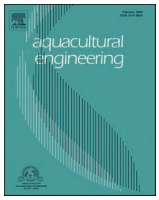
Aquacultural Engineering 99 (2022) 102299

Contents lists available at ScienceDirect

Aquacultural Engineering

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



Impact evaluation of deep learning on image segmentation for automatic bluefin tuna sizing

P. Muñoz-Benavent [a,*], J. Martínez-Peiró [a], G. Andreu-García [a], V. Puig-Pons [b], V. Espinosa [b], I. Pérez-Arjona [b], F. De la Gándara [c], A. Ortega [c]

[a] Institute of Control Systems and Industrial Computing (AI2). Universitat Politècnica de València (UPV), 46022 València, Spain
[b] Institut d'Investigació per a la Gestió Integrada de Zones Costaneres (IGIC). Universitat Politècnica de València (UPV), 46730, Gandia, Spain
[c] Instituto Español de Oceanografía (IEO-CSIC), 30740, San Pedro del Pinatar, Murcia, Spain

ARTICLE INFO

Keywords:
Underwater stereo vision
Computer vision
Fishery management
Automatic fish sizing
Biomass estimation
Convolutional neural networks

ABSTRACT

This paper evaluates the impact of using deep learning techniques in an automatic fish sizing process. Automatic fish sizing with a non-invasive approach involves working with different views of the fish's body and changing environments, being the stage of extraction of individuals in the image and the quality of the segmentation essential to obtain good sizing measurements. The goal of this work is to improve the results and functionality achieved in our previous studies with conventional segmentation methods based on local thresholding, where different limitations were observed, mainly the necessity of parameters tuning and a high computational cost. The number of detections must also increase significantly to increase the reliability of the statistical results. An approach using convolutional neural networks is proposed for fish detection and segmentation in videos acquired under real conditions, which eliminates the engineering procedure of parameter adjustment and generalises the solution for fish segmentation to deal with different environmental conditions (illumination and water turbidity) and background variability. The results show that the fish sizing procedure is enhanced thanks to the improvement in fish image instance segmentation. In particular, the number of fish measurements increases by up to 2.45 times when using Mask R-CNN and the PointRend module, thus increasing the accuracy of the fish length estimation, and the number of measurements per minute of computing time increases by up to 3.5 times. Our proposal obtains highly accurate fish length estimations in juvenile bluefin tuna based on a stereoscopic vision system and a deformable model of the fish's silhouette, both from the ventral and dorsal perspectives. An important improvement is achieved by applying CNN, as demonstrated by the number of segmented instances, the time required to segment an instance, and the accuracy of the fish sizing achieved.

1. Introduction

The huge size of the oceans and seas make observing and monitoring the marine environment a titanic task. Even so, many countries are proposing sustainability policies and doing work related to specific ecosystems that require great human and technological effort. Recently, fish farmers, ecologists and governments have also expressed an urgent need to accurately estimate the biomass of both schools and individual fish in their natural environment (Føre et al., 2018; Zhang et al., 2020). To do so, collecting a large amount of accurate data on size or age without the need to physically handle live fish has been identified as an essential requirement. Indeed, traditional methods based on manual measurements are invasive, expensive and stressful for animals, which entails a limitation on the amount of data collected, reducing the validation of the tests performed. Nevertheless, a quantitative estimation of fish biomass forms the basis of scientific fishery management and conservation strategies (Saberioon and Císař, 2018). Indeed, biomass estimations provide an important input to design adequate models to assess fisheries, explore growth stages, define growth models and evaluate the state of health of the fish. Biologists could also define growth models for different species of fish, but periodic systematic monitoring is required (Li et al., 2021). Fortunately, aquaculture farms could be a good

* Corresponding author.
E-mail addresses: pmunyoz@upv.es, pmunyoz@disca.upv.es (P. Muñoz-Benavent), joama14j@upv.es (J. Martínez-Peiró), gandreu@upv.es (G. Andreu-García), vipuipon@upv.es (V. Puig-Pons), vespinos@upv.es (V. Espinosa), iparjona@upv.es (I. Pérez-Arjona), fernando.delagandara@ieo.es (F. De la Gándara), aurelio.ortega@ieo.es (A. Ortega).

https://doi.org/10.1016/j.aquaeng.2022.102299
Received 13 June 2022; Received in revised form 13 October 2022; Accepted 20 October 2022
Available online 21 October 2022

environment for such monitoring, where species such as tuna and salmon are the most commonly farmed species due to their market acceptance and rapid growth (Shortis, 2015). In fact, Bluefin tuna (BFT) is one of the most in-demand fish species in the world. Currently, the need to strengthen management of it has increased research on BFT aquaculture production in many countries (Abe et al., 2021).

Computer vision is a non-invasive technique and stereoscopic systems (two cameras in a side-by-side arrangement providing depth information) have been considered a good tool to obtain biometric measurements of individuals, which has attracted the interest of researchers. In this sense, some specific applications have been reported: fish sizing (Álvarez-Ellacuría et al., 2020; Fernandes et al., 2020; Voskakis et al., 2021; Williams and Lauffenburger, 2016); fish counting and sizing (Costa et al., 2009; Rosen et al., 2013); and fish sizing in combination with acoustic techniques (Espinosa et al., 2011; Kloser et al., 2011). Although such contributions are significant, many challenges still exist and their authors combine different technologies and methods with the aim of providing a semi-automatic or fully automatic tool that allows them to estimate fish sizing. Working with stereoscopic video requires the joint processing of two frames for each instant of time acquired. In addition, the fish, our subject to be detected and measured, are swimming freely, which means it is not known when they are in front of the acquisition system. Furthermore, it is obviously unknown if it is acquired in the best pose to estimate their size, either. Therefore, long-term continuous monitoring of fish is necessary to obtain an adequate number of frames to take measurements. The above aspects highlight the need to develop fully automatic processing tools for underwater video processing.

An arduous sequence of steps and essential procedures such as camera calibration, acquisition of hours of video, video pre-processing, object detection, instance segmentation, stereoscopic correspondence in pairs of images and 3D triangulation must usually be performed to estimate individual measurements (Muñoz-Benavent et al., 2018; Puig-Pons et al., 2019). However, the detection of individuals, which is directly related to the segmentation of instances in the image, is one of the most relevant steps to provide accurate biometric estimates and an important factor in developing automated systems. Traditional approaches are based on searching for good hand-designed features and sophisticated descriptors. The resulting feature set is subject to a selection process that allows redundant and non-useful information to be eliminated (Guyon and Elisseeff, 2003), and then a learning or identification algorithm is used in these feature spaces. Developments in this field include local descriptors such as HOG (Histogram of Oriented Gradients) and SIFT (Scale Invariant Characteristics Transformation), which are later grouped with approximations such as bag-of-visual-words and the Fisher Vector to obtain the identification of objects (Sánchez et al., 2013). In general, feature engineering is a complex and tedious process that must be reviewed each time the problem or the associated image dataset changes. Applications based on these techniques have great limitations in adapting to change and are very much geared towards solving a narrow problem with a strict image dataset.

An unrestricted natural environment makes the segmentation step a very difficult task. Processing underwater images throws up numerous challenges, and classic image processing methods are adversely affected by them. Underwater scene frames often have complex backgrounds, changes in illuminance, limited visibility, low resolution and low contrast caused by attenuated light and turbidity. Furthermore, the fish swim freely, causing occlusions, overlapping and even distortions. Fish segmentation underwater requires the correct detection of all the objects in an image, as well as accurate segmentation of each instance (He et al., 2020). Instance segmentation combines object detection (where the goal is to classify individual objects and localise each one using a bounding box) and semantic segmentation (where the goal is to classify each pixel into a fixed set of categories without differentiating object instances). Hence, instance segmentation may be defined as the technique of simultaneously solving the problem of object detection as well as that of semantic segmentation (Hafiz and Bhat, 2020).

The need to process a large volume of data with the requirement of getting fully automatic tools is a motivation to work using methods based on deep learning (DL). Of course, another motivation is the considerable progress shown in computer vision tasks related to identification and recognition in different environments using DL (Krizhevsky et al., 2017). Moreover, the variability and challenging conditions of underwater images suggest the need to work with tools equipped with a great capacity for adaptation and learning. DL methods are composed of multiple layers to learn features of data with multiple levels of abstraction (Lecun et al., 2015). These models can learn visual fish features, are insensitive to environmental changes and variations, and could be used to extract fish from images collected in an unconstrained underwater environment. One of the most important advantages of using DL is the removal of the need for feature engineering, because these models find the significant features by themselves through training. However, the extremely high training times and the need for large datasets with thousands of labelled images representative of all the classes and variations to identify are two of the main disadvantages of these methods. Nevertheless, the combination of current technologies makes it easier to work with DL: the high performance of parallel computing with GPUs accelerates the training process, whereas transfer learning saves a significant amount of labelling effort, while data augmentation techniques artificially enlarge the number of training images (Liu et al., 2019). Convolutional neural networks (CNNs) play an important role in DL. Classic CNNs are composed of two main parts: the first one includes the convolutional process and the max pooling operation, whereas in the second part a fully connected layer takes the input from the output of the previous layers and performs the classification task. CNNs use automatic feature extractors hierarchically to map the value of the visual features into the vector space.

In this study, the most up-to-date CNN proposals—Faster R-CNN (Ren et al., 2017), YOLO in its v5 version (Jocher et al., 2020), Mask R-CNN (He et al., 2020) and the PointRend module (Kirillov et al., 2019)—have been applied for fish detection and segmentation in videos acquired under real conditions. The results in Section 3 show that the fish sizing procedure is enhanced thanks to the improvement in fish detection and segmentation. The main contributions of this work include the following:

(1) Generation of a new ground truth dataset containing 1000 labelled stereoscopic images of BFT acquired in real conditions.
(2) Application of different CNN proposals in a challenging underwater environment for fish detection and segmentation using data augmentation and transfer learning.
(3) Evaluation of the impact of the improved fish segmentation in the fish sizing procedure, both from ventral and dorsal perspectives, and the variability associated with the measurements.

2. Materials and methods

This study focuses on improving the segmentation of fish instances by using DL techniques, in particular CNNs, to increase the number and accuracy of fish measurements. The proposed DL techniques for image segmentation attempt to substitute the image segmentation and blob-filtering steps used in our previous research (Muñoz-Benavent et al., 2018) with the aim of obtaining an efficient procedure able to adapt to different environmental conditions automatically. Fig. 1 summarizes the processing algorithms involved in the procedure of BFT sizing and the place where the new procedure could be carried out.

2.1. Data acquisition

The recordings were taken at the Infrastructure for Atlantic Bluefin Tuna Aquaculture (ICAR), belonging to the IEO (Spanish Oceanographic

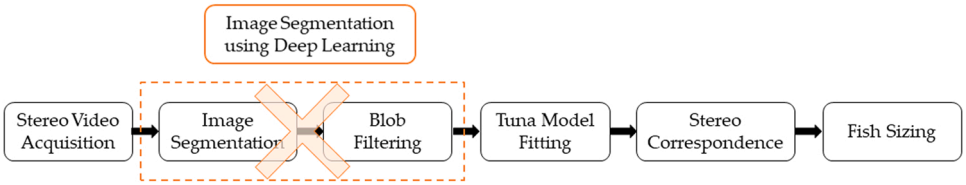


Fig. 1. Sequence of stereo video processing algorithms involved in the process of Bluefin Tuna sizing. The proposed procedure for image segmentation using deep learning replaces the previous image segmentation and blob-filtering steps.

Institute). The ICAR is a unique scientific and technical infrastructure (ICTS) devoted to studying the complete aquaculture of BFT. The equipment used to record the fish is shown in Fig. 2.

A sensor platform (Fig. 2a) was placed in different months in a tank with sea water measuring 20 m in diameter, 10 m in depth and 3500 m^3 in volume, and containing 77 BFT juveniles ranging from 40 to 140 cm in Snout Fork Length (SFL). The platform was equipped with sensors including a stereoscopic camera comprised of two Gigabit Ethernet cameras with a 2048 × 1536-pixel resolution and framerate of 35 fps. The cameras were mounted in an underwater housing, with a baseline of 85 cm and inward convergence of 5°. Camera synchronisation was achieved using the IEEE 1588 Precision Time Protocol (PTP). The system is rated for a depth of 40 m and has an umbilical cable that supplies Power over Ethernet to the cameras and transfers images to a logging computer (Fig. 2b), which encodes left and right videos using GPU encoding. The stereoscopic system was previously calibrated using a check pattern and the MATLAB® Stereo Calibration Application.

The camera was placed in the tank to record the fish from two different perspectives, dorsal and ventral: firstly, floating on the water surface and looking towards the floor of the tank to have a dorsal perspective of the fish; and secondly, lying on the bottom of the tank and looking towards the water surface in order to have a ventral perspective of them. This work is part of an ambitious project whose main goal is to monitor the tuna population and estimate growth models, but the paper itself focuses on analysing the influence of the image background on the segmentation procedure and hence on the number of measurements and accuracy of the fish sizing. For this purpose, four videos of 10 min each (a total of 40 min) were selected, each of them with different image backgrounds. In the case of the dorsal perspective, the floor of the tank constitutes the image background, but its original colour uniformity was altered due to an accumulation of sediments. To overcome this problem, different approaches were tested. In the first video, artificial back panels (background #1, Fig. 3a) were placed on the floor of the tank, but they were discarded since they altered the fish's behaviour, which stopped eating and became stressed, possibly increasing their mortality. In the second video, the floor was cleaned slightly with a cleaning robot (background #2, Fig. 3b) and then cleaned deeply with a diver for the third video (background #3, Fig. 3c). In the case of the ventral perspective (fourth video), white artificial back panels (background #4, Fig. 3d) were added, since the building's ceiling had poor illumination and the fish were almost impossible to distinguish from the ceiling. Table 1 summarizes the recordings, the month of acquisition and the description of the different image backgrounds.

2.2. Image segmentation using deep learning

Fish are not clearly visible in underwater images due to low contrast, light scattering and high noise in the environment. In general, it is difficult to segment objects in underwater images without losing their details. In a previous study (Muñoz-Benavent et al., 2018), where it was necessary to segment tuna in images while they swam freely, the image segmentation was accomplished using local thresholding, a region-based technique for extracting compact regions (blobs) in each video frame, followed by morphological operations. The segmented blobs were then geometrically characterised and filtered using shape (aspect ratio), pixel density and dimensional filters. Although these conventional image processing techniques worked well in the reported experiments, one main limitation was observed: the accuracy of the resulting segmentation is strongly dependent on the block size and local threshold parameters, which had to be adapted to the different environmental conditions (image backgrounds, illumination, water turbidity and others). The variability and adverse conditions of underwater images suggested the need to work with tools equipped with a great capacity for adaptation and learning. DL models have such an ability to learn the characteristics of objects, which could provide robustness to segment these objects, even in images where changes and variations appear in their backgrounds (Yang et al., 2021). These models must be trained with a large number of images that represent all the possible variabilities of the problem to be solved. Thus, in this study we developed and evaluated two approaches based on DL to segment fish in underwater images.

The first approach used neural networks for object detection and a clustering algorithm for segmentation. Object detection involves detecting instances of semantic objects of certain classes (fish in our case) and drawing a bounding box around each object of interest in the image. The most up-to-date CNN proposals have been tested: Faster R-CNN (Ren et al., 2017) (with ResNet-50 as backbone), YOLO in its v5 version (Jocher et al., 2020) and Mask R-CNN (He et al., 2020) (with ResNet-50 as backbone). When the fish are framed inside their bounding boxes, their silhouettes are segmented using a k-means clustering algorithm (Arthur and Vassilvitskii, 2006), with $k = 2$ clusters, the foreground (fish) and the background. Finally, the contour definition of the

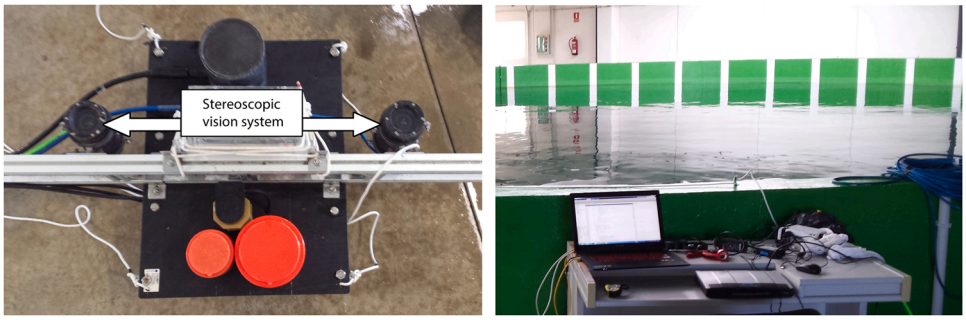


Fig. 2. (a) Sensor platform equipped with a stereoscopic camera and other sensors. (b) Logging computer used to encode stereoscopic videos.

Fig. 3. Snapshots of the videos of Bluefin Tuna juveniles with different image backgrounds. Dorsal perspective: a) artificial background panels, b) slightly cleaned floor, c) deeply cleaned floor. Ventral perspective: d) white artificial background panels.

Table 1
Dataset of recordings in ICAR tanks with a description of the different image backgrounds.

Video	View	Month of acquisition	Image background
Video #1	Dorsal view. Cameras looking towards floor of the tank.	April 2019	Background #1 (Fig. 3a): artificial back panels placed on the floor of the tank.
Video #2		May 2019	Background #2 (Fig. 3b): floor of the tank slightly cleaned.
Video #3		December 2019	Background #3 (Fig. 3c): floor of the tank deeply cleaned.
Video #4	Ventral view. Cameras looking towards water surface.	December 2019	Background #4 (Fig. 3d): artificial white back panels placed above the tank.

fish's silhouette is enhanced by applying an active contours algorithm (Chan and Vese, 2001), an iterative region-growing technique.

The second approach used neural networks for object instance segmentation, directly segmenting fish from the background. In this case, the most up-to-date CNN proposals have been tested, including Mask R-CNN (with ResNet-50 as backbone) and the PointRend module (Kirillov et al., 2019).

2.3. A brief description of the CNN models explored

Next, a brief description of each CNN model mentioned above is introduced, while the experiments and results are described in Section 3.

Region-based convolutional neural networks (R-CNN) were proposed (Girshick et al., 2013) to address the problem of object detection in images, i.e. the problem of classifying individual objects and localising each object in the image using a bounding box. With this method, region proposals are generated with the selective search algorithm, and convolutional networks are evaluated independently on each region. The same author proposed an improvement of the R-CNN called Fast R-CNN (Girshick, 2015), consisting of generating a feature map from the convolution operation applied to each image, instead of applying the convolutional network to each region proposal. In Faster R-CNN (Ren et al., 2017), instead of using a selective search algorithm on the feature map to identify the region proposals, a separate network was proposed to predict them. Finally, Mask R-CNN (He et al., 2020) extended Faster R-CNN by adding a branch for predicting segmentation masks on each region proposal, in parallel with the existing branch for classification and bounding box regression. The PointRend module, added on top of mask segmentation networks, such as Mask R-CNN, is able to detect uncertain points that require higher definition and refine the segmentation in those points by means of bilinear interpolation and a multilayer perception. This allows segmentations to be obtained with higher resolution than those obtained with Mask-RCNN. YOLO or You Only Look Once (Redmon et al., 2015) is an object detection algorithm different from the region-based algorithms. In YOLO, a single convolutional network predicts the bounding boxes and the class probabilities for these boxes. It is faster than the R-CNN, but its main limitation is that it struggles with small objects.

2.4. Evaluation metrics

The metric proposed in order to evaluate the performance of the neural networks is the Average Precision (AP) used in the well-known Pascal VOC challenge (Everingham et al., 2010) and in Microsoft COCO (Lin et al., 2014), which is defined as the area under the precision-recall curve:

$$AP = \int_0^1 p(r)dr$$

Precision (p) is defined as the number of true positives (TP) divided by the sum of true positives and false positives (FP). Recall (r) is defined as the number of true positives divided by the sum of true positives and false negatives (FN), i.e. the number of labelled ground truth objects.

$$p = \frac{TP}{TP + FP}; r = \frac{TP}{TP + FN};$$

Detections are considered positive when the Intersection over Union (IoU) is above a threshold, which is typically greater than 0.5. IoU is defined as the area of the intersection divided by the area of the union of a predicted bounding box (BboxP) and ground truth label (BboxT). A perfect match occurs when IoU = 1, whereas IoU = 0 if the bounding boxes do not intercept each other.

$$IoU = \frac{area(BboxP \cap BboxT)}{area(BboxP \cup BboxT)}; \; IoU > 0.5, \ldots, 0.75, \ldots, 0.9, 0.95;$$

AP can be calculated for different IoU thresholds, usually indicated with a subscript: AP_{50} represents the AP for an IoU accuracy greater than 0.50; AP_{75} for IoU accuracy greater than 0.75; and AP_{90} for IoU accuracy greater than 0.90. The final AP, which matches with the mAP (mean Average Precision) in one-class detections, is calculated by taking the mean AP over all IoU thresholds between 0.50 and 0.95 in increments of 0.05.

$$mAP = \frac{\sum_{IoU} AP_{IoU}}{n}; \; IoU \in [0.5, 0.95]; \Delta_{IoU} = 0.05$$

where n is the total number of IoU thresholds.

In the proposed procedure, the detection and segmentation must be as accurate as possible to then apply the fitting of the deformable model of the fish's silhouette and have an accurate sizing, so AP_{90} will be used as the evaluation metrics together with the commonly used AP.

In the case of object detection, IoU is calculated comparing the bounding boxes, whereas in object instance segmentation it is calculated comparing segmentation masks in a pixel-to-pixel manner. These methods are denoted as "bb" and "mask" in the AP's superscript, respectively. See (Padilla et al., 2021) for further details in metrics.

The proposed methods will only work when the entire fish is observed in the image, without any occlusion or overlapping. This would be a limitation in environments with a high density of fish, but that is not the case for the fish in the ICAR tanks, as can be seen in Fig. 3. Further improvements will be made to tackle the problem in other environments.

2.5. Dataset generation for deep learning

The training and test datasets are generated using samples of the four different image backgrounds. Samples are extracted from setup videos and manually labelled using VGG Image Annotator (Dutta and Zisserman, 2019). The setup videos have a duration of 2 min and were acquired prior to the 10-minute video made up of the recordings described in Table 1 and in the same scenario. The fact that samples are selected for network training in the same scenario where the networks would be evaluated may limit their applicability. If these trained networks were to be used in new scenarios, new samples should be manually labelled to increase the training dataset in order to get satisfactory and accurate results. However, in this study, the recordings are made in an indoor facility and, despite working with four different backgrounds, the variability associated with this scenario is limited, so we consider that the same trained networks would serve for future recordings without needing to retrain them.

Apart from the manual labelling, we developed a semi-automatic tool to considerably increase the number of samples without having to manually label them. The semi-automatic labelling uses a method similar to the first approach described in Section 2.2, but instead of using a neural network to detect objects, the bounding box is manually drawn by an operator. Then, the fish are automatically segmented using a k-means clustering algorithm and active contours. The operator supervises the segmentation, discarding samples if the resulting segmentation is not accurate. Thus, this semi-automatic labelling method achieves a large number of labelled samples easily and in a short time.

As a result, we have 250 labelled samples for each image background, made up of training and test datasets of 900 and 100 images respectively, as summarized in Table 2.

2.6. Stereo vision data processing for fish sizing

When the fish have been segmented in the image, an edge detection algorithm is then applied and a minimisation algorithm is used to fit a deformable model of the tuna's silhouette. A Fitting Error Index (FEI) based on the quadratic distance between the model points and target edge points is used to quantify how well the model fits. Samples with poor FEIs are discarded, since the index reveals wrong segmentations, occlusions of significant parts of the fish, poor model fittings and more. Fish are deformable due to the swimming motion and, consequently, measurements taken from a single frame may not be reliable (Shortis et al., 2013). Two main options are used in the literature to reduce the effect of swimming motion on length measurement: i) take measurements in all frames and deduce straight body length from a sinusoid-like pattern (Shortis et al., 2013); ii) account for body bending by adding contiguous linear segments (Williams and Lauffenburger, 2016). In our case, the swimming length problem is resolved using the tuna model bending angle, by identifying as valid samples the ones whose vertebral points form a straight line and discarding the others. The tracking algorithm presented in (Muñoz-Benavent et al., 2020), allows us to obtain reliable size measurements based on the repetition of several measurements of the same fish. Fig. 4 illustrates the segmentation, model fitting and tracking procedures applied to a sample image.

The results for left and right videos, obtained separately, are merged to calculate fish sizes. The stereoscopic system is previously calibrated using a checkerboard pattern of known size to estimate the intrinsic and extrinsic camera parameters. The image plane information, i.e. snout and fork pixel coordinates, is transformed into 3D measurements using the calibration parameters of the stereoscopic vision system and 3D triangulation. Fish lengths are computed as the 3D Euclidean distances between snout and fork. Samples are discarded if the stereo correspondence is not met for the first and last model vertebrae, that is, if the distance from the points to the epipolar lines is greater than a threshold. See Muñoz-Benavent et al. (2018), for further details on the computer vision algorithms.

Note that the proposed tracking, based on overlapping fish silhouettes in subsequent video frames, is a basic procedure used to measure the same fish several times in order to obtain more accurate measurements. However, more advanced tracking procedures are needed to deal with occlusions and fish counting, especially in environments with higher density of fish. In this sense, the emergence of deep learning has created new approaches for object tracking. For example, DeepSORT algorithm (Wojke et al., 2018), which combines Kalman filter, Mahalanobis distance, Hungarian algorithm, and appearance feature vector could be evaluated to track the fish.

3. Results

Many of the fully automated processes that involve underwater imaging, such as biometric measurements, species identification, biomass

Table 2
Ground truth dataset for CNN model training. Contains 1000 labelled images divided into training and test datasets, with the different image backgrounds equally represented.

Ground truth	Image background	#1	#2	#3	#4
dataset of	Training (90%)	225	225	225	225
1000 images	Test (10%)	25	25	25	25

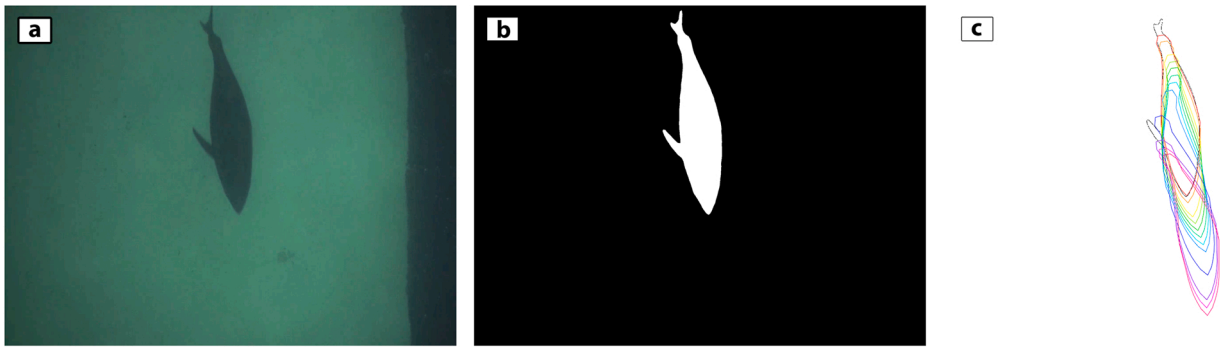


Fig. 4. Image processing algorithms involved in the process of Bluefin Tuna sizing. (a) original image, (b) segmented image using deep learning, (c) tuna model fitting and visual tracking.

estimation or fish counting, require segmentation of individual and whole fish. The number of segmented fish, as well as the accuracy of the segmentation, are very relevant for the subsequent processes and for reaching the ultimate objective successfully. In this section, techniques based on deep learning are compared with conventional segmentation techniques from the points of view of segmentation accuracy, computing time, number of measurements and fish sizing accuracy. Moreover, the variability associated with the measurement by the automatic system is presented, to emphasise the importance of having as many measurements as possible of each fish to increase the accuracy of the fish sizing.

3.1. Deep Learning models for fish detection and segmentation

The dataset described in Section 2.5 is used to train, validate and test the different deep learning models explained in Sections 2.2 and 2.3: Faster R-CNN, YOLO v5 and Mask R-CNN networks for object detection, and Mask R-CNN and PointRend for instance segmentation.

A 5-fold cross-validation technique was applied to train the models. As a result, the validation set size for each iteration of the validation process consists of 200 images. The training hyperparameters are shown in Table 3. Note that hyperparameters for YOLO v5 are different from the rest of the models due to the different frameworks used for training. Detectron2 framework, used for Faster R-CNN, Mask R-CNN and PointRend, is bsaed on iteration-based schedules, while YOLO v5′s framework uses epoch-based schedules. For both frameworks, the best set of hyperparamaters has been studied. Image augmentation techniques have been applied to training data in both frameworks. The augmentations applied and their probabilities are shown in Table 4.

The comparison of the different models and the attained inference is presented in Table 5 in terms of accuracy (with the metrics presented in Section 2.4) and inference time (measured in 2048 × 1536 pixel inputs using a single NVIDIA RTX 3090 with 24 GB of VRAM and an AMD Ryzen 9 3900X @3.8 GHz CPU). As shown in Table 5, Faster R-CNN, YOLO v5 and Mask R-CNN fish detection networks have good and similar accuracy (AP^{bb}), slightly better for Mask R-CNN in the high accuracy index (AP^{bb}_{90}). As regards the inference time, YOLO v5 is approximately 5 times faster than the others. However, in the current implementation, the computing time devoted to segmenting the fish inside the bounding box using k-means clustering and active contours (100 ms/fish) is bigger than the inference time devoted to finding the bounding box (7.4–36.9 ms/image). For all object detection networks, bounding boxes are accepted as true positives when the confidence exceeds a threshold of 0.8. For the case of fish instance segmentation, PointRend gives better results, especially in the high accuracy index AP_{90}, with a little increase of 5 ms per image with respect to Mask R-CNN. Since the CNN models in instance segmentation directly segment fish from the background, no further segmentation steps are needed.

Table 3
Training hyperparameters used for Faster R-CNN, Mask R-CNN, PointRend and YOLO v5 models. (*) Steps for learning rate scheduler specifies in which iterations the learning rate is modified, being reduced by a factor of 10. The rest of hyperparameters are well known in networks training.

Faster R-CNN, Mask R-CNN and PointRend		YOLO v5	
Number of iterations	9000	Number of epochs	50
Initial learning rate	0005	Initial learning rate	0,02
Steps for learning rate scheduler(*)	(6000, 8000)	Batch size	16
Images per batch	10		
Optimizer	SGD	Optimizer	SGD
Pre-training weights	ImageNet	Pre-training weights	ImageNet
Image augmentation	Custom (see Table 4)	Image augmentation	Custom (see Table 4)

Table 4
Image augmentations applied during training.

Augmentation	Probability (%)
Horizontal flip	30
Rotation	50
Contrast	50
Brightness	50
Saturation	50
Colour Temperature	20
Mosaic (only YOLO v5)	100

The conventional image processing techniques based on local thresholding have also been analysed with the test dataset and the proposed metrics. The results show that the segmentation of instances using Mask R-CNN and PointRend reaches higher AP metrics than using conventional image processing techniques, whereas the segmentation time per image is approximately 5 times faster when these DL techniques are applied. The improvement can be seen in Fig. 5, which shows four snapshots and the different fish segmentations achieved when applying conventional image processing techniques and PointRend. One can see both an increase in segmented individuals and a better definition of the fish's silhouette, which is very important when trying to estimate sizes.

When the objective is to estimate the fish's size, after the fish segmentation it is necessary to detect their silhouette and fit a deformable geometric model of tuna. The results for left and right videos are merged to compute 3D sizing using the calibration parameters of the stereoscopic vision system and 3D triangulation, as mentioned in Section 2.6. It is important to note that, although there should be a direct correspondence between the number of segmented instances and the number of sizing measurements achieved, it is necessary to detect the instances of the same fish in the images of both stereoscopic videos to estimate its length. In addition, sizing consecutive samples of the same fish reduces the error and increases the precision, as will be demonstrated in Section 3.3.

Table 5
Comparison of accuracy and inference time for the different networks and conventional image processing techniques. AP (Average Precision) metrics, with the Intersection over Union (IoU) threshold in the subscript and the method (bb for bounding boxes in object detection and mask for segmentation masks in instance segmentation) in the superscript.

	Fish detection		
	AP^{bb}	AP^{bb}_{90}	Inference time (ms/image)
Faster R-CNN	0.806	0.445	36.9
YOLOv5x	0.850	0.529	7.4
Mask R-CNN	0.855	0.663	36.4
	Fish instance segmentation		
	AP^{mask}	AP^{mask}_{90}	Inference time (ms/image)
Conventional image processing techniques	0.62	0.4182	185
Mask R-CNN	0.856	0.708	36.4
PointRend	0.870	0.792	41.6

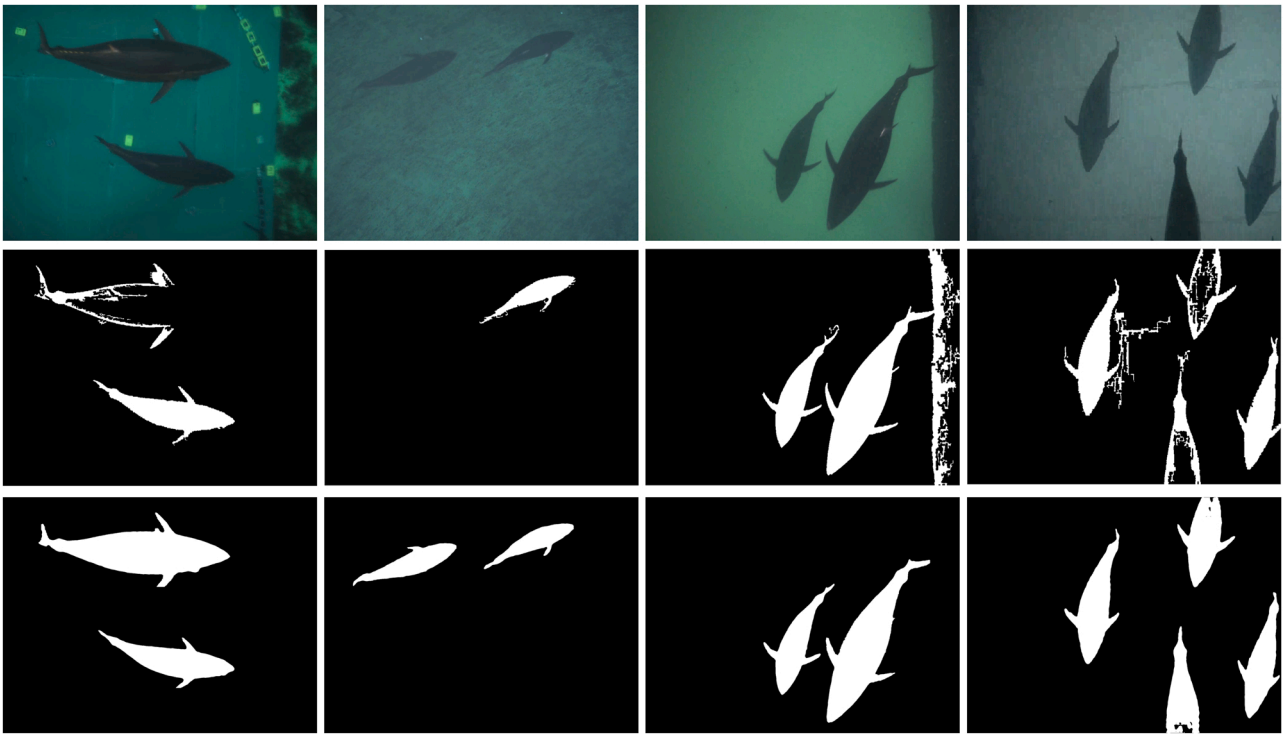


Fig. 5. Comparison of instance segmentation methods: Top row: original image; middle row: instance segmentation using conventional image processing techniques; bottom row: instance segmentation with PointRend.

3.2. Fish sizing

After applying the different models to the test dataset and according to the results obtained in Section 3.1, the methods and models with higher APs were selected to size the fish in the videos of Table 1: Mask R-CNN for object detection (followed by k-means clustering algorithm and active contours for segmentation) and PointRend for instance segmentation. The resulting fish sizing was compared with the previous automatic procedure based on conventional image processing techniques in terms of the number of sizing measurements (NM) and computing time (CT). As can be seen in Fig. 6, the improvement in NM when using segmentation based on DL is substantial. More than double the NMs are achieved in pooled data for all the considered DL methods, and the NM/CT relationship increases 1.7 times with Mask R-CNN and 3.5 times with PointRend.

Our videos were recorded on different days and with the fish swimming freely, so it is impossible to know, for each of the videos, the number of individuals that passed through the camera's field of view and the occlusions and overlapping produced. In short, we cannot know a priori how many individuals could have been segmented and therefore measured in each video, so a conclusive comparison between the results obtained with different backgrounds does not seem adequate. However, it is important to note that the behaviour of the different methods with respect to each of the image backgrounds is similar, i.e. all of the methods reach higher NM by processing images with background #4 and their lowest NM with background #1. This similar behaviour helps us to ensure that the conditions of the experiments for the different methods are comparable and means that the results obtained are worthy.

For the dorsal perspective, background #2 is the one preferred by the ICAR managers, since the installation of artificial back panels (background #1) is avoided, which can stress and alter the behaviour of the fish in the tanks, and costly cleaning is not needed (background #3). In the particular case of background #2, the NM is multiplied by 5.3 using

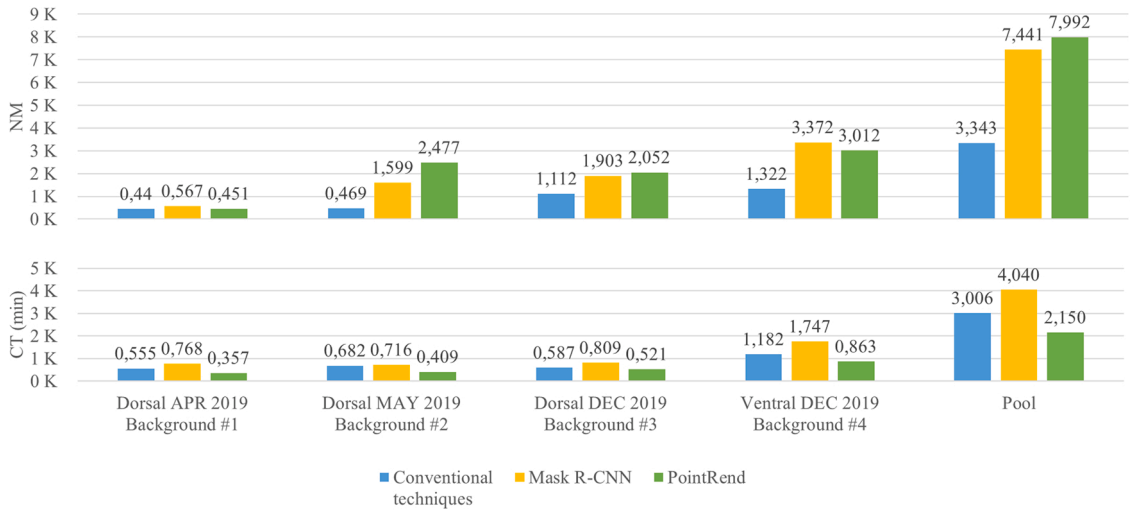


Fig. 6. Comparison of the number of sizing measurements (NM) and computing time (CT, in minutes) for the different deep learning models and the conventional image processing techniques in the different recordings and pooled data.

PointRend, and by 3.4 when Mask R-CNN is used. Differences are also observed in the NM/CT relationship, since the conventional techniques produce 0.7 measurements per minute, Mask R-CNN reaches 2.1 and PointRend reaches 6.9. A periodic cleaning has to be done to achieve background #2, so the floor of the tank does not get very dark along time.

Ground truth of fish distribution is impossible to obtain, since fish in ICAR tanks were not harvested immediately after the study and have never been physically sized, since they are extremely sensitive to handling and manipulation to compare the fish sizing with the different techniques in terms of accuracy, a new dataset was created with 396 fish from the pooled data, comprising the fish measured with all the techniques. As the fish sizing procedure with conventional techniques was validated in Muñoz-Benavent et al. (2018), with data from harvests, we assumed that the techniques based on DL would be equally validated if there is no statistically significant difference between the length frequency distributions shown in Fig. 7. The differences in distributions were analysed with the two-sample Kolmogorov-Smirnov test (Massey, 1951). As Table 6 shows, the test gives p-values higher than the 5% significance level, thereby validating the measurements obtained with DL techniques. Furthermore, these techniques obtain approximately twice as many measurements as conventional ones. Using the tracking techniques, we can deduce that conventional techniques measure 6.8 times the same fish on average, whereas DL techniques manage to measure each one 12.9–13.6 times on average.

Another interesting result is the application of our deformable tuna model to recordings from the dorsal view (cameras looking towards the bottom of the cage or tank). In our previous studies (Muñoz-Benavent et al., 2018; Puig-Pons et al., 2019), the tuna model was always applied to fish recorded from the ventral view (cameras looking towards the surface). In this study, recordings of the same growth tank from both the dorsal and ventral views were made (backgrounds #3 and #4 respectively) in December 2019, in order to compare the results from the two views. As Fig. 8 shows, the two distributions are very similar and the Kolmogorov-Smirnov test gives p-value = 0.389584 (higher than the 5% significance level), thereby validating the measurements from the dorsal view. Discrepancies between histograms occur because the fish population is randomly swimming through the cameras' field of view. To overcome this issue, we are currently working on a tagging system able to identify each individual.

Table 6
Statistical comparison of fish sizing measurements between conventional and deep learning techniques using a 396-fish dataset.

	Conventional	Mask R-CNN	PointRend
Total number of measurements	2677	5381	5128
Average number of measurements of each fish	6.8	13.6	12.9
Kolmogorov-Smirnov test p-value		0.7422	0.9999

3.3. Variability of the measurement error

The visual tracking described in Section 2.6 allows us to carry out a study on the variability associated with our automatic measurements. For fish measured more than once, the relative error e can be defined as the error of each individual measurement with respect to the median of all the measurements of the same fish in consecutive frames ($\widetilde{SFL}$):

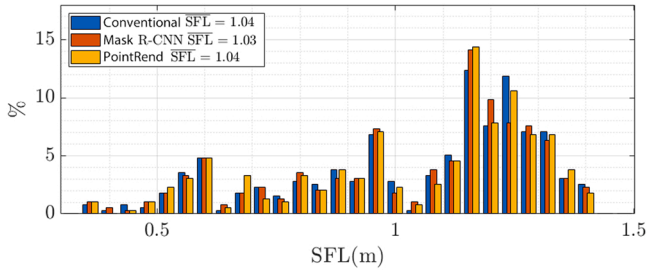


Fig. 7. Snout Fork Length (SFL) frequency histograms comparing conventional and deep learning techniques using a 396-fish dataset.

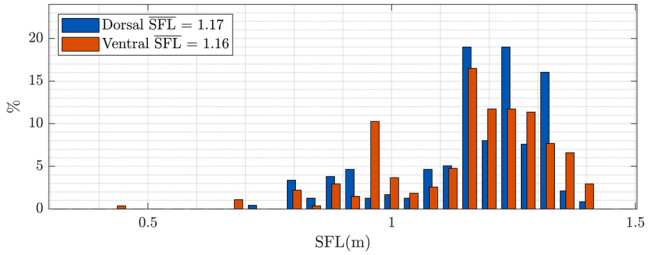


Fig. 8. Snout Fork Length (SFL) frequency histograms for recordings of the fish population in the growth tank from both dorsal and ventral views.

$$e_i(\%) = \frac{(SFL_i - \widetilde{SFL})}{\widetilde{SFL}} \cdot 100$$

and the mean relative error $\widetilde{e}$ is defined as follows:

$$\widetilde{e}(\%) = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

where n is the number of times a fish is measured.

As Fig. 9 shows, the mean relative error $\widetilde{e}$ decreases with the number of measurements per fish. For each box in the boxplot, the bottom of the central rectangle represents the 25th percentile, whereas the top represents the 75th percentile. The red segment inside the rectangle shows the median error. The whiskers represent the minimum and maximum mean relative error $\widetilde{e}$. Therefore, it can be seen that for n = 2, i.e., for fish measured two times in consecutive frames, the median of the mean relative error is around 0.2%, 75% of the fish are measured with a mean relative error lower than 0.6% and the maximum error is 1.3%. However, for n = 8 i.e., for fish measured eight times in consecutive frames, 75% of the fish are measured with a mean relative error lower than 0.3% and the maximum error is 0.5%. From these results, it can be concluded that it is important to have as many measurements as possible of each fish to increase the accuracy of the fish sizing.

4. Conclusions and further work

Conventional methods based on local thresholds enable automatic segmentation of fish but require a prior engineering process to deal with the adjustment of a series of parameters to adapt to the environment of each moment, as well as models of their silhouettes designed depending on the perspective from which the images are obtained. Although highly satisfactory results have been achieved with these methods, it is difficult to get rid of this parameter adjustment when, for example, the turbidity of the water blurs the image or the perspective of the silhouette changes from dorsal to ventral.

A new approach based on DL has been proposed for fish detection and segmentation in videos acquired under real conditions. One of the most important advantages is the elimination of the adaptive engineering process, because these models find the significant features by themselves through training. The results show that the fish sizing procedure is enhanced thanks to the improvement in segmentation of fish instances. In particular, the number of fish measurements increases up to 2.45 times when using the PointRend module. This increase in the number of measurements allows the accuracy of the fish sizing to be improved, since having as many measurements as possible of each fish reduces the variability and the error associated with each measurement, as demonstrated in Section 3.3. Looking at Fig. 5, an improvement in the definition of the silhouette of the fish can also be seen, which leads to a better estimation of measurements. Moreover, the number of measurements per minute of computing time increases up to 3.5 times with PointRend.

The proposed procedure could be a significant contribution towards a commercial system for fully automatic fish sizing using stereoscopic vision, since it increases the number of measurements and decreases the computing time compared to our previous developments. The automatic system has been used for fish sizing on juvenile BFT, but the procedure could be applied to other species, retraining the neural networks for instance segmentation with new images and adapting the geometric model. We intend to replicate the study to other environments with a higher density of fish (such as in-shore aquaculture facilities).

The International Commission for the Conservation of Atlantic Tunas (ICCAT) establishes a catch reporting system which covers the full chain of the Atlantic Bluefin Tuna fishery process from capture to sale. The use of a stereoscopic system to estimate catch quotas is estabAlished in 2015. To control fishing quotas, the authorities carry out a biomass assessment during fish transfers, counting transferred tuna and sizing at least 20% of the stock. The improvement in segmentation of fish instances and thus in the number of fish measurements could lead us to reach an accurate sizing of at least 20% of the fish in a wider range of conditions. We are currently working on optimising the implementation to reduce the computing time and have a system able to operate effectively in real time in a commercial operating environment. We plan to apply long short-term memory (LSTM) neural networks and DeepSort for fish tracking and counting. Moreover, we will focus on semantic segmentation methods using, for example, U-Net (Weng and Zhu, 2015) or DeepLab (Chen et al., 2018), which provide better performance in most applications compared to deep learning methods based on R-CNN. The use of night cameras to avoid interfering in the aquariums will also be studied.

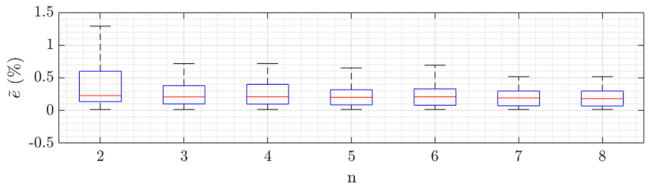


Fig. 9. Mean relative error $\widetilde{e}$ depending on the number of measurements per fish (n).

Funding

This study forms part of the ThinkInAzul programme and was supported by MCIN with funding from European Union NextGenerationEU (PRTR-C17. I1) and by Generalitat Valenciana (THINKINAZUL/2021/.007). It was also supported by funding from AICO/2021/016 (Generalitat Valenciana), PAID-10-19 (UPV) and IDIFEDER/2018/025 (EU-FEDER Comunitat Valenciana 2014–2020).

CRediT authorship contribution statement

P. Muñoz-Benavent: Methodology, Writing – original draft, Investigation, Software, **J. Martínez-Peiró:** Writing – original draft, Data curation, Software, **G. Andreu-García:** Project administration, Funding acquisition, Conceptualization, Validation, Writing – review & editing, **V. Puig-Pons:** Methodology, Investigation, **V. Espinosa:** Project administration, Funding acquisition, Conceptualization, Investigation, **I. Pérez-Arjona:** Funding acquisition, Conceptualization, F. De la Gándara: Funding acquisition, Validation, Supervision, Resources, **A. Ortega:** Funding acquisition, Methodology, Supervision, Resources.

Declaration of Competing Interest

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

Data Availability

The data that has been used is confidential.

Acknowledgements

This project has been possible thanks to the collaboration of the members of the Infrastructure for Atlantic Bluefin Tuna Aquaculture (ICAR), belonging to the IEO (Spanish Oceanographic Institute).

References

Abe, S., Takagi, T., Torisawa, S., Abe, K., Habe, H., Iguchi, N., Takehara, K., Masuma, S., Yagi, H., Yamaguchi, T., Asaumi, S., 2021. Development of fish spatio-temporal

identifying technology using SegNet in aquaculture net cages. Aquac. Eng. 93, 102146 https://doi.org/10.1016/j.aquaeng.2021.102146.
Álvarez-Ellacuría, A., Palmer, M., Catalán, I.A., Lisani, J.L., 2020. Image-based, unsupervised estimation of fish size from commercial landings using deep learning. ICES J. Mar. Sci. 77 (4), 1330–1339. https://doi.org/10.1093/ICESJMS/FSZ216.
ICCAT. (2015). Recommendation by ICCAT amending the recommendation 13–07 by ICCAT to establish a multi-annual recovery plan for Bluefin Tuna in the eastern Atlantic and Mediterranean. Rec [14–04]. In 2015 Compendium management recommendations and resolutions adopted by ICCAT for conservation of Atlantic tunas and tuna-like species (pp. 47–82).
Arthur, D., Vassilvitskii, S. (2006). k-means++: The Advantages of Careful Seeding. ⟨http://ilpubs.stanford.edu:8090/778⟩.
Chan, T.F., Vese, L.A., 2001. Active contours without edges. IEEE Trans. Image Process. 10 (2), 266–277. https://doi.org/10.1109/83.902291.
Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184.
Costa, C., Scardi, M., Vitalini, V., Cataudella, S., 2009. A dual camera system for counting and sizing Northern Bluefin Tuna (Thunnus thynnus; Linnaeus, 1758) stock, during transfer to aquaculture cages, with a semi automatic Artificial Neural Network tool. Aquaculture 291 (3–4), 161–167. https://doi.org/10.1016/j.aquaculture.2009.02.013.
Dutta, A., Zisserman, A. (2019). The VIA annotation software for images, audio and video. MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia, 2276–2279. https://doi.org/10.1145/3343031.3350535.
Espinosa, V., Soliveres, E., Cebrecos, A., Puig, V., Sainz-Pardo, S., & de la Gándara, F. (2011). Growing Monitoring in Sea Cages: Ts Measurements Issues. Proceedings of the 34th Scandinavian Symposium on Physical Acoustics, Geilo, Norway, 30 January – 2 February, 2011.
Everingham, M., van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. 88 (2), 303–338. https://doi.org/10.1007/s11263-009-0275-4.
Fernandes, A.F.A., Turra, E.M., de Alvarenga, É.R., Passafaro, T.L., Lopes, F.B., Alves, G. F.O., Singh, V., Rosa, G.J.M., 2020. Deep Learning image segmentation for extraction of fish body measurements and prediction of body weight and carcass traits in Nile tilapia. Comput. Electron. Agric. 170, 105274 https://doi.org/10.1016/j.compag.2020.105274.
Føre, M., Frank, K., Norton, T., Svendsen, E., Alfredsen, J.A., Dempster, T., Eguiraun, H., Watson, W., Stahl, A., Sunde, L.M., Schellewald, C., Skøien, K.R., Alver, M.O., Berckmans, D., 2018. Precision fish farming: A new framework to improve production in aquaculture. Biosyst. Eng. 173, 176–193. https://doi.org/10.1016/J.BIOSYSTEMSENG.2017.10.014.
Girshick, R., Donahue, J., Darrell, T., Malik, J., 2013. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 580–587. https://doi.org/10.1109/CVPR.2014.81.
Girshick, R. (2015). Fast R-CNN. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1440–1448. https://github.com/rbgirshick/.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182. ⟨https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf?ref=driverlayer.com/web⟩.
Hafiz, A.M., Bhat, G.M., 2020. A survey on instance segmentation: state of the art. Int. J. Multimed. Inf. Retr. 9 (3), 171–189. https://doi.org/10.1007/s13735-020-00195-x.
He, K., Gkioxari, G., Dollár, P., Girshick, R., 2020. Mask R-CNN. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 386–397. https://doi.org/10.1109/TPAMI.2018.2844175.
Jocher, G.,Changyu, L., Hogan, A., 于力军L. Y., changyu98, Rai, P., & Sullivan, T. (2020). ultralytics/yolov5:Initial Release. https://doi.org/10.5281/ZENODO.3908560.
Kirillov, A., Wu, Y., He, K., & Girshick, R. (2019). PointRend: Image Segmentation as Rendering. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 9796–9805. http://arxiv.org/abs/1912.08193.
Kloser, R.J., Ryan, T.E., Macaulay, G.J., Lewis, M.E., 2011. In situ measurements of target strength with optical and model verification: a case study for blue grenadier, Macruronus novaezelandiae. ICES J. Mar. Sci. 68 (9), 1986–1995. https://doi.org/10.1093/icesjms/fsr127.
Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. https://doi.org/10.1145/3065386.
Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature. Nature Publishing Group, pp. 436–444. https://doi.org/10.1038/nature14539.
Li, H., Chen, Y., Li, W., Wang, Q., Duan, Y., Chen, T., 2021. An adaptive method for fish growth prediction with empirical knowledge extraction. Biosyst. Eng. 212, 336–346. https://doi.org/10.1016/J.BIOSYSTEMSENG.2021.11.012.
Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C.L. (2014). Microsoft COCO: Common objects in context. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8693 LNCS(PART 5), 740–755. https://doi.org/10.1007/978–3-319–10602-1_48.
Liu, P., Wang, G., Qi, H., Zhang, C., Zheng, H., Yu, Z., 2019. Underwater image enhancement with a deep residual framework. IEEE Access 7, 94614–94629. https://doi.org/10.1109/ACCESS.2019.2928976.
Massey, F.J., 1951. The Kolmogorov-Smirnov test for goodness of fit. J. Am. Stat. Assoc. 46 (253), 68–78.
Muñoz-Benavent, P., Andreu-García, G., Valiente-González, J.M., Atienza-Vanacloig, V., Puig-Pons, V., Espinosa, V., 2018. Enhanced fish bending model for automatic tuna sizing using computer vision. Comput. Electron. Agric. 150, 52–61. https://doi.org/10.1016/j.compag.2018.04.005.
Muñoz-Benavent, P., Puig-Pons, V., Andreu-García, G., Espinosa, V., Atienza-Vanacloig, V., Pérez-Arjona, I., 2020. Automatic bluefin tuna sizing with a combined acoustic and optical sensor. Sensors 20 (18), 1–17. https://doi.org/10.3390/s20185294.
Padilla, R., Passos, W.L., Dias, T.L.B., Netto, S.L., da Silva, E.A.B., 2021. A comparative analysis of object detection metrics with a companion open-source toolkit. Electronics 10 (3), 1–28. https://doi.org/10.3390/electronics10030279.
Puig-Pons, V., Muñoz-Benavent, P., Espinosa, V., Andreu-García, G., Valiente-González, J.M., Estruch, V.D., Ordóñez, P., Pérez-Arjona, I., Atienza, V., Mèlich, B., de la Gándara, F., Santaella, E., 2019. Automatic Bluefin Tuna (Thunnus thynnus) biomass estimation during transfers using acoustic and computer vision techniques. Aquac. Eng. 85, 22–31. https://doi.org/10.1016/j.aquaeng.2019.01.005.
Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2015). You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 779–788. http://arxiv.org/abs/1506.02640.
Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.
Rosen, S., Jörgensen, T., Hammersland-White, D., Holst, J.C., Grant, J., 2013. DeepVision: a stereo camera system provides highly accurate counts and lengths of fish passing inside a trawl. Can. J. Fish. Aquat. Sci. 70 (10), 1456–1467. https://doi.org/10.1139/cjfas-2013-0124.
Saberioon, M., Císař, P., 2018. Automated within tank fish mass estimation using infrared reflection system. Comput. Electron. Agric. 150, 484–492. https://doi.org/10.1016/j.compag.2018.05.025.
Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J., 2013. Image classification with the fisher vector: theory and practice. Int. J. Comput. Vis. 105 (3), 222–245. https://doi.org/10.1007/s11263-013-0636-x.
Shortis, M., 2015. Calibration techniques for accurate measurements by underwater camera systems. Sensors 15 (12), 30810–30827. https://doi.org/10.3390/s151229831.
Shortis, M., Ravanbakskh, M., Shaifat, F., Harvey, E.S., Mian, A., Seager, J.W., Culverhouse, P.F., Cline, D.E., & Edgington, D.R. (2013). A review of techniques for the identification and measurement of fish in underwater stereo-video image sequences. Proc. SPIE 8791, Videometrics, Range Imaging, and Applications XII; and Automated Visual Inspection, 87910G. https://doi.org/10.1117/12.2020941.
Voskakis, D., Makris, A., Papandroulakis, N. (2021). Deep learning based fish length estimation. An application for the Mediterranean aquaculture. Oceans Conference Record (IEEE), 2021-September. https://doi.org/10.23919/OCEANS44145.2021.9705813.
Weng, W., Zhu, X., 2015. U-Net: convolutional networks for biomedical image segmentation. IEEE Access 9, 16591–16603. https://doi.org/10.48550/arxiv.1505.04597.
Williams, K., Lauffenburger, N., 2016. Automated measurements of fish within a trawl using stereo images from a Camera-Trawl device (CamTrawl. Methods Oceanogr. 17, 138–152. https://doi.org/10.1016/j.mio.2016.09.008.
Wojke, N., Bewley, A., & Paulus, D. (2018). Simple online and realtime tracking with a deep association metric. Proceedings - International Conference on Image Processing, ICIP, 2017-September, 3645–3649. https://doi.org/10.1109/ICIP.2017.8296962.
Yang, X., Zhang, S., Liu, J., Gao, Q., Dong, S., Zhou, C., 2021. Deep learning for smart fish farming: applications, opportunities and challenges. Reviews in Aquaculture. Wiley-Blackwell, pp. 66–90. https://doi.org/10.1111/raq.12464.
Zhang, L., Wang, J., Duan, Q., 2020. Estimation for fish mass using image analysis and neural network. Comput. Electron. Agric. 173, 105439 https://doi.org/10.1016/j.compag.2020.105439.