




## Contribution to the Themed Section: 'Applications of machine learning and artificial intelligence in marine science'

### Original Article

# Automatic segmentation of fish using deep learning with application to fish size measurement

Rafael Garcia <sup>1,2\*</sup>, Ricard Prados<sup>2</sup>, Josep Quintana<sup>3</sup>, Alexander Tempelaar<sup>3</sup>, Nuno Gracias<sup>1</sup>, Shale Rosen<sup>4</sup>, Håvard Vågstøl<sup>5</sup>, and Kristoffer Løvall<sup>5</sup>

<sup>1</sup>Computer Vision and Robotics Institute, University of Girona, Campus Montilivi, Edif. P4, ES17003, Girona, Spain

<sup>2</sup>Girona Vision Research SL, Science and Technology Park of the University of Girona, c/ Pic de Peguera 11, Edif. Giroemprèn, ES17003, Girona, Spain

<sup>3</sup>Coronis Computing SL, Science and Technology Park of the University of Girona, c/ Pic de Peguera 11, Edif. Giroemprèn, ES17003, Girona, Spain

<sup>4</sup>Institute of Marine Research, P.O. Box 1870 Nordnes, NO-5817 Bergen, Norway

<sup>5</sup>Scantrol Deep Vision, Sandviksboder 1C, NO-5035 Bergen, Norway

\*Corresponding author: tel: + 34 676 511 024; e-mail: [rafael.garcia@udg.edu](mailto:rafael.garcia@udg.edu).

Garcia, R., Prados, R., Quintana, J., Tempelaar, A., Gracias, N., Rosen, S., Vågstøl, H., and Løvall, K. Automatic segmentation of fish using deep learning with application to fish size measurement. – ICES Journal of Marine Science, 77: 1354–1366.

Received 10 May 2019; revised 11 July 2019; accepted 14 August 2019; advance access publication 22 October 2019.

One of the leading causes of overfishing is the catch of unwanted fish and marine life in commercial fishing gears. Echosounders are nowadays routinely used to detect fish schools and make qualitative estimates of the amount of fish and species present. However, the problem of estimating sizes using acoustic systems is still largely unsolved, with only a few attempts at real-time operation and only at demonstration level. This paper proposes a novel image-based method for individual fish detection, targeted at drastically reducing catches of undersized fish in commercial trawling. The proposal is based on the processing of stereo images acquired by the Deep Vision imaging system, directly placed in the trawl. The images are pre-processed to correct for nonlinearities of the camera response. Then, a Mask R-CNN architecture is used to localize and segment each individual fish in the images. This segmentation is subsequently refined using local gradients to obtain an accurate estimate of the boundary of every fish. Testing was conducted with two representative datasets, containing in excess of 2600 manually annotated individual fish, and acquired using distinct artificial illumination setups. A distinctive advantage of this proposal is the ability to successfully deal with cluttered images containing overlapping fish.

**Keywords:** deep learning, fish sizing, trawl camera system

## Introduction

According to the UN Food and Agriculture Organization, 33% of commercially important marine fish stocks worldwide are overfished (FAO, 2018). One of the causes of overfishing is that, in addition to targeted species, the fishing gears often catch other unwanted fish and marine life. Globally, nearly 11% of total catches are discarded because they are not the proper species or sizes (Pérez Roda *et al.*, 2019). In some cases, the quantity of this by-catch can exceed that of the targeted species. Excessive by-

catch is an immediate problem for fishers as it slows their catch sorting operations considerably, increases fuel consumption and wear on their fishing gear. Under management systems utilizing by-catch caps or closures to protect juveniles, fishing opportunities may be curtailed. In the long term, high levels of by-catch can contribute to overfishing jeopardize the long-term sustainability of the fishery.

Some countries and regions have enacted prohibitions on discarding unwanted catches. The most recent revision to the EU

Common Fisheries Policy (EU regulation 1380/2013) institutes a landing obligation requiring all catches of regulated commercial species to be landed and counted against quota. This includes catches of undersized individuals, which can be utilized to avoid waste, but not for direct human consumption or at a profit which could result in the establishment of markets.

Most fishermen use echosounders to detect fish schools and make qualitative estimates of the amount of fish and species present. Advanced “split beam” echosounders can give an indication of fish size, and characteristics such as frequency-response and school geometry can be used to differentiate between some species (Korneliusen *et al.*, 2009). However, systems to provide quantitative real-time species identification and measurement during fishing are largely in the demonstration phase (Pobitzer *et al.*, 2015; Berges *et al.*, 2018). As a result of this uncertainty, vessels relying on acoustics to target-specific species may catch undersized individuals or other species.

This paper proposes a novel fish sizing method when capturing fish using a trawl. The proposal is based on the use of the existing Deep Vision system (Rosen and Holst, 2013), directly placed in the trawl, to acquire stereo image pairs at a fixed frequency of five or ten images per second. The images are saved in a solid-state unit capable of storing  $\sim 1$  million image pairs, equivalent to 60 h of data collection. In this paper, the images have been processed offline, but we aim at processing them onboard the Deep Vision system in the near future which will make real-time active sorting possible. This will enable more sustainable fishing activities by reducing catches of undersized individuals and unwanted species.

## Material and methods

### Data acquisition

Data were obtained on two testing cruises in the North Atlantic, the first in the North Sea onboard the Norwegian R/V “Dr Fridtjof Nansen” during March of 2017 (hereafter dataset 1), and the second in the Norwegian Sea with the chartered fishing vessel M/S “Vendla” during May of 2017 (hereafter dataset 2). Both vessels used an 832-m circumference pelagic trawl designed for surveys of small pelagic species in the Northeast Atlantic. Dataset 1 included images of saithe (*Pollachius virens*), blue whiting (*Micromesistius poutassou*), redfish (*Sebastes* spp.), Atlantic mackerel (*Scomber scombrus*), velvet belly lanternshark (*Etmopterus spinax*), and Norway pout (*Trisopterus esmarkii*), while dataset 2 included images of Atlantic mackerel, blue whiting, and Atlantic herring (*Clupea harengus*).

Acquisition of stereo image pairs of fish in the trawl was done using the Deep Vision system which is currently used to provide fisheries survey operations with information about depth and position of fish entering the sampling trawl. Using Deep Vision, it is also possible to conduct surveys which retain images rather than the actual fish. This lessens the environmental impact of the sampling and the workload of handling and measuring the catch. At the same time it provides images and metadata that can be used for length measurements and species classification. Combined with acoustic measurements this information provides higher confidence data used as input for stock assessment.

The Deep Vision system is divided into a subsea system and a topside system. The subsea system has a stereo camera, strobe lights, battery, and an enclosing studio frame designed for optimal image quality and consistency. The studio frame is integrated into the trawl to ensure smooth flow of catch through the system,

and protects the electronic components from the rigours of trawl handling and operations (see Figure 1).

The topside system provides a graphical user interface for size measurement and species classification, through a combination of manual and more automated processes. The output from the analysis software is combined with the data from the subsea system into an annotated dataset that can be used to produce statistical data.

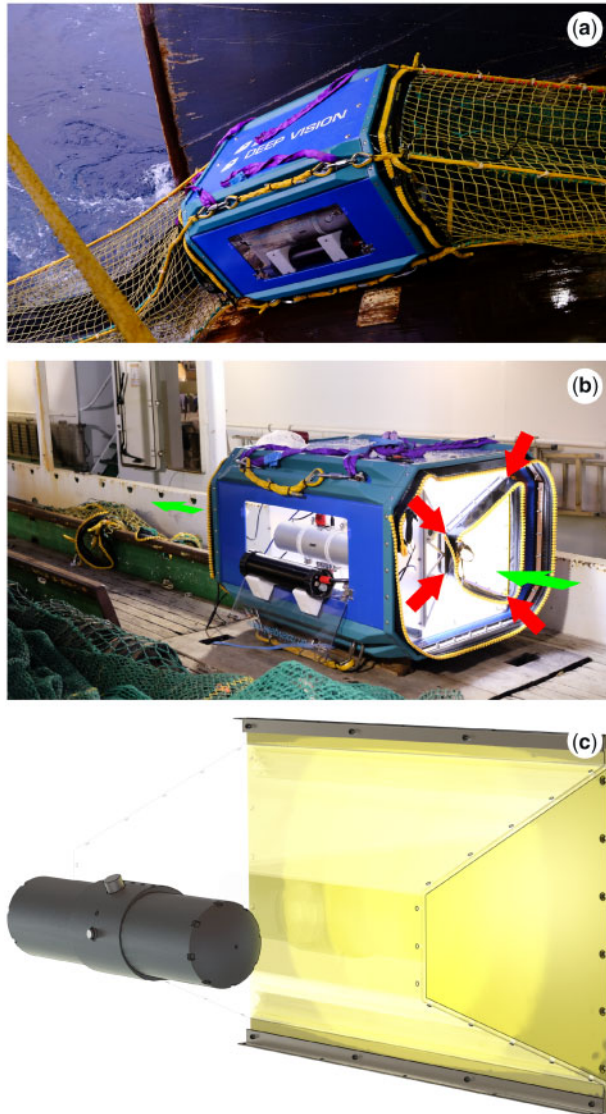
During both surveys, the stereo image pairs were recorded at 5 fps, in JPG format, with an image resolution of  $1392 \times 1040$  pixels. Lighting was provided by two synchronized strobes producing  $\sim 18000$  lumen each at a colour of 4100 K. Although the lights were pointed to the ceiling and floor of the studio frame to provide diffused lighting, their angle varied slightly between cruises resulting in slight differences in reflection and illuminance inside the volume where objects pass through the Deep Vision canal (Figure 4). In addition, the user was allowed to make changes to camera exposure time, gain and gamma correction, introducing an additional source of inconsistency in image appearance. The impact of this uneven appearance on further image analysis prompted a full mechanical redesign of the lights to a production model with both higher total light output and fixed angle (Figure 1).

All the acquired images are analysed using the processing pipeline illustrated in Figure 2. First, images are pre-processed to correct nonlinearities and non-uniform lighting effects. Next, we use a Mask R-CNN architecture to localize and segment every individual fish in the image. The obtained segmentation is then refined in the next step using the local gradient to estimate the boundary of every fish. Finally, the length of the fish is computed exploiting stereo information. The different processing phases are detailed below.

### Image pre-processing

Image pre-processing aims at correcting non-uniform lighting to produce images with a similar contrast between the fish and the background regardless of the location of the fish in the image. To carry out this correction, we should first linearize the image (Prados *et al.*, 2017).

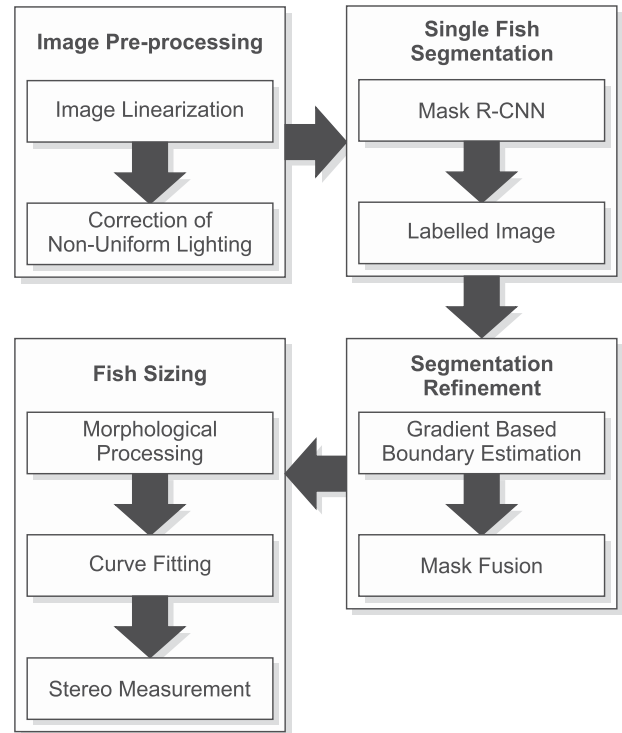
Linearization is a desirable pre-processing step since cameras provide RGB values that are non-proportional with the incoming light energy. This is so because the human visual system has a nonlinear response (Burton, 1973). If an image encodes light in a  $[0,255]$  interval, a value of 128 is perceived as half the lightness by the human eye, but in reality that point is reflecting ( $\sim$ ) 25% of the light. That is, the camera response functions for all the colour channels are adapted to the human eye, and therefore they are nonlinear, especially if images have been stored using the JPG format, as it is often the case to minimize disc space to store large datasets. Therefore, since most processing algorithms assume that the value of a pixel is proportional to the amount of light collected by that specific pixel, linearizing the image would provide a better-conditioned set of pixel values for further processing. Moreover, using linearized images ensures providing the processing algorithms with a more accurate representation of the measured spectra, and consequently its behaviour and outputs become more consistent. In our case, images are linearized using the camera linearization method described in Debevec and Malik (1997). After this process, the RGB values become proportional with the irradiance on the sensor pixels, and the image is ready to



**Figure 1.** Deep Vision subsea system. The system is placed inside a trawl net (a) and contains a stereovision camera set and indirect lighting source. The arrows in the middle figure (b) define the “studio” section, corresponding to the area where the catch flows, and which can be seen in detail in the bottom schematic (c). Fish cross through a trapezoidal plexiglass section which ensures they maintain at least 20 cm distance from the cameras and lights and are within the field of view of the cameras.

undergo further linear operations, such as the correction of the non-uniform lightning. All subsequent operations are performed in linear RGB values.

Although the Deep Vision system provides images with a good overall illumination, the amount of light on the central area of the images is higher than that at the corners of the image. Therefore, once the images are linearized, we also correct the images for non-uniform lighting. To do this, we first convert the images from RGB to HSV (Hue, Saturation, Value), where  $V$  corresponds to the image luminance (Schwarz et al., 1987). The luminance channel is the only component that will be used to correct the illumination effect. The illumination correction is performed by modelling the background, i.e. we compute the



**Figure 2.** Automatic fish measurement pipeline. The process starts with the pre-processing of the image, and then a CNN localizes every individual fish. The CNN also provides a segmentation mask for the fish. Next, these masks are refined using local contrast information to delineate the boundary of every fish, and finally the length of the specimen is measured based on stereo cues.

median of a sufficiently large set of images of the scene (typically 300). The high power of the lighting system makes any external lighting contribution negligible, and consequently the illumination can be assumed as constant during the whole trawl. Ideally the images are selected at the beginning of the trawl haul before fish begin entering the field of view, although the only requirement is that, for the volume of 300 images, every pixel coordinate should not contain fish in slightly more than half the images ( $>150$ ). The median value for each image pixel will be later on computed. If a given coordinate show no fish most of the time, the appropriate background value will be kept for this pixel location by the median measure. Once the median image has been computed from the  $V$  component of the set of images, we obtain a background luminance image that allows us to infer the illumination of the scene. The estimated background image is then inverted and applied as a non-uniform illumination compensation pattern to correct the luminance ( $V$ ) of every image of the sequence. The RGB values of the final images are recovered from the HSV representation, ensuring that the correlation between the RGB channels is preserved, i.e. the original colours are kept.

It should be noted that working directly on the RGB colour space using channel-wise processing, as is commonly done in several image processing algorithms, may lead to a loss of the correlation between the values of the RGB triplets, thus shifting the original colours acquired by the camera.

Compensating the non-uniform illumination on all the images has proved to better condition the data to perform the subsequent fish segmentation (Prados et al., 2014).

### Single fish detection

Our aim is to be able to segment individual fish in the images, so that measuring the fish once it has been segmented becomes a trivial task. Figure 3 illustrates the problem we want to solve. Figure 3a shows a situation in which fish segmentation is quite easy since the background of the Deep Vision system can be modelled, and everything that is not background could be assumed to be a fish. However, Figure 3b shows a more challenging situation in which the fish to be measured are overlapping, making it difficult to determine their outline. In these situations in which we are not able to formalize an algorithm to recognize an object (e.g. a fish), using of machine learning methods has shown to be the best alternative. Among machine learning, deep convolutional neural networks (CNNs) have proved to be capable of achieving the best results on challenging datasets using supervised learning (Krizhevsky *et al.*, 2017). CNNs have also demonstrated good accuracy in automatic classification of species using simulated Deep Vision images (Allken *et al.*, 2019).

One of the state-of-the-art CNN-based deep learning object detection approaches is *Region-CNN* (or *R-CNN*). *R-CNN* provides a solution to the fast detection of regions of interest (RoI) within an image. Based on this approach, more complex architectures have recently appeared such as *Faster R-CNN* (Girshick, 2015) for faster speed object detection, as well as *Mask R-CNN* (He *et al.*, 2017) for object segmentation. In this paper, we use a *Mask R-CNN* architecture for fish detection and segmentation. *Mask R-CNN* combines *Faster R-CNN* for object detection in which the number of objects may vary from image to image, and fully convolutional networks (FCNs) for segmentation to establish what pixels in the image belong to what object. This step of detecting and delineating the boundaries of every individual object in an image is called “semantic segmentation,” and allows us to differentiate individual fish when two or more instances of a fish overlap in the image, as illustrated in Figure 3b.

*Faster R-CNN* performs individual fish detection in two stages. First, it determines the bounding boxes (i.e. RoIs) using the region proposal network (RPN) standard. The RPN is basically a lightweight neural network that scans the image in a sliding-window fashion to find regions that contain objects. Second, for each RoI it determines the class label of the object through RoI pooling. Therefore, *Mask R-CNN* incorporates these two stages, but it performs RoI pooling in such a way that there is no loss in stride quantization due to rounding when pooling is performed, as opposed to the rounding performed by *Faster R-CNN* (Ren *et al.*, 2015). Moreover, the sliding window is handled by the convolutional nature of the RPN, which allows it to scan all regions in parallel exploiting the GPU architecture.

FCNs are used to predict the mask for every RoI. Convolutional layers retain spatial orientation and this information is crucial for location-specific tasks such as creating a mask for every individual fish (He *et al.*, 2017). This is a clear advantage with respect to fully connected layers, in which the spatial orientation of pixels with respect to each other is lost as they are squeezed together to form a feature vector (Long *et al.*, 2015).

Our *Mask R-CNN* architecture was initially pre-trained for the COCO dataset (Lin *et al.*, 2014). Then, the last layer was modified to classify between fish and background and we re-trained the last layers using our fish training data for 20 iterations. This fine-tuning strategy allows us to reduce the training time and the

needed amount of data compared to training from scratch. Next, the full network was trained with our trawling data. In all cases, during training we tried to reduce overfitting on image data by artificially enlarging the dataset using data augmentation, which included image translations, horizontal and vertical reflections, rotations, and shear transformations.

### Segmentation refinement

The mask computed by *Mask R-CNN* has been obtained using a low-resolution image. Thus, the mask that segments the fish has a lower accuracy than those that can be obtained from the full-resolution original images. Therefore, a final stage of mask refinement is applied to obtain a much finer spatial layout of the fish, i.e. a more accurate segmentation.

The blobs estimated by *Mask R-CNN* are first scaled and transferred to the full-resolution image ( $1228 \times 1027$  pixels). Then, the gradients of the *V* channel on the original image are computed. This results in an image where the boundary of the objects is clearly distinguishable. The gradient magnitudes are thresholded to keep only the higher values, that is, the most prominent boundaries. Finally, both the *Mask R-CNN* masks, resulting in most cases in conservative segmentation, and the gradient-based boundary refinement masks, are fused into a single one for each image object. Empty inner areas are filled using binary morphological operators.

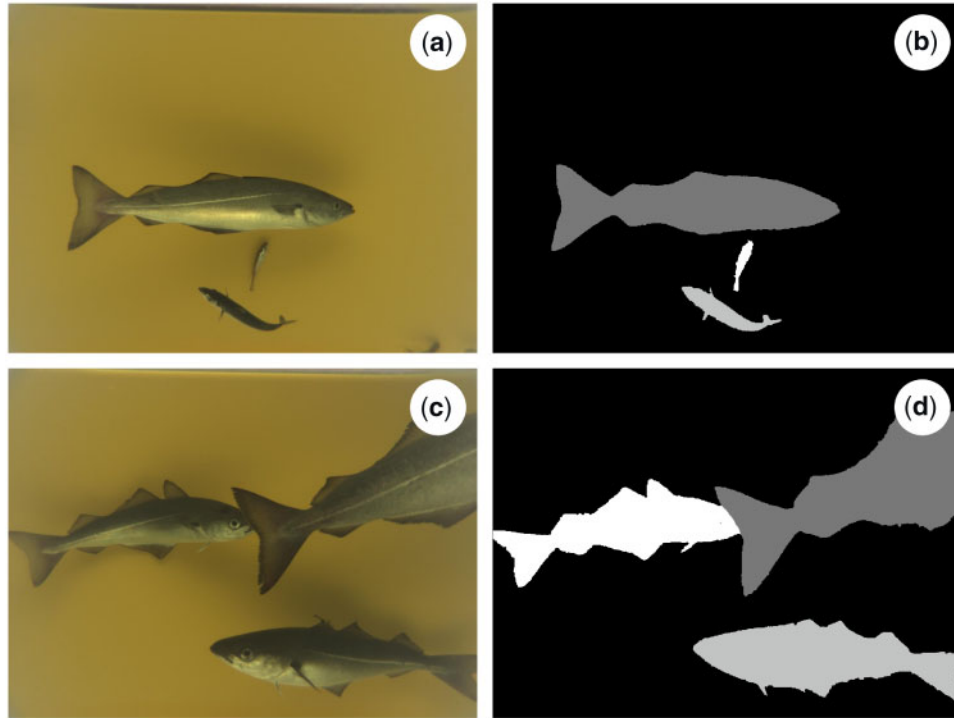
In case of overlapping fish, *Mask R-CNN* masks are used to guess where the boundaries of every specimen should be placed, given that the gradient-based refinement cannot distinguish among different objects. To determine which pixel belongs to each fish, *Mask R-CNN* masks are dilated using a customized multi-label dilate operation, which stops growing in a given direction when another neighbouring object is growing in the opposite direction and colliding with the first. The result of this dilate operation is used to determine the contribution of the gradients image to each fish mask.

### Segmentation performance

To evaluate the performance of the masks obtained by our processing pipeline, a detection accuracy measure is required. A standard set of metrics [intersection over union (IoU) and pixel accuracy] is used to quantify the segmentation results, since they are the *de facto* evaluation metrics used in object detection. IoU, also referred as Jaccard index, is an evaluation metric used to measure the accuracy of object segmentation on a particular dataset. IoU is often computed using the bounding box predicted by the CNN detector and the ground-truth (i.e. hand labelled) bounding box. In our case, since our detector generates a pixel region (mask) containing the pixels that correspond to a given fish, and the ground-truth is also a hand-labelled pixel region, IoU is computed using these two regions. The final score is obtained by dividing the area of overlap of the predicted region and the ground-truth region by the area of union of both the predicted region and the ground-truth region:

$$\text{IoU} = \frac{\text{ground-truth} \cap \text{prediction}}{\text{ground-truth} \cup \text{prediction}}$$

However, the measure of pixel accuracy corresponds to the percentage of pixels in the image which were correctly classified.



**Figure 3.** Fish segmentation. In simple cases such as (a), fish can be segmented into individual specimens simply by background subtraction (b). However, we need a cognitive understanding of the image to be able to segment the three fish instances in (c) shown in (d).

Usually it is presented for each class and the mean of all classes is provided. In our case both values are the same as we only have the “fish” class.

For this metric we need to introduce the notions of TP, TN, FP, and FN. True positive (TP) represents a pixel that is correctly predicted to belong to the given class whereas a true negative (TN) represents a pixel that is correctly identified as not belonging to the given class. False positives (FP) and false negatives (FN) are defined accordingly. The accuracy metric is then computed as

$$\text{accuracy} = \sum \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

### Length estimation

Once the specimens have been properly segmented, the final stage consists of finding a line that accurately describes the length of the fish. For this purpose, we estimate the fish skeleton using morphological operations applied to the labelled image, but it should be noted that the actual length of the fish should be estimated taking into account its 3D pose. The thinning morphological operation involves eroding the segmented region until skeleton level (Dougherty, 1992), i.e. shrinking the region corresponding to the individual fish until the blob becomes 1 pixel wide. This typically leads to a line centred along the main axis of the fish. Before performing morphological skeletonization, the binary masks resulting from the segmentation of the previous section are smoothed by applying a “closing” morphological operation. In this way, a continuous and typically smooth line is obtained, representing the main axis of the fish.

The next step is the estimation of a curve following the trajectory described by the pixels of the skeleton. Once the points

defining the skeleton have been obtained, a cubic polynomial is estimated using RANSAC (Fischler and Bolles, 1981). In this way, the points of the skeleton are classified in inliers and outliers, and after a number of iterations, a consensus solution is computed by least squares fit of the largest set of inliers, obtaining the final estimation of the curve.

Once the curve equation is derived, the starting and ending points defining the length of the fish are determined as the intersection between the estimated curve and the boundaries of the smoothed fish blob. Since the stereo system has been calibrated and the images rectified (Hartley and Zisserman, 2003), these points can then be easily transferred from the right to the left image of the stereo pair by applying the axis constraints determined by the stereo rectification. Then, once front and back points have been established in both images of the stereo pair, a set of uniformly distributed points along the curve are selected in the right image. These points are transferred to the left image following the same uniform distribution, using the image rectification to determine its Y location. Finally, the set of measurement point pairs from the right and left images is used to compute the distances of the segments connecting them using epipolar geometry, thanks to the calibration of the stereo system.

### Results

A total of 1805 manually annotated images (corresponding to the left camera of the stereo pairs) have been used to validate the pipeline proposed in this paper, with a total of 2629 fish annotations. These images have been acquired in two different cruises. Dataset 1, including 1605 annotated images, was acquired by R/V Dr Fridtjof Nansen on March 2017. This dataset represents a small subset of all the images acquired during the survey, and includes frames from three different hauls (138 055 stereo pairs).

Dataset 2 was acquired by F/V Vendla on May 2017 and it includes 200 annotated images, all of them from the same haul (28 117 stereo pairs). Both surveys consist of thousands of images, but only small samples containing fish suitable for an appropriate labelling (a large percentage of images contain no fish at all) can be used. The annotation effort is significant, taking into account that the labelling procedure implies a precise manual segmentation of each specimen, not a simpler approximate bounding box specification.

Figure 4 illustrates the appearance of the images of both datasets, as well as the result of correcting non-uniform illumination. It should be noted that the appearance of the images in both datasets is different due to the change of lighting arrangement and camera parameters (with a gain factor of 1.2 in case of dataset 1 and gain factor of 2 in case of dataset 2). In dataset 2, the central part of the image is considerably brighter than in dataset 1, and as a consequence, the margins of the image are darker than in the first dataset. After applying the strategy to compensate the non-uniform lighting, using a specific per-haul pattern to maximize precision, the images of both datasets become better suited for posterior processing. The frames attain a more even appearance, with uniform light distribution, making the contained data better conditioned for the subsequent steps.

Two different sets of experiments have been conducted. In the first experiment, we aimed at evaluating the capability of the architecture to generalize the problem of fish detection by training using the 1605 images of dataset 1, and then testing on the 200 annotated images of dataset 2, in which lighting conditions and camera settings have changed.

It should be noted that the two datasets also present different characteristics in terms of the type of fish present. Saithe dominated in the first cruise, which also included blue whiting, redfish, Atlantic mackerel, velvet belly lanternshark, and Norway pout. The second cruise included images of Atlantic mackerel, blue whiting, and Atlantic herring. In addition to these fish, the second dataset also included northern krill, *Meganyctiphanes norvegica*, in most images. Moreover, the average number of fish per image is also much larger in the second dataset.

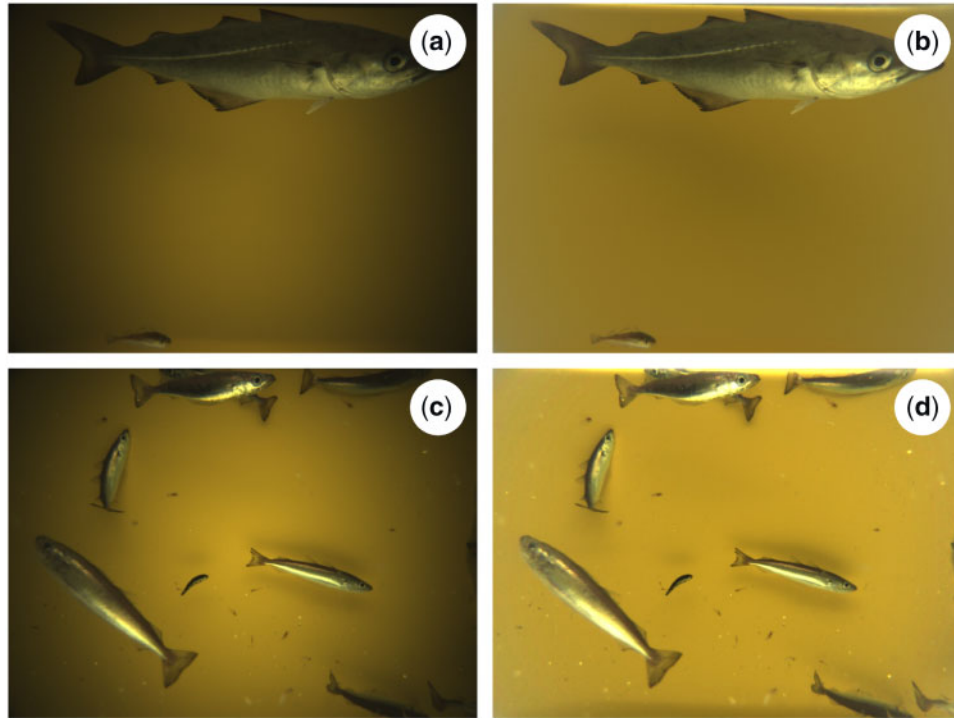
The Mask R-CNN was trained with the images of dataset 1, acquired by the R/V “Dr Fridtjof Nansen,” but applying the data augmentation techniques described above. The original dataset was split into 80% for training and 20% for validating.

After finishing this training we applied the obtained weights on 200 annotated images from the second dataset acquired by F/V “Vendla.” This dataset is completely independent from the images used for training and validation. Test images were previously segmented by hand, creating a ground-truth to compare all methods. Fifty of these images contain overlapping fish while the other 150 contain one or more fish, but with no overlap. Table 1 illustrates the results obtained in this first trial.

Analysing the values of Table 1, the reader would think that the CNN is doing a good job. We differentiate between “single fish,” which is the detection of fish when the masks corresponding to the fish are not connected to each other (see Figure 3a), and “overlapping fish,” which corresponds to the cases in which these masks overlap (see Figure 3b, central fish). In Table 1, IoU is ranging between 0.84 for “single fish” detection, and 0.82 for “overlapping fish.” And the accuracy is even higher with values of  $>0.98$  in both cases. Therefore, at first glance, the Mask R-CNN architecture seems to have done a good job to generalize the problem of fish detection.

It should be noted, however, that in our case we want to segment every isolated fish to enable its later sizing. In the case of overlapping fish (see Figure 3b), applying IoU out of the box would only take into account if a pixel that was predicted as class “fish” belongs to a fish in the ground-truth. However, this is not what we need in our application. Consider the example of Figure 5. The ideal ground-truth masks are shown in Figure 5a, with the red fish labelled as 1 and the blue fish with label 2. Figure 5c shows a fish segmentation in which the two overlapping fish are detected as a single fish. This would be considered as a very good segmentation in the standard IoU metric frequently used in the literature, e.g. (He *et al.*, 2017), but in our case we consider this a bad result since it is missing the detection of fish 2, and over-segmenting fish 1. Therefore, we introduce a new metric, namely IoU\*, to measure IoU on a slightly different way that better serves our purposes. This measurement of IoU\* will work as follows. An IoU\* measurement will be computed for every fish in the ground-truth. The IoU\* corresponding to the red fish as the area of intersection between the red region in Figure 5a and the red area in Figure 5c, and that value will be divided by the union of the same two regions. In this way, the detection of fish 1 will have a low IoU, as we will divide by a large area of union. Equally, for fish 2 we will divide the area of intersection by the total area of union of Figure 5b plus the blue area of Figure 5a, also producing a low IoU\* value since it will have also a large number in the denominator. Using this metric, large values of IoU\* guarantee that only one fish has been detected, while low values indicate that two or more overlapping fish in the ground-truth have been predicted as a single fish in the detection phase. Experimentally, this threshold has been set as 0.7.

The results of this new metric are given in Table 2. Again, we distinguish between the previous two cases depending on whether fish are overlapping to have a better insight of the performance of the system under this critical situation. In the first two columns the table details the number of images of the second dataset, and the total number of fish manually annotated in those images. The third column states how many of these fish are detected with an IoU\* with a value of  $>0.7$ , which intuitively means that the detection is good, i.e. two fish in the ground-truth are detected as two fish in the trial, and not as a single, larger fish. For the case of single fish (non-overlapping) we observe that 334 fish are correctly detected out of the 368 fish in the ground-truth. This is really a good performance if we take into account that several of the fish manually annotated in the datasets correspond to partially visible fish that are entering or leaving the field of view of the camera. However, for the images in which fish are overlapping, only 154 out of 272 fish are detected with an IoU\*  $>0.7$ . And 94 fish are detected with IoU\*  $<0.7$ , i.e. one fish is detected when  $>1$  fish appeared in the ground-truth. It can be observed that, as opposed to what it seemed in Table 1 using the standard IoU metric, the performance of Mask R-CNN in this first trial is not so great, especially in the case of overlapping fish. The next two columns present the number of false negatives, i.e. fish not detected at all, and false positive. In this dataset the false positives normally correspond to the prediction of fish in areas of the image that correspond to northern krill, present in all the images of sequence 2. Finally, the last column corresponds to the average IoU\* measurement, giving a value of 0.76 for the single fish case, and 0.58 in the case of overlapping fish. It should be noted that this average is computed from all the IoU\* values of all the



**Figure 4.** Correction of non-uniform illumination in dataset 1 (top) and dataset 2 (bottom). (a) Image from the Dr Fridtjof Nansen March 2017 dataset. (b) Image after non-uniform illumination compensation. (c) Image corresponding to the Vendla May 2017 dataset. Note the different appearance of the image with respect to (a). The centre of the image is brighter, while the boundary areas are still significantly dark. (d) Image after non-uniform illumination compensation.

**Table 1.** Results obtained by Mask R-CNN.

	IoU	Accuracy
Single fish	0.845	0.994
Overlapping fish	0.824	0.984

The network was trained using dataset 1, and the test has been quantified using the images of dataset 2. The results suggest a very good generalization capability of the network for detecting fish.

images in the corresponding dataset. We average all IoU\* values for every fish in the ground-truth, but we also accumulate and account for 0 if FN or FP occur in the test images. Therefore, our average IoU\* metric strongly penalizes false detections.

The last two rows of Table 2 detail the results of taking the fish detection masks obtained in this first trial by Mask R-CNN and applying the gradient mask refinement to them. We notice that gradient refinement is not able to improve fish detection, although it raises IoU\* to 0.80 and 0.61, respectively. This basically means that the segmentation mask is more accurate after gradient refinement.

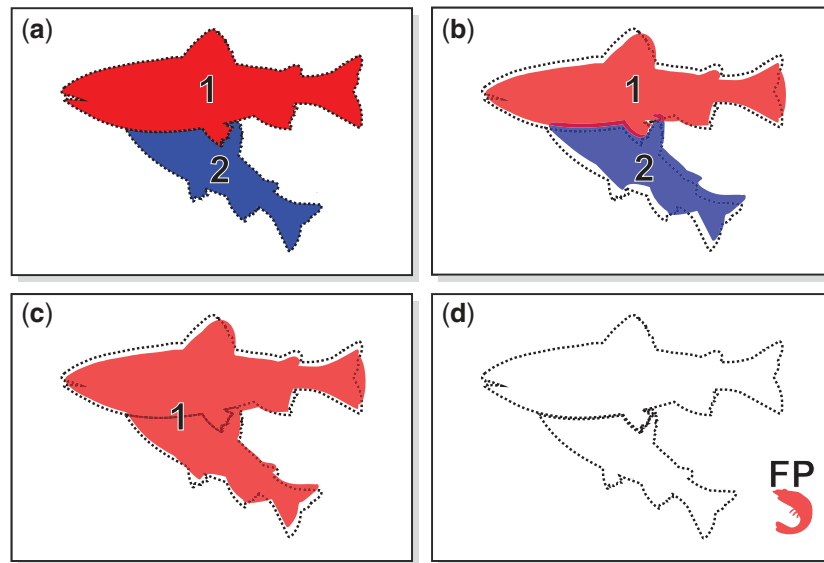
Table 3 reports the results of the second experiment. In this case, both datasets were used to create the train, validation, and test sets. Out of the total number of images (1805), roughly a 10% is used to evaluate the final model fit on the training dataset (test set), and the remaining 90% of the images were further divided into 80% for training and 20% for validation to tune the hyperparameters of the Mask R-CNN. Again, to better understand the performance of the network, we divided the test set images between (a) single fish and (b) overlapping fish situations.

For the single fish scenario, as expected, we see that the performance of the detection is better than in the first experiment, since the training data includes images of datasets 1 and 2. More than 96% of the fish are correctly detected when there is no overlapping fish, i.e. 225 correct detections from 233 annotated fish. This percentage goes down to roughly 79% when fish are occluded by other fish. These results with overlapping fish drastically improve the results of experiment 1, with 57% of correct detections of overlapping fish. It can also be observed that the number of FN and FP has also been drastically reduced with respect to the previous trial. Finally, the last column of Table 3 includes IoU\* average values of 0.89 and 0.79 for non-overlapping and overlapping fish, respectively. These values are slightly improved by the gradient refinement technique, on 0.01 in every case. This is a sign that the masks generated by Mask R-CNN in the second experiment are more accurate than the ones predicted in the first trial, but can still be improved through gradient refinement. Some sample results of the second experiment can be shown in Figures 6 and 7.

Figure 7 shows intermediate qualitative results of the proposed pipeline. It can be observed how the individual fish segmentation algorithm provides a much better fish delineation with respect to the labelled image provided by Mask R-CNN.

## Discussion and conclusions

Fish length estimation and catch composition are among the most crucial information collected in fisheries research. The Deep Vision system allows fishing vessels to collect stereo imagery, and proper processing of these data enables gaining critical information about average fish size and catch composition during the trawling operation.



**Figure 5.** Fish masks. (a) Ground-truth hand annotation. (b) Example of masks detected by the CNN. The dashed lines show the corresponding ground-truth. The coloured area outside the dashed region corresponds to a false positive area, the white area inside the dashed region defines a false negative. (c) Example of an incorrect segmentation in which the CNN detects as a single instance the two fish of (a). (d) False detection of a non-existent fish, giving rise to another false positive.

**Table 2.** Experiment 1: results obtained by Mask R-CNN after training with dataset 1 (D#1) and testing with dataset 2 (D#2).

		No. of images	Total no. of annotated fish	No. of detected fish with $\text{IoU}^* > 0.7$	No. of detected fish with $\text{IoU}^* < 0.7$	FN	FP	$\text{IoU}^*$
Mask R-CNN train and valid. on D#1 + test on D#2	Single fish	150	368	334	15	19	25	0.76
	Overlapping fish	50	272	154	94	24	16	0.58
Gradient refinement	Single fish	150	368	333	16	19	24	0.80
	Overlapping fish	50	272	156	95	21	15	0.61

Performance taking into account the new metric  $\text{IoU}^*$  that penalizes detection of a single fish when two or more fish instances are labelled in the ground-truth.

Several works in the literature have tried to segment fish in underwater video sequences. Some achieve fish detection based on matrix decomposition (Qin *et al.*, 2014) or exploiting texture and shape features that characterize fish with respect to the background (Spampinato *et al.*, 2010). Other works rely on salient features (Fernandes *et al.*, 2016), carefully selected double thresholds (Chuang *et al.*, 2016), or the guided filter (Sanchez-Torres *et al.*, 2018). In many cases, the approach involves a static camera that allows modelling the background to then isolate the fish to carry out monocular detection or stereo measurements (Costa *et al.*, 2006; Pérez *et al.*, 2018), while other works train-specific Deep Learning architectures for fish classification (Qin *et al.*, 2016). However, in all cases the detected fish were not overlapping with other fish in the field of view of the camera. Proper delineation of individual fish in overlapping situations still remains a challenge.

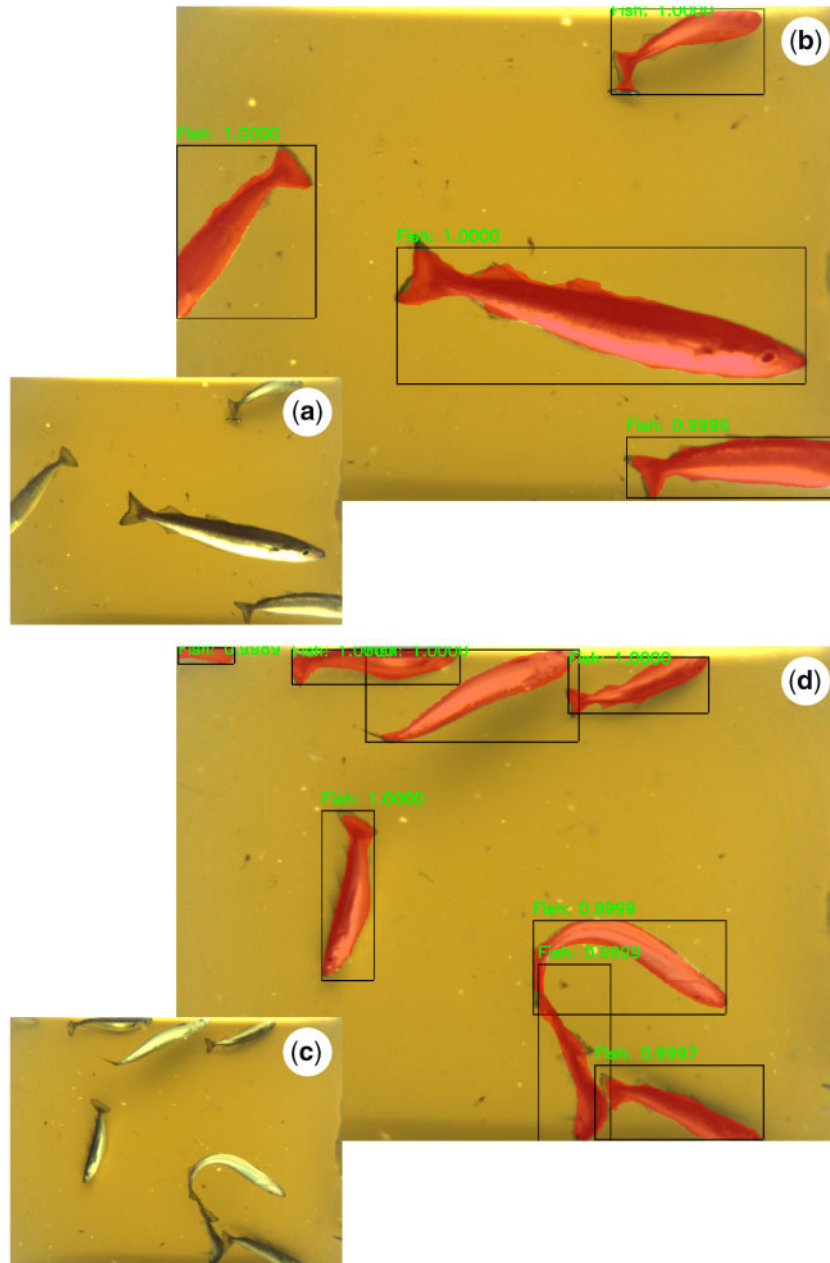
Stereo imaging is often employed to obtain depth information, and depth cues can be used to segment RoI in some well-conditioned situations. However, traditional stereo matching techniques such as *Semi Global Matching* (Hirschmuller, 2005) or *Block Matching* (Konolige, 1998) fail to reliably detecting the fish boundaries in cluttered situations, as depicted in Figure 8. Depth cues from stereo alone can potentially be used to separate fish standing at clearly different distances, such as in the case of

Figure 8a and b. On the contrary, we find in our datasets many cases in which multiple fish stand at approximately the same distance while overlapping, or are imaged while being significantly rotated from the ideal fronto-parallel configuration (such as in Figure 8e and f). In these situations, stereo matching fails to provide enough information to successfully and robustly separate the fish (Figure 8g and h). Figure 9 illustrates the result of our approach for this particular complicated case. While the result is not perfect in Figure 9b, it can nonetheless be considered as a successful detection and separation.

The processing pipeline proposed in this paper is able to provide accurate segmentations of individual fish in images acquired during standard fisheries surveys using the Deep Vision commercially available system. The pipeline involves three main phases: pre-processing, CNN-based segmentation, and gradient refining. Each phase contributes decisively to the performance of the overall system.

*Pre-processing* aims at exploiting the fact the imaging acquisition setup is well defined and constrained in terms of optical sensors, illumination characteristics, and background. By performing adequate modelling of the camera response and background illumination field, the variability of the visual appearance is reduced across different datasets and surveys. This, in turn, promotes the performance of the CNN, and, to





**Figure 6.** Fish detection and semantic segmentation performed by Mask R-CNN. (a) and (c) correspond to the original images. (b) and (d) illustrate the outcome of the algorithm. Note how Mask R-CNN is also able to detect overlapping fish, as shown in (d).

a lesser extent, also benefits the gradient refinement step at the end.

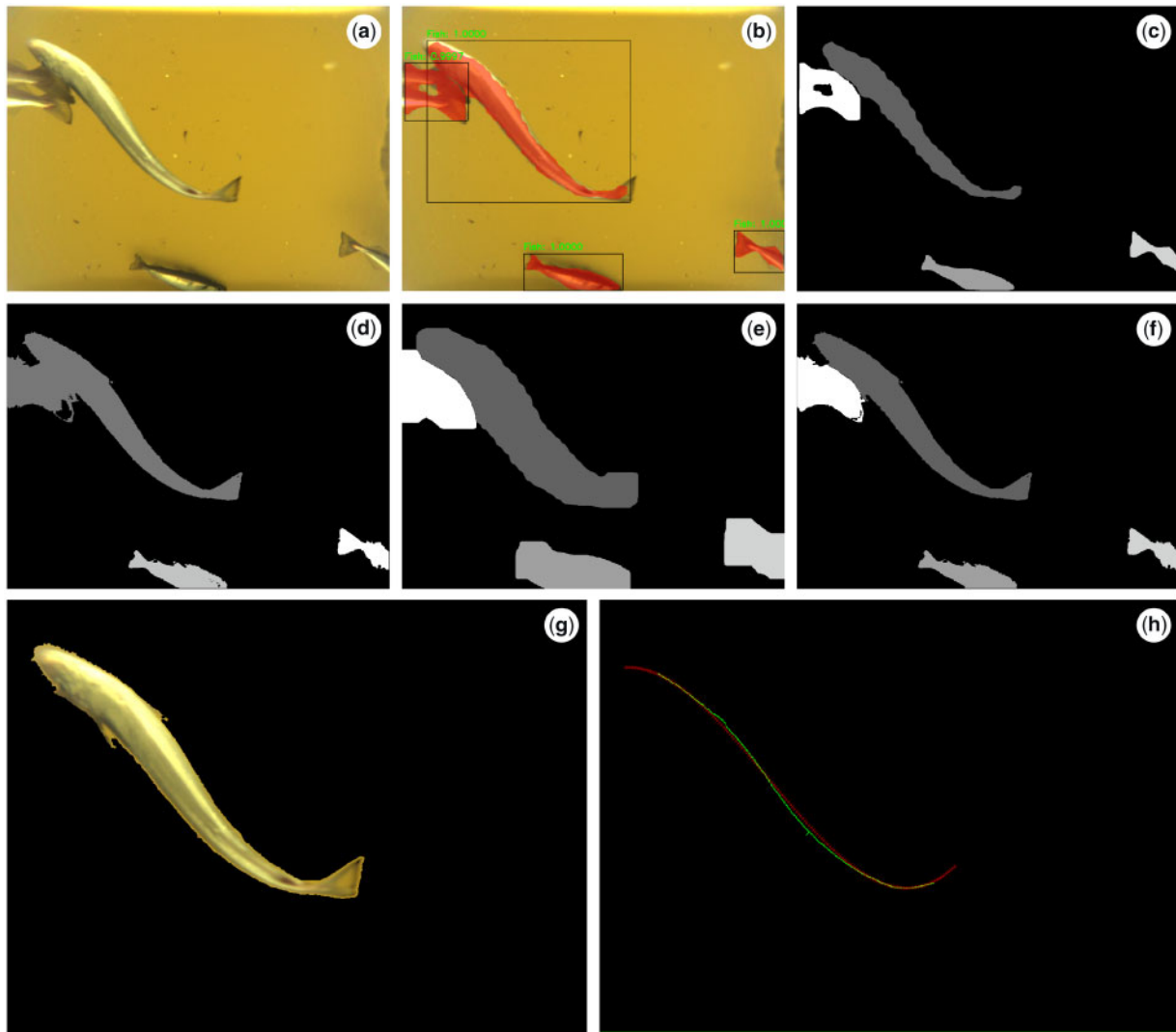
The Mask R-CNN architecture was selected for the *CNN-based segmentation*. A central reason behind this choice was its superior performance reported by He *et al.* (2017), when compared to closely related instance-aware alternatives such as Multi-task Network Cascades (Dai *et al.*, 2016) and Fully Convolutional Semantic Segmentation (Li *et al.*, 2017).

Finally, the gradient refining phase improves the delineation of the fish by using local contour cues. The impact of this step is clearly visible on Tables 2 and 3 regarding the IoU\* measurement, where there was a noticeable improvement. The

improved delineation is also of clear benefit for fish sizing accuracy.

In this study, we have also proved that standard IoU values are not adequate to quantify the performance of segmentation of individual fish in the overlapping situations in which specimens are occluded by other fish. A modification of the previous metric has been proposed (IoU\*) as a statistic that can effectively be used for gauging the similarity of the detected masks with respect to the hand-labelled ground-truth masks.

The approach in this paper has been developed with the operational goal of achieving real-time execution on dedicated hardware inside the Deep Vision imaging system. The testing



**Figure 7.** Automatic fish detection and length estimation. (a) Original image. (b) Fish detection and semantic segmentation through the Mask R-CNN processing. Note that the system is able to correctly detect the central fish, although it fails to detect the two tails on the left as two separate fish. (c) Labelled image as provided by Mask R-CNN. (d) Fish boundary gradient refinement mask. Note that, in this case, the segmentation is not able to distinguish among touching fish. (e) Multi-label dilate morphological operation of the Mask R-CNN segmentation. (f) Fish mask resulting of the combination of both gradient refinement and multi-label dilate. (g) Final segmented fish. (h) Skeleton pixels (in green) of the segmented fish and measurement points (in red) of the estimated fish-shape curve used to perform an automatic size measurement.

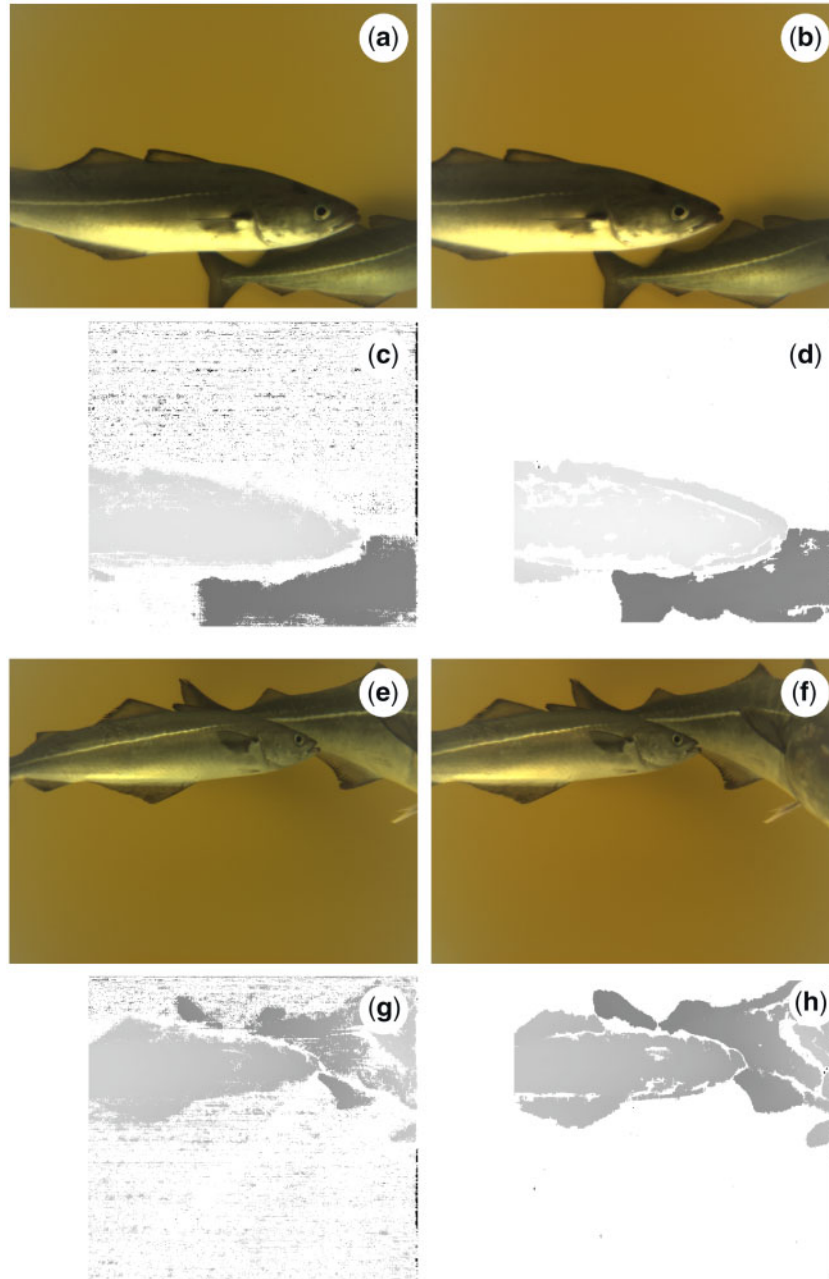
**Table 3.** Experiment 2: results obtained by Mask R-CNN after training with randomly selected 90% images from dataset 1 (D#1) and dataset 2 (D#2), the other 10% is reserved for testing.

		No. of images	Total no. of annotated fish	No. of detected fish with $IoU^* > 0.7$	No. of detected fish with $IoU^* < 0.7$	FN	FP	$IoU^*$
Mask R-CNN train and valid. on 90% (D#1 + D#2), test in 10% (D#1 + D#2)	Single fish	170	233	225	7	1	10	0.89
	Overlapping fish	26	104	82	16	6	5	0.79
Gradient refinement	Single fish	170	233	224	8	1	10	0.90
	Overlapping fish	26	104	84	14	6	4	0.80

Performance taking into account the new metric  $IoU^*$  that penalizes detection of a single fish when two or more fish instances are labelled in the ground-truth.

reported in this paper was conducted offline on a high-end desktop computer with a NVIDIA TITAN V GPU. The segmentation was run on the GPU at a frame rate was 2.67 images per second. The refinement in the current state is not optimized for speed.

A number of extensions to this work is planned in the near future. The validation of the size measurements is currently being pursuit with the intent of using fish specimens or accurate fish shape reproductions of known dimensions. The testing is to be conducted in water, to take into account the



**Figure 8.** Traditional stereo matching techniques fail to segment overlapping fish due to lack of sufficient salient features and visual texture. The (a, b) and (e, f) images correspond to a pair of stereo images presenting fish that are partially occluded by other fish. (c, d) and (g, h) show the resulting disparity maps using two standard stereo processing techniques: (left) Semi Global Matching (Hirschmuller, 2005) and (right) Block Matching (Konolige, 1998).

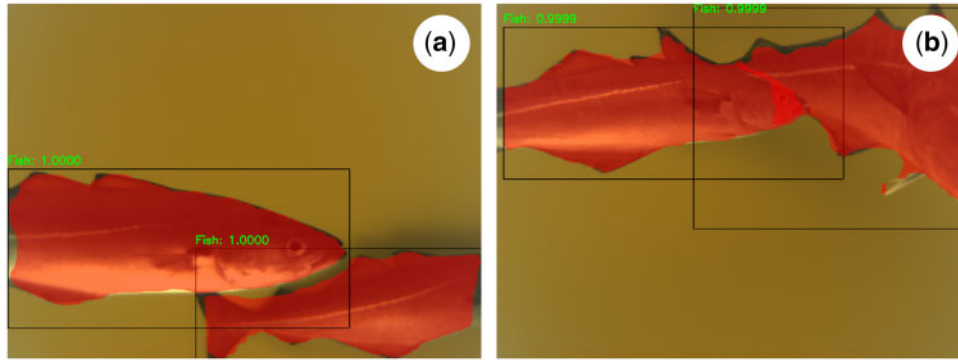
refraction effects of the flat-port camera housing and how it affects the stereo geometry.

A second extension is directed towards achieving an execution frame rate in the order of 10 fps, on the target embedded processing hardware. This hardware is based on NVIDIA Jetson AGX Xavier modules and will be deployed with Deep Vision imaging system. The intended frame rate will allow performing tracking of fish across time, given that multiple instances of the same fish are likely to occur when images are acquired at 10 fps or higher, for nominal trawling speeds. This will enable the ability of estimating in real time the amount of fish in the trawl as well as the average

size. Finally, as more data becomes annotated, future development will extend this work to use Mask R-CNN for automatic fish species identification.

### Funding

Development of Deep Vision technology has been supported through the Research Council of Norway's Industrial PhD Programme and Innovation Norway's program for development of environmental technology (project 100424). R. Garcia and N. Gracías were partly funded by the Spanish Ministry of Education, Culture, and Sport under project CTM2017-83075-R. Data



**Figure 9.** (a) Result of Mask R-CNN for the images of Figure 8b. (b) Instance segmentation of the two fish of Figure 8f: note the small error in the detection of the fish on the right. Although far from achieving ideal results, Mask R-CNN outperforms the state-of-the-art stereo processing techniques of Figure 8, even when the fish are not completely visible.

collection onboard R/V “Dr Fridtjof Nansen” was supported by the Institute of Marine Research under the CRISP centre for research innovation (Research Council of Norway project 203477) and vessel time onboard M/S “Vendla” was provided by the REDUS project with funding from the Norwegian Ministry of Trade, Industry, and Fisheries. The authors would like to thank Roger Portas for his assistance with this project.

## References

- Allken, V., Olav, N., 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.
- Berges, B., Sakinan, S., and van Helmond, E. 2018. Practical Implementation of Real-time Fish Classification from Acoustic Broadband Echo Sounder Data—RealFishEcho Progress Report. Wageningen Marine Research (University & Research Centre), Wageningen. Wageningen Marine Research Report, C062/18. 42 pp.
- Burton, G. J. 1973. Evidence for non-linear response processes in the human visual system from measurements on the thresholds of spatial beat frequencies. *Vision Research*, 13: 1211–1225.
- Chuang, M., Hwang, J., and Williams, K. 2016. Automatic fish segmentation and recognition for trawl-based cameras. *In Computer Vision and Pattern Recognition in Environmental Informatics*, pp. 79–106. Ed. by J. Zhou, X. Bai, and T. Caelli. IGI Global, Hershey, PA.
- Costa, C., Loy, A., Cataudella, S., Davis, D., and Scardi, M. 2006. Extracting fish size using dual underwater cameras. *Agricultural Engineering*, 35: 218–227.
- Dai, J., He, K., and Sun, J. 2016. Instance-aware semantic segmentation via multi-task network cascades. *In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3150–3158. DOI: 10.1109/CVPR.2016.343.
- Debevec, P. E., and Malik, J. 1997. Recovering high dynamic range radiance maps from photographs. *In Proceedings of the 24th annual conference on Computer graphics and interactive techniques (SIGGRAPH '97)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, pp. 369–378.
- Dougherty, E. 1992. *An Introduction to Morphological Image Processing*. SPIE Optical Engineering Press. ISBN0-8194-0845-X.
- FAO. 2018. *The State of World Fisheries and Aquaculture. Meeting the Sustainable Development Goals*, Rome, Italy. <http://www.fao.org/3/i9540en/I9540EN.pdf>.
- Fernandes, P. G., Copland, G., Garcia, R., Nicosevici, T., and Scoulding, B. 2016. Additional evidence for fisheries acoustics: small cameras and angling gear provide tilt angle distributions and other relevant data for mackerel surveys. *ICES Journal of Marine Science*, 73: 8.
- Fischler, M. A., and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24: 381–395.
- Girshick, R. 2015. Fast R-CNN. *In IEEE International Conference on Computer Vision (ICCV)*, Santiago, 2015, pp. 1440–1448. DOI: 10.1109/ICCV.2015.169.
- Hartley, R., and Zisserman, A. 2003. *Multiple View Geometry in Computer Vision*. 2nd edn, Cambridge University Press, New York, NY.
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. 2017. Mask R-CNN. *In IEEE International Conference on Computer Vision (ICCV)*, Venice, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322.
- Hirschmuller, H. 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 807–814. IEEE, San Diego, CA.
- Konolige, K. 1998. Small vision systems: hardware and implementation. *In Proceedings of the 8th International Symposium in Robotic Research*, Springer, London, pp. 203–212.
- Korneliussen, R. J., Heggelund, Y., Eliassen, I. K., and Johansen, G. O. 2009. Acoustic species identification of schooling fish. *ICES Journal of Marine Science*, 66: 1111–1118.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60: 84–90.
- Li, Y., Qi, H., Dai, J., Ji, X., and Wei, Y. 2017. Fully convolutional instance-aware semantic segmentation. *In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2359–2367.
- 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 European Conference on Computer Vision (ECCV)*, Springer International Publishing, pp. 740–755. DOI: 10.1007/978-3-319-10602-1.
- Long, J., Shelhamer, E., and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 3431–3440. DOI: 10.1109/CVPR.2015.7298965.
- Pérez, D., Ferrero, F. J., Alvarez, I., Valledor, M., and Campo, J. C. 2018. Automatic measurement of fish size using stereo vision. *In IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pp. 1–6.
- Pérez Roda, M. A. (ed.), Gilman, E., Huntington, T., Kennelly, S. J., Suuronen, P., Chaloupka, M., and Medley, P. 2019. A third

- assessment of global marine fisheries discards. FAO Fisheries and Aquaculture Technical Paper, 633. FAO, Rome. 78 pp.
- Pobitzer, A., Ona, E., Macaulay, G., Korneliussen, R., Totland, A., Heggelund, Y., and Eliassen, I. K. 2015. Pre-catch sizing of herring and mackerel using broadband acoustics. *In* ICES Symposium on “Marine Ecosystem Acoustics (Some Acoustics)—Observing the Ocean Interior in Support of Integrated Management”, pp. 25–28. Nantes, France.
- Prados, R., Garcia, R., Gracias, N., Neumann, L., and Vågstøl, H. 2017. Real-time Fish Detection in Trawl Nets. *In* Proc. of the MTS/IEEE OCEANS 2017 Conference, Aberdeen, UK, pp. 1–5.
- Prados, R., Garcia, R., and Neumann, L. 2014. Image Blending Techniques and Their Application in Underwater Mosaicing, Springer, Heidelberg. ISBN: 978-3-319-05557-2.
- Qin, H., Li, X., Liang, J., Peng, Y., and Zhang, C. 2016. DeepFish: accurate underwater live fish recognition with a deep architecture. *Neurocomputing*, 187: 49–58.
- Qin, H., Peng, Y., and Li, X. 2014. Foreground extraction of underwater videos via sparse and low-rank matrix decomposition. *In* ICPR Workshop on Computer Vision for Analysis of Underwater Imagery, Stockholm, 2014, pp. 65–72. DOI: 10.1109/CVAUI.2014.16.
- Ren, S., He, K., Girshick, R., and Sun, J. F. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39: 1137–1149.
- Rosen, S., and Holst, J. C. 2013. DeepVision in-trawl imaging: sampling the water column in four dimensions. *Fisheries Research*, 148: 64–73.
- Sanchez-Torres, G., Ceballos-Arroyo, A., and Robles-Serrano, S. 2018. Automatic measurement of fish weight and size by processing underwater hatchery images. *Engineering Letters*, 26: 461–472.
- Schwarz, M. W., Cowan, W. B., and Beatty, J. C. 1987. An experimental comparison of RGB, YIQ, LAB, HSV, and opponent color models. *ACM Transactions on Graphics*, 6: 123–158.
- Spampinato, C., Giordano, D., Di Salvo, R., Chen-Burger, Y.-H. J., Fisher, R. B., and Nadarajan, G. 2010. Automatic fish classification for underwater species behavior understanding. *In* Proceedings of the First ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams, ACM, Firenze, Italy, pp. 45–50.

*Handling editor: Cigdem Beyan*