. Moreover, detailed information about the composition and number of discards will allow for more precise estimation of individual fisher's contribution to the overall fishing pressure. This is needed for pragmatic fisheries administration and sustainable management under the current EU Common Fisheries Policy.

The contributions of this paper can be summarized as follows:

- We presented a complete system for image-based documentation of discards on the conveyor belt. The system included the dedicated image-acquisition hardware, which provided high-resolution images.
- We proposed an integrated end-to-end neural network approach, specifically a non-trivial modification of YOLOv5 with an additional output for integrated fish detection and weight estimation. Four training routines were tested to find an optimal policy, assessing the impact of using: (i) a dataset, specifically designed for FD, for pre-training our models and (ii) using different sets of hyperparameters and training stages.
- We introduced a novel publicly available dataset, the Fish Detection and Weight Estimation (FDWE) dataset, composed of more than 1000 images containing nine fish species. The annotations of our dataset not only presented information related to the localization and species of the fish but also its weight and level of occlusion.
- To assess the occlusion influence on the detection and weight estimation, we evaluated the performance of the

proposed approach according to the defined four levels of occlusion.

Material and methods

In this section, we describe: (i) data collection, data pre-processing, and annotation methods; (ii) the description of modified YOLOv5 architecture and training routines as well as evaluation methods; and (iii) evaluation methods of the proposed approach.

Data collection

In this subsection, we present the dedicated image-acquisition system developed to record the discards on the conveyor belt. Further, we present a pipeline for image pre-processing and a description of the collected dataset together with the annotation.

Image-acquisition system

The image-acquisition system comprises a metal box with dimensions (*Width* \times *Height* \times *Length*) of $60 \times 59 \times 60$ cm with four attachment points for secure fix at the conveyor belt (Supplementary Figure S1). A module containing a camera, illumination, and a computer was protected from the outside environment, i.e. humidity and salt, in the sealed top compartment of the metal box. A Karbon 410 Intel Elkhart Lake Compact Rugged Computer (www.onlogic.com/nl-nl/k410/) with a passive cooling system was coupled with the camera in the top compartment. To ensure sufficient and consistent illumination, LEDs were tilted to avoid reflections from the wet surfaces (Supplementary Figure S1). Raw images were taken with FramosTM D415e (www.framos.com/en/industrial-depth-cameras), focal length 1.88 mm, and with an image resolution 1920×1080 pixels.

Datasets description

The proposed computer vision method performed two tasks: (i) discards detection and (iii) weight estimation. To train and test the method, two datasets were used. The first one, referred to as Fish Detection (FD), is an open-source dataset published by van Essen *et al.* (2021). The FD dataset contained ten species that are typically discarded in Dutch demersal beam trawlers operating in the North Sea. The set of images was collected by the custom image-acquisition system placed on top of the conveyor belt. System specifications are described in van Essen *et al.* (2021). The images were extracted from the video recordings; therefore, the presence of the same fish in consecutive frames was typical in the dataset. In our training pipeline, the FD dataset was used only in the first training stage to initialize the weights of the neural network for FD, since it did not contain the record of individual fish weight. Thus, the FD dataset was not used for testing.

In addition to solving the object-detection task, we aimed at predicting weight per detection. To train and test the integrated end-to-end detection and weight estimation method, we collected the FDWE dataset. This dataset contained 1086 images depicting nine common species presented in the discards (Figure 1). To the best of our knowledge, this is the first dataset fulfilling this requirement, specifically containing annotated images with bounding box, class label, and weight annotation per individual fish. By using these two diverse

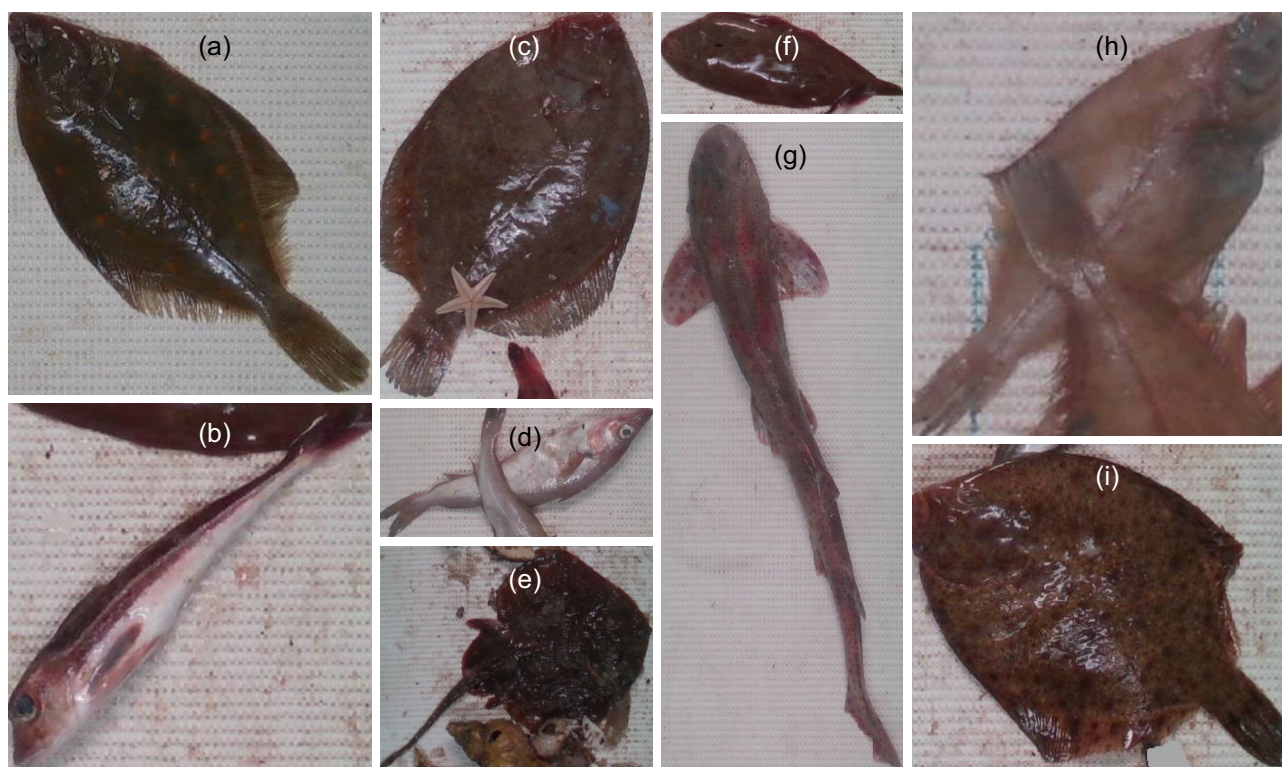


Figure 1. Examples of the fish species cropped from the original images of the Fish Detection and Weight Estimation (FDWE) image dataset according to the corresponding bounding box annotation. The species are listed as follows: (a) *Pleuronectes platessa*, (b) *Eutrigla gurnardus*, (c) *Scophthalmus maximus*, (d) *Merlangius merlangus*, (e) *Amblyraja radiata*, (f) *Solea solea*, (g) *Scyliorhinus canicula*, (h) *Limanda limanda*, and (i) *Scophthalmus rhombus*.

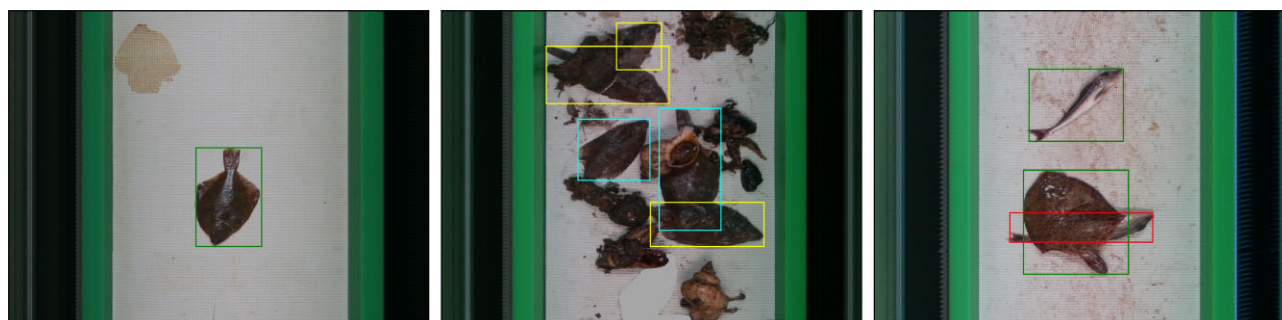


Figure 2. Examples of images of discarded fish on the conveyor belt with the four defined occlusion levels. The colour of the bounding box indicates the level of occlusion, as follows: green: 0%, cyan: 1–30%, yellow: 31–60%, and red: 61–90%.

datasets to train and test our method, we accounted for variation in fish species appearance, sizes, and orientation.

Linescan images generation

The computer placed in the top compartment of the image-acquisition system was used to pre-process raw RGB images. This pre-processing step merged the consecutive frames into semi-linescan images (Supplementary Figure S2). The approach allowed avoiding the issue with the same fish being visible in the sequential frames. This was accomplished in two steps: (i) determining the displacement of the conveyor belt, and (ii) stitching the images. The conveyor belt speed could be controlled by the fisher during the catch sorting process. Besides, during the fishing operation, the belt was stopped, while the camera was still on. For this reason, prior to linescan image generation, it was important to estimate the con-

veyor belt speed based on the raw input images. To reliably do this, we used the block matching technique. The block was defined with 240×240 pixels region of interest (ROI) selected in the two consecutive frames, further the absolute difference between the two ROIs was calculated (Bradski, 2000). The absolute difference between the ROIs in the two frames was calculated 50 times after both ROIs were shifted by one pixel in the direction of the conveyor belt movement. The smallest absolute difference calculated between the two ROIs defined the height of the image region that was then cropped from the consecutive frame and pasted. The resulting linescan images were saved as separate files with the height of 1920 rows. The second step included the reduction of the partially visible fish that were cut between the frames. To tackle this issue, the generation of the new linescan image has been initialized from the middle of the previous frame. In the final processing step,

Table 1. Summary of Fish Detection (FD) image dataset indicating the number of instances per species split into train, validation, and test subsets.

Species	#Annotations		
	Training	Validation	Test
<i>Amblyraja radiata</i>	1073	330	302
<i>Callionymus lyra</i>	25	–	–
<i>Eutrigla gurnardus</i>	187	48	31
<i>Limanda limanda</i>	1885	540	322
<i>Merlangius merlangus</i>	3548	988	529
<i>Microstomus kitt</i>	978	227	111
<i>Pleuronectes platessa</i>	4413	1153	739
<i>Scyliorhinus canicula</i>	254	62	44
<i>Scophthalmus maximus</i>	109	47	15
<i>Trisopterus luscus</i>	69	7	14
<i>Solea solea</i>	3484	901	452

Table 2. Summary of the Fish Detection and Weight Estimation (FDWE) image dataset indicating the number of instances per species split into train, validation, and test subsets.

Species	#Annotations			Mean weight \pm SD (g)
	Training	Validation	Test	
<i>Amblyraja radiata</i>	40	13	6	323 \pm 22
<i>Eutrigla gurnardus</i>	224	61	25	127 \pm 70
<i>Limanda limanda</i>	122	35	15	80 \pm 28
<i>Merlangius merlangus</i>	59	20	8	108 \pm 31
<i>Pleuronectes platessa</i>	876	206	101	138 \pm 72
<i>Scyliorhinus canicula</i>	20	5	5	560 \pm 123
<i>Scophthalmus maximus</i>	35	28	11	659 \pm 542
<i>Scophthalmus rhombus</i>	49	7	3	467 \pm 165
<i>Solea solea</i>	170	49	23	134 \pm 70

white masks were manually applied to the remaining partially visible fish. The resulting images' dimensions were 1280 by 1280 pixels.

Image annotation

In the case of the FD dataset, both images and their annotations were used directly from the publicly available dataset (van Essen *et al.*, 2021). In the case of FDWE dataset, images were annotated from scratch via Darwin v7 Software (<https://darwin.v7labs.com>). The annotation process included marking every object of interest with a polygon with its class label and the ground truth weight value of individual fish. Annotation process consisted of an initial annotation and a review of the annotation, which has been done by two separate human annotators. The resulting annotations were then converted to YOLO bounding box format. Thus, each annotation (d_i) of the FDWE dataset contained seven values: $d_i = \{c_i, x_i, y_i, w_i, b_i, w_i, o_i\}$, where c_i was the class ID, x_i, y_i were the coordinates of the bounding box centre, w_i, b_i were width and height of the bounding box, w_i was weight, and o_i was the occlusion level. The resulting number of annotations in the FD and the FDWE datasets are summarized in Tables 1 and 2, respectively. Additionally, the FDWE dataset was collected by simulating diverse levels of object occlusion. Specifically, four levels of occlusion estimated by human annotator were defined: 0%, 1 – 30%, 31 – 60%, and 61 – 90% (Figure 2). The number of fish under the defined occlusion levels in each of the subsets of the FDWE dataset is presented in Table 3.

Proposed method for FD and weight prediction

In this section, we describe modifications applied to a deep convolutional neural network, which, in addition to solving

object detection task, enabled weight prediction per detection. We further present training routines of the network.

Modified YOLOv5 with additional regression output

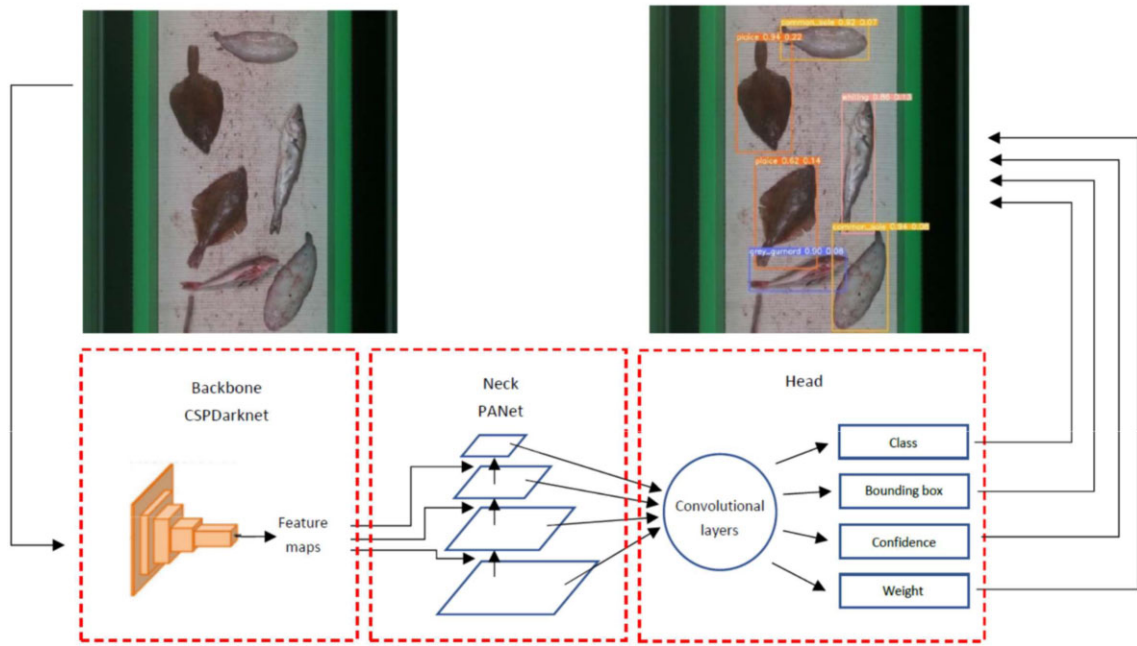
In this study, a modified version of YOLOv5 (Jocher *et al.*, 2022) was proposed to accomplish two tasks in an integrated end-to-end approach: (i) FD and (ii) fish weight estimation. The state-of-the-art architecture has been extended with an additional output for prediction of fish weight per detection (Figure 3). This implied changing the number of outputs per anchor from five to six, accordingly, to allow learning weight regression in addition to class probabilities, four bounding box coordinates, and confidence. The eventual three outputs for weight, class, and bounding box were produced in parallel, given the 2D RGB input image. Therefore, the weight prediction did not rely on the bounding box itself. The YOLOv5 architecture was available in a range of sizes corresponding to the different number of convolutional layers in the model. In this study, the YOLOv5 small version was used due to its highest inference speed, which is essential for the real-world implementation.

In addition to changing the number of outputs per anchor, the loss function needed to be modified to enable weight-regression learning during training. Thus, along with the bounding box regression loss (L_{box}), the confidence loss (L_{obj}), and the classification loss (L_{cls}) (Jocher *et al.*, 2022), we included the weight-regression loss (L_{weight}), calculated as mean squared errors loss:

$$MSE = \frac{1}{N} \sum_{i=1}^N (\hat{weight}_i - weight_i)^2. \quad (1)$$

Table 3. Number of instances in train, validation, and test sets in the Fish Detection and Weight Estimation (FDWE) image dataset per occlusion level.

Subset	Occlusion level			
	0%	1–30%	31–60%	61–90%
Train	877	259	354	105
Validation	234	71	88	31
Test	108	41	35	13

**Figure 3.** Overview of the YOLOv5 architecture with an additional output for weight regression.

Thus, the complete loss function was defined as follows:

$$Loss = \lambda_{box} \cdot L_{box} + \lambda_{obj} \cdot L_{obj} + \lambda_{cls} \cdot L_{cls} + \lambda_{weight} \cdot L_{weight} \quad (2)$$

Contribution of the weight loss was regulated by the λ_{weight} coefficient, which has been empirically defined.

Training of the models

Ground truth values for the fish weight were labour- and time-consuming to obtain. Therefore, we decided to use two datasets efficiently and estimate the contribution of the pre-training step on the FD dataset to the overall model performance. However, this raises the question about how to use both datasets efficiently during training. To answer this question, we explored four training routines (Figure 4). Routine 1 was trained and tested only on the FDWE dataset with default values for all hyperparameters. Routine 2 comprised an additional training step of the Weight model resulting from Routine 1 with increased λ_{weight} value from 0.015 to 0.05 corresponding to increased penalty for the erroneous weight prediction. At the same time, to ensure model convergence, the learning rate was reduced from 0.01 to 0.001. To minimize a bias caused by image augmentations on the weight-regression learning, only flipping from left to right was used with a 50% probability as well as translation with 10% probability. During the training Routines 3 and 4, the FD dataset was used for pre-training the models with the goal of assessing its influence on the final model performance. Since the FD dataset did

not have individual fish weight data, the λ_{weight} was set to 0. The first training step included training with the same hyperparameters as in Routine 1. The aim of the training Routine 3 was simultaneous learning of both detection and weight, with λ_{weight} parameter set to 0.015. The Routine 4 included an additional fine-tuning step on the FDWE dataset to emphasize weight learning, similarly as in the Routine 2, stipulated by the increased λ_{weight} to 0.05 and learning rate reduction down to 0.001.

Evaluation

Evaluation of detection performance

Performance of object detection was assessed using the F1-macro (Equation 4) and F1-weighted (Equation 5) scores. These two metrics were derived from the class-dependent F1 score, which was calculated as follows:

$$F1 = \frac{TP_c}{TP_c + \frac{1}{2}(FP_c + FN_c)} \quad (3)$$

where TP_c was the number of true positive detections, i.e. the detections with correct class and bounding box predicted, which was defined by the confidence score threshold of 0.25 and IoU threshold set to 0.45, the default values used by Ultralytics (Jocher *et al.*, 2022). FP_c was the number of false positive detections, i.e. the predictions with the wrong class and an IoU value below the threshold; and FN_c was the number of false negative detections, corresponding to ground truth ob-

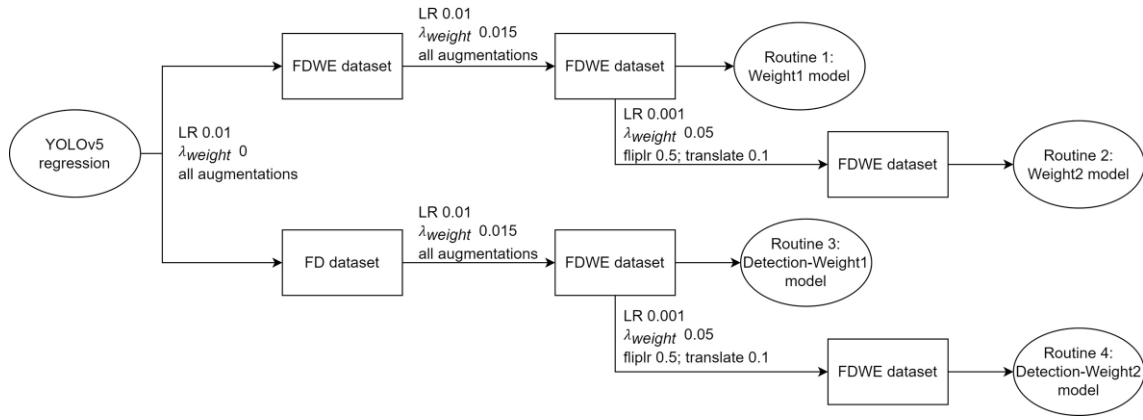


Figure 4. Flowchart illustrating training stages and hyperparameters for the four training routines of the YOLOv5 with an additional output for weight regression. Arrows correspond to training the model for 300 epochs.

jects that were not detected. The F1-macro and F1-weighted were calculated as:

$$\text{F1-macro} = \frac{1}{C} \sum_{c=1}^C \text{F1}_c, \quad (4)$$

$$\text{F1-weighted} = \frac{\sum_{c=1}^C N_c \cdot \text{F1}_c}{\sum_{c=1}^C N_c}, \quad (5)$$

where F1_c is the F1-score for class c , C is the number of classes, and N_c is the number of fish of class c . The F1-macro treats equally every class, while F1-weighted takes the class size into account.

The detection performance was evaluated and compared for the four training routines to investigate the effectiveness in detecting different fish species. The performance was evaluated for all the fish in the test subset and for each species separately to investigate if there were any species-specific differences in the detection performance. For the non-maximum suppression, the thresholds for the confidence (0.4) and IoU (0.6) were defined using a grid search executed on the original validation set.

Comparison of the four training routines using 5-fold cross-validation

The comparison between the four training routines was carried out through 5-fold cross-validation (Bishop, 2006). For this analysis, the training set of the FDWE dataset was split into five subsets. During 5-fold cross-validation, the model from the last epoch was used to compare the four training routines. That model was chosen because it was not influenced by the validation set during the training process, which was the case of the “best” model. After the 5-fold cross-validation procedure, the best training routine was defined.

Evaluation of weight prediction

To estimate the weight prediction performance, we used three metrics: (i) mean absolute error (MAE) [g] (Equation 6), (ii) mean absolute percentage error (MAPE) [%] (Equation 7), and (iii) root mean square error (RMSE) [g] (Equation 8).

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|, \quad (6)$$

$$\text{MAPE} = \frac{100\%}{N} \sum_{i=1}^N \left| \frac{\hat{y}_i - y_i}{y_i} \right|, \quad (7)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2}, \quad (8)$$

where \hat{y} was the prediction of weight in grams, y was the true weight in grams, and N was the total number of fish. The MAE gave the absolute prediction error in grams, the MAPE provided the error as a percentage proportional to the weight of the fish, and the RMSE measured the root-mean-squared error in grams, which was more influenced by outliers due to the square of the error used. The weight estimation errors were calculated for only true positive detections.

Results

In this section, we presented the results of the proposed approach performance to detect and estimate the weight of the discarded fish. We began with the comparison of the models resulting from the four training routines evaluated with 5-fold cross-validation. The best-performing model was then evaluated on the species-level in terms of detection and weight prediction. Finally, the best model performance was estimated according to different occlusion levels.

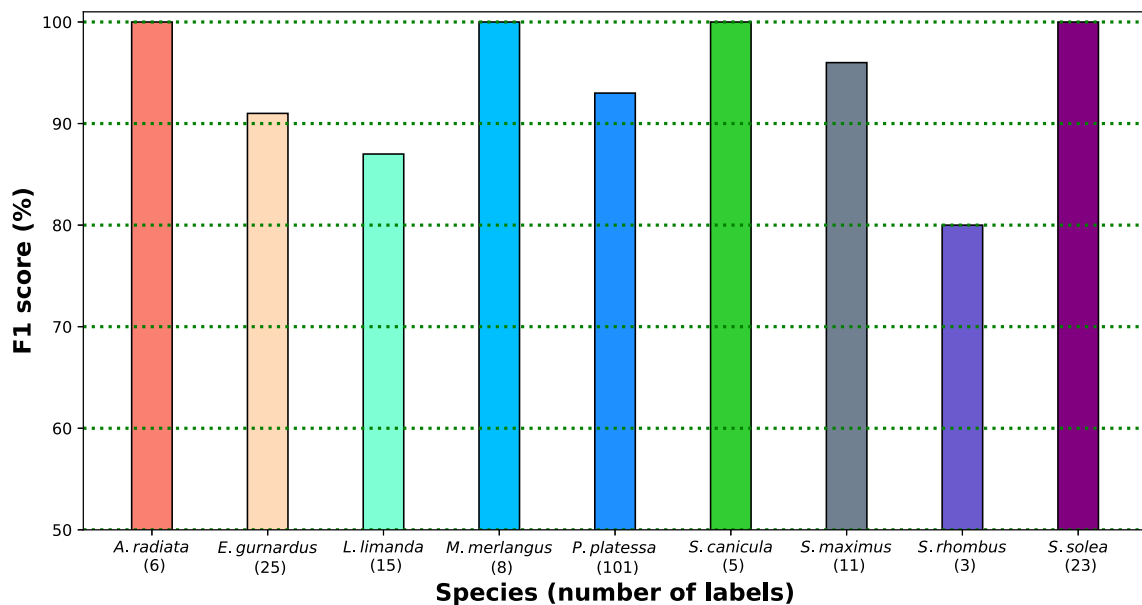
Discards detection and weight prediction

All the models scored above 93 and 92%, in F1-macro and F1-weighted, respectively. The relative weight estimation error (MAPE) was below 32% for all the models (Table 4). The best-performing model was obtained via the 4th training Routine (Figure 4) and referred to as Detection-Weight2 model. This model had the highest values for both F1-macro and F1-weighted. There was a distinctive difference in the F1 scores between Weight2 and Detection-Weight2 models. Specifically, the latter had, on average, a 5% increase in both F1-macro and F1-weighted scores compared to the other three models. The Detection-Weight2 model outperformed Detection-Weight model by 5.21% for F1-macro and 4.27% for the F1-weighted; and by 5.47 and 4.58% comparing to the F1-macro and F1-weighted, respectively, of the Weight2 model. The effect of pre-training stage on the FD dataset, however, did not contribute to a substantial improvement in model performance resulting from the training Routine 3 compared to the Weight models.

Table 4. Detection performance metrics and errors in weight prediction obtained via 5-fold cross-validation of the YOLOv5 with an additional output for weight regression.

Model	Detection		Weight prediction		
	F1-macro (%)	F1-weighted (%)	MAE (g)	MAPE (%)	RMSE (g)
Weight	93.93 \pm 1.36	92.60 \pm 1.24	40.35 \pm 2.35	27.94 \pm 1.92	67.91 \pm 7.07
Weight2	93.04 \pm 3.78	92.45 \pm 1.66	31.88 \pm 2.21	21.99 \pm 2.11	51.98 \pm 3.52
Detection-Weight	93.30 \pm 2.90	92.76 \pm 0.78	43.17 \pm 3.79	31.04 \pm 2.00	69.06 \pm 7.51
Detection-Weight2	98.51 \pm 0.92	97.03 \pm 0.80	24.61 \pm 1.16	17.72 \pm 1.71	37.59 \pm 2.12

The mean detection performance metrics and errors in weight prediction are presented with the corresponding \pm standard deviations.

**Figure 5.** F1 scores per species in the test subset of the Fish Detection and Weight Estimation (FDWE) dataset for the Detection-Weight2 model. The number of individuals in a test set is indicated in brackets.

The error in weight prediction drops after the second fine-tuning step on the FDWE dataset, indicated by the values comparison between Weight and Weight2 models and Detection-Weight and Detection-Weight2 models. The Weight2 model reduces the error of Weight model by 8.47 (g) for MAE, by 5.95% for the MAPE, and by 15.93 (g) for the RMSE. In the case of Detection-Weight models, the last training stage leads to a more considerable difference. Specifically, 18.56 (g) in MAE, 13.32% in MAPE, and 31.47 (g) in RMSE.

Based on the 5-fold cross-validation, Detection-Weight2 was chosen as the best performing model among the four training routines. To report the results on the original test set of the FDWE dataset, we trained on the whole training set using the best training routine. Afterwards, the original test set was used to obtain performance metrics. Thus, F1-macro and F1-weighted scores equaled to 94.10 and 93.88%, respectively; errors in weight prediction equaled to as follows: MAE to 29.74 (g), MAPE to 23.78%, and RMSE to 44.69 (g).

Per species analysis

The analysis of the detection and weight prediction showed variation in performance between species. Detection accuracy per species is summarized in the two figures: Figure 5 indicates F1 scores per species and Figure 6 is a confusion matrix for

the Detection-Weight2 model. For four out of nine species, F1 score reached 100%. The lowest F1 score was obtained for *Scophthalmus rhombus*, which was also not well represented in the dataset, due to its low occurrence frequency in the real catches data. This species was often confused with *S. maximus*, to which they were similar in appearance (Figure 6). Notably, this class together with *A. radiata*, *Merlangius merlangus* and *Scyliorhinus canicula* contributed to the minority species of the FDWE dataset. The lower F1 score for *S. rhombus* indicated that with such a limited number of instances the model was not able to learn enough features to reliably detect this class. Moreover, for this class there was the fewest number of instances in the test set, which may have led to bias in model performance. On the contrary, the other three minority species were not biased during the detection, which could be explained by their distinct appearance from the other species (Figure 1).

Figure 7 presents weight prediction errors for each species in the FDWE dataset. The lowest error values in weight prediction evaluation signaled more accurate weight prediction. Thus, the minimum MAE and RMSE were observed for *M. merlangus*, whereas the lowest Mean absolute percentage error (MAPE) was obtained for *S. canicula*. The highest weight prediction errors were observed for *S. rhombus*.

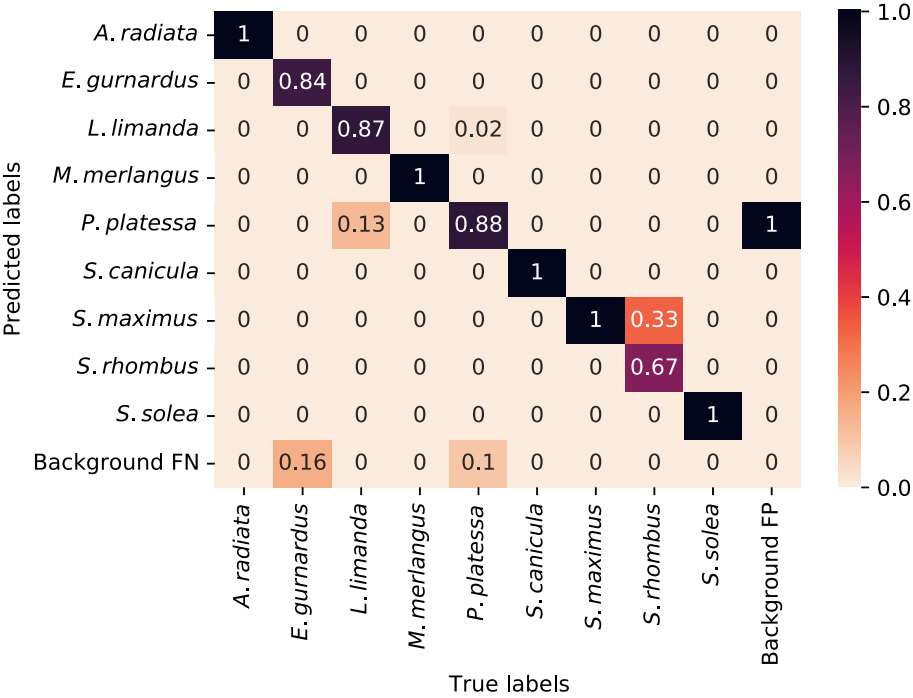


Figure 6. Confusion matrix for the Detection-Weight2 model evaluated on the test subset of the Fish Detection and Weight Estimation (FDWE) dataset.

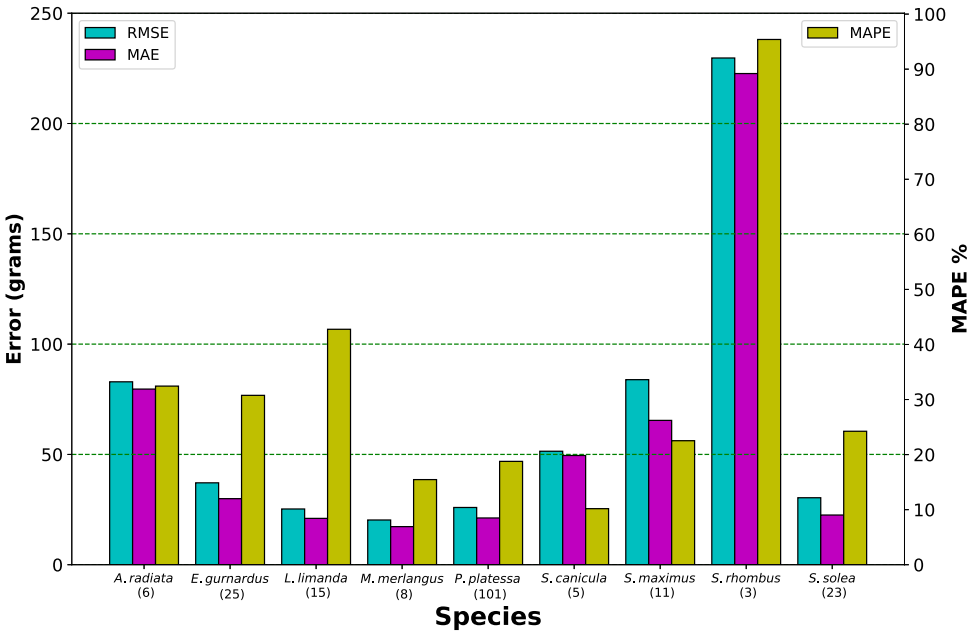


Figure 7. Root mean square error (RMSE), Mean absolute error (MAE), and Mean absolute percentage error (MAPE) for weight prediction per species for the Detection-Weight2 model evaluated on the test subset of the Fish Detection and Weight Estimation (FDWE) dataset. The number of individuals in a test set is indicated in brackets.

Discards documentation under different occlusion levels

We have also evaluated the impact of occlusion on both detection and weight prediction performance of the Detection-Weight2 model. For this, each fish instance of our dataset has an associated occlusion level, i.e. 0% (fully visible), 1–30%, 31–60%, and 61–90% occluded. Detection performance was measured using recall, since it indicated the number of false

negative detections, which was typically the case when the fish were not fully visible. As we can see in Figure 8, most of the species were depicted under 0% and 1–30% occlusion levels. A few instances of *Eutrigla gurnardus*, *Limanda limanda*, *Pleuronectes platessa*, and *Solea solea* were captured by the camera under 31–60% of occlusion. Images with fish under the highest occlusion level (61–90%) were represented only by two species: *P. platessa* and *E. gurnardus*, which appeared

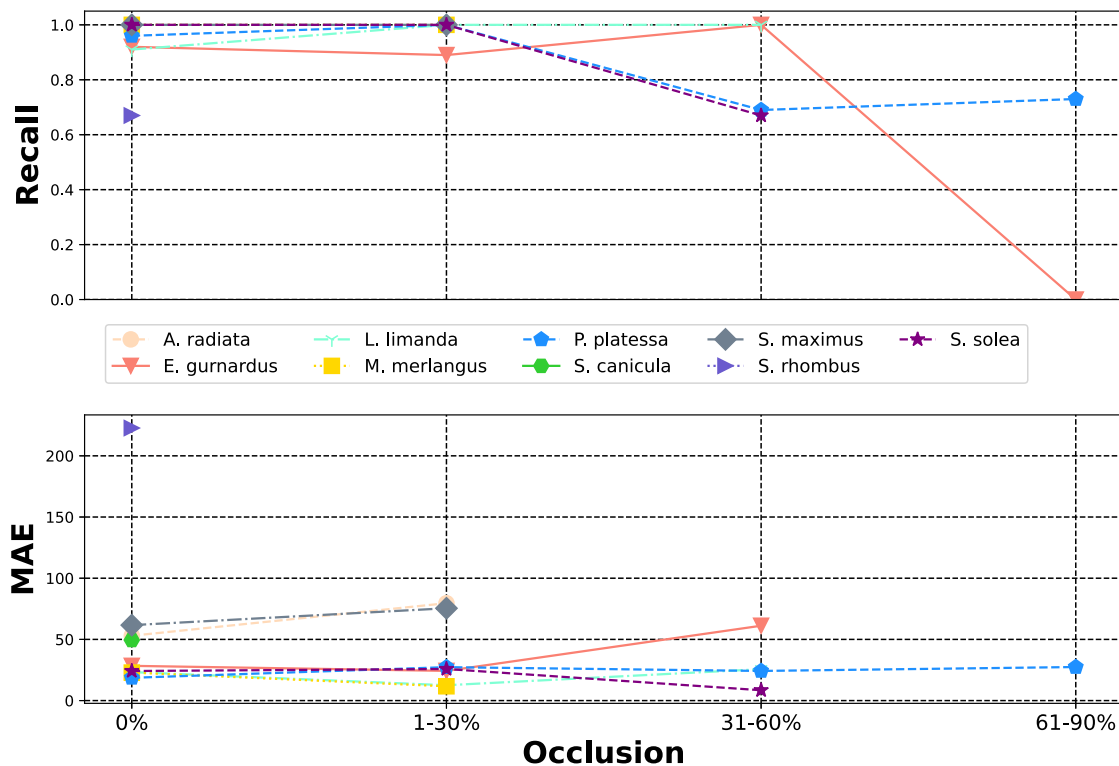


Figure 8. Recall (first row) and Mean absolute error (MAE) (second row) per species by occlusion level. The recall and MAE per species are separated by the colour code and different symbols at each of the defined occlusion levels. *Amblyraja radiata* and *Scyliorhinus canicula*, have a recall of 1 in the first two occlusion levels.

most frequently in the test set. Yet, our model was able to correctly detect some *P. platessa* instances even with an occlusion level of 61–90%, which was indicated by the recall score of 0.73 (Figure 8). In the case of *E. gurnardus*, represented by two instances under 61–90% of occlusion level, none were detected by the model. In both cases, our approach only kept detecting the fully visible fishes covering the occluded instances (see Supplementary Figure S3). Overall, the recall stays high for 1–30% occlusion; however, it drops with higher levels of occlusion.

In most of the cases, we can see that the greater the occlusion level, the greater the MAE becomes (Figure 8). In the case of *P. platessa*, the only species with true positives in all occlusion levels. For this species, the lower MAE was obtained among the fully visible instances (0% occlusion), in the remaining scenarios, the MAE was remarkably similar, but the last scenario (31–60%) was the one with the highest MAE. On the contrary, *L. limanda* obtained lower MAE with fishes occluded in the range of 1–30% than fully visible fishes. One of the reasons for this result is that most of the fully visible cases were placed in backgrounds with debris, which may have made the task more challenging. Then, as expected, the greater MAE (25.65) was reported with the higher occlusion level—31–60%. Moreover, in the case of *S. solea*, our neural network could obtain a MAE of 8.46 from three fishes with occlusion level of 31–60%, lower than fully visible fishes and fishes with an occlusion level between 1–30%.

Discussion

In this work, we proposed a dedicated discard registration system that automatically predicts species composition and fish weight on the conveyor belt. In this section, the results will be critically discussed and put in perspective of related work.

Automated discards detection and weight prediction

Our final YOLOv5 model with weight regression output reached F1-macro 94.10 and 93.88%. These results were obtained via a training routine, which included fine-tuning the network on the FD dataset and double fine-tuning on the FDWE dataset with different hyperparameters. The proposed training approach led to high detection performance and significant reduction in weight prediction error. The obtained results were achieved using two dedicated datasets of discards, choice of hyperparameters, and image augmentation techniques. Such image augmentations as colour and geometric transformations may have compromised weight prediction by altering the appearance of fish. Therefore, the weight prediction error has been reduced by not implementing these augmentations during the second fine-tuning on the FDWE dataset.

A direct comparison of our results with related studies on FD is difficult to make. First, different methods have been used to acquire images and different compositions of species.

Second, different approaches have been taken, such as object detection and instance segmentation. Finally, the evaluation methods are often dataset-dependent. In the following section, we will compare our results to those reported in the literature. However, it is important to consider the limitations mentioned in the comparison.

The detection performance varied per species. In our case, *S. rhombus* had the lowest F1 score; however, we only had few instances for this species, making it hard to reliably train the network and test the predictions for this class. One instance of *S. rhombus* was falsely identified as *S. maximus*, which is a similar-looking species belonging to the same genus—*Scophthalmus*. In the study of French *et al.* (2020), a similar result with *M. merlangus* and *Melanogrammus aeglefinus* was observed, as well as with *L. limanda* and *P. platessa*.

We obtained a higher detection performance compared to our previous work (van Essen *et al.*, 2021). In that study, the YOLOv3 network has been used to perform discards detection, resulting in an F1-macro score of 70% and an F1-weighted score of 80%, compared to 94.10% and 93.88% achieved in our current work. In van Essen *et al.* (2021), the FD dataset was used to train the method, whereas in our work, we used this dataset for pre-training and fine-tuned the method using the FDWE dataset, thus using more data. Moreover, we used YOLOv5, an improved version of YOLO.

The other studies devoted to discards detection report lower classification and detection accuracy (French *et al.*, 2020; Ovalle *et al.*, 2022). In these studies, however, the images were obtained with the CCTV cameras fixed above the conveyor belt, without a dedicated illumination. Also, the images were more complex due to people and objects in the camera field of view. Besides, multiple algorithms were used to solve classification, object detection, and segmentation tasks separately. This contrasts our approach, where a single integrated end-to-end algorithm solved multiple tasks simultaneously, in particular, discards detection, classification, and weight prediction. To our knowledge, we are the first to propose a single deep neural network as an integrated approach to process images to perform these tasks.

During the analysis of the weight prediction accuracy, we have considered only the true-positive detections, as for false-positive detections, no error can be predicted. In a real-world application, the method will predict weight for every detected fish, irrespective whether the class is correct or not. This will result in a lower accuracy of the weight prediction of the catch composition.

Influence of the occlusion level on discards registration

Our results show an overall decrease in detection performance in the images with the highest occlusion level. Nonetheless, a high detection rate remains for the fish with up to 30% occlusion levels. We performed an analysis per species, showing that recall decreases with the higher occlusion levels. However, not all species were represented by instances in all occlusion levels, meaning that a full analysis of all species over all occlusion levels could not be made. Additionally, the number of represented instances is uneven for the different levels, with sometimes only few instances for a specific occlusion level, making it impossible to untangle if changes in performance were due to the occlusion level or due to low number of train-

ing samples for that specific occlusion level. Similar results have been obtained by van Essen *et al.* (2021), who showed that the detection performance decreased drastically for the images under the highest occlusion levels. The authors of the study focused on the F1 score as a metric indicating detection performance, while we have considered recall specifically, since the increase in occlusion level typically results in false negatives.

Our study is the first one where the influence of occlusion levels on weight prediction has been evaluated. We expected a drastic increase of the MAE as a function of the occlusion level. And although the results showed an increase for some species, for others, the weight could be estimated equally well for the more occluded fish. However, the evaluation on the higher occlusion levels was based on a few instances in a test set. Yet, we suggest that this shows the advantage of our integrated end-to-end approach, directly estimating fish weight from the images. A method that would first segment the fish to then perform base weight prediction on the fish segment would suffer from a partial segment due to occlusion.

It is likely that the method can learn to deal better with occlusion if more training data would be collected. Also, specific data augmentations to artificially create more occluded training samples would benefit the method. Another approach is to prevent occlusions in the first case. This could, for instance, be done by adding mechanical solutions to better spread individuals on the conveyor belt before being captured by the camera. However, under the significant space limitations onboard beam trawlers, the implementation of such solutions is challenging.

Real-world application of the discards-registration system

The approach was developed and evaluated on the dataset that included typical species encountered in the North Sea. Although the dataset is large compared to related work and covers several different trips at various locations and dates, it is still small in comparison to the huge variation in the appearance of discards in different parts of the North Sea and at different moments in time. Especially in the case of system application in other areas where new species are present or where the species distribution is significantly different, extension of the dataset is required. Given enough data, our method can detect and predict the weight of any other bycatch or discarded species. With the presented image-acquisition system, it is feasible to collect large image datasets. However, the collection of high-quality ground-truth annotations is time consuming and costly. To improve the selection of training instances that most benefit the deep neural network, one can make use of techniques like active learning (Gal and Ghahramani, 2016; Blok *et al.*, 2022). Moreover, specific data-augmentation techniques tailored to the fish data and the application of unsupervised and self-supervised pre-training methods will boost performance.

Considering our results and future endeavours to collect more training data, we deem it possible to achieve acceptable performance of the onboard discards-registration system. We believe it has the potential to transform the Landing Obligation to a Registration Obligation, releasing fishers of the burden to land all discards in the future. This is possible if an

acceptable performance of the onboard discards-registration system is achieved.

Conclusion

In this study, we described an efficient discards-registration system. For the first time, we demonstrated an integrated end-to-end deep learning approach for simultaneous detection and weight prediction of the discarded species from a single colour image. Additionally, we defined a training procedure to elevate the detection performance while minimizing weight prediction error. For detection, the best model reached F1-macro of 94.10% and F1-weighted of 93.88% on the test set. The weight prediction errors were 29.74 grams, 44.69 grams and 23.78% for MAE, RMSE, and MAPE, respectively. The overall detection performance starts to decrease in the images with the occlusion level exceeding 30%; however, species-specific variations from this trend were observed. The weight prediction errors show a trend of overall increase with the higher occlusion levels.

Additionally, we presented the FDWE dataset of images, which contained instances from nine fish species. Unlike typical datasets for object detection, our dataset also contains the weight and level of occlusion in addition to class and bounding box annotation. We made this dataset publicly available with the goal of contributing to the community and to foster the proposal of new machine-learning approaches for FDWE.

Successful implementation of the discards registration system on the fishing vessels can grant an exemption from the Landing Obligation. Besides, the information about catch composition and metrics can be used for fish monitoring and stock assessment purposes. The proposed integrated end-to-end deep neural network is a general-purpose method that can be trained to detect and estimate the weight of other species given enough training data and ground truth measurements are provided.

Supplementary material

Supplementary material is available at the *ICESJMS* online version of the manuscript.

Acknowledgements

Authors would like to acknowledge skippers and crew of the vessels involved in the Fully Documented Fisheries project for providing discarded fish. We also thank Wageningen Marine Research for their help in data collection and preparation.

Author contributions

- Conceptualization: all authors
- Data curation: Michiel Mans, Henk Nap, Manuel Cordova, Maria Sokolova
- Formal Analysis: Michiel Mans, Henk Nap, Manuel Cordova, Maria Sokolova
- Funding acquisition: Aloysius van Helmond, Angelo Mencarelli, Gert Kootstra
- Hardware: Arjan Vroegop, Angelo Mencarelli
- Methodology: Michiel Mans, Henk Nap, Manuel Cordova, Maria Sokolova, Gert Kootstra
- Project administration: Aloysius van Helmond, Angelo Mencarelli, Gert Kootstra

- Software: Michiel Mans, Henk Nap, Manuel Cordova, Maria Sokolova
- Supervision: Gert Kootstra
- Writing – original draft: Maria Sokolova
- Writing – review and editing: all authors

Funding

This study was done under the Fully Documented Fisheries project initiated by the Dutch Ministry of Agriculture, Nature and Food Quality and funded by the European Maritime and Fisheries Fund (grant agreement number 16302).

Conflict of interest

The authors have no conflicts of interest to declare.

Data availability

The data underlying this article are available in at <https://dx.doi.org/10.4121/a6d5a40e-0358-47cf-9ec1-335df0e4a3c3>.

References

- Balaban, M. O., Chombeau, M., Cırbacı, D., and Gümüş, B. 2010. Prediction of the weight of alaskan pollock using image analysis. *Journal of Food Science*, 75: E552–E556.
- Benoît, H. P., and Allard, J. 2009. Can the data from at-sea observer surveys be used to make general inferences about catch composition and discards? *Canadian Journal of Fisheries and Aquatic Sciences*, 66: 2025–2039. doi:10.1139/F09-116.
- Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer, 233 Spring Street, New York, NY 10013, USA.
- Blok, P. M., Kootstra, G., Elghor, H. E., Diallo, B., van Evert, F. K., and van Henten, E. J. 2022. Active learning with MaskAL reduces annotation effort for training Mask R-CNN on a broccoli dataset with visually similar classes. *Computers and Electronics in Agriculture*, 197: 106917. doi:10.1016/j.compag.2022.106917.
- Bradski, G. 2000. The OpenCV Library. *Dr. Dobbs' Journal of Software Tools*.
- French, G., Fisher, M., Mackiewicz, M., and Needle, C. 2015. Convolutional Neural Networks for Counting Fish in Fisheries Surveillance Video. *Workshop on Machine Vision of Animals and their Behaviour*.
- French, G., Mackiewicz, M., Fisher, M., Holah, H., Kilburn, R., Campbell, N., and Needle, C. 2020. Deep neural networks for analysis of fisheries surveillance video and automated monitoring of fish discards. *ICES Journal of Marine Science*, 77: 1340–1353. doi:10.1093/icesjms/fsz149.
- Gal, Y., and Ghahramani, Z. 2016. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In *International Conference on Machine Learning* 1050–1059pp. PMLR.
- Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., and Xie, T., *NanoCode012*, et al. 2022. *ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference*.
- Kennelly, S. J., and Broadhurst, M. K. 2021. A review of bycatch reduction in demersal fish trawls. *Reviews in Fish Biology and Fisheries*, 31: 289–318. doi:10.1007/s11160-021-09644-0.
- Konovalov, D. A., Saleh, A., Efremova, D. B., Domingos, J. A., and Jerry, D. R. 2019. Automatic Weight Estimation of Harvested Fish from Images. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*. 1–7pp. IEEE.
- Lado, E. P. 2016. *The Common Fisheries Policy: the Quest for Sustainability*. John Wiley and Sons, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, United Kingdom.

- Ovalle, J. C., Vilas, C., and Antelo, L. T. 2022. On the use of deep learning for fish species recognition and quantification on board fishing vessels. *Marine Policy*, 139: 105015. doi:10.1016/j.marpol.2022.105015.
- Saberioon, M., and Cisar, P. 2018. Automated within tank fish mass estimation using infrared reflection system. *Computers and Electronics in Agriculture*, 150: 484–492. doi:10.1016/j.compag.2018.05.025.
- Tseng, C.-H., and Kuo, Y.-F. 2020. Detecting and counting harvested fish and identifying fish types in electronic monitoring system videos using deep convolutional neural networks. *ICES Journal of Marine Science*, 77: 1367–1378. doi:10.1093/icesjms/fsaa076.
- Underwood, M. J., Rosen, S., Engås, A., and Eriksen, E. 2014. Deep vision: an in-trawl stereo camera makes a step forward in monitoring the pelagic community. *PLoS One*, 9: e112304. doi:10.1371/journal.pone.0112304.
- van Essen, R., Mencarelli, A., van Helmond, A., Nguyen, L., Batsleer, J., Poos, J.-J., and Kootstra, G. 2021. Automatic discard registration in cluttered environments using deep learning and object tracking: class imbalance, occlusion, and a comparison to human review. *ICES Journal of Marine Science*, 78: 3834–3846. doi:10.1093/icesjms/fsab233.
- van Helmond, A. T., Mortensen, L. O., Plet-Hansen, K. S., Ulrich, C., Needle, C. L., Oesterwind, D., Kindt-Larsen, L. *et al.* 2020. Electronic monitoring in fisheries: lessons from global experiences and future opportunities. *Fish and Fisheries*, 21: 162–189. doi:10.1111/faf.12425

Handling editor: Howard Browman

4 Technical report

Author: A. Mencarelli

4.1 Overview

The detection system, the Catch Wageningen Autonomous Monitoring System (CatchWAM), was developed specifically to acquire images and process these images using the algorithms developed in chapters 2 and 3 onboard a vessel during fishing activity.

To achieve those two activities, the CatchWAM system is designed using the concept of distributed hardware to concentrate the component's hardware and software for the image acquisition in the working area where the fish is sorted; while the image processing activity is in the wheelhouse of the vessel, see a draw of this concept in Figure 4.1.

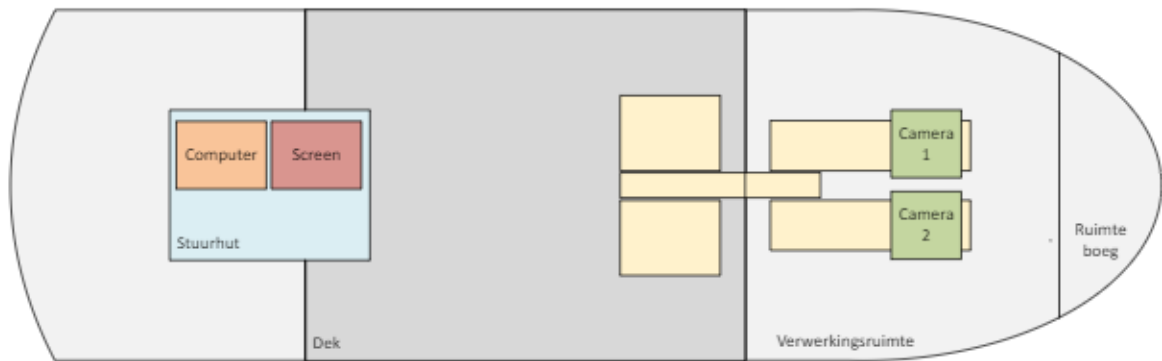


Figure 4.1. Overview of the distribution of the hardware on the vessel: the two green boxes represents the CatchWAM image acquisition system on the conveyor belts; the orange box the main computer in the wheelhouse.

The main reason for this choice is due to the rough environmental conditions that are present in the workspace area and the limited space in that area (Needle et al., 2015) that hampers the use of part electronics and computing power. Furthermore, the presence of working personnel in the area together with the mechanical stress and suboptimal light conditions imposed a radical design of the image acquisition component of the CatchWAM, for more details see the *Image-acquisition System* paragraph. To handle the image quality against the presence of dirtiness, a detection system based on a Laplacian Operator to detect the blurriness of the images during the acquisition process was implemented in the onboard processing unit inside the image-acquisition system.

To send the images from the image acquisition area to the image processing area we used an ethernet platform based on ZeroMQ protocol (Hintjens. 2013), see also the *Communication System* paragraph. A GUI interface was developed to handle the information and to visualize the captured images by the Image-acquisition system and sent to the image-processing computer in the wheelhouse, *Image-processing system* paragraph. Finally, to handle the different parallel process threads present in the different components of the CatchWAM the parallel framework module was developed and deployed in the image-acquisition system and image-processing computer, *Parallel Framework* paragraph.

4.2 Image-Acquisition System

4.2.1 General

The CatchWAM image-acquisition system is positioned in the working area at the end of the conveyor belt, as shown in Figure 4.2 left. As introduced in the *Overview of the CatchWAM System*, the image-acquisition system is the result of a research process in which the quality of the image acquisition is the fundamental goal against very demanding environmental conditions, limited workspace dimensions, and, last but not least, General Data Protection Regulation (GDPR) compliance issues.

Furthermore, due to the difference between the different vessels, the system is designed to be easily mounted, calibrated, and maintained according to the different vessels' work area specifications, e.g. the width of the conveyor belt and the vertical dimension of the lateral sides of the conveyor belt (as it is possible to notice from the green plastic sides in the Figure 4.2 left). For further details see also the sub-paragraph *Mechanical Design*.

To get a good image resolution, minimizing the influence of the dirtiness, the occlusions due to the presence of the personnel, and the poor light conditions, the camera for the image acquisition is positioned at a distance of circa 50 cm from the conveyor belt. The camera is enclosed in a sealed module of dimensions (Width × Height × Length) of 58 × 67 × 25 cm, Figure 4.2 right. The sealed module, see also the *Mechanical Design* sub-paragraph, protects the camera from salt, humidity, the mechanical stress allowing optimal illumination and avoiding the presence of the personnel inside the field of view (FoV) of the camera itself. The limited dimension of the module along the length of the conveyor belt allows the personnel to work and move in the workspace with the least possible nuisance.

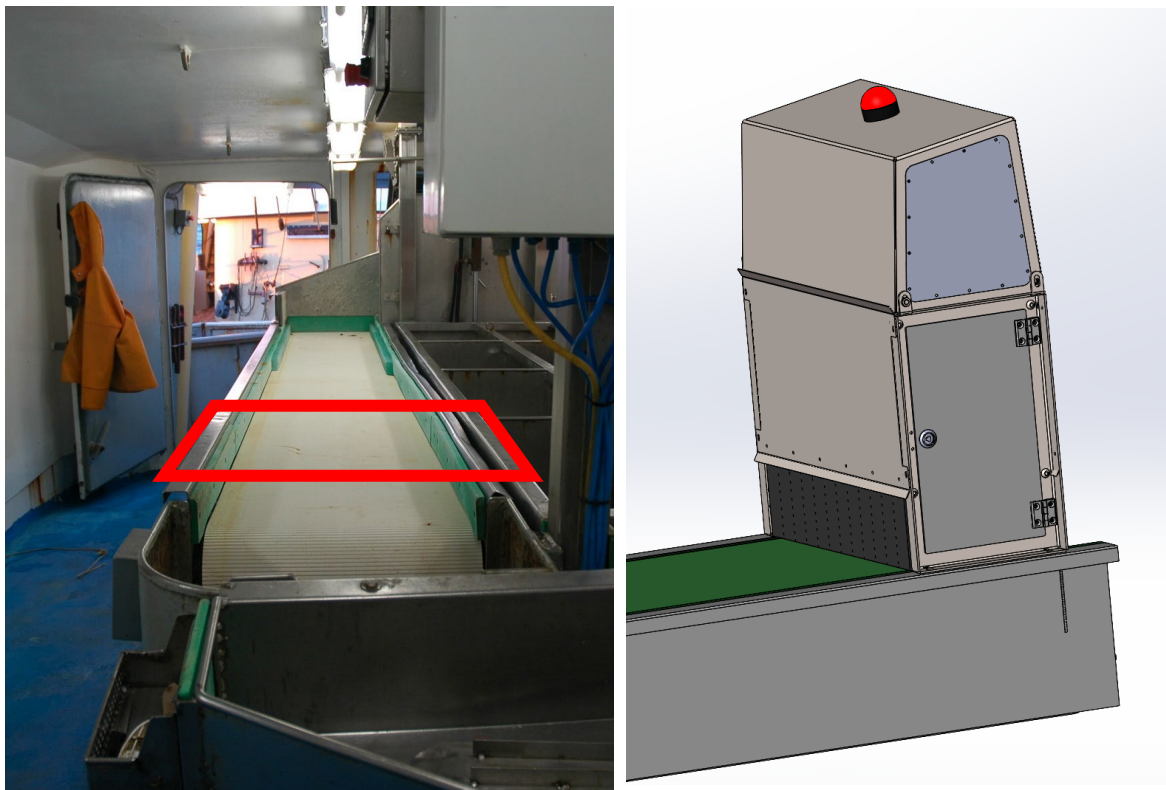


Figure 4.2. Left, view of the starboard conveyor belt in the workspace area. The red-marked area defines the zone on the conveyor belt on which the CatchWAM is placed. Right, the digital design of the image-acquisition module.

In the following three sub-paragraphs, it will be given a more detailed description of the hardware and the software design behind the image-acquisition system.

4.2.2 Mechanical Design

The image-acquisition system comprises two detachable units: the camera-computer unit and the frame on the conveyor belt, Figure 4.3.a. The camera-computer unit, Figure 4.3.b. is a sealed marine-grade metal compartment that contains the camera for the image acquisition, the illumination setup, and the fanless computer to handle the image acquisition, see also the *Line Scan Design* sub-paragraph.

The camera used for the image acquisition is the Framos TM D415e industrial camera (FRAMOS Technologies d.o.o., Čakovec, Croatia), equipped with an RGB camera and b/w stereo cameras, focal length 1.88 mm, and with an image resolution 1920 × 1080 pixels. Rugged, IP64, LuxaLight LED-strip White 5500K Protected (24 Volt, 140 LEDs, 2835, IP64) (Luxalight B.V. Eindhoven The Netherlands), mounted close to the Framos camera on a metallic plate for passive warm dissipation, provide the illumination inside the image-acquisition system, Figure 4.3.c. On top, a Werma 240, multicolor beacon (WERMA Signaltechnik GmbH + CO.KG, Rietheim-Weilhem, Germany) for visual communication of the status of the system.

A Karbon 410 Intel Elkhart Lake Compact Rugged Computer (OnLogic, Oosterhout, The Netherlands) with a passive cooling system hosts the code for the image acquisition, see *Line Scan Design* sub-paragraph, the communication with the image-processing system, the dirtiness detection, the color code for the multicolor beacon, and the digital electronics. A Milk Plexiglas panel with a transparent window for the Framos camera, Figure 4.3.c., seals the bottom side of the camera-computer unit allowing diffuse illumination and image acquisition.

The frame on the conveyor belt, Figure 4.3.a, is the mechanical interface of the camera-computer unit and the conveyor belt. It is a marine-grade metal frame with two white plastic lateral doors for easy access to the conveyor belt and to the Plexiglas panel for maintenance and camera window cleaning. The frame is designed to be easily custom-built to fit the different widths of the conveyor belts. The stainless frontal and backward walls and the white doors ensure light diffusion inside the system.

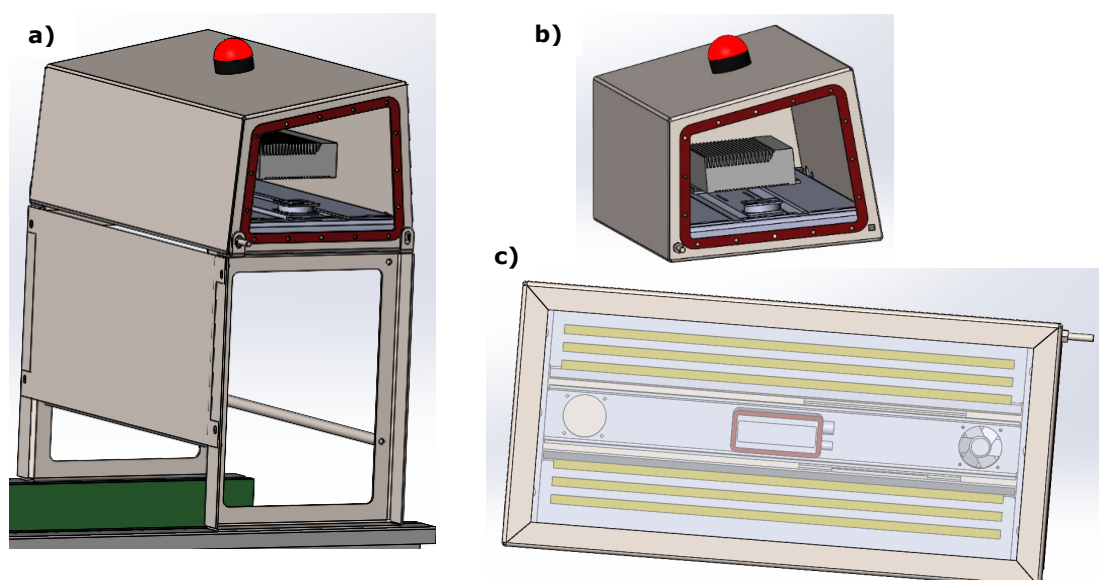


Figure 4.3. a) the bare digital design of the image-acquisition system that comprises two detachable units; b) the camera-computer unit without the lateral panels; c) the bottom side of the camera-computer unit sealed by the milky Plexiglas panel. Notice the Framos camera in the center of the panel and the LuxaLight LED-strips setup.

In Figure 4.4, the image-acquisition system with the black rubber curtains (only on the backside) is shown. It was developed to limit the external light and the dirtiness/water droplets from the sink when the sea conditions and the vessel orientation create bottom-up splashes of seawater.

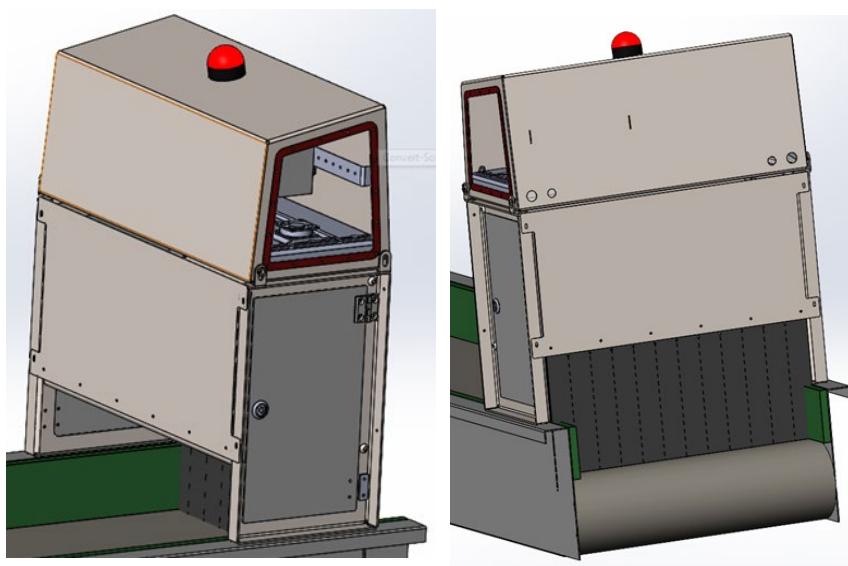


Figure 4.4. From the frontal and back sides of the image-acquisition system, it is possible to see the lateral doors and the black rubber curtain that limits the external light and the water droplets from the sink in case of bottom-up splashes.

The concept of two detachable units allows a user-friendly custom deployment and easy maintenance.

In Figure 4.5, the complete setup of the image-acquisition system deployed in the workspace of the UK 246 is shown.



Figure 4.5. The image-acquisition system deployed onboard the UK 246.

4.2.3 Line Scan Design

The Famos camera acquires RGB and Depth images with a rate of circa 20 frames per second. In Figure 4.6, left it is shown one of the main drawbacks of the use of a matrix image i.e. the presence of the same fish in a sequence of images. Although this issue can be solved using image tracking methods, see Chapter 2, there are substantial other benefits beyond the above-mentioned problem, in the use of a line scan image instead of a matrix one. In line-scan images, the reflections due to the illumination on a wet surface can be reduced with an appropriate configuration of the lighting setup. Furthermore, the perspective deformations are limited to a small amount of pixels. In this case, however, two of the most interesting benefits are the reduction of the FoV of the camera along the length of the conveyor belt allows a reduction of the workspace occupied by the image-acquisition system; and the reduction of the frequency of the images sent in the communication system, see the *Communication System* paragraph.

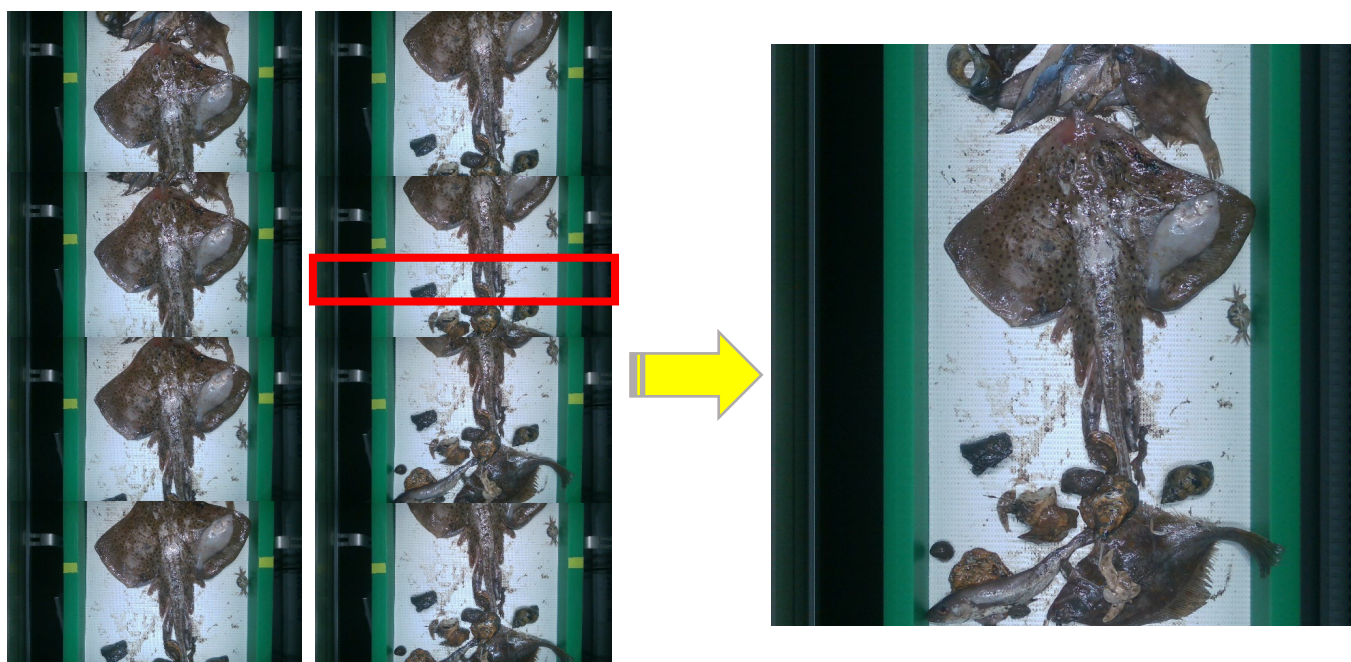


Figure 4.6. Left, an example of the frame rate of the Famos camera and the problem of a sequence of images of the same fish; the red-marked rectangle is a no-in-scale region of interest (ROI). Right, the reconstructed semi-line scan image depicts the full size fish of the left image.

Using the Compact Rugged Computer to preprocess the RGB and the Depth images from the Famos camera the consecutive frames are merged into a semi-line scan image, Figure 4.6. Due to the fact that there is no encoder for these vessel conveyor belts, the first step is to determine the conveyor belt displacement and its speed using a block matching technique, see also Chapter 3.

A region of interest (ROI) is selected in two consecutive frames and their absolute difference is calculated. The smallest absolute difference computed defines the number of lines in a frame that are stitched together. This produces a theoretical infinite image that in this application is chunked in files of 1920 x 1920 pixels.

4.2.4 Parallel Framework

The parallel framework is an user-friendly application developed in-house to handle multiple threads. Designed to be portable, especially for Edge computers and embedded systems, in the CatchWAM system it handles the different parallel processes during the image acquisition, image-preprocessing, image sending, and image-detection pipeline, avoiding race conditions.

In Figure 4.7., the basic architecture of the Parallel Framework is depicted. The application presents two layers (global variables and shared memory) completely transparent for the coded processes (e.g. image acquisition) that manage the internal traffic of the threads when parallel processes are implied. The code developed for an application (e.g. vision detection of fish) can be deployed in this framework without worries about the handling of the parallel threads. A GUI interface can applied on top to handle the program by the user. A communication layer assures the communication of the deployed code with the other components, e.g. an external computer.

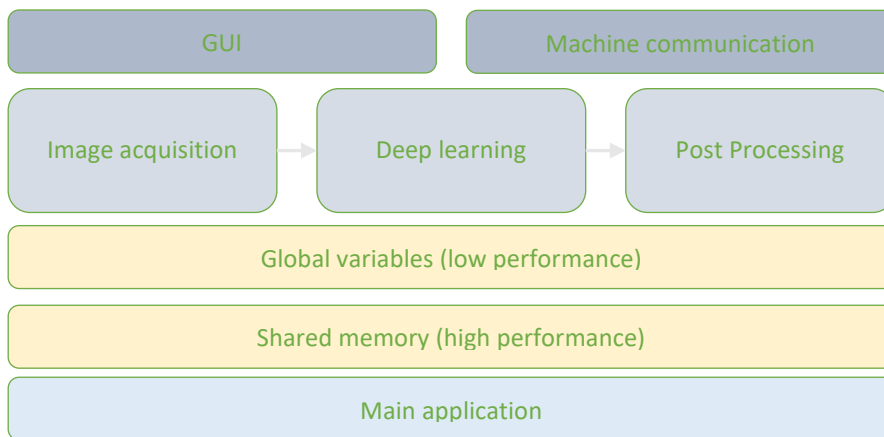


Figure 4.7. Parallel Framework architecture. The diagram of the application presents external interfaces, one for the user (GUI) and one for the communications with the hardware components (Machine communication). At the lower level, there are two layers (Global variables and Shared memory) that handle the parallel processes in a transparent way for the code developer. The layer in which the developer writes his own code without worries about the handling of the parallel threads is depicted by the three typical blocks for an application of visual detection.

The parallel framework is deployed in the Compact Rugged Computer of the image-acquisition system and the hardware of the Image-processing system. It handles the vision detection pipeline, the applications, and the communications of the whole CatchWAM system.

4.3 Image-Processing System

The image-processing system is the terminal part of the CatchWAM system in which the line scan image, Figure 4.8 right, are processed applying algorithms developed in Chapter 3. This is also the system where the images are stored during the fishing trips to be used for further research.



Figure 4.8. Left, the GUI that shows in real-time the situation inside the Image-acquisition system and the status of the system. Right, the line scan image that the waterproof GPU computer handles for processing and that it stores in its internal hard disk for further use.

As described in the above sub-paragraph *Parallel Framework*, here there is a GUI interface that monitors in real-time the situation inside the image-acquisition system and the status of the system, Figure 4.8.Left. The GUI allows monitoring of the raw RGB and Depth images and checks the RGB and Depth line scan images after the preprocessing in the Compact Rugged Computer of the image-acquisition system. Furthermore, in the GUI it is embedded also the calibration application to check the Roll, Pitch, and Yaw of the system during the montage of the system on the conveyor belt and the maintenance routine. It provides remote access by means of an external router when the vessel is docking. Remote access allows the system to check and the code update.

The image-processing system is located in the wheelhouse in the location for the computer and the other electronic components of the vessel, Figure x9.Left. It is hosted in a IP67 waterproof GPU computer supporting NVIDIA®Tesla, the SEMIL-1724GC-A2K i7 (Intemo B.V., Helmond, The Netherlands), Figure 4.9. right.



Figure 4.9. Left, Location in the wheelhouse where the computation computer is allocated. Right, Details of the deployment of the computation computer.

4.4 Communication System

The communication system is the core of the concept of distributed hardware that allows sending data and images in real-time from the acquisition to the detection component of the system, Figure 4.10. The ZeroMQ Push/Pull method (Hintjens. 2013) is used to communicate between the parallel frameworks. Here an indirect benefit of the line scan is exploited to avoid a jamming in the communication on the ethernet line; in fact, the frame rate of the line scan produced by the preprocessing in the Compact Rugged Computer is around two frames per second instead of 20 f/s of the raw images.

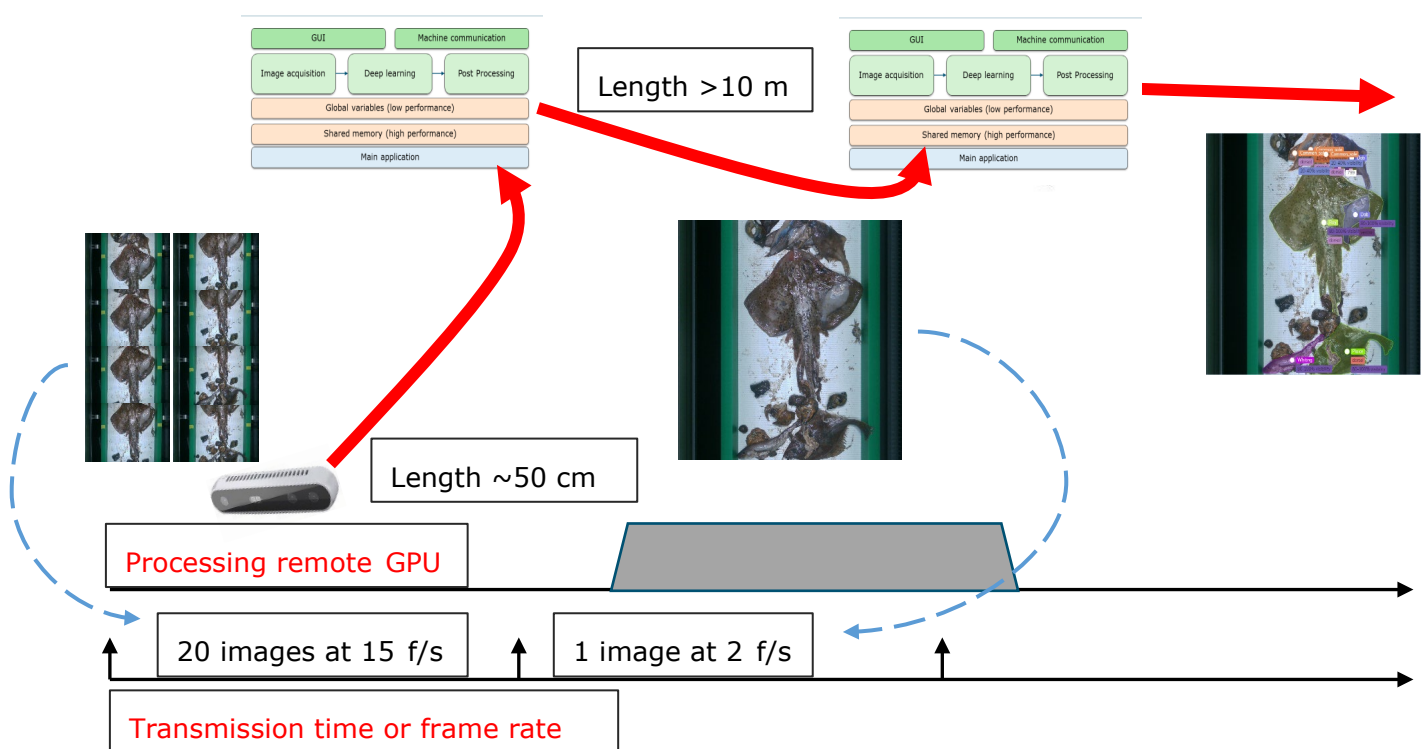


Figure 4.10. The schema of the communication between the different components and the transmission time on the function of the rate of the sending data.

5 Multi-stage image-based approach for fish detection and weight estimation

Manuel Cordova¹, Maria Sokolova¹, Aloysius van Helmond², Angelo Mencarelli³, Gert Kootstra¹

¹ Wageningen University and Research, Farm Technology Group, Wageningen, 6700 AA, The Netherlands

² Wageningen University and Research, Wageningen Marine Research, IJmuiden, 1970 AB, The Netherlands

³ Wageningen University and Research, Greenhouse Horticulture Unit, Wageningen, 6700 AP, The Netherlands

Multi-stage image-based approach for fish detection and weight estimation

Manuel Cordova^a, Maria Sokolova^a, Aloysius van Helmond^b, Angelo Mencarelli^c, Gert Kootstra^a

^a*Agricultural Biosystems Engineering Group, Wageningen University and Research, 6700 AA Wageningen, The Netherlands*

^b*Wageningen Marine Research, Wageningen University and Research, 1970 AB, IJmuiden, The Netherlands*

^c*Greenhouse Horticulture Unit, Wageningen University and Research, 6700 AP, Wageningen, The Netherlands*

Abstract

Challenges with sustainable use of aquatic resources stimulated the development of tools that help raising marine awareness. These developments are promoted by an increasing regulatory pressure on fisheries, which require fine-grained data from the individual fishing vessels. To tackle these challenges, fisheries observer programmes and video monitoring have been implemented. However, the former is an expensive and exhaustive task with low coverage that is performed physically fish by fish. The latter, despite increasing the coverage, recorded videos are processed manually by human expert which also implies extra time, costs, and labor. To support the automated process of those videos, we introduce a novel multi-stage machine-learning approach for fish detection and weight estimation, which is composed by three stages: localization, classification, and weight estimation. In addition, several state-of-the-art approaches are evaluated in each one of the stages, and we also assessed the impact of using a general species-agnostic classifier vs species-specific regressors. Collecting data for detection is much easier than data for weight estimation, the latter needs physical measurement. Unlike end-to-end approaches that require the detection and weight information to train the whole pipeline, our multi-stage approach allow using different data in each stage. Hence, more annotated data could be collected to improve fish detection, whereas techniques for estimating the weight of fish using less data could be used. Experimental results demonstrated the effectiveness of the multi-stage approach to deal with fish detection and weight estimation. Our best configuration (YOLOv7-e6e + EfficientNetB3 + SVR) using a general species-agnostic regressor achieved a F_1 -weighted of 95.12% along with a mean absolute error, mean absolute percentage error and root mean squared error of 27.183 grams, 20.81%, and 47.12 grams, respectively, outperforming the end-to-end approach on FDWE dataset.

Keywords: multi-stage approach; machine learning; computer vision; fish detection; weight estimation

1. Introduction

Seafood is one of the key components for feeding the world's population. Moreover, fisheries in specific plays an essential role in the economy of coastal countries. However, there exist some concerns related to fisheries, such as illegal fishing, overfishing, unsustainable fishing activities, poor management, among others [1, 2]. Aiming at increasing transparency of fishing activities, the European Union implemented the landing obligation which implies that fish can not be discarded overboard, instead, the fishers need to land the catch and declare the complete catch including the unwanted fish (Article 15 of Regulation (EU) No 1380/2013). For this, huge amounts of data need to be inspected to check the compliance of regulations. Eventually, this information is needed, for managers and control agencies, to estimate the fishing yield and to predict the amount of fish to be extracted in the following years.

Trying to support the compliance of this regulation, fisheries observer programmes and video monitoring have been implemented onboard fishing vessels for the purpose of registering the catch and estimating the catch composition [3, 4]. The former is performed in a manual manner, i.e., information of fish is collected one by one, it is an expensive task and with low coverage (less than 1% of fishing activities). The latter, increases the fleet coverage; nevertheless, it requires the need for video analysis, which is currently performed manually by a human expert [5]. The manual analysis of all video recordings is not feasible because of the vast amount of data collected from several fishing vessels. This creates a bottleneck in receiving the information regarding the required catch composition. Therefore, there is a need for automated analysis of video data.

Computer-vision techniques offer an alternative, implying automated analysis of the collected data, which proved to be efficient in other fields, where video monitoring is used [6, 7, 8]. In the particular case of fisheries, the successful implementation of computer-vision techniques is challenged by the quality of the collected data, and the limited availability of labeled data due to time-consuming and expert-dependent way to obtain ground-truth annotations. Among the visual challenges present in the collected images are: limited visibility of the targets caused by occlusions or blur, complex backgrounds given by the presence of debris, different levels of illumination, natural variation in the appearance of the fish, and variation in pose and composition of multiple fish. Additionally, the gathered data presents a large class-imbalanced due to uneven distribution of the fish species present in the catch.

Automated fish detection supports a variety of tasks and helps researchers to monitor biodiversity [9, 10], fish populations sizes and migration [11, 12], catch composition onboard the fishing vessels [13], spread of diseases [14], and fish weight [15, 16, 17]. The data is being obtained both underwater [12, 9, 11, 10] and on land or onboard fishing vessels [13, 15, 17]. Large amounts of collected data require automated processing to efficiently extract the task-specific information. Several studies [11, 18, 19, 12, 9, 10, 20] demonstrated the application

of well-established deep neural networks, such as Single Shot MultiBox Detector [11], Region-Based Convolutional Neural Networks [18, 19], and most of them used methods from the YOLO-family approaches [12, 9, 10, 20, 19, 13, 19, 21, 14]. In contrast, there are few studies dealing with fish detection along with weight estimation. Existing methods also used convolutional neural networks for fish detection followed by an estimation of the weight of the detected fish instances based on weight-from-area [16], weight-from-perimeter [22], or weight-from-length [15]. Those approaches used conventional image-processing methods to estimate the weight, not leveraging on the recent advantages of deep learning. Unlike the previous works, an end-to-end neural network approach to perform fish detection and weight estimation was proposed in [17].

Although some machine-learning methods have been proposed for fish detection, one of the main limitations in the fisheries field is the lack of annotated data containing fish weight information, which restricts the potential of deep-learning-based methods. Specifically, methods such as end-to-end approaches that require the detection and weight information as part of the ground truth to train the whole pipeline. However, ground truth needed for performing detection (bounding boxes and species) is easier to collect than registering the weight of each fish. For detection, data could be annotated afterwards using the recorded videos even with the support of machine-learning-based tools; on the other hand, weight information is collected manually fish by fish.

To overcome this limitation, the overall task of detection and weight estimation can be split up on multiple subtasks. In this work, we propose a multi-stage machine-learning approach based on task decomposition to deal with fish detection and weight estimation. Our proposal is composed by three stages, each one of them responsible for a specific task, i.e., localization, classification and regression. Based on late-fusion strategies, results from the three stages are concatenated to produce the final output. In the first stage, our approach uses a single-class detector to localize fish. In the second stage, species are predicted for the fish detected in the previous stage, and then, in the final stage, the regressor estimates the weight of the fish. We present two variations of our multi-stage approach, the difference among them is the regressor, we compare the use of a general species-agnostic regressor vs species-specific regressors.

Our approach allows the use of different data in each one of the tasks. Therefore, more data could be collected to improve the localization and classification of fish, whereas techniques for weight estimation could be trained on smaller dataset. Another advantage of our proposal is that its modular structure increases the flexibility and versatility of the method. For example, given that new approaches will continue emerging in the computer vision area, the proposed methodology allows to incorporate newer methods in an independent way. Moreover, new down-stream tasks, such as length estimation, quality assessment or survival rate, can be included in a plug-and-play manner without the need for complete retraining.

In order to define which methods will be used in the three-stage pipeline, several well-known and state-of-the-art deep learning architectures for fish detection and classification, such as ResNet [23], EfficientNet [24], YOLOv5 [25],

YOLOv7 [26], and YOLOv8 [27], were assessed.

In this paper, we compare multi-stage approaches to the one-stage end-to-end neural network proposed in [17]. To evaluate the performance of the methods, we used the Fish Detection and Weight Estimation dataset (FDWE) [17] that contains both detection and weight annotations. FDWE represents typical challenges of fish discards registration onboard fishing vessels. Specifically, variations of fish appearance, class imbalance, and even different levels of visibility because of overlapping fish or the presence of debris.

In summary, the main contributions of this paper are:

1. We present a multi-stage image-based approach for fish detection and weight estimation.
2. We evaluate several deep learning object detectors as single-class fish detectors.
3. We compare the effectiveness of well-known classifiers in the context of fish classification.
4. We evaluate the impact of using a general species-agnostic regressor vs species-specific regressors (one regressor per specie).
5. We perform a comparison of per species and per level of occlusion between our multi-stage approach and end-to-end approach.

2. Materials and Methods

Here, we present the FDWE dataset (Section 2.1), our multi-stage approach for fish detection and weight estimation (Section 2.2). Next, the experimental protocol (Section 2.3) describing each one of the components used in our methodology. Last, we present the metrics used to assess the effectiveness of the approaches (Section 2.4).

2.1. Dataset

In order to evaluate the performance of the evaluated methods, we used the Fish Detection and Weight Estimation (FDWE) dataset presented in [17]. This dataset contains images of the discarded fish filmed on an experimental conveyor belt, typical for demersal beam trawlers. In addition to fish species, location, and weight in kilograms, the dataset includes information about the occlusion level of each fish. Four levels of occlusion were defined in this dataset: 0% (fully visible), 1 – 30%, 31 – 60%, and 61 – 90%, which are produced by either overlapped fishes or debris.

The images of FDWE were extracted frame by frame from video recordings, and then, those images were filtered to avoid repeated individuals in consecutive frames. Table 1 lists some details related to the number of images and annotations and Table 2 describes the species and the number of instances in each set. Figure 1 shows some visual examples.

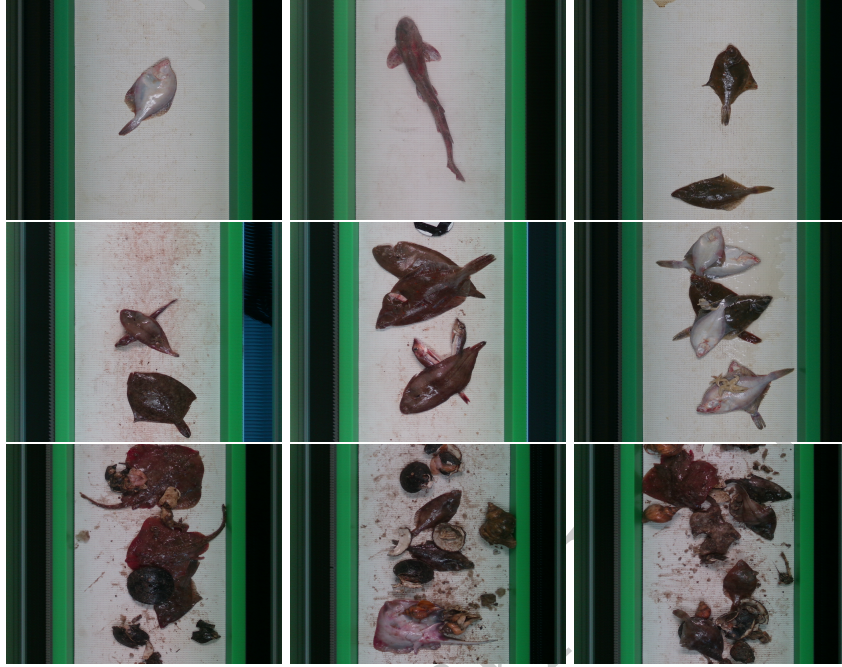


Figure 1: Images from FDWE [17] dataset.

Table 1: FDWE [17] dataset information.

#Images			#Annotations		
Training	Validation	Test	Training	Validation	Test
783	204	99	1,595	424	197

2.2. Multi-stage approach

In this section, we introduce our multi-stage approach to deal with fish detection and weight estimation. Our goal is to introduce flexible approaches aiming their use as part of automatic systems for video analysis from fisheries, for example, for discards registration onboard fishing vessels. Unlike end-to-end approaches, where a single component is responsible for performing all tasks simultaneously [17], herein, we propose a three-stage pipeline based on task decomposition that use different components working on specific tasks: localization, classification, and weight estimation. At the end, their outputs are merged to present the final result.

Our multi-stage proposal allows to update its components for future changes in an independent way. Furthermore, its modular structure makes feasible the use of different data for training in each one of the three stages. Nowadays, there exist some tools to support the process of image annotation for detection; however, collecting annotations with fish weight information is a costly and time-

Table 2: Number of fishes by species in the FDWE dataset.

Species	# Annotations		
	Training	Validation	Test
<i>Pleuronectes platessa</i>	876	206	101
<i>Eutrigla gurnardus</i>	224	61	25
<i>Solea solea</i>	170	49	23
<i>Limanda limanda</i>	122	35	15
<i>Merlangius merlangus</i>	59	20	8
<i>Scophthalmus rhombus</i>	49	7	3
<i>Amblyraja radiata</i>	40	13	6
<i>Scophthalmus maximus</i>	35	28	11
<i>Scyliorhins canicula</i>	20	5	5

consuming labor that needs physical measurement. Based on this fact, more data could be collected to improve the detector and classifier, whereas techniques for weight estimation could be trained on smaller dataset. Additionally, new components, for example, to estimate the length of the fish, quality assessment, or survival rate, could be added to the multi-stage approach in a plug-and-play manner without the need of retraining the full pipeline.

We present two variants of our multi-stage pipeline in Figures 2 and 3. As we can see, the first stages (localization and classification) are the same for both pipelines. First, a single-class detector is used to localize fish. In this first stage, independent of the different species present in the dataset, all fish are labeled as belonging to the general “fish” class. The goal behind this strategy is to avoid the conflict between the localization and classification losses aiming to improve the effectiveness of the fish detector, increasing the capability of this component to detect as much fish as possible. In this scenario, the fish detector will put full attention on how to distinguish fish from the background (clean background and background with debris) without concerns about species.

The fish detector is the basis for the whole pipeline, all the detected fish will be processed in the following stages. As output of this stage, bounding boxes are predicted which describe the localization of the detected fish in the image. Finally, using those bounding boxes, crops are extracted from the original image.

In the next stage, the classifier is responsible for predicting the specie of each one of the detected fish. Given that all the evaluated classifiers are convolutional neural networks, the crops obtained in the previous step are directly used as input. In the last stage, the regressor is responsible for estimating the weight of the fish. This component takes as input the features extracted by the backbone of the classifier used in the previous stage. To estimate the weight of the fish, we evaluated two perspectives: i) the use of one general species-agnostic regressor (Figure 2) and ii) the use of specie-specific regressors (Figure 3).

On one hand, the general regressor is species-agnostic, i.e., it predicts the weight of the fish without any information about its specie. On the other hand, the species-specific regressors use the specie predicted by the classifier to define which regressor will be used. We train one regressor per specie. This approach evaluates whether the regressor could specialize its behaviour per specie. More-

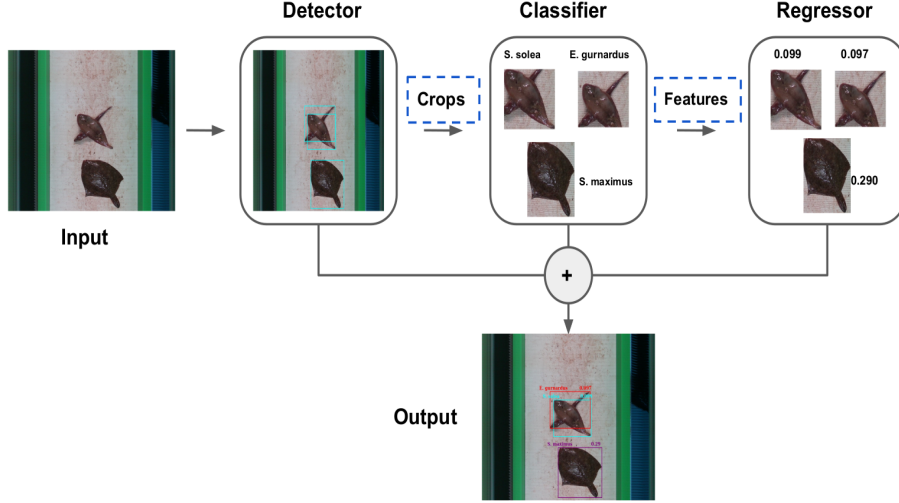


Figure 2: Multi-stage approach with a general regressor.

over, this variant facilitates retraining regressors when new data for a specific species is collected. In addition, it also allows adding new regressors when new species appear in the future.

Finally, for both pipelines, the bounding boxes predicted by the detector, the species given by the classifier, and the weights estimated by the regressor are concatenated to produce the final output.

2.3. Experimental Setup

In all the experiments, we used the original training, validation, and test partitions of FDWE [17]. Details about FDWE are presented in Section 2.1. Next, we describe the machine-learning tasks along with their training protocols and the metrics used to assess the effectiveness of the multi-stage approach.

2.3.1. Stage I: Fish detection

The fish detection is of paramount importance for the whole pipeline, all the detected fish will be considered in the following stages. Based on this fact, we evaluated several state-of-the-art object detectors [25, 26, 27] to accomplish this crucial task.

YOLO-family approaches are state-of-the-art object detectors [25, 26, 27]. The main principle of YOLO (You Only Look Once) is to use grid-cells to split the input image, then, in each grid-cell a vector is produced with information about the bounding box, its objectness score, and the likelihood of the detected object belonging to each one of the classes. In our work, we compared the performance of three YOLO-family methods: YOLOv5_release6.1 [25] (for comparison purposes with [17]), YOLOv7 [26], and YOLOv8 [27]. We used compact (e.g.,

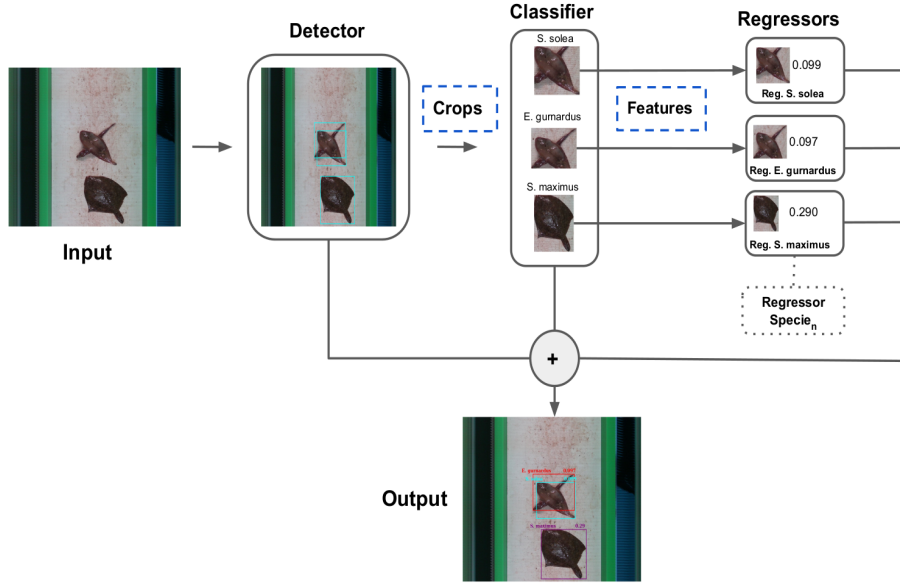


Figure 3: Multi-stage approach with species-specific regressors. In this example, three regressors are used based on the three species predicted by the classifier; however, as the dotted lines depict, there exist one regressor trained per species.

YOLOv7-tiny and YOLOv8s) and deeper (e.g., YOLOv7-6e6 and YOLOv8x) neural networks aiming to assess their effectiveness to detect fish.

All experiments were executed using default parameters from the original implementations. By default, all YOLO versions receive an image of 640×640 as input except for YOLOv7-e6e that uses a 1280×1280 image instead.

2.3.2. Stage II: Fish classification

For classification, we assessed some well-known deep learning classifiers: ResNet [23], EfficientNet [24], YOLOv5 [25], and YOLOv8 [27]. EfficientNets [24] are a set of novel neural networks designed by architecture search algorithms with the goal of balancing depth, width, and resolution. Another widely used neural networks are the ResNets [23] which were proposed with the goal of reducing the vanishing gradient problem through the inclusion of shortcut connections. In the case of YOLOv5 [25] and YOLOv8 [27], apart of being mainly designed for object detection, the authors also released specific implementations for classification. For experiments, to distinguish between YOLO detectors and YOLO classifiers, we used the prefix “C_YOLOv5” and “C_YOLOv8” for the YOLOv5 and YOLOv8 classifiers. All the evaluated classifiers receive images of 224×224 as input, for the remaining parameters, we used the default values.

To select the best performing classifier, we used the training and validation sets of FDWE. However, for classification, we need single fish images instead of

the whole original images. Thus, during training crops were extracted from the original images using the ground truth annotations.

2.3.3. Stage III: Fish weight estimation

For estimating the weight of the fish, we used the Support Vector Regressor (SVR) [28, 29] algorithm. SVR is an adaptation of the widely used Support Vector Machine for regression. This algorithm contains different kernels to approximate functions, such as linear, polynomial, sigmoid, and radial basis function. The hyperparameters of the SVR were defined through a grid-search strategy using the training and validation sets.

Taking advantage of the aforementioned classifiers, already trained on FDWE dataset, the SVR uses as input the features extracted by the backbone of the best performing classifier. For this, given the different sizes of the crops, we applied resizing and padding on those crops in order to generate new images with the same size and preserving the original aspect ratio of the crops.

As described before, we compare the effectiveness of a general species-agnostic regressor vs species-specific regressors (one regressor per specie). The SVR parameters for the species-agnostic regressor were: $C = 1$, $\epsilon = 0.001$, $\gamma = 0.01$, and the polynomial function as a kernel. In the case of the species-specific regressors, information about the parameters is presented in Table 3.

Table 3: Species-specific regressor: SVR parameters.

Species	kernel	# Parameters		
		C	epsilon	gamma
<i>Pleuronectes platessa</i>	rbf	10	0.01	0.001
<i>Eutrigla gurnardus</i>	rbf	20	0.0001	0.001
<i>Solea solea</i>	linear	0.1	0.001	0.001
<i>Limanda limanda</i>	poly	10	0.01	0.001
<i>Merlangius merlangus</i>	linear	1	0.0001	0.001
<i>Scophthalmus rhombus</i>	linear	1	0.0001	0.001
<i>Amblyraja radiata</i>	rbf	0.1	0.001	0.01
<i>Scophthalmus maximus</i>	poly	1	0.1	0.01
<i>Scyliorhinus canicula</i>	poly	0.1	0.0001	0.01

2.4. Metrics

For evaluation, in the case of detection, the experiments measured the effectiveness of the applied methods in two scenarios. In the first stage, we performed a single-class fish detection, hence, we used the F_1 -score (Eq. 1) to compare the methods. Then, after concatenating the results from the detector and classifier, we computed the final multiclass detection results using the F_1 -weighted (Eq. 3). For this, we considered a bounding box as a true positive if the intersection over union (IoU) (Equation 2) with the ground truth is equal to or greater than 0.5.

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \quad (2)$$

$$F_1\text{-weighted} = \frac{\sum_{i=1}^C N_i \cdot F_{1i}}{\sum_{i=1}^C N_i} \quad (3)$$

In the equations, C refers to the total number of classes, and N_i is the number of fish belonging to class i . For classification, we used the accuracy to assess the effectiveness of the methods (Eq. 4).

$$\text{Accuracy} = \frac{\#\text{correct predictions}}{\#\text{total predictions}} \quad (4)$$

Finally, for regression, we used three metrics: i) Mean Absolute Error (MAE) (Eq. 5), Mean Absolute Percentage Error (MAPE) (Eq. 6), and Root Mean Square Error (Eq. 7). In the regression metrics, \hat{y} is the weight in kilograms predicted by the method, y is the ground truth weight, and N is the total number of fish.

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (5) \quad \text{MAPE} = \frac{100\%}{N} \sum_{i=1}^N \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (6)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (7)$$

3. Results and Discussion

First, in order to define which methods will be used in each one of the components of our pipeline, we evaluated the performance of different algorithms for fish detection (Section 3.1) and classification (Section 3.2). Then, we select the best ones to execute the whole pipeline and discuss the effectiveness of the two variants of the multi-stage approach against the end-to-end approach (Section 3.3).

3.1. Fish Detection

As described before, in this stage we performed a single-class fish detection, i.e., independent of the specie, all fish instances are labeled as belonging to the general “fish” class. To compare the effectiveness of the evaluated detectors, a 5-fold cross validation procedure was applied on the training set and the average precision, recall, and F_1 -score are reported in Figure 4. Moreover, considering the model size of the detectors, we used light versions of the neural networks, such as YOLOv7-tiny and YOLOv8s, and also deeper versions, such as YOLOv7-6e6 and YOLOv8x, in order to get insights about the performance of these kind of models when detecting fish.

As we can see, all detectors performed well reaching F_1 -scores above 90%, the best performing method was YOLOv7-e6e with a F_1 -score of 96.1% followed

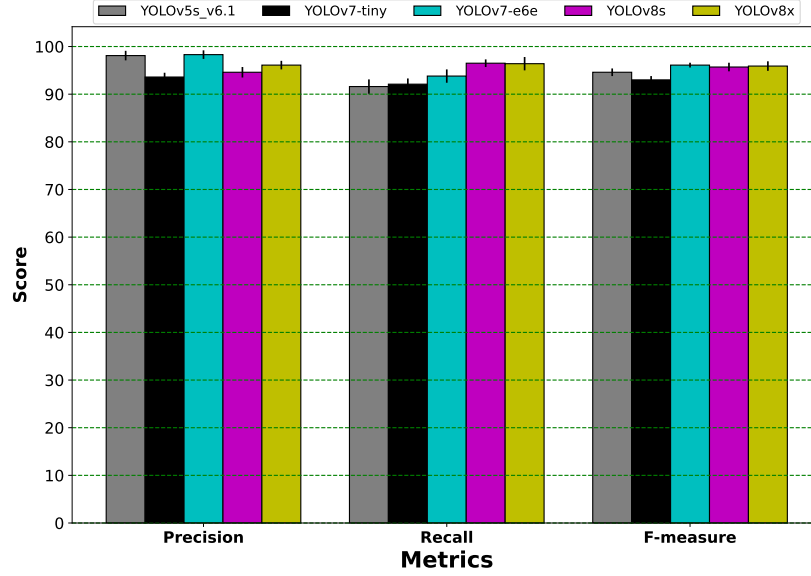


Figure 4: Single-class detection: 5-fold cross validation results.

closely by YOLOv8x (95.9%) and YOLOv8s (95.7%). In most cases, overlapping situations were the main reason of missed detections.

Considering the detection results, YOLOv7-e6e was selected to be used as fish detector in the first stage of our approach. In addition, YOLOv5s will also be used for comparison purposes with the end-to-end approach [17].

3.2. Fish Classification

Figure 5 shows the classification results by specie and the overall accuracy on the validation set. C-YOLOv8 models did not perform well, getting the lowest performance among the evaluated classifiers. The remaining methods reached an accuracy greater than 92%; nevertheless, they presented some limitations with some species. For instance, in the case of *L. limanda*, the accuracy of C-YOLOv8 models dropped to 57.10%, but still above 80% for most of the classifiers. The worst scenario was related to *S. rhombus*, 7 out of 8 classifiers were not capable of classifying correctly any of the instances of this specie.

Considering the overall accuracy, EfficientNet-B3 [24] was the best performing classifier, most of the instances were correctly classified, but there also exist some missclassified fishes. For example, all *S. rhombus* instances were erroneously assigned to the specie *S. maximus*. This behaviour could be related to the high visual similarities between these two species and the small number of samples in the dataset. Furthermore, considering all species, most errors were observed in occluded fish or fish depicted with ventral side up. In the case of occluded fish, nearly all given predictions corresponded to the specie of the other partly visible fish present in the crop. Additionally, most of the ventral side

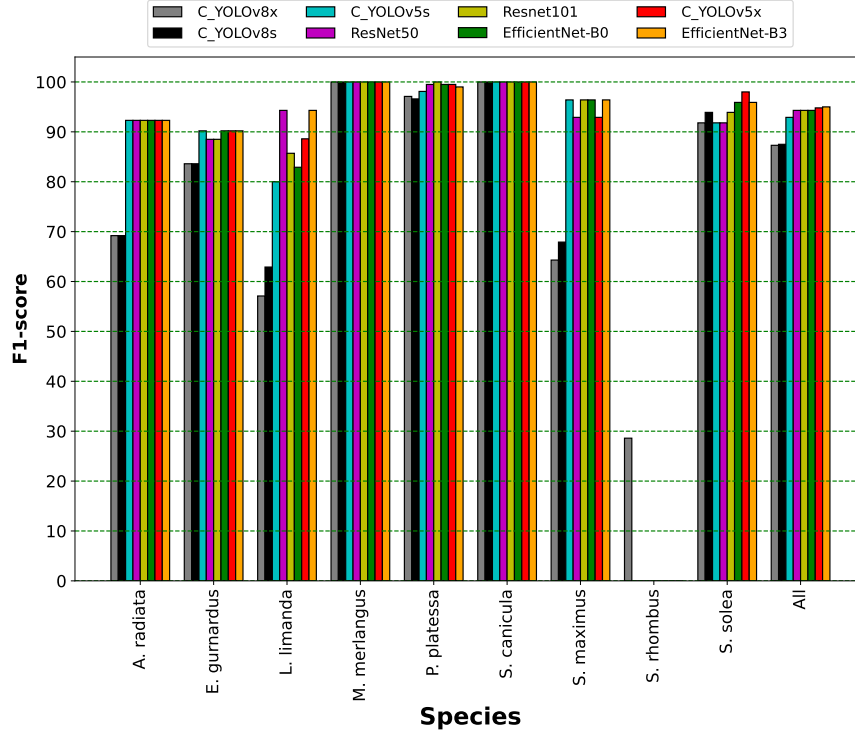


Figure 5: Classification results on the validation set of FDWE [17].

instances of flatfish species were confused with the majority class, *P. platessa*, which contains several ventral side instances in the training set.

Based on the results, EfficientNet-B3 [24] was chosen to be responsible for classifying fish in the second stage of our multi-stage approach.

3.3. Multi-stage vs End-to-end approaches

In this section, we compare the two versions of our multi-stage approach against the end-to-end Weight2 model presented in [17]. We chose the Weight2 model for comparison given that our approaches and the Weight2 model were trained only on the FDWE dataset. The reader may refer to [17] for a detailed description of the Weight2 model.

The multi-stage approach used the best performing methods from the previous sections. YOLOv7-e6e [26] as detector, EfficientNet-B3 [24] as classifier, and SVR [29] as regressor which receives as input the features extracted by the backbone of EfficientNet-B3. The end-to-end approach [17] is based on YOLOv5s.6.1 [25], for that reason, we also used this specific version of YOLO as fish detector for fair comparison.

As explained before, the main difference between the two versions of our multi-stage approach is the regressor, thus, the results corresponding to the

detection and classification stages are the same for both of them. It is worth mentioning that for evaluating our proposals on the test set, we trained the models on the whole training set. Additionally, in the case of detectors and regressors, we performed a grid search procedure using the validation set to define the best parameters. Details about the parameters of the regressors are presented in Table 3. For detection, the IoU for non-maximum suppression was set to 0.7 for YOLOv5s and 0.8 for YOLOv7-e6e, and the confidence score thresholds were set to 0.6 for YOLOv5s and 0.85 for YOLOv7-e6e.

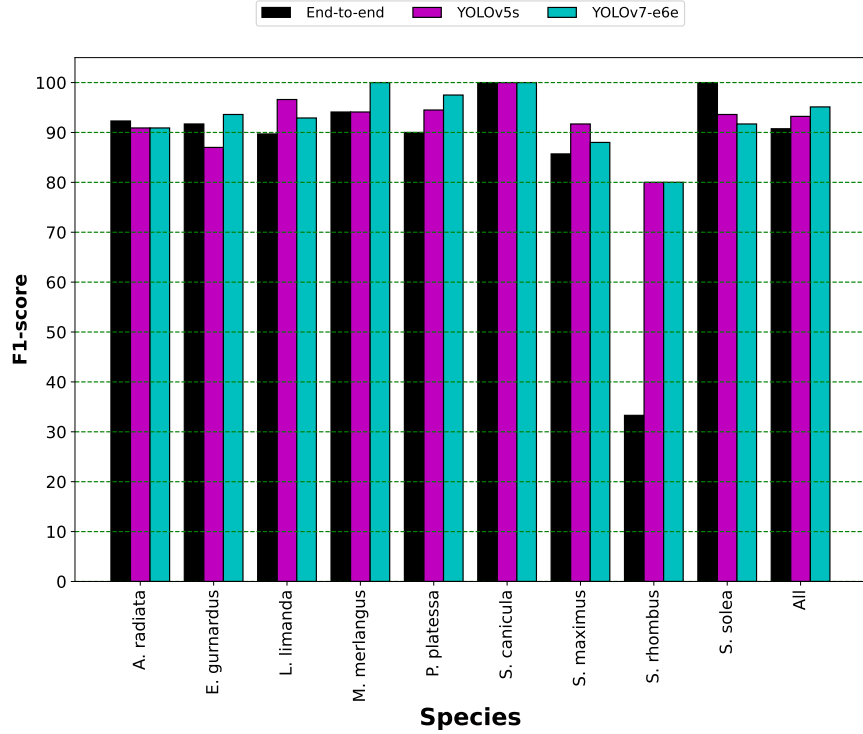


Figure 6: F₁-score per species and the overall F₁-weighted on the test set of FDWE [17]. “End-to-end” refers to the results obtained by the Weight2 model presented in [17].

Figure 6 presents the multiclass fish detection results per species and the overall F₁-weighted. In the experiments, the multiclass detection results are given by the bounding boxes predicted by the detectors (YOLOv5s and YOLOv7-e6e) along with the species predicted by the EfficientNet-B3 classifier.

Our multiclass fish detection pipeline (fish detector + classifier) surpassed the end-to-end approach. In terms of F₁-weighted, results demonstrated that even using the YOLOv5s_r6.1 as detector, our approach improved the result by 2.48 percentage points, reducing the number of false negatives by 20% and the false positives by 37.5% compared to the end-to-end approach. Moreover, among the evaluated neural networks, YOLOv7-e6e was the best performing

approach on the test set with a F_1 -weighted of 95.12%.

In addition, we measured the detection task considering: i) effectiveness of YOLOv7-e6e as a species-agnostic fish detector, and ii) effectiveness as multi-class fish detector (YOLOv7-e6e + EfficientNet-B3). We found that the species-agnostic fish detector was capable of detecting 189 fishes out of 197, reaching a F_1 -score of 97.67%. We can see that the effectiveness of the multiclass fish detection pipeline drops to 95.12% because of wrong predictions made by the classifier, most of them due to occluded instances. Based on this fact, the performance of the detector and classifier could probably improve using more data for training, involving more examples of instances with higher levels of occlusion and balancing the dataset in terms of number of instances per specie.

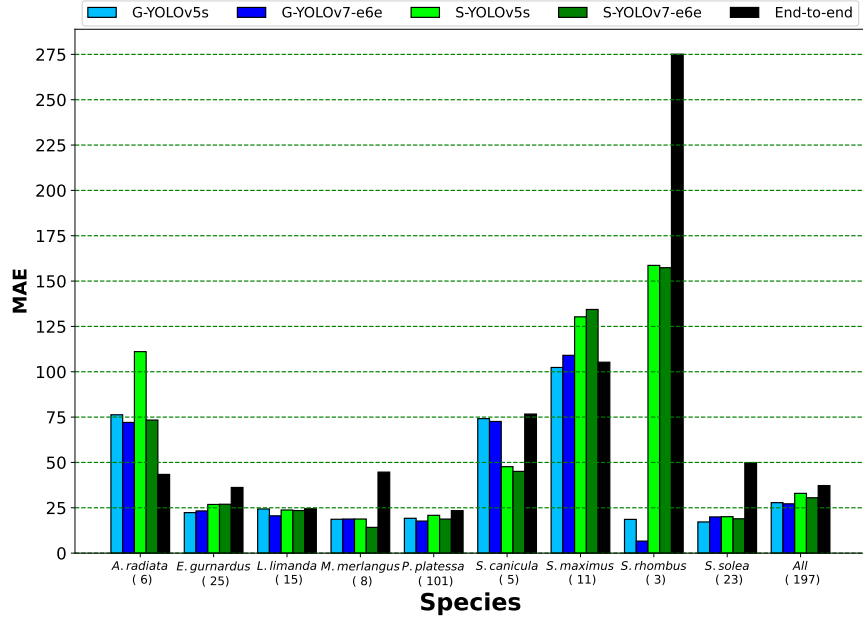


Figure 7: MAE per species on the test set. “G” refers to multi-stage approach using a general species-agnostic regressors (blue bars), “S” to multi-stage approach using species-specific regressors (green bars).

In regard to weight estimation, Figure 7 shows the mean average error (MAE) per specie and also the overall MAE. Based on the overall MAE, both versions of the multi-stage approach obtained lower error rates compared to the end-to-end approach, the same behaviour was reported per specie in 7 out of 9 species. Among our approaches, the general species-agnostic regressor was most effective than the species-specific regressors. The effectiveness of the species-specific regressors may be restricted by the few instances per species for training.

Figure 8 presents the MAE per specie and occlusion level. As we can see, not all species have samples in all occlusion levels which makes it difficult to

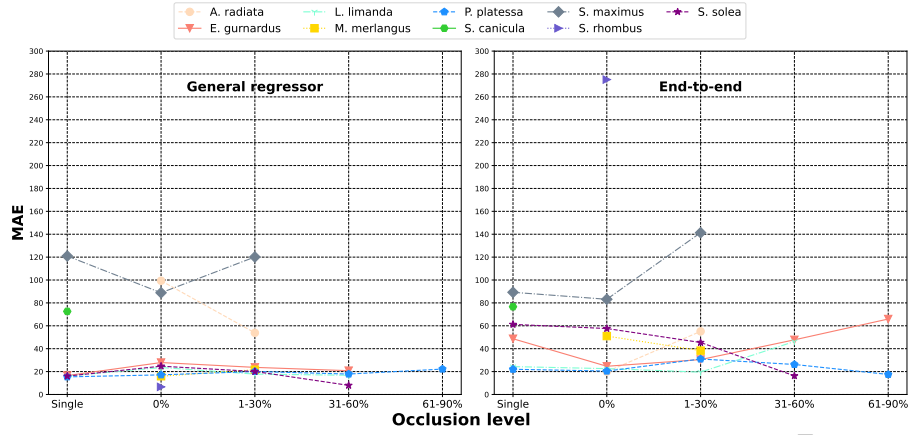


Figure 8: MAE per species and occlusion level on the test set. “Single” refers to crops containing single fish.

compare the performance of the regressors on those scenarios. In general, our best approach that uses a species-agnostic regressor showed a more stable behaviour among the different occlusion levels. On the other hand, the end-to-end approach presented some fluctuations in the results; nevertheless, the trend for some species is the higher the occlusion level, the higher the weight estimation error. Moreover, as expected, in most of the cases of the general regressor, lower errors were obtained with crops with the presence of single fish (“Single” level).

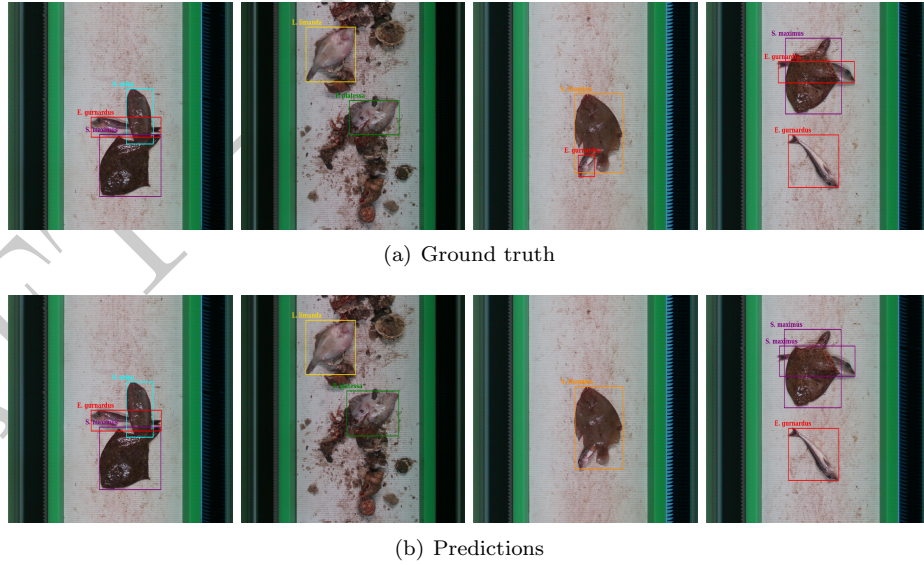


Figure 9: Visual results on the test set of FDWE dataset using our multi-stage approach.

The best performing configuration of our multi-stage approach was using YOLOv7-e6e as detector, EfficientNet-B3 as classifier, and a general species-agnostic regressor. This combination reached a F_1 -weighted of 95.12% along with a MAE, MAPE, and RMSE of 27.18, 20.81%, and 47.12, respectively.

Finally, we present some visual results in Figure 9. Our approach can handle some overlapped scenarios and images with the presence of debris with accurate results for both localization and classification (first two images); nonetheless, some cases require further investigation in order to improve the effectiveness. For example, in the penultimate image of the Figure 9, an instance of *E. gurnardus*, that was occluded more than 60%, was not detected. Moreover, in the last image, all fishes were detected; however, a *E. gurnardus* fish was missclassified as *S. maximus*. As we can see, the main reason for this missclassification is also related to the high occlusion level, specifically, the *S. maximus* fish is on top covering most of the *E. gurnardus*.

Experiments demonstrated promising results of our multi-stage approach. The general fish detection component increased the number of correct detections compared to the end-to-end approach. In real-world applications, increasing the number of detected fishes allows to consider those instances not only in the classification and weight estimation stages, but also, for example, to deal with prominent errors where the uncertainty of the classifier is high, or even to identify new species that were not seen during training.

Additionally, given the costly and time-consuming labor of collecting data, specially the weight of fish, one of the main advantages of our multi-stage pipeline is the flexibility of using different data for training each component (localization, classification and regression). In addition, this pipeline try to alleviate the full ground truth (bounding boxes, species and weight) dependency of end-to-end approaches. Hence, more data could be collected and annotated to improve the fish localization and classification, whereas techniques for estimating the weight of fish using smaller dataset could be explored. Another advantage is its modular structure, each component could be updated in an independent way and even new components could be incorporated in a plug-and-play manner, for example, to predict the length of the fish, quality assessment, or their survival rate. On the other hand, one of the main limitations of our approach is related to the fish detector dependency, fish not detected in the first stage will not be considered in the following stages. Another limitation could be related to the training time, three components must be trained independently.

4. Conclusions

In this work, we proposed two versions of a multi-stage approach based on task decomposition to deal with fish detection and weight estimation. Our approaches are composed by three stages: a general fish detector for localizing fish, a classifier for determining the species, and a regressor responsible for estimating the weight of fish. To select the methods that were used as part of our multi-stage approach, we evaluated several state-of-the-art algorithms [25, 26, 27, 23, 24].

The experimental results on FDWE dataset [17] showed the effectiveness of the two versions of the multi-stage approach surpassing the end-to-end approach proposed in [17] both detecting fish and also estimating their weight. Our multi-stage approach using YOLOv7-e6e as detector, EfficientNet-B3 as classifier, and SVR as a general species-agnostic regressor, was the best performing configuration. This combination reached a detection and classification performance with a F_1 -weighted of 95.12% along with a weight estimation performance measured by MAE, MAPE, and RMSE of 27.18 grams, 20.81%, and 47.12 grams, respectively. Focusing on the weight estimation, the general species-agnostic regressor reached better results than the species-specific regressor; however, having one regressor per species could be a promising alternative when trained with sufficient number of instances per specie.

In this study, we demonstrated that our proposed approach arises as a promising flexible pipeline to deal with fish detection and weight estimation and could be used, for example, for application in real-world conditions on-board fishing vessels for efficient catch documentation. Moreover, our proposals try to alleviate the full ground truth dependency of end-to-end methods, facilitating the independent training process of their components and even allowing the incorporation of new components in the future to perform new tasks.

Finally, given the few instances per species and the data imbalance of the FDWE dataset, future research efforts will be focused on smart data augmentation strategies. Furthermore, given that most missed detections and wrong species predicted by the classifier were related to occluded fish, we will investigate how to improve the results in occluded scenarios. Another research direction is the recognition of new species that can appear in different fishing regions. The detector and classifier outputs could be used to detect unknown species and incorporate them on the fly. In those scenarios, strategies, such as open-set recognition [30], human-in-the-loop [31] along with active learning [32], could be explored to boost the performance of the approaches.

Acknowledgements

This study was funded by the European Maritime and Fisheries Fund (contract number 16302) under the Fully Documented Fisheries project. Authors would like to acknowledge skippers and crew of the vessels involved in Fully Documented Fisheries project for providing discarded fish. We also thank Wageningen Marine Research for their help in data collection and preparation.

References

- [1] Rajakannu Amuthakkannan, K. Vijayalakshmi, Saleh Al Araiimi, and Maa-mar Ali Saud Al Tobi. A review to do fishermen boat automation with artificial intelligence for sustainable fishing experience ensuring safety, security, navigation and sharing information for omani fishermen. *Journal of Marine Science and Engineering*, 11(3), 2023. ISSN 2077-1312. doi: 10.3390/jmse11030630. URL <https://www.mdpi.com/2077-1312/11/3/630>.

- [2] Sara Orofino, Gavin McDonald, Juan Mayorga, Christopher Costello, and Darcy Bradley. Opportunities and challenges for improving fisheries management through greater transparency in vessel tracking. *ICES Journal of Marine Science*, 80(4):675–689, 02 2023. ISSN 1054-3139. doi: 10.1093/icesjms/fsad008. URL <https://doi.org/10.1093/icesjms/fsad008>.
- [3] C. Vilas, L.T. Antelo, F. Martin-Rodriguez, X. Morales, R.I. Perez-Martin, A.A. Alonso, J. Valeiras, E. Abad, M. Quinzan, and M. Barral-Martinez. Use of computer vision onboard fishing vessels to quantify catches: The iobserver. *Marine Policy*, 116:103714, 2020. ISSN 0308-597X. doi: <https://doi.org/10.1016/j.marpol.2019.103714>.
- [4] Aloysius T.M. van Helmond, Lars O. Mortensen, Kristian S. Plet-Hansen, Clara Ulrich, Coby L. Needle, Daniel Oesterwind, Lotte Kindt-Larsen, Thomas Catchpole, Stephen Mangi, Christopher Zimmermann, Hans Jakob Olesen, Nick Bailey, Heidrikur Bergsson, Jørgen Dalskov, Jon Elson, Malo Hosken, Lisa Peterson, Howard McElderry, Jon Ruiz, Johanna P. Pierre, Claude Dykstra, and Jan Jaap Poos. Electronic monitoring in fisheries: Lessons from global experiences and future opportunities. *Fish and Fisheries*, 21(1):162–189, 2020. doi: <https://doi.org/10.1111/faf.12425>.
- [5] Aloysius TM van Helmond, Lars O Mortensen, Kristian S Plet-Hansen, Clara Ulrich, Coby L Needle, Daniel Oesterwind, Lotte Kindt-Larsen, Thomas Catchpole, Stephen Mangi, Christopher Zimmermann, et al. Electronic monitoring in fisheries: lessons from global experiences and future opportunities. *Fish and Fisheries*, 21(1):162–189, 2020.
- [6] F.E.T. Schöller, M. Blanke, M.K. Plenge-Feidenhans’, and L. Nalpan-tidis. Vision-based object tracking in marine environments using features from neural network detections. *IFAC-PapersOnLine*, 53(2):14517–14523, 2020. ISSN 2405-8963. doi: <https://doi.org/10.1016/j.ifacol.2020.12.1455>. URL <https://www.sciencedirect.com/science/article/pii/S240589632031867X>. 21st IFAC World Congress.
- [7] Weisheng Lu and Junjie Chen. Computer vision for solid waste sorting: A critical review of academic research. *Waste Management*, 142: 29–43, 2022. ISSN 0956-053X. doi: <https://doi.org/10.1016/j.wasman.2022.02.009>. URL <https://www.sciencedirect.com/science/article/pii/S0956053X22000678>.
- [8] Keval Doshi and Yasin Yilmaz. Multi-task learning for video surveillance with limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3889–3899, June 2022.
- [9] Ahsan Jalal, Ahmad Salman, Ajmal Mian, Mark Shortis, and Faisal Shafait. Fish detection and species classification in underwater environments using deep learning with temporal information. *Ecological Infor-*

- matics*, 57:101088, 2020. ISSN 1574-9541. doi: <https://doi.org/10.1016/j.econinf.2020.101088>.
- [10] Kristian Muri Knausgård, Arne Wiklund, Tonje Knutsen Sjørdalen, Kim Tallaksen Halvorsen, Alf Ring Kleiven, Lei Jiao, and Morten Goodwin. Temperate fish detection and classification: A deep learning based approach. *Applied Intelligence*, 52(6):6988–7001, apr 2022. ISSN 0924-669X. doi: 10.1007/s10489-020-02154-9.
 - [11] Simegnew Yihunie Alaba, M M Nabi, Chiranjibi Shah, Jack Prior, Matthew D. Campbell, Farron Wallace, John E. Ball, and Robert Moorhead. Class-aware fish species recognition using deep learning for an imbalanced dataset. *Sensors*, 22(21), 2022. ISSN 1424-8220. doi: 10.3390/s22218268.
 - [12] Hussam El-Din Mohamed, Ali Fadl, Omar Anas, Youssef Wageeh, Noha ElMasry, Ayman Nabil, and Ayman Atia. Msr-yolo: Method to enhance fish detection and tracking in fish farms. *Procedia Computer Science*, 170: 539–546, 2020. ISSN 1877-0509. doi: <https://doi.org/10.1016/j.procs.2020.03.123>. The 11th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 3rd International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops.
 - [13] Rick van Essen, Angelo Mencarelli, Aloysius van Helmond, Linh Nguyen, Jurgen Batsleer, Jan-Jaap Poos, and Gert Kootstra. Automatic discard registration in cluttered environments using deep learning and object tracking: class imbalance, occlusion, and a comparison to human review. *ICES Journal of Marine Science*, 78(10):3834–3846, 2021.
 - [14] Zhen Wang, Haolu Liu, Guangyue Zhang, Xiao Yang, Lingmei Wen, and Wei Zhao. Diseased fish detection in the underwater environment using an improved yolov5 network for intensive aquaculture. *Fishes*, 8(3), 2023. ISSN 2410-3888. doi: 10.3390/fishes8030169.
 - [15] Juan Carlos Ovalle, Carlos Vilas, and Luís T. Antelo. On the use of deep learning for fish species recognition and quantification on board fishing vessels. *Marine Policy*, 139:105015, 2022. ISSN 0308-597X. doi: <https://doi.org/10.1016/j.marpol.2022.105015>.
 - [16] Dmitry A Konovalov, Alzayat Saleh, Dina B Efremova, Jose A Domingos, and Dean R Jerry. Automatic weight estimation of harvested fish from images. In *2019 Digital image computing: Techniques and applications (DICTA)*, pages 1–7. IEEE, 2019.
 - [17] Maria Sokolova, Manuel Cordova, Henk Nap, Michiel Mas, Arjan Vroegop, Angelo Mencarelli, and Gert Kootstra. Automated bycatch detection and weight estimation onboard beam trawlers: a case study. *ICES Journal of Marine Science*, 1(2):00–00, 2023.

- [18] Ahmad Salman, Shoaib Ahmad Siddiqui, Faisal Shafait, Ajmal Mian, Mark R Shortis, Khawar Khurshid, Adrian Ulges, and Ulrich Schwanecke. Automatic fish detection in underwater videos by a deep neural network-based hybrid motion learning system. *ICES Journal of Marine Science*, 77(4):1295–1307, 02 2019. ISSN 1054-3139. doi: 10.1093/icesjms/fsz025.
- [19] Vishnu Kandimalla, Matt Richard, Frank Smith, Jean Quirion, Luis Torgo, and Chris Whidden. Automated detection, classification and counting of fish in fish passages with deep learning. *Frontiers in Marine Science*, 8, 2022. ISSN 2296-7745. doi: 10.3389/fmars.2021.823173.
- [20] Abdullah Al Muksit, Fakhurul Hasan, Md. Fahad Hasan Bhuiyan Emon, Md Rakibul Haque, Arif Reza Anwary, and Swakkhar Shatabda. Yolo-fish: A robust fish detection model to detect fish in realistic underwater environment. *Ecological Informatics*, 72:101847, 2022. ISSN 1574-9541. doi: <https://doi.org/10.1016/j.ecoinf.2022.101847>.
- [21] Ari Kuswantori, Taweepol Suesut, Worapong Tangsrirat, Gerhard Schleininger, and Navaphattra Nunak. Fish detection and classification for automatic sorting system with an optimized yolo algorithm. *Applied Sciences*, 13(6), 2023. ISSN 2076-3417. doi: 10.3390/app13063812.
- [22] Xiaoning Yu, Yaqian Wang, Jincun Liu, Jia Wang, Dong An, and Yaoguang Wei. Non-contact weight estimation system for fish based on instance segmentation. *Expert Systems with Applications*, 210:118403, 2022. ISSN 0957-4174. doi: <https://doi.org/10.1016/j.eswa.2022.118403>.
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [24] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.
- [25] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Jiacong Fang, imyhxy, Kalen Michael, Lorna, Abhiram V, Diego Montes, Jebastin Nadar, Laughing, tkianai, yxNONG, Piotr Skalski, Zhiqiang Wang, Adam Hogan, Cristi Fati, Lorenzo Mammana, AlexWang1900, Deep Patel, Ding Yiwei, Felix You, Jan Hajek, Laurentiu Diaconu, and Mai Thanh Minh. ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, February 2022.
- [26] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 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 (CVPR)*, pages 7464–7475, June 2023.

- [27] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, January 2023. URL <https://github.com/ultralytics/ultralytics>.
- [28] Mariette Awad and Rahul Khanna. *Support Vector Regression*, pages 67–80. Apress, Berkeley, CA, 2015. ISBN 978-1-4302-5990-9. doi: 10.1007/978-1-4302-5990-9_4.
- [29] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [30] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.
- [31] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56:3005–3054, 2022. doi: 10.1007/s10462-022-10246-w.
- [32] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Comput. Surv.*, 54(9), oct 2021. ISSN 0360-0300. doi: 10.1145/3472291.

6 Improving automated discards registration using active learning technique for object detection

Maria Sokolova¹, Pieter M. Blok^{1,2}, Angelo Mencarelli¹, Aloysius van Helmond¹, Arjan Vroegop¹, and Gert Kootstra¹

¹ Wageningen University and Research

² Graduate School of Agricultural and Life Sciences, The University of Tokyo, Japan

Improving automated discards registration using active learning technique for object detection

Maria Sokolova¹, Pieter M. Blok^{1,2}, Angelo Mencarelli¹, Edwin van Helmond¹, Arjan Vroegop¹, and Gert Kootstra¹

¹Wageningen University and Research

²Graduate School of Agricultural and Life Sciences, The University of Tokyo, Japan

December 14, 2023

Abstract

Video monitoring systems aim to support fisheries management and sustainable exploitation of the aquatic natural resources. This is being achieved by providing the information about the catch composition and quantity, which allows estimating fishing effort and fraction of the non-target catch with higher precision. Camera systems embedded at the conveyor belt of the fishing vessel enable complete catch recording every time the vessel is out fishing. This results in a vast amount of video data the full analysis of which is infeasible to perform manually, by human review. The catch obtained by demersal trawls is typically mixed and includes high catch rates of unmarketable and undersized fish. This creates an extra challenge for accurate identification of the catch in this fishery, both manual and, consequently, automated. In recent years, powerful data-driven deep learning techniques have been developed that allow effective processing of visual data. Like in many other applications, these techniques have been successfully applied for catch registration. However, these methods are dependent on the labelled data, which is time-consuming, labour-intensive and expensive. Additionally, specific expertise are needed to correctly identify species of fish. In this study, we present an active learning technique, which allows epistemic certainty estimation of the object detection model. Based on the certainty estimations the most informative training images from the pool are selected and used to fine-tune the model. Specifically, we propose BoxAL, an active learning technique that estimates the certainty of the predictions of Faster RCNN. To evaluate the method, we used an open source image dataset obtained with a dedicated discards-registration image acquisition system developed for commercial trawlers targeting demersal species. We have demonstrated, that our approach allows reaching the same object detector performance as with the random sampling with 400 fewer labelled images.

Keywords: Remote electronic monitoring, sustainable fisheries, convolutional neural networks, training optimization

1 Introduction

Sustainable use of marine aquatic resources became of a larger concern since the 1980s (). In the EU, new fisheries regulation, Landing Obligation, has been implemented to document all the discarded fish and, therefore, stimulate prevention of obtaining unwanted catch and have a better estimation of the amount of unwanted catch when it is obtained (). This regulation, has been introduced in the 2013 and has become fully implemented in the member states in the 2019. The aim of the regulation is to make fisheries transparent and to document the amount of fish that has been discarded, since the

amount of landed fish is noted at the harbor at the fish auction (). To control the execution of the regulation, human observers join fishing trips onboard commercial vessels to sample the discards; the samples are then analysed manually by the experts (). The issue with this type of control system is low coverage of the fleet as well as high costs of sending the observers onboard and analysing the samples. To assist the human observers and to increase the fleet coverage, electronic monitoring systems are being implemented. These systems include the video cameras placed above the conveyor belt and are dedicated to recording the catch. This system increases the fleet coverage, however introduces several other types of challenges. Specifically, the amount of data obtained by the cameras cannot be fully analysed in manual manner, therefore, an automated component has to be implemented to analyse this vast amount of data. Convolutional neural networks (CNNs) are widely used analyze visual type of data, however their performance greatly depends on the input data quality. In the specific case of onboard monitoring in fisheries, observation conditions can be highly variable due to unstable illumination, high density of catch, presence of people and non-target objects obscuring the catch (). Such conditions make it difficult to analyse the videos manually and, consequently, label the data, which is needed to train the CNNs efficiently. In this study, we use the open-source dataset, which has been collected with a dedicated discards registration camera with a stable illumination and isolated from the outside environment by the metal box van Essen et al. (2021). While the challenge with observation conditions can be partially solved with a dedicated camera system, the efficient training of the CNN remains dependent on a significant contribution from expert annotators.

In the current study, we propose BoxAL architecture, which is based on an active learning framework for instance segmentation (Blok et al. (2022)). We have shown that using active learning framework for object detection allows improving fish detection with fewer annotated images.

2 Materials and Methods

2.1 Dataset

We have used an open-source dataset depicting the discarded fish from the Dutch beam trawlers (van Essen et al. (2021)). The summary of the dataset split into train, validation and test is presented in Table 1. The dataset contains typical species that are commonly discarded during beam trawling fishing operations. The dataset is highly unbalanced and dominated by the three flatfish species (*Pleuronectes platessa*, *Solea solea*, *Limanda limanda*) and whiting (*Merlangus merlangus*). Additionally, the fish typically overlap between each other and/or covered by debris. These factors complicate the annotation process and make it more time-consuming.

2.2 BoxAL

In this study, Faster R-CNN has been used as a base model (Ren et al. (2015)). This two-stage object detection convolutional neural network has been chosen due to its higher detection performance in the case of large amount of small and densely distributed objects in the image comparing to one-stage detectors (Zou et al. (2023)). The head of the model has been modified with the dropout layers implemented during inference mode. The dropout probability has been set to 0.75 with 15 montecarlo iterations.

2.3 Certainty estimation

Resulting certainty values were obtained from the predictions on the instance set. The instance set resulted from the multiple forward passes of BoxAL with the 75% probability for dropout in the heads. Three types of certainty have been estimated: semantic certainty (C_{sem}), spatial certainty (C_{spl}) and

Species	# Annotations		
	Training	Validation	Test
<i>Amblyraja radiata</i>	1073	330	302
<i>Callionymus lyra</i>	25	-	-
<i>Eutrigla gurnardus</i>	187	48	31
<i>Limanda limanda</i>	1885	540	322
<i>Merlangius merlangus</i>	3548	988	529
<i>Microstomus kitt</i>	978	227	111
<i>Pleuronectes platessa</i>	4413	1153	739
<i>Scylliorhynchus canicula</i>	254	62	44
<i>Scophthalmus maximus</i>	109	47	15
<i>Trisopterus luscus</i>	69	7	14
<i>Solea solea</i>	3484	901	452

Table 1: Dataset summary indicating number of instances per species split into train, validation and test subsets.

occurrence certainty (C_{occ}).

2.4 Training procedure

All images from the training subset were considered as training pool, totalling 2830 images. Initially, BoxAL was trained on 100 randomly selected images from the training pool. The best trained model was saved based on its fit on a test set. Further, the obtained model checkpoint was used to run in inference mode on the remaining training pool with the dropout layers implemented in the four fully connected layers of the box head. The 'severeness' of the dropout has been set to 75%, where 0% implied no dropout probability and 100% corresponded to complete disconnection of the neurons. The predictions have been completed 15 times during each sampling iteration. This setup allowed collecting 15 prediction sets per instance that were used for certainty calculation (2.3). After the certainty was complete, the 100 images with minimum certainty value were added to the training set. Then, the training procedure repeated similarly for 10 training iterations (Figure 1). After each training iteration the mean average precision (mAP) has been saved. To estimate the efficiency of the active learning framework the same training procedure was repeated with random sampling. Both training procedures have been repeated three times to estimate the effect of variation in model performance caused by initial selection of the images and introduced by the dropout layer in box head.

2.5 Relationship between prediction performance and certainty estimation

In deep active learning the queries are selected based on the selected strategy and include ... these frameworks, however do not account for the correlation between model certainty estimation and actual errors made by the model.

3 Results

- the model trained with minimum certainty sampling strategy reaches higher mAP with fewer training examples comparing to random sampling strategy
- Noteworthy, average certainty sampling training strategy results in the model worse performance comparing to the one based on random sampling strategy

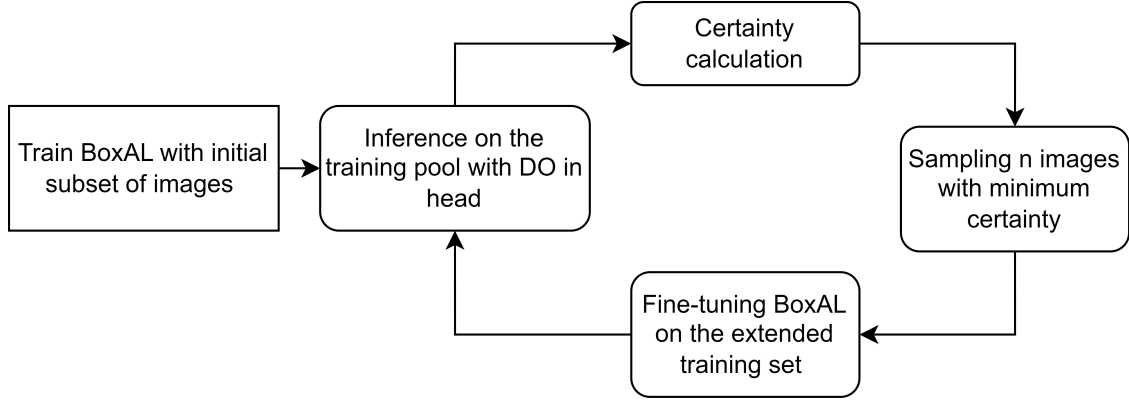


Figure 1: Iterative training procedure of BoxAL

3.1 Minimum certainty sampling vs random sampling

Active sampling of images based on the minimum certainty sampling showed to significantly faster improvement of the mAP.

3.2 Correlation of certainty values and prediction performance

4 Discussion

- Sampling based on minimum certainty outperforms the average certainty sampling, and random sampling. In comparison with the other applications in agriculture (Blok et al. (2022))
- Considering the amount of data and the need to manual processing and annotation this is a significant contribution to saving time during data preparation
- The proposed approach relies on the closed class approach, meaning all expected species in the catch have to be a part of the initial training set, thus, the approach is suitable for adapting the model for its accuracy increase in the case of minority species presence or the presence of the appearance diversity of the well-represented species.
- Alternative methods for uncertainty estimation, i.e. region-level uncertainty sampling (Laielli et al. (2021)).

Given the complexity of the images depicting discarded fish on the conveyor belt, which complicates the annotation process in terms of time that need to be spend on marking often covered fish and expert knowledge to identify the species of the individuals where only small part of the fish is visible, active learning technique can play an important role in speeding up the process of data preparation

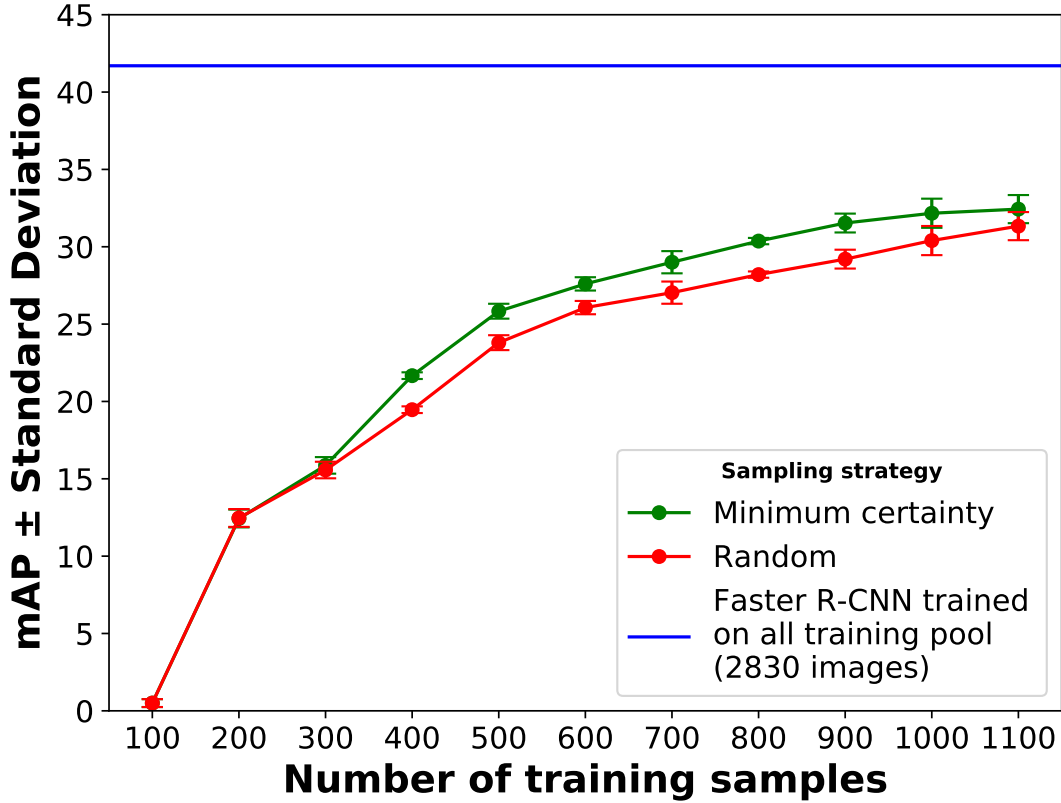


Figure 2: Comparison of the mAP for BoxAL trained with random sampling and active learning sampling based on minimum certainty

and retraining the models with the selected images. Additionally, as it has been shown in the results section, the images with lower F1 score have been selected. This fact proves that the images in which the model makes more mistakes are chosen by the model for retraining and therefore beneficial to improve the detection performance. Besides, the images that are selected by the method and indicate the lower F1 score can mean that the new species occur in the image, which will be confirmed by the human-in-the-loop during the annotation of the selected images. Active learning select most beneficial images for annotation, which is shown to be beneficial, however this does not substitute the manual annotation process. Instead it reduces the effort. Considering the current state of the electronic monitoring onboard fishing vessels, which is done manually and in a sample-based manner, our proposed approach is a substantial step towards automation of the monitoring data analysis. The annotation process of one image in the occluded environment can take up to one hour, in this paper, we have demonstrated that it is possible to reach the same detection performance with 200 less images that were sampled randomly. Taking into account that it can take almost an hour to annotate one image, our proposed approach can save 200 work hours, which also involves high costs.

4.1 Related work

A large variety of application domains where object detection is used faces a bottleneck in terms of labelled data availability. For this reason, a range of methods to minimize the human labour for labeled data preparation has been proposed. For instance, Abramson and Freund (2006) Another relevant work

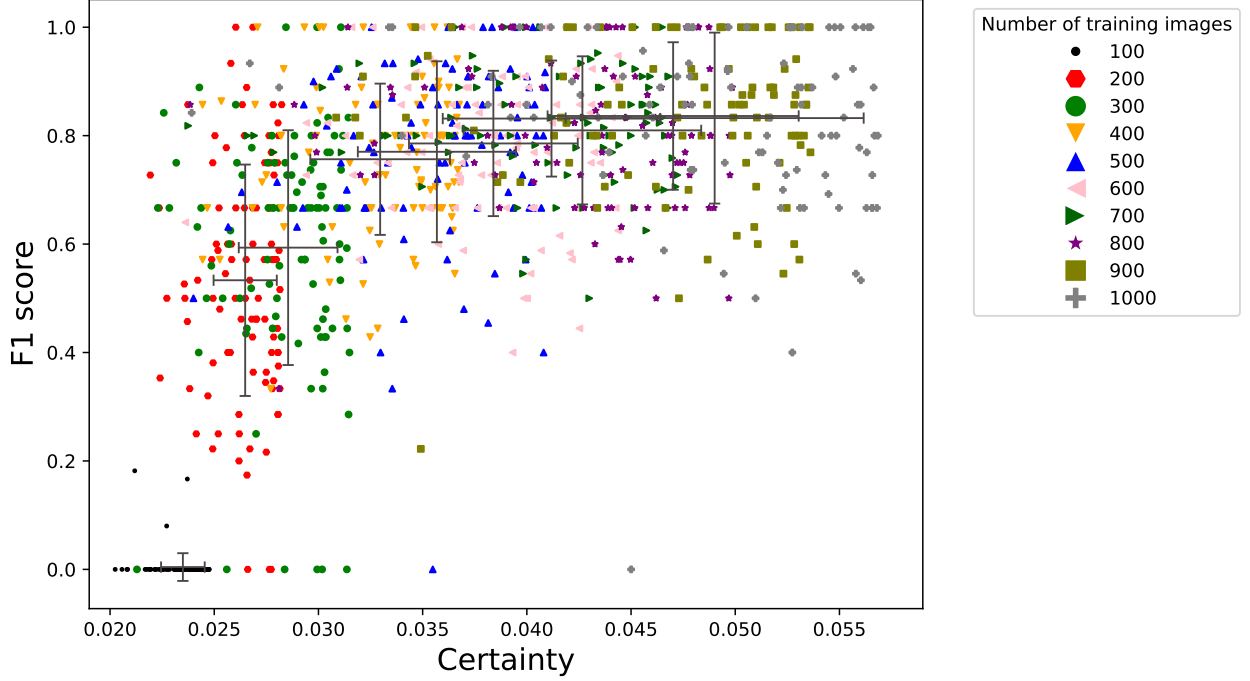


Figure 3: Relationship between F1 scores and certainty values for the training images sampled during ten iterations. Horizontal error bars correspond to certainty standard deviation; vertical error bars correspond to F1 score standard deviation.

that has employed object detection-based active learning is the Lightly YOLOv8-based active learning. This method was evaluated on the Lincolnbeet dataset and showed promising reductions in annotation effort.

4.2 Methodology implementation

Our approach take the full image as an input. Considering the high number of fish that is usually present in an image, it is often the case that the majority species is frequently present in the images together with less frequently occurring fish species. This is taken into account by sampling images with minimum certainty, where the lowest certainty value is taken from the list of all calculated certainties.

4.3 Active learning in aquatic science

AL was used in plankton classification problem, where it is as well difficult to get the dataset collected and annotated (Bochinski et al. (2019)).

5 Conclusion

- minimum sampling strategy results in reaching the same mAP as the random sampling strategy with 200 fewer training examples
- Considering the amount of data and the need to manual processing and annotation this is a significant contribution to saving time during data preparation

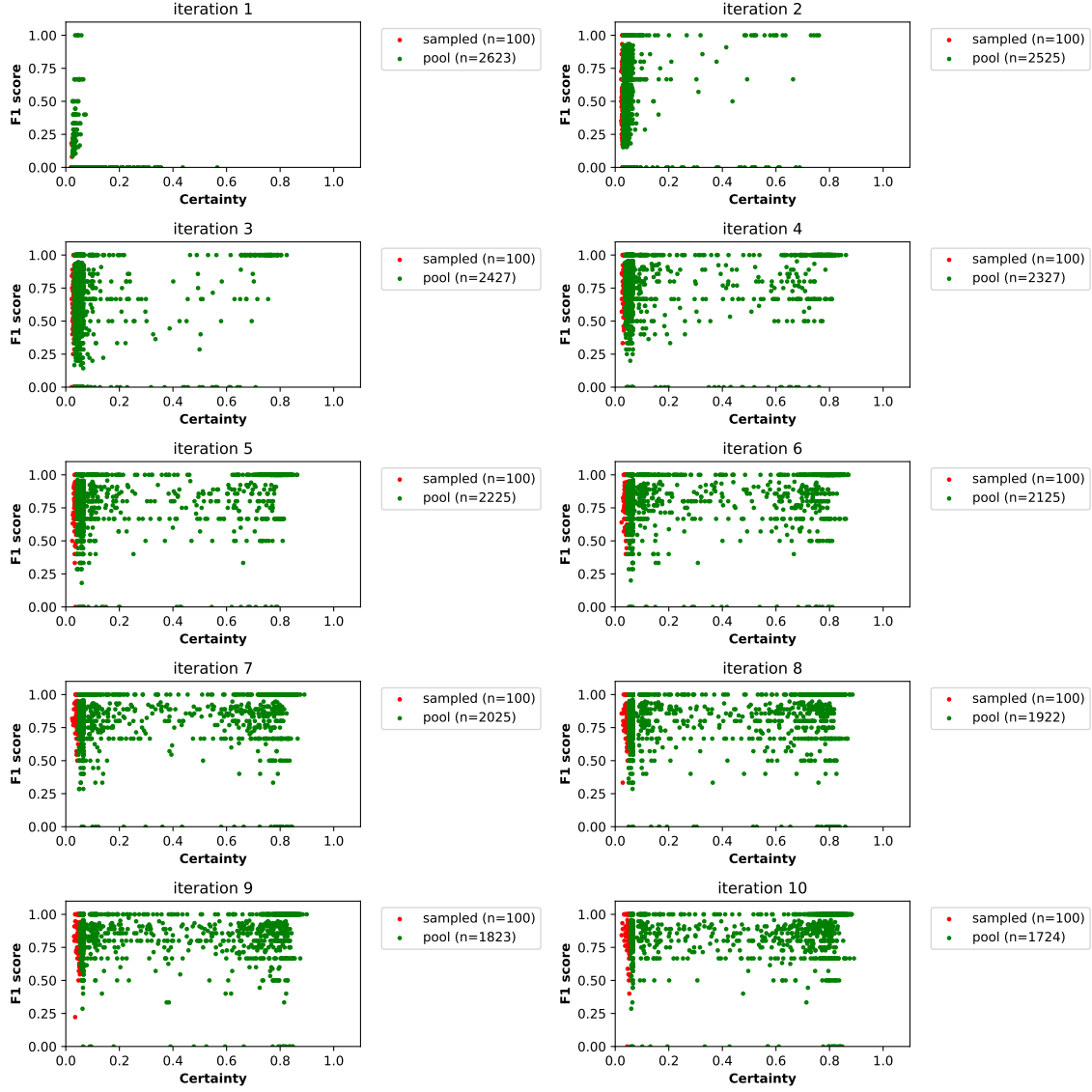


Figure 4: F1 scores and certainty values for the sampled images based on minimum certainty and the images that remain in the training pool for each of ten iterations

References

- Abramson, Y. and Freund, Y. (2006). Active learning for visual object detection.
- Blok, P. M., Kootstra, G., Elghor, H. E., Diallo, B., van Evert, F. K., and van Henten, E. J. (2022). Active learning with maskal reduces annotation effort for training mask r-cnn on a broccoli dataset with visually similar classes. *Computers and Electronics in Agriculture*, 197:106917.
- Bochinski, E., Bacha, G., Eiselein, V., Waller, T. J., Nejstgaard, J. C., and Sikora, T. (2019). Deep active learning for in situ plankton classification. In *Pattern Recognition and Information Forensics: ICPR 2018 International Workshops, CVAUI, IWCF, and MIPPSNA, Beijing, China, August 20-24, 2018, Revised Selected Papers 24*, pages 5–15. Springer.
- Laielli, M., Biamby, G., Chen, D., Loeffler, A., Nguyen, P. D., Luo, R., Darrell, T., and Ebrahimi, S. (2021). Region-level active learning for cluttered scenes. *arXiv preprint arXiv:2108.09186*.

- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
- van Essen, R., Mencarelli, A., van Helmond, A., Nguyen, L., Batsleer, J., Poos, J.-J., and Kootstra, G. (2021). Automatic discard registration in cluttered environments using deep learning and object tracking: class imbalance, occlusion, and a comparison to human review. *ICES Journal of Marine Science*, 78(10):3834–3846.
- Zou, Z., Chen, K., Shi, Z., Guo, Y., and Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*.

7 General discussion

Establishing an efficient discards registration system on board demersal mixed fisheries is challenging. Controlling the conditions that negatively influence image quality seems to be the key element in successfully deploying AI for discard registration. Although it is necessary to cover the fisher's working area, it was quickly recognized by the experts that the footage collected with a regular EM setup, e.g. four to five CCTV cameras overlooking the sorting process on board, would not be sufficient to accurately estimate discards on board demersal trawlers. It is difficult to obtain sufficiently accurate estimates of catch compositions due to low image quality: illuminance variation, camera dirtiness, and parts of the catch that are hidden by crew processing the catch, camera resolution and distance from the conveyor often do not allow to capture species-specific feature among other sources of image quality corruption (French et al., 2015; French et al., 2020). This leads to severe reduction in precision and accuracy. By integrating a high resolution camera in combination with 3-D images in a system where the image acquisition scene is semi-closed to ensure stable lighting conditions, prevent dirt on lenses, and blocked view (e.g. hands of crew) we are able to automatically record length-and-weight-by species of discarded catch on bottom trawlers.

During the first phase of the project, an on-board-like conveyor belt setup was constructed. This system, an exact copy of the situation on board, made testing newly developed algorithms possible, without being dependent of fishing vessels at sea (see Figure 1 in chapter 2). Working on land made it considerably easier to prepare ground truth of the collected images. The ground truth preparation consisted of two steps. The first step consisted of measurement of individual fish weight and species recognition by the expert and recording the fish using the developed image acquisition system. The second step consisted of annotating individuals location and species in the image and assigning the corresponding weight. To conduct trials, discarded catches of beam trawlers in the North Sea were collected, besides all fish species, these samples included debris, e.g. benthic species (sea stars, crabs, etc.), wood, stones, and other materials present in the catch, making it possible to mimic the sorting process on board with different levels of catch occlusions. These occlusion, e.g. cluttered catch, appear because fish lay on top of each other and are covered by debris. Such conditions considerably challenge annotation process as well as the detection performance, as the visible parts of fish are often not discriminative enough for both, accurate species detection and further check by human supervisor.

The automatic fish registration method developed in the FDF-project is a promising method to increase the monitoring coverage on demersal European fisheries. The ability of the developed algorithms to handle cluttered catch on the sorting belts enables recording of the complete catch without disturbing the process on board. But, the higher the level of occlusion and debris in the catch on the conveyor belt, the lower the detection performance. However, there is a correlation between the level of occlusion and the number of training examples. By adding more data (either real or synthetic) for heavily occluded fish, the performance could be improved. Unfortunately, this still does not solve the verification of the detection step in the case the detected fish is highly occluded. Although a larger training set may improve the system performance, the possibility to use some mechanical solutions, such as a precision pacing conveyor, can physically spread and separate the stacked by-catch to simplify the scene. Such a mechanical solution is expected to enable the most gain in performance. However it takes up more of the limited space available on board fishing vessels, and is, therefore, not the preferred solution, when implementing a system for the entire European fishing fleet. Alternatively, other cameras or sensor types than RGB camera, i.e. camera's using visible light, red, green and blue (RGB), can increase visibility of fish. Recent developments of sensor technology offer, i.e. X-ray line scan cameras to record the objects that are not visible by the RGB (<https://www.hamamatsu.com>). Such sensors are of the similar size as the industrial RGB cameras, however deviated exploitation terms may imply. Fishes with their dorsal side (from Latin dorsum 'back') facing the camera showed better detection, of course this only effects species with different dorsal and ventral (from Latin venter 'belly') sides, such as flat fish and ray species. In case of the

FDF-project, working with catches of beam trawlers, targeting flatfish, this was a factor to consider (Ulleweit et al. 2010). Nevertheless, during trials in the project algorithms performed better in identifying ray species based on the ventral side than human reviewers. The system transformation from a two-stage detection and tracking system, with colour images to a line scan camera system improved the accuracy of fish counting (chapter 4, technical report). Instead of recording the whole image at once, the line scan builds an image line by line, creating an infinite image that is “cut off” in separate pictures, that are, eventually, analysed by the algorithms. The use of the line scan technology eliminates the need for a tracking method and thereby removes one source of errors in counting, e.g. in situations where the tracker is failing to follow a fish in the subsequent picture frame, and, consequently, identifying it as a new fish, the same fish is counted double. An additional advantage of this transformation was that less space is needed for image acquisition. Since only a line of a several pixels wide is needed the box including the cameras and the lamps can be smaller and taking less space of the conveyor belt. This was considered to be a significant improvement, because space needed to process and sort the catch on board fishing vessels is limited (Needle et al., 2015).

The development of machine learning technology is an ongoing process of applying more advanced algorithms and train these algorithms with improved data sets, which is also shown in the improved performance levels between our first and second publication (chapters 2 and 3). The latest outcome, a novel integrated end-to-end simultaneous detection and weight prediction pipeline based on the state-of-the-art deep convolutional neural network, is able to identify species and weights of individual fish in the discarded catch with high accuracy levels (chapter 3). However, discard composition is very variable in space and time (Catchpole et al., 2005; Uhlmann et al., 2014). To be able to anticipate on the expected variation in catch composition of demersal trawlers a continuous learning method is being developed by constantly processing data sets to select relevant training samples (Gal & Ghahramani, 2016; Blok et al., 2022). Dealing with the fast amount of variation in the fish composition and appearance over the season and at different fishing grounds, it is important to further improve the generalization of the trained computer-vision methods. To this end, future research needs to explore a number of different methods:

- 1) Active learning: It is important to have a large and varied annotated data set to properly train the deep neural networks. Chapter 6 explored the use of active learning to select new images to be annotated by a human expert. Although the method showed a significant improvement over randomly selecting new training data, there is plenty of room for improvement. Future work needs to focus on a better estimation of the uncertainty of the neural network and on the selection based on diversity sampling.
- 2) Open-set recognition: In practice, it can occur that fishes appear in a completely new way or even that new species occur, not included in the training set. Chapter 5 explored the use of a DNN to detect individual fish without considering the species. Based on that, new methods can be developed that classify the individual fish as one of the known species or as an unknown species. In the latter case, the image can be presented to a human expert for annotation.
- 3) Methods that make better use of the available training data need to be explored, such as advanced data augmentation methods and domain randomization.
- 4) Getting annotated data is time-consuming and costly. However, we have already vast amounts of unlabeled images and with the presented camera system (Chapter 4), it is easy to get much more unlabeled data. Recently, new AI methods have been developed that make use of the unlabeled data. Future work, therefore, needs to explore methods such as self-supervised and unsupervised pre-training and domain adaptation.
- 5) Methods to assist image annotation: Recently so-called foundation models have been developed that are trained on extremely large general image datasets. Models such as “segment anything model” are able to segment general objects in images, only needing a single mouse click, which will speed-up annotation. There are now even AI models available that are linked to large-language models, such as “grounding DINO” that can detect objects based on a language query. Although these methods are far from perfect at the moment, they can be used to automate part of the annotation process.

Besides the technical challenges there is also a practical side to the implementation of EM on fishing vessels. Communication with the fishing industry, but certainly with the fishers participating in the FDF-project is a key element for a successful implementation of the technology. The guidelines developed in the FDF-project provide the practical context of what is needed to implement EM

technology is a fishery. The biggest challenges on a fleet wide implementation are (1) differences between vessels, and probably fleets, a one-size fits all approach is not going to work, and (2) the level of resources needed for maintenance and repair, which also requires involvement of the fishers and their crew. The impact of the conditions at sea on the, mainly electronic, equipment is enormous, and the options for maintenance and repair is limited, given the relatively short time window vessels are in the harbour.

In the second phase of the FDF-project (FDF 2.0) the technology should be prepared for a wider implementation in the European fleet, possibly include demersal trawlers from other EU members states, i.e. Denmark and Belgium. In the second phase we aim to provide high quality labelled data to facilitate an easy start for algorithm development in other regions, e.g. vessels fishing in different fishing grounds, vessels from other member states. Another important aspect is the operationalization of the computer vision technology within existing EM service provider workflows, software, user interfaces (GUI's), and general EM system technology. Integration with commercial EM systems is essential to enable a wider use of the developed technology. Many fisheries worldwide already implemented EM in the national fisheries management schemes, therefore, changing to a different technology is an undesirable option. Embedding the developed of the automated catch registration tool into existing commercial EM technology will move the developed tool from proof of concept to a market-ready technology.

References

- Allken, V., Handegard, N.O., Rosen, S., Schreyeck, T., Mahiout, T., Malde, K., 2019. Fish species identification using a convolutional neural network trained on synthetic data. *ICES Journal of Marine Science* 76, 342–349. <https://doi.org/10.1093/icesjms/fsy147>
- Allken, V., Rosen, S., Handegard, N.O., Malde, K., 2021. A real-world dataset and data simulation algorithm for automated fish species identification. *Geoscience Data Journal*. <https://doi.org/10.1002/gdj3.114>
- Ames, R.T. (2005) The efficacy of electronic monitoring systems: a case study on the applicability of video technology for longline fisheries management. Scientific Report No. 80. International Pacific Halibut Commission, Seattle, Washington.
- Baker, M.S., Von Harten, A., Batty, A. and McElderry, H. (2013) Evaluation of electronic monitoring as a tool to quantify catch in a multispecies reef fish fishery. In: 7th International Fisheries Observing and Monitoring Conference. Vina del Mar, Chile, p Presentation.
- Blok, P. M., Kootstra, G., Elghor, H. E., Diallo, B., van Evert, F. K., & van Henten, E. J. (2022). Active learning with MaskAL reduces annotation effort for training Mask R-CNN on a broccoli dataset with visually similar classes. *Computers and Electronics in Agriculture*, 197, 106917.
- Catchpole, T.L., Frid, C.L.J., Gray, T.S. (2005). Discards in North Sea fisheries: causes, consequences and solutions. *Marine Policy* 29, 421–430.
- French, G., Fisher, M., Mackiewicz, M., Needle, C., 2015. Convolutional Neural Networks for Counting Fish in Fisheries Surveillance Video. *British Machine Vision Association and Society for Pattern Recognition*. <https://doi.org/10.5244/c.29.mvab.7>
- French, G., Mackiewicz, M., Fisher, M., Holah, H., Kilburn, R., Campbell, N., & Needle, C. (2020). Deep neural networks for analysis of fisheries surveillance video and automated monitoring of fish discards. *ICES Journal of Marine Science*, 77(4), 1340–1353.
- Gal, Y., & Ghahramani, Z. (2016, June). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059). PMLR.
- Hintjens, P., 2013. ZeroMQ: messaging for many applications. " O'Reilly Media, Inc."
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.
- Mangi, S.C., Dolder, P.J., Catchpole, T.L., Rodmell, D. and de Rozarieux, N. (2015) Approaches to fully documented fisheries: Practical issues and stakeholder perceptions. *Fish and Fisheries* 16, 426–452.
- McElderry, H., Schrader, J. and Illingworth, J. (2003) The Efficacy of Video-Based Monitoring for the Halibut Longline. Victoria, Canada. Canadian Research Advisory Secretariat Research Document 2003/042. 80 pp.
- Michelin, M., Elliott, M., Bucher, M., Zimring, M., Sweeney, M. (2018) "Catalyzing the Growth of Electronic Monitoring in Fisheries." California Environmental Associates, September 10. 63pp.

Michelin, M., Zimring, M. 2020. Catalyzing the growth of electronic monitoring in fisheries. California Environmental Associates (CEA) Consulting, The Nature Conservancy.
<https://www.fisheriesem.com/pdf/Catalyzing-EM-2020report.pdf>

Mohri, M., Rostamizadeh, A., Talwalkar, A., 2012. Foundations of Machine Learning. 7-9.

Mortensen, L.O., Ulrich, C., Olesen, H.J., Bergsson, H., Berg, C.W., Tzamouranis, N. and Dalskov, J. (2017) Effectiveness of fully documented fisheries to estimate discards in a participatory research scheme. *Fisheries Research* 187, 150–157.

Msomphora, M.R. and Aanesen, M. (2015) Is the catch quota management (CQM) mechanism attractive to fishers? A preliminary analysis of the Danish 2011 CQM trial project. *Marine Policy* 58, 78–87.

Needle, C.L., Dinsdale, R., Buch, T.B., Catarino, R.M.D., Drewery, J. and Butler, N. (2015) Scottish science applications of Remote Electronic Monitoring. *ICES Journal of Marine Science* 72, 1214–1229.

Plet-Hansen, K.S., Bergsson, H., Mortensen, L.O., Ulrich, C., Dalskov, J., Jensen, S.P. and Olesen, H.J. (2015) Final Report on Catch Quota Management and choke species – 2014. Technical Report DOI: 10.13140/RG.2.2.11883.95524

Plet-Hansen, K.S., Eliassen, S.Q., Mortensen, L.O., Bergsson, H., Olesen, H.J. and Ulrich, C. (2017) Remote electronic monitoring and the landing obligation – some insights into fishers’ and fishery inspectors’ opinions. *Marine Policy* 76, 98–106.

Stanley, R.D., McElderry, H.I., Mawani, T. and Koolman, J. (2011) The advantages of an audit over a census approach to the review of video imagery in fishery monitoring. *ICES Journal of Marine Science* 68, 1621–1627.

Tseng, C.H., Kuo, Y.F., 2020. Detecting and counting harvested fish and identifying fish types in electronic monitoring system videos using deep convolutional neural networks. *ICES Journal of Marine Science* 77, 1367–1378. <https://doi.org/10.1093/icesjms/fsaa076>

Uhlmann, S.S., Helmond, A.T.M. Van, Stefánsdóttir, E.K., et al. (2014) Discarded fish in European waters: general patterns and contrasts. *ICES Journal of Marine Science* 71, 1235–1245.

Ulleweit, J., Stransky, C., and Panten, K. (2010) Discards and discarding practices in German fisheries in the North Sea and Northeast Atlantic during 2002–2008. *Journal of Applied Ichthyology* 26, 54–66.

Ulrich, C., Olesen, H.J., Bergsson, H., et al. (2015) Discarding of cod in the Danish Fully Documented Fisheries trials. *ICES Journal of Marine Science* 72, 1848–1860.

van Helmond, A.T.M., Chen, C. and Poos, J.J. (2015) How effective is electronic monitoring in mixed bottom-trawl fisheries? *ICES Journal of Marine Science* 72, 1192–1200.

van Helmond, A.T.M., Chen, C. and Poos, J.J. (2017) Using electronic monitoring to record catches of sole (*Solea solea*) in a bottom trawl fishery. *ICES Journal of Marine Science* 74, 1421–1427.

van Helmond, A.T.M., Mortensen L.O., Plet-Hansen K.S., Ulrich C., Needle C.L., Oesterwind D., Kindt-Larsen L., Catchpole T., Mangi S., Zimmermann C., Olesen H.J., Bailey N., Bergsson H., Dalskov J., Elson J., Hosken M., Peterson L., McElderry H., Ruiz J., Pierre J.P., Dykstra C., and Poos J.J. (2020). Electronic monitoring in fisheries: Lessons from global experiences and future opportunities. *Fish and Fisheries* 21: 162–189.

van Helmond, A.T.M. (2021). Research for PECH Committee – Workshop on electronic technologies for fisheries – Part II: Electronic monitoring systems, European Parliament, Policy Department for Structural and Cohesion Policies, Brussels

Wang, X., Girdhar, R., Yu, S. X., & Misra, I. (2023). Cut and learn for unsupervised object detection and instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3124-3134).

Justification

Report C076/23

Project Number: 4311400031

The scientific quality of this report has been peer reviewed by a colleague scientist and a member of the Management Team of Wageningen Marine Research

Approved: Allard van Mens
Researcher

Signature:

Date: 20/12/2023

Approved: Tammo Bult
Director

Signature:

Date: 20/12/2023

Wageningen Marine Research
T +31 (0)317 48 7000
E: marine-research@wur.nl
www.wur.eu/marine-research

Visitors' address

- Ankerpark 27 1781 AG Den Helder
- Korringaweg 7, 4401 NT Yerseke
- Haringkade 1, 1976 CP IJmuiden

With knowledge, independent scientific research and advice, **Wageningen Marine Research** substantially contributes to more sustainable and more careful management, use and protection of natural riches in marine, coastal and freshwater areas.



Wageningen Marine Research is part of Wageningen University & Research. Wageningen University & Research is the collaboration between Wageningen University and the Wageningen Research Foundation and its mission is: 'To explore the potential for improving the quality of life'
