IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 October 2023, accepted 5 November 2023, date of publication 8 November 2023,
date of current version 14 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3330968

RESEARCH ARTICLE

Leveraging Bounding Box Annotations for Fish Segmentation in Underwater Images

JOSEP S. SÁNCHEZ[1], JOSE-LUIS LISANI[1], (Member, IEEE), IGNACIO A. CATALÁN[2], AND AMAYA ÁLVAREZ-ELLACURÍA[2]
[1]Ciències Matemàtiques i Informàtica, Universitat de les Illes Balears, 07122 Palma de Mallorca, Spain
[2]IMEDEA, UIB-CSIC, 07190 Esporles, Spain

Corresponding author: Jose-Luis Lisani (joseluis.lisani@uib.es)

This work has been partially sponsored and promoted by the Comunitat Autonoma de les Illes Balears through the Direcció General de Recerca, Innovació i Transformació Digital and the Conselleria de Economia, Hisenda i Innovació via Plans complementaris del Pla de Recuperació, Transformació i Resiliència (PRTR-C17-I1) and by the European Union- Next Generation UE (BIO/002A.1/2). Nevertheless, the views and opinions expressed are solely those of the author or authors, and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission are to be held responsible. The second author is also partially sponsored by grant BIO/022B.1.

ABSTRACT The use of Deep learning techniques in the field of Marine Science has become popular in recent years. For instance, many works propose the application of instance segmentation neural networks (in particular, Mask R-CNN) for detection and classification of fish in underwater images. The performance of these learning-based approaches depends heavily on the volume of data used for training, which, in the case of instance segmentation models for fish detection, implies that human experts must label and mark the shapes of all the fish appearing in a vast amount of underwater images. This is an enormously time-consuming task that we seek to alleviate in this paper. We propose a training strategy that combines manual and semi-automatic annotations. The latter are obtained in a weakly-supervised manner: the bounding box that contains the fish is manually selected, but its shape is automatically obtained thanks to a pretrained encoder-decoder segmentation network. Several popular architectures for this encoder-decoder network are examined. This strategy permits to reduce drastically the annotation cost for instance segmentation, at the expense of a small drop in performance with respect to the use of fully manual annotations. We show that a balance can be achieved between the segmentation performance and the time used to collect the training data by using the proposed strategy.

INDEX TERMS Deep learning, fish analysis, instance segmentation, bounding box, weakly-supervised learning, encoder-decoder network.

I. INTRODUCTION

Management of fisheries is of paramount importance to guarantee the sustainability of marine ecosystems. The study of fish abundance, size and biodiversity can be performed by analyzing underwater marine images [1], [2], [3], [4]. This process can be automated with the help of computer vision techniques. In particular, deep learning methods have been recently applied to fish detection and classification [5], [6], [7], [8], coral classification [9], [10], coastal sediment transport and morphodynamics [11], assessment of marine pollution [12], [13], etc.

If the goal is the classification of fish and the obtention of shape measurements the use of instance segmentation techniques is required. Instance segmentation methods are able to detect and localize objects (e.g. fish) in the image and assign the same label to all the pixels that describe them. The result is a *segmentation mask* that is unique for each detected object (see Figure 1-bottom). This mask can be analyzed and descriptors of the shape and size of the object can be obtained afterwards.

Several recent works show the advantages of using instance segmentation networks for fish analysis (see next section). These networks are trained by providing images and the segmentation masks of all the objects that we seek to detect. These masks need to be carefully annotated by human

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Huang.

experts, so that all the parts of the fish stand out with respect to the background. In the case of submarine images this can be a daunting task due to the limited visual quality of these images. Some pre-processing is often required to facilitate the annotation task [14]. On the other hand, the number of fish that may appear in a single image can be big and some of them overlap each other. According to [15], it can take up to 2 minutes per fish to acquire accurate segmentation labels. Besides, the diversity of the images and the number of instances of each type of fish must be big enough in the training set so that the network is able to generalize and perform well on unseen images. As a consequence, the number of expert hours needed to annotate a dataset in order to attain the desired performance of the network can be very high.

The burden of the annotation task can be alleviated with the use of weak annotations. These annotations may consist in the use of bounding boxes, or one or several clicks on the objects of interest, from which the segmentation masks can be automatically obtained. By using this type of annotations, the time required to label each fish can be reduced from minutes to seconds. Besides speeding up the annotation of new data, the use of bounding boxes permits to take advantage of a large amount of existing data already labeled using this format.

In this paper we focus on the use of bounding boxes for weak annotation. Figure 1 illustrates the proposed approach. In the top image the positions of the fish have been manually marked with bounding boxes. The corresponding segmentation masks (bottom) are automatically inferred from the bounding boxes using an encoder-decoder network. These segmentation masks are then used to train an instance segmentation model (Mask R-CNN [16]) for fish detection and classification.

Our contributions can be summarized as follows:

- We propose a framework that leverages bounding box annotations and obtains accurate segmentations of fish in the wild.
- We test several encoder-decoder network architectures to obtain segmentation masks from bounding box annotations.
- We present experimental results that show that the proposed approach permits to struck a balance between the performance of the instance segmentation network and the time needed to gather the training data.

The paper is organized as follows: in the next section we review recent works that apply deep learning techniques for fish analysis; in Section III the proposed method for the extraction of segmentation masks from bounding boxes is described; several experiments illustrating the performance of an instance segmentation network (Mask R-CNN) trained with the extracted masks are presented in Section IV; finally, some conclusions are drawn in Section V.

II. RELATED WORK

Several articles have been published in recent years proposing the use of object detection and image segmentation networks for the analysis of fish populations. Object detection models

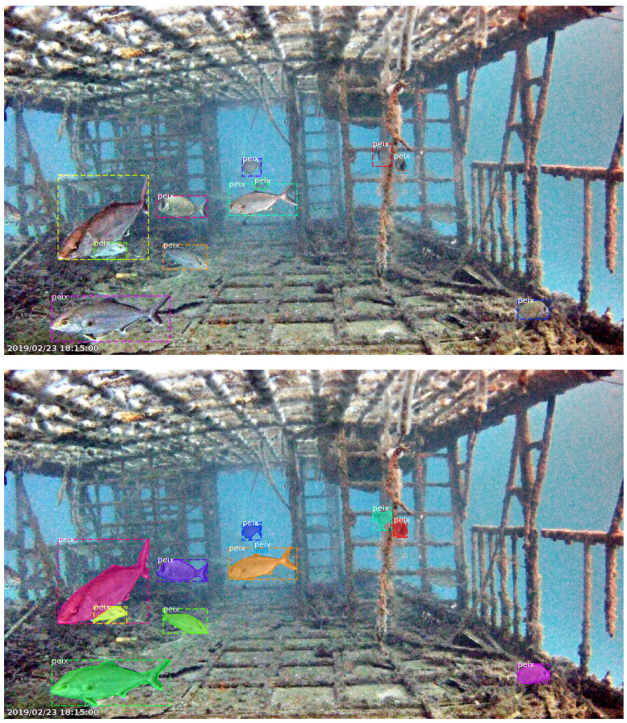


FIGURE 1. Manually annotated bounding boxes (top) and segmentation masks (bottom) automatically obtained using the proposed approach.

can detect, locate and classify the fish, while segmentation models provide information about their shape. Instance segmentation models are able to perform both tasks simultaneously and they permit to detect, classify and locate all the fish in the image, assigning labels at the pixel level.

Popular object detection architectures such as Faster R-CNN [17] and YOLO, in its different versions [18], [19], [20], have been used for fish detection and abundance estimation [21], [22], [23], [24], [25], [26]. Other works [27] propose the use of a feature pyramid architecture (FPN) to extract robust features at different scales and improve the detection performance. Qi et al. [28] apply a deformable convolutional pyramid structure for detection of small objects. In [29] the authors propose to train two YOLOv5 networks simultaneously so that they learn from each other based on a selection of the cleaner samples. Tseng et al. [30] train two networks to detect head and tail of fish, and estimate their length based on the distance between the detections. A recent modication of YOLO, YOLOACT [31], which permits fast instance segmentation has been proposed for fish identification in [32]. Rather than focusing on the architecture, [33] analyzes the importance in the selection of an adequate dataset for robust training of the object detector. Recent reviews of the literature on underwater object detection can be found in [34], [35], and [36].

Segmentation networks have been used in several works. Thampi et al. [37] apply the popular U-Net architecture [38] to the segmentation of five different fish species. In [39] two networks are used to produce the final segmentation,

one of them for optimal feature extraction and the other for multi-level feature accumulation that improves the pixel-wise prediction. Recent advances in network architectures such as attention modules and Transformers have been used to capture long-range dependencies between the image pixels and improve the segmentation results [40], [41].

One of the most popular architectures for instance segmentation, Mask R-CNN [16] have been profusely used for classification and estimation of fish size and other morphological features [25], [42], [43], [44], [45], [46]. Following this trend, we use this architecture in the current paper.

Weakly supervised methods for image segmentation have become popular since they permit to decrease the time cost of gathering a labeled training set. They are based on the use of weak annotations, from which a strong predictive model can be trained. There are four main types of weak annotations for object detection/segmentation: image-level annotations (a label assigned to the whole image, indicating the presence or absence of the sought object), point-level annotations (labels assigned to one or several pixels of the object), scribbles (labels assigned to a set of pixels along the object), and bounding boxes (a rectangle around the object). In the general computer vision literature several recent articles propose the use of these types of annotations, specially in the field of medical image segmentation. Scribble annotations are used in [47], [48], [49], and [50], points in [51], [52], and [53] and bounding boxes in [54], [55], [56], and [57]. In Marine Science applications the literature on weakly supervised methods is scarce. Laradji et al. [15] use single point annotations for fish segmentation. They design a network with activation and affinity branches merged through a random walk to diffuse labels from the selected point. However, their results lack accuracy in the fine details of the fish (tails or fins). Saleh et al. in [41] and [58] propose two different approaches for unsupervised fish segmentation that leverage spatial and temporal variations in video data.

In addition to the weakly supervised approach, other methods have been proposed in the literature to reduce the annotation cost. Active learning methods [59], [60], [61] advocate for reducing the number of images to annotate by selecting the ones that are expected to yield the largest increase in the model's performance. This requires the definition of an ''acquisition function'' adapted to the underlying data distribution. Interactive Learning methods [62], [63] add human interaction in the segmentation process: positive and negative clicks are used to define the foreground and the background in the scene. Recently [64], a combination of both techniques has been proposed.

In our work, most of the data used for training have annotations in the form of bounding boxes. We have decided to use all the available data instead of sampling this training set using an active learning method. Moreover, we don't want to increase the annotation time by using an interactive method to refine the segmentation results. Since we want to take full advantage of the available annotations we propose the use of bounding boxes as weak annotations for fish instance segmentation. Inspired by transfer learning techniques we propose to use an existing image segmentation architecture to segment the contents of each bounding box. These segmentations can then be used to train an instance segmentation network for the detection of fish in any image. To the best of our knowledge, this is the first time this strategy for fish instance segmentation has been proposed in the literature. The method is described in the next section.

III. FROM BOUNDING BOXES TO SEGMENTATION MASKS

Our goal is to develop a method capable of transforming datasets with bounding box type annotations into valid annotations for segmentation. The process is illustrated in Figure 2. First, given an image containing several objects of interest (in our case fish) framed by bounding boxes, these bounding boxes are extracted in the form of sub-images. Each of these sub-images is segmented using a neural network and the masks obtained are drawn over the original image, after post-processing using classical techniques. Details on the neural network used for the segmentation and on the post-processing of the masks are given below.

A. SEGMENTATION NETWORK

In order to segment the fish contained in the sub-images associated to the bounding boxes, we will make use of a segmentation model.

Many of such models have been proposed in the literature. On the the most popular is U-Net [38], originally proposed for the segmentation of medical images, but that has found application in several fields such as satellite imagery [65], ecological monitoring [66], autonomous driving [67], etc.

The U-Net consist of a contraction path (encoder) composed by four blocks, and an expansion path (decoder) with also four blocks. Both paths are connected through skip connections. Figure 3 displays the architecture.

Due to its simplicity and its proven efficacy in the segmentation of different types of images we have used U-Net in our tests. Moreover, we have tested two state-of-the-art alternatives: U-Net++ [68], [69] and DeepLabv3+ [70]. We do not use recent segmentation methods based on Visual Transformers due to their high computational requirements and the vast amount of data needed to train them.

UNet++ consists of an encoder and a decoder connected through a series of nested dense convolutional blocks (see Figure 4). The goal is to bridge the semantic gap between the feature maps of encoder and decoder prior to fusion.

DeepLabv3+ employs also an encoder-decoder architecture where the rich contextual information is encoded using DeepLabv3 [71] and a simple decoder manages to recover the object boundaries (see Figure 5). DeepLabv3 employs atrous convolutions to capture multi-scale context, which permits to handle the problem of segmenting objects at multiple scales.

In our experiments we have implemented U-Net following the description of the original paper. Instead of the conventional scaling that can be found in many of the public

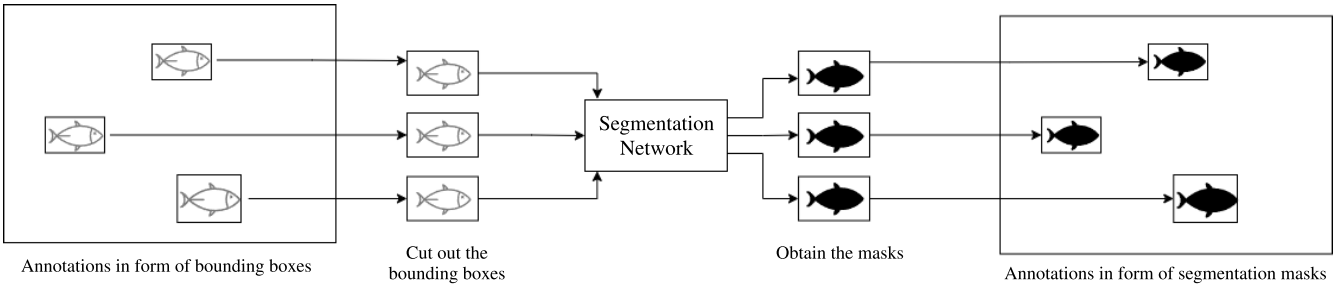


FIGURE 2. Diagram of the proposed approach.

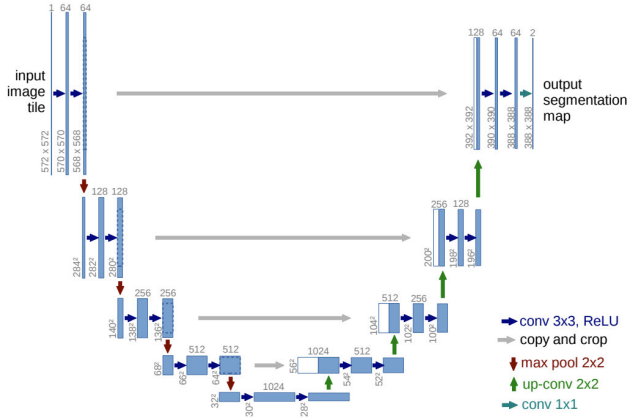


FIGURE 3. U-Net architecture [38].

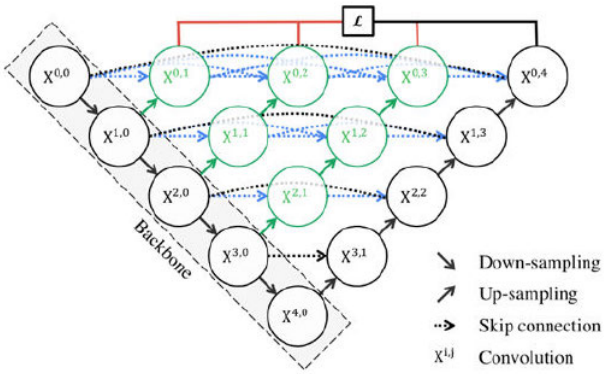


FIGURE 4. U-Net++ architecture (from [68]).

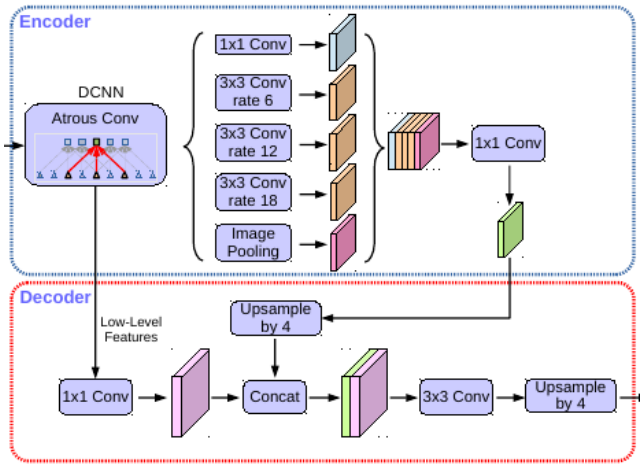


FIGURE 5. DeepLabv3+ architecture (from [70]).

implementations we use learnable inverse convolutions in the upscaling layers and, in order to prevent overfitting, we apply the dropout technique [72].

Regarding U-Net++ and DeepLabv3+, we have used public implementations of the models, available in GitHub.[1]

B. TRAINING

The three segmentation models described in the previous section have been trained with 2480 sub-images cropped from an original set of 600 larger images (see Figure 2).

Each of these sub-images displays a fish whose shape has been manually segmented and constitutes the ground truth for the training. In addition, 756 sub-images, with their corresponding ground truths, have been used to validate the network after training. All these sub-images have sizes bigger than 25 × 25 pixels.

The models have been trained using one NVIDIA GeForce RTX 2080 Ti card with 11GB of VRAM. Different configurations of each model, leading to different numbers of learnable parameters, have been tested. For the U-Net we have used either 32 or 64 filters in the first layer of the encoder, which implies 8.6M or 34.5M parameters. Similarly, for U-Net++ we have made experiments with 16 and 32 filters in the first layer, leading to 24.1M and 41.8M parameters. Finally, 256 or 512 filters have been used in the convolutional blocks of DeepLabv3+ (11.8M and 18.4M parameters, respectively).

The U-Net has been trained from scratch using Xavier initialization of the parameters. For U-Net++ we use a VGG16 backbone and for DeepLabv3+ a ResNet50, both initialized with weights learnt on Imagenet. We have then fine-tuned the models with the training images. Regarding the loss function, we use the binary cross-entropy, and the Adam optimizer to minimize it. The initial learning rates have been 0.0001 for U-Net and U-Net++ and 0.001 for DeepLabv3+. Different numbers of epochs have been used for each model: 60 epochs for U-Net and 40 epochs for U-Net++ and DeepLabv3+. For each model, the parameters providing the minimum value of the loss function over the validation set during the training process have been saved.

[1] https://github.com/MrGiovanni/UNetPlusPlus/tree/master/keras
https://keras.io/examples/vision/deeplabv3_plus/

Moreover, we have tested two different sizes for the input images: 64×64 and 128×128. The batch size has been set to the maximum value that fits the memory of our GPU, which depends of the number of parameters of the model and on the size of the input images (in practice, values 32 or 16 have been used in all the experiments). The average times needed for each training have been 25 min. for the U-Net and 15 min. for the U-Net++ and DeepLabv3+.

For each model and configuration 5 trainings have been performed and evaluated on the images of the validation set. The metric used to measure each network performance is the intersection over union, commonly known as IoU (intersection over union).[2] This metric defines a number in the range $[0, 1]$ where 0 means that there is no overlap between the output of the network and the ground truth and 1 means a perfect match.

For each image in the validation set the IoU of the segmentation result, for each training, has been computed. The averages and standard deviations of these values are displayed in Table 1.

TABLE 1. IoU values over the validation set for the three tested architectures, for different number of parameters and input resolutions. Average and standard deviation values (in brackets) are displayed. The best result is highlighted.

		Input resolution	
Architecture	#Parameters	64×64	128×128
U-Net	8.6M	0.764 (0.010)	0.768 (0.010)
	34.5M	0.796 (0.005)	0.798 (0.004)
U-Net++	24.1M	0.830 (0.006)	0.826 (0.005)
	41.8M	**0.832 (0.004)**	0.828 (0.004)
DeepLabv3+	11.8M	0.696 (0.162)	0.744 (0.042)
	18.4M	0.692 (0.023)	0.724 (0.023)

We observe that the best performances are obtained with the U-Net++ model, followed by U-Net. The U-Net++ configuration with 41.8M parameters and input size 64×64 obtains, in average, the best results, but closely followed by the configuration with 24.1M. We have chosen this less computationally demanding configuration to obtain segmentation masks from bounding boxes in Section IV.

The images in Figure 6 permit a qualitative evaluation of the results obtained with each model. The configurations used to obtain these results are: U-Net 34.5M parameters and 64×64 input, U-Net++ 24.1M parameters and 64×64 input and DeepLabv3+ 18.4M parameters and 128×128 input.

The images displayed in Figure 6 illustrate the performance of the methods under different scenarios. The first row displays a good qualiy image with well contrasted fish. The image in the second row has low quality and part of the background has a similar color than the fish. In the third row the fish is only partially visible, while it is partially occluded by another fish in the fourth row. Finally, the shape of the fish in the last row is different from the shape of most other fish in the dataset.

We observe that, in general, DeepLabv3+ is unable to recover the correct shape of the fish, specially when partial occlusions are present. U-Net and U-Net++ do a similar job, but U-Net++ recovers slighly better fine shape details as the tails of the fish.

C. MASK ENHANCEMENT

The output of the segmentation network is a pixel map with values in the range $[0, 1]$. These values indicate the probability that the pixel belongs to the segmented object.

The standard procedure to obtain the segmentation mask is to binarize this map with threshold 0.5. The masks displayed in Figure 6 have been obtained this way.

However, a method for the selection of this threshold based on the conformal risk paradigm has been recently proposed and applied to the segmentation of medical images [73]. The authors use a conformal risk control algorithm to pick the threshold value λ guaranteeing that the average fraction of missed foreground pixels (the *false negative rate*) is below some fixed parameter α. The algorithm works with an already trained network and only needs a set of images with known ground truth to estimate λ. Some images are randomly selected from this set (the *calibration points*) and used to estimate the lowest λ that guarantees that the empirical false negative rate of the selected set is below α. The process is repeated several times and the final value of λ is computed as the average of the obtained results.

We have adopted the same approach, with $\alpha = 0.1$, applied to the validation set of our experiment (756 images, 300 of which have been randomly selected as calibration points). After 10 runs of the algorithm we have found that the average value of the threshold is, approximately, 0.64. This value has then been selected as the binarization threshold to obtain the segmentation masks in the rest of the paper. Figure 7 shows an example of the application of this threshold to the result of U-Net++. The obtained mask can be compared to the one computed with the standard threshold. We observe that the number of missclassified background pixels decreases.

Additionally some classic post processing is applied to the thresholded output: first a closing operation [74] with a circular kernel of size 3×3 is used to fill small holes and connect nearby connected regions, and then the largest connected component is taken as the final output of the processing (see an example in Figure 7(f)).

IV. INSTANCE SEGMENTATION WITH MANUAL AND AUTOMATIC ANNOTATIONS

In this section we make experiments on instance segmentation of fish in underwater images, comparing the results obtained after training the instance segmentation network with masks manually annotated and with the masks obtained from bounding boxes with the U-Net++ method proposed in Section III.

[2] Given two segmentation masks A and B, $IoU_{A,B} = \frac{|A \cap B|}{|A \cup B|}$, where $|\cdot|$ denotes the area operator.

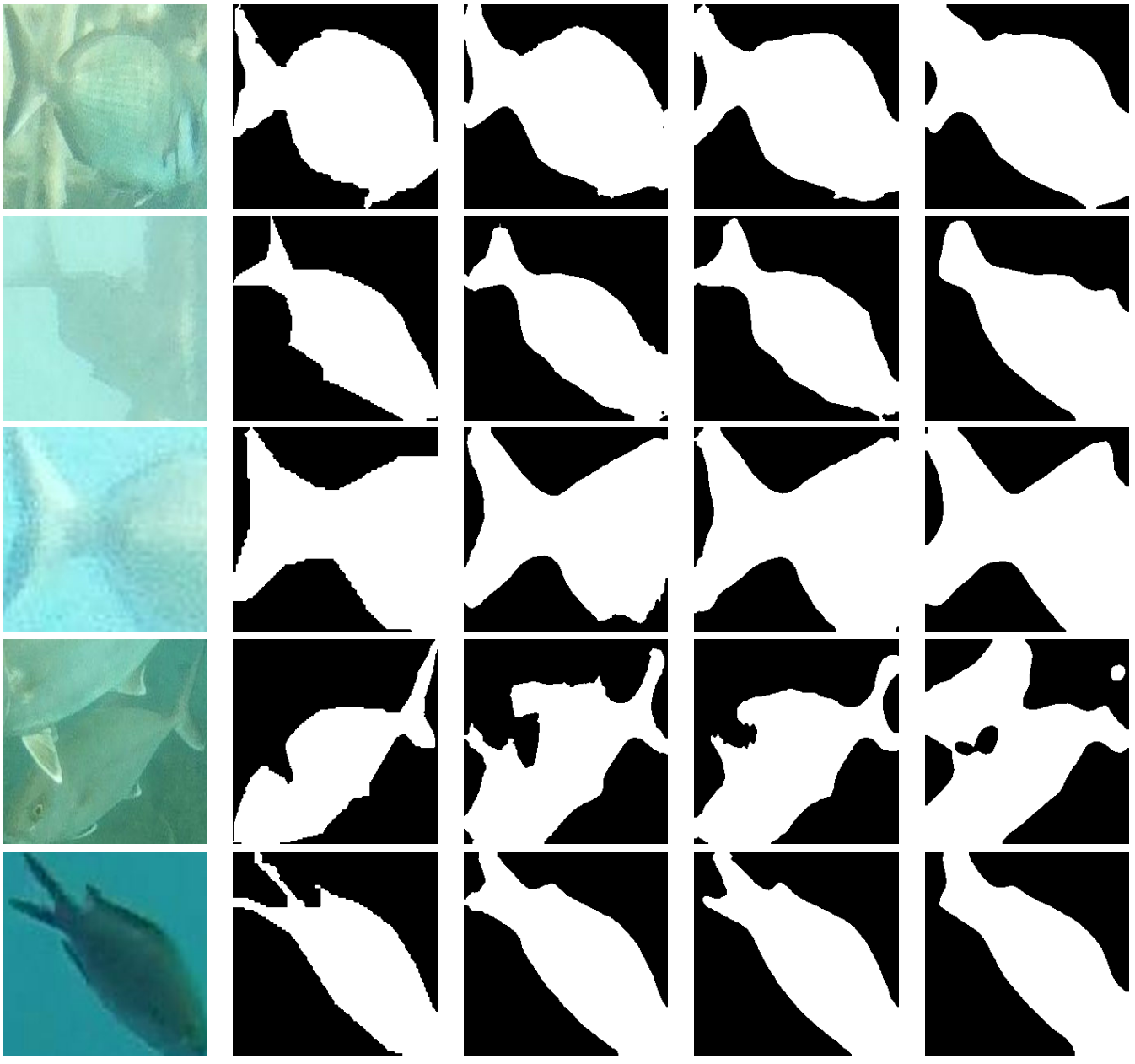


FIGURE 6. Examples of segmentation masks obtained with the three tested architectures for different input images. The masks are obtained after applying the standard binarization threshold (0.5) to the output of each network.

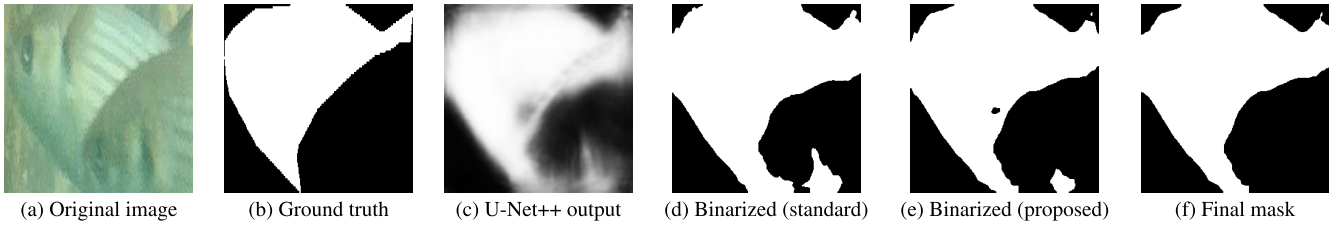


FIGURE 7. Illustration if the post-processing applied to the output of the UNet++: a binarization threshold computed after conformal risk control analysis is applied (e) (compare with result of applying standard threshold 0.5 (d)); the final mask is obtained after a closing operation (which removes small holes and connects nearby connected regions) and selection of the biggest connected component.

Although any instance segmentation architecture could have been chosen for our experiments, we have opted for Mask R-CNN [16], which is widely used by the Marine Sciences community (see Section II). However, more

recent models could have been used (e.g. YOLOACT [31], YOLOACT++ [75], Mask2Former [76]).

We use the MMDetection[3] implementation of the network, with ResNet 50 backbone and pre-trained weights, which we fine-tune[4] to detect a single class of object (*'fish'*). The network has been trained with 815 images containing several instances of fish (of different species). The only data augmentation method we have used is image flipping. Other techniques such as rotation, shifting or cropping did not improve the results. The performance of the network has been assessed on a validation set with 418 images. The training and validation sets contain about 6600 and 2500 fish, respectively. The manually annotated segmentation masks for all these fish are available as ground truth. Only fish larger than 40× 40 pixels were annotated by the human experts. It must be remarked that the U-Net++ model described in the previous section was trained with sub-images extracted from a set of images different from the ones used to train the Mask R-CNN network.

Mask R-CNN has been trained 7 times with different settings for the training data. Recall that this network needs annotations in the form of segmentation masks for training. For each training setting the images are the same, but the associated annotation files are a mixture of manual and semi-automatic annotations. The semi-automatic annotations are obtained from the bounding boxes that delimit each fish: the U-Net++ model from Section III is applied to the sub-image contained in the bounding box and the segmentation mask is obtained as output.

The different settings of our experiment use different percentages of manual annotations: 100% (all manual), 80%, 60%, 50%, 40%, 20% and 0% (all automatic). Mask R-CNN has been trained for 12 epochs with each setting, using one NVIDIA GeForce RTX 2080 Ti card with 11GB of VRAM. The average time needed for each training has been 40 min. In inference, the 418 images in the validation set are segmented in approximately 55 seconds.

For each training setting the performance of the network is evaluated using the mean average precision (mAP[5]), which summarizes the relationship between precision and recall for different values of the score of the detected objects. Precision measures the ratio of correct detections to total detections, and recall measures the ratio of correct detections to total objects in the ground truth. The decision on whether a detection is correct is made based on the IoU between the detected object and the ground truth. In our experiments we require a minimum IoU of 0.5 to consider a detection as correct. This IoU can be calculated in two ways, based on the intersection/union of the bounding boxes, or based on the intersection/union of the segmentation masks.[6] We apply this second definition since the obtained mAP is related to the quality of the instance segmentation.

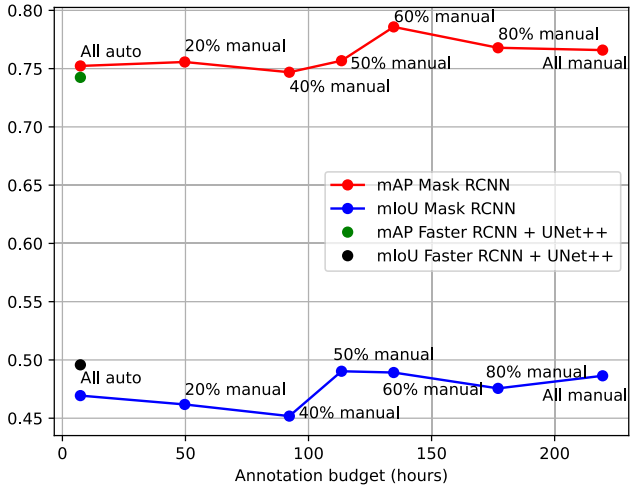


FIGURE 8. mAP and average IoU values over the validation set for different training strategies (the closer to one the better). Several mixes of manual and semi-automatic annotations have been tested: from fully manual annotations (100%) to fully automatic (0% manual). The estimated time needed to annotate the data associated with each strategy is shown in the horizontal axis. See comments in the text.

Figure 8 summarizes the results of our experiments. This graphic displays, in red the mAP values (the closer to one the better) over the validation set obtained for different training settings. Besides, the IoU for each segmented image[7] is computed and the averages are displayed in the blue curve. The horizontal axis of the graphic shows an estimation of the time needed to obtain the annotated data in each setting. We consider an average of 2 minutes per fish to manually obtain the segmentation masks [15] and 4 seconds to draw the bounding boxes used for the automatic estimation (recall that the training set contains about 6600 fish). If only bounding boxes are used for annotation (all automatic setting) the annotation time is 7.3 hours, in contrast with the 219.5 hours needed to manually annotate all the segmentation masks.

In general, as the number of manual annotations increases, the performance of the network increases (both in terms of mAP and mIoU), however, it is interesting to observe that a balanced mix of both types of annotation (around 50% or 60% manual), produces the best results. The images in Figure 9 help to explain this result.

This figure compares the ground truth of a given image with the segmentations obtained under different training settings.[8] We observe that some fish are missing in the segmented images, but also that the ground truth is not perfect since one fish at the top of the image was not annotated. This fish has been however detected by the

[3]https://github.com/open-mmlab/mmdetection

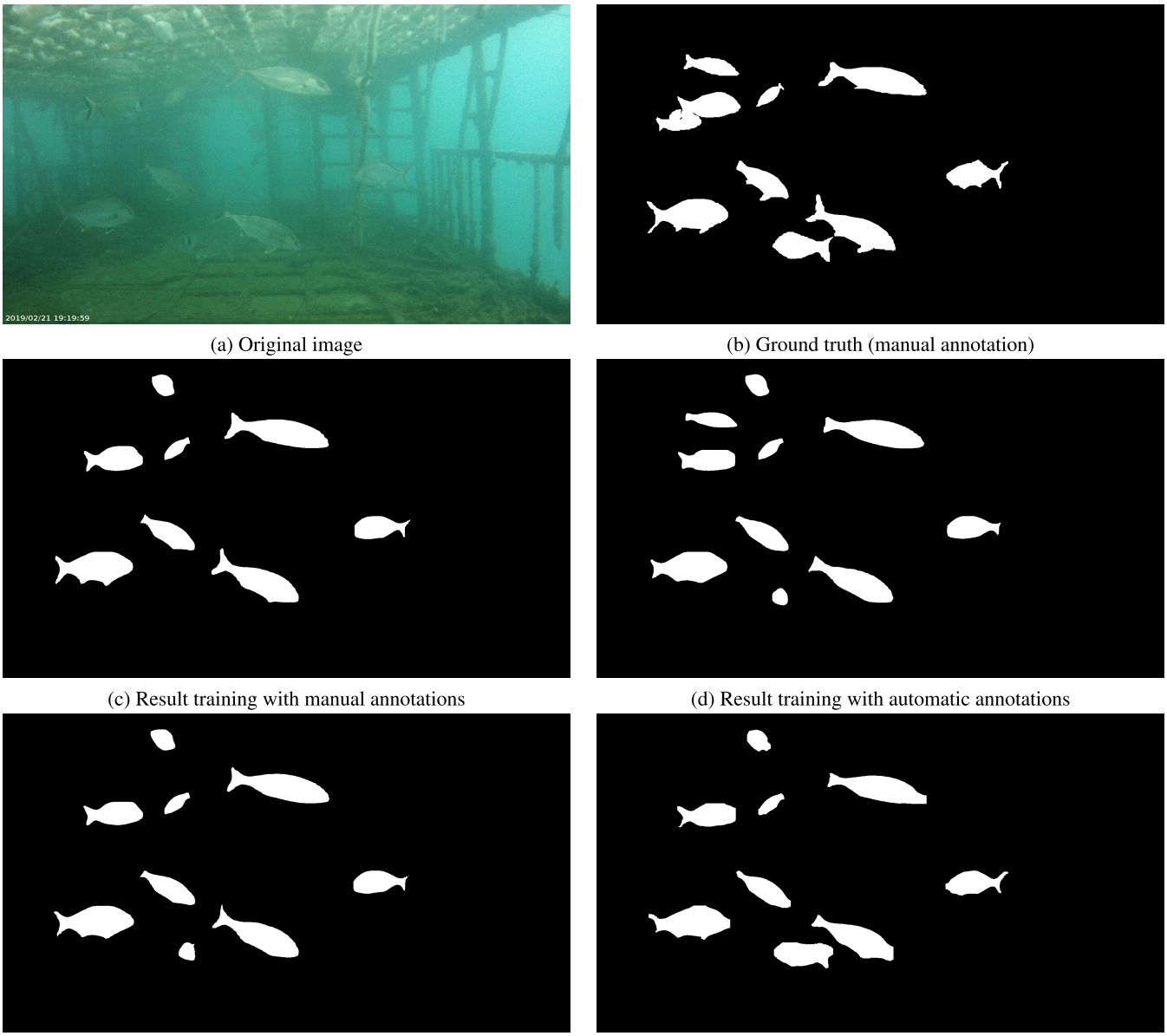
[4]MMdetection respects the convention used in Detectron (https://github.com/facebookresearch/detectron2) to freeze the ResNet layers, by default and the configuration we used freezes the stem and the first stage of ResNet. Tests performed by increasing the number of trained stages did not positively affect our results.

[5]https://cocodataset.org/#detection-eval

[6]In this case the ground truth consists of the manually annotated segmentation masks.

[7]The segmentation result displays the segmentation masks of all the fish whose detection score is above 0.7.

[8]The images display the segmentation masks of all the fish whose detection score is above 0.7.



(a) Original image

(b) Ground truth (manual annotation)

(c) Result training with manual annotations

(d) Result training with automatic annotations

(e) Result training with 60% manual and 40% automatic annotations

(f) Result Faster RCNN followed by UNet++ segmentation

FIGURE 9. **Original image (a) and ground truth segmentation (manual annotation) (b). Examples of prediction results of Mask RCNN with several training settings: all manual annotations (c), all automatic annotations (d), and 60% manual and 40% automatic (e). Result of applying the U-Net++ segmentation network trained in Section III to the bounding boxes detected with Faster RCNN (f).**

network. We also observe that the main bodies of the fish are correctly detected, but most of the fine details (tails, fins) are missing. This is specially true when using only automatic annotations (Figure 9(d)). When using only manual annotations more details are recovered (Figure 9(c)), and when using a 60%-40% mix we obtain intermediate results (Figure 9(e)). However, in this last case one additional fish is partially detected (bottom-most detection in image (e)). This may explain why the performance results in Figure 8 are better than the rest with this training setting.

One could wonder if the method described in Section III to obtain segmentation masks from bounding boxes could be used at the *output* of an object detection network, instead of using it to train an instance segmentation network. That is, could we apply the U-Net++ trained in Section III to the bounding boxes obtained with an objet detector and obtain segmentation masks of similar quality that the ones obtained with the instance segmentation network? In order to test this possibility, we have trained a popular object detector (Faster R-CNN [17]) with the same training set as Mask R-CNN, but only using the bounding box information.

As for Mask RCNN, we have also used the MMdetection framework to train Faster RCNN. Since both architectures are closely related, we use the same Resnet 50 backbone and pre-trained weights which we fine-tune to detect a single class object ('fish').

The trained model has been applied to detect and locate the fish in the validation set and the sub-images associated to the detected bounding boxes have been segmented using U-Net++ (with the post-processing described in Section III-C).

Figure 9(f) shows an example of the obtained results and the green and black points in Figure 8 display, respectively, the mAP and average IoU values over the validation set. We observe that, in terms of mAP, the results of this approach are worse than using Mask RCNN. However, the average IoU is quite high. This may be explained by the fact that, in some cases, Faster RCNN is able to correctly detect fishes that are missed by Mask RCNN (see bottom-most detection Figure 9(f)), while U-Net++ recovers better the fine details in the tails and fins of the fish.

V. CONCLUSION

In this article we have proposed a strategy that allows training an instance segmentation network (Mask RCNN) for fish detection in a fraction of the time normally needed for the task.

We show that the combination of weak annotations in the form of bounding boxes and a U-Net++ encoder-decoder architecture may be used to obtain segmentation masks that, in whole or in part, can replace manually annotated masks.

Our experimental results show that good instance segmentation results can be obtained by combining manual and automatic annotations, with the additional benefit of reducing the time required to obtain the training data.

A wealth of datasets containing still images of fish, annotated using bounding boxes (e.g. [77], [78]) presents an untapped opportunity for our proposed methodology. The potential to automate the segmentation of these labeled objects would lower human labor costs and bolster our capacity to obtain precise estimates of fish size, shape, and health metrics. This advancement stands to facilitate more informed decision-making, optimize breeding practices [6], and improved ecosystem management, ultimately fostering sustainable fisheries and healthier aquatic environments [79].

REFERENCES

[1] J. Aguzzi, C. Doya, S. Tecchio, F. C. De Leo, E. Azzurro, C. Costa, V. Sbragaglia, J. Del Río, J. Navarro, H. A. Ruhl, J. B. Company, P. Favali, A. Purser, L. Thomsen, and I. A. Catalán, "Coastal observatories for monitoring of fish behaviour and their responses to environmental changes," *Rev. Fish Biol. Fisheries*, vol. 25, no. 3, pp. 463–483, Sep. 2015, doi: 10.1007/s11160-015-9387-9.

[2] C. Díaz-Gil, S. L. Smee, L. Cotgrove, G. Follana-Berná, H. Hinz, P. Marti-Puig, A. Grau, M. Palmer, and I. A. Catalán, "Using stereoscopic video cameras to evaluate seagrass meadows nursery function in the Mediterranean," *Mar. Biol.*, vol. 164, no. 6, p. 137, Jun. 2017, doi: 10.1007/s00227-017-3169-y.

[3] G. Follana-Berná, P. Arechavala-Lopez, E. Ramirez-Romero, E. Koleva, A. Grau, and M. Palmer, "Mesoscale assessment of sedentary coastal fish density using vertical underwater cameras," *Fisheries Res.*, vol. 253, Sep. 2022, Art. no. 106362, doi: 10.1016/j.fishres.2022.106362.

[4] M. Francescangeli, V. Sbragaglia, J. del Rio Fernandez, E. Trullols, J. Antonijuan, I. Massana, J. Prat, M. Nogueras Cervera, D. M. Toma, and J. Aguzzi, "Long-term monitoring of diel and seasonal rhythm of *Dentex dentex* at an artificial reef," *Frontiers Mar. Sci.*, vol. 9, Mar. 2022, Art. no. 837216, doi: 10.3389/fmars.2022.837216.

[5] M. Goodwin, K. T. Halvorsen, L. Jiao, K. M. Knausgård, A. H. Martin, M. Moyano, R. A. Oomen, J. H. Rasmussen, T. K. Sørdalen, and S. H. Thorbjørnsen, "Unlocking the potential of deep learning for marine ecology: Overview, applications, and outlook," *ICES J. Mar. Sci.*, vol. 79, no. 2, pp. 319–336, Mar. 2022, doi: 10.1093/icesjms/fsab255.

[6] D. Li and L. Du, "Recent advances of deep learning algorithms for aquacultural machine vision systems with emphasis on fish," *Artif. Intell. Rev.*, vol. 55, no. 5, pp. 4077–4116, Jun. 2022, doi: 10.1007/s10462-021-10102-3.

[7] S. Mittal, S. Srivastava, and J. P. Jayanth, "A survey of deep learning techniques for underwater image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 6968–6982, Oct. 2023, doi: 10.1109/TNNLS.2022.3143887.

[8] A. Saleh, M. Sheaves, and M. Rahimi Azghadi, "Computer vision and deep learning for fish classification in underwater habitats: A survey," *Fish Fisheries*, vol. 23, no. 4, pp. 977–999, Jul. 2022, doi: 10.1111/faf.12666.

[9] A. Lumini, L. Nanni, and G. Maguolo, "Deep learning for plankton and coral classification," *Appl. Comput. Informat.*, vol. 19, nos. 3–4, pp. 265–283, Jun. 2023, doi: 10.1016/j.aci.2019.11.004.

[10] T. N. T. Arsad, E. A. Awalludin, Z. Bachok, W. N. J. H. W. Yussof, and M. S. Hitam, "A review of coral reef classification study using deep learning approach," *AIP Conf. Proc.*, vol. 2484, no. 1, Mar. 2023, Art. no. 050005, doi: 10.1063/5.0110245.

[11] E. B. Goldstein, G. Coco, and N. G. Plant, "A review of machine learning applications to coastal sediment transport and morphodynamics," *Earth-Sci. Rev.*, vol. 194, pp. 97–108, Jul. 2019, doi: 10.1016/j.earscirev.2019.04.022.

[12] X. Li, B. Liu, G. Zheng, Y. Ren, S. Zhang, Y. Liu, L. Gao, Y. Liu, B. Zhang, and F. Wang, "Deep-learning-based information mining from ocean remote-sensing imagery," *Nat. Sci. Rev.*, vol. 7, no. 10, pp. 1584–1605, Oct. 2020, doi: 10.1093/nsr/nwaa047.

[13] M. Wolf, K. van den Berg, S. P. Garaba, N. Gnann, K. Sattler, F. Stahl, and O. Zielinski, "Machine learning for aquatic plastic litter detection, classification and quantification (APLASTIC-Q)," *Environ. Res. Lett.*, vol. 15, no. 11, Nov. 2020, Art. no. 114042, doi: 10.1088/1748-9326/abbd01.

[14] J.-L. Lisani, A.-B. Petro, C. Sbert, A. Álvarez-Ellacuría, I. A. Catalán, and M. Palmer, "Analysis of underwater image processing methods for annotation in deep learning based fish detection," *IEEE Access*, vol. 10, pp. 130359–130372, 2022, doi: 10.1109/ACCESS.2022.3227026.

[15] I. H. Laradji, A. Saleh, P. Rodriguez, D. Nowrouzezahrai, M. R. Azghadi, and D. Vazquez, "Weakly supervised underwater fish segmentation using affinity LCFCN," *Sci. Rep.*, vol. 11, no. 1, Aug. 2021, Art. no. 17379, doi: 10.1038/s41598-021-96610-2.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[19] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[20] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Proc. Comput. Sci.*, vol. 199, pp. 1066–1073, Jan. 2022, doi: 10.1016/j.procs.2022.01.135.

[21] M. Sung, S.-C. Yu, and Y. Girdhar, "Vision based real-time fish detection using convolutional neural network," in *Proc. OCEANS*, Jun. 2017, pp. 1–6, doi: 10.1109/OCEANSE.2017.8084889.

[22] R. Mandal, R. M. Connolly, T. A. Schlacher, and B. Stantic, "Assessing fish abundance from underwater video using deep neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–6, doi: 10.1109/IJCNN.2018.8489482.

[23] Y. Wageeh, H. E.-D. Mohamed, A. Fadl, O. Anas, N. ElMasry, A. Nabil, and A. Atia, "YOLO fish detection with Euclidean tracking in fish farms," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 1, pp. 5–12, Jan. 2021, doi: 10.1007/s12652-020-02847-6.

[24] A. A. Muksit, F. Hasan, M. F. H. B. Emon, M. R. Haque, A. R. Anwary, and S. Shatabda, "YOLO-Fish: A robust fish detection model to detect fish in realistic underwater environment," *Ecol. Informat.*, vol. 72, Dec. 2022, Art. no. 101847, doi: 10.1016/j.ecoinf.2022.101847.

[25] P. Muñoz-Benavent, J. Martínez-Peiró, G. Andreu-García, V. Puig-Pons, V. Espinosa, I. Pérez-Arjona, F. De la Gándara, and A. Ortega, "Impact evaluation of deep learning on image segmentation for automatic bluefin tuna sizing," *Aquacultural Eng.*, vol. 99, Nov. 2022, Art. no. 102299, doi: 10.1016/j.aquaeng.2022.102299.

[26] H. Malik, A. Naeem, S. Hassan, F. Ali, R. A. Naqvi, and D. K. Yon, "Multi-classification deep neural networks for identification of fish species using camera captured images," *PLoS ONE*, vol. 18, no. 4, Apr. 2023, Art. no. e0284992.

[27] F. Xu, H. Wang, J. Peng, and X. Fu, "Scale-aware feature pyramid architecture for marine object detection," *Neural Comput. Appl.*, vol. 33, no. 8, pp. 3637–3653, Apr. 2021, doi: 10.1007/s00521-020-05217-7.

[28] S. Qi, J. Du, M. Wu, H. Yi, L. Tang, T. Qian, and X. Wang, "Underwater small target detection based on deformable convolutional pyramid," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2784–2788, doi: 10.1109/ICASSP43922.2022.9746575.

[29] S. Cai, G. Li, and Y. Shan, "Underwater object detection using collaborative weakly supervision," *Comput. Electr. Eng.*, vol. 102, Sep. 2022, Art. no. 108159, doi: 10.1016/j.compeleceng.2022.108159.

[30] C.-H. Tseng, C.-L. Hsieh, and Y.-F. Kuo, "Automatic measurement of the body length of harvested fish using convolutional neural networks," *Biosyst. Eng.*, vol. 189, pp. 36–47, Jan. 2020, doi: 10.1016/j.biosystemseng.2019.11.002.

[31] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9156–9165.

[32] N. E. Garcia-D'Urso, A. Galan-Cuenca, P. Climent-Pérez, M. Saval-Calvo, J. Azorin-Lopez, and A. Fuster-Guillo, "Efficient instance segmentation using deep learning for species identification in fish markets," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2022, pp. 1–8.

[33] I. A. Catalán, A. Álvarez-Ellacuría, J.-L. Lisani, J. Sánchez, G. Vizoso, A. E. Heinrichs-Maquilón, H. Hinz, J. Alós, M. Signarioli, J. Aguzzi, M. Francescangeli, and M. Palmer, "Automatic detection and classification of coastal Mediterranean fish from underwater images: Good practices for robust training," *Frontiers Mar. Sci.*, vol. 10, Apr. 2023, Art. no. 1151758, doi: 10.3389/fmars.2023.1151758.

[34] P. Sarkar, S. De, and S. Gurung, "A survey on underwater object detection," in *Intelligence Enabled Research—DoSIER*. Singapore: Springer, 2022, pp. 91–104, doi: 10.1007/978-981-19-0489-9_8.

[35] X. Yang, S. Zhang, J. Liu, Q. Gao, S. Dong, and C. Zhou, "Deep learning for smart fish farming: Applications, opportunities and challenges," *Rev. Aquaculture*, vol. 13, no. 1, pp. 66–90, Jan. 2021, doi: 10.1111/raq.12464.

[36] S. Xu, M. Zhang, W. Song, H. Mei, Q. He, and A. Liotta, "A systematic review and analysis of deep learning-based underwater object detection," *Neurocomputing*, vol. 527, pp. 204–232, Mar. 2023.

[37] L. Thampi, R. Thomas, S. Kamal, A. A. Balakrishnan, T. P. M. Haridas, and M. H. Supriya, "Analysis of U-Net based image segmentation model on underwater images of different species of fishes," in *Proc. Int. Symp. Ocean Technol. (SYMPOL)*, Dec. 2021, pp. 1–5, doi: 10.1109/SYMPOL53555.2021.9689415.

[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Munich, Germany: Springer, Oct. 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[39] A. Haider, M. Arsalan, J. Choi, H. Sultan, and K. R. Park, "Robust segmentation of underwater fish based on multi-level feature accumulation," *Frontiers Mar. Sci.*, vol. 9, Oct. 2022, Art. no. 1010565, doi: 10.3389/fmars.2022.1010565.

[40] W. Zhang, C. Wu, and Z. Bao, "DPANet: Dual Pooling-aggregated Attention Network for fish segmentation," *IET Comput. Vis.*, vol. 16, no. 1, pp. 67–82, Feb. 2022, doi: 10.1049/cvi2.12065.

[41] A. Saleh, M. Sheaves, D. Jerry, and M. R. Azghadi, "Transformer-based self-supervised fish segmentation in underwater videos," 2022, *arXiv:2206.05390*.

[42] A. Álvarez-Ellacuría, M. Palmer, I. A. Catalán, and J.-L. Lisani, "Image-based, unsupervised estimation of fish size from commercial landings using deep learning," *ICES J. Mar. Sci.*, vol. 77, no. 4, pp. 1330–1339, Jul. 2020, doi: 10.1093/icesjms/fsz216.

[43] G. French, M. Mackiewicz, M. Fisher, H. Holah, R. Kilburn, N. Campbell, and C. Needle, "Deep neural networks for analysis of fisheries surveillance video and automated monitoring of fish discards," *ICES J. Mar. Sci.*, vol. 77, no. 4, pp. 1340–1353, Jul. 2020, doi: 10.1093/icesjms/fsz149.

[44] R. Garcia, R. Prados, J. Quintana, A. Tempelaar, N. Gracias, S. Rosen, H. Vågstøl, and K. Løvall, "Automatic segmentation of fish using deep learning with application to fish size measurement," *ICES J. Mar. Sci.*, vol. 77, no. 4, pp. 1354–1366, Jul. 2020, doi: 10.1093/icesjms/fsz186.

[45] C. Yu, X. Fan, Z. Hu, X. Xia, Y. Zhao, R. Li, and Y. Bai, "Segmentation and measurement scheme for fish morphological features based on mask R-CNN," *Inf. Process. Agricult.*, vol. 7, no. 4, pp. 523–534, Dec. 2020, doi: 10.1016/j.inpa.2020.01.002.

[46] E. M. Ditria, S. Lopez-Marcano, M. Sievers, E. L. Jinks, C. J. Brown, and R. M. Connolly, "Automating the analysis of fish abundance using object detection: Optimizing animal ecology with deep learning," *Frontiers Mar. Sci.*, vol. 7, p. 429, Jun. 2020, doi: 10.3389/fmars.2020.00429.

[47] Z. Al-Huda, D. Zhai, Y. Yang, and R. N. A. Algburi, "Optimal scale of hierarchical image segmentation with scribbles guidance for weakly supervised semantic segmentation," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 35, no. 10, Aug. 2021, Art. no. 2154026, doi: 10.1142/s0218001421540264.

[48] X. Luo, M. Hu, W. Liao, S. Zhai, T. Song, G. Wang, and S. Zhang, "Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 528–538, doi: 10.1007/978-3-031-16431-6_50.

[49] K. Zhang and X. Zhuang, "CycleMix: A holistic strategy for medical image segmentation from scribble supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11646–11655.

[50] K. Zhang and X. Zhuang, "ZScribbleSeg: Zen and the art of scribble supervised medical image segmentation," 2023, *arXiv:2301.04882*.

[51] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vazquez, and M. Schmidt, "Proposal-based instance segmentation with point supervision," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 2126–2130, doi: 10.1109/ICIP40778.2020.9190782.

[52] R. Dorent, S. Joutard, J. Shapey, A. Kujawa, M. Modat, S. Ourselin, and T. Vercauteren, "Inter extreme points geodesics for end-to-end weakly supervised image segmentation," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*. Strasbourg, France: Springer, Sep. 2021, pp. 615–624, doi: 10.1007/978-3-030-87196-3_57.

[53] Z. Chen, Z. Chen, J. Liu, Q. Zheng, Y. Zhu, Y. Zuo, Z. Wang, X. Guan, Y. Wang, and Y. Li, "Weakly supervised histopathology image segmentation with sparse point annotations," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1673–1685, May 2021, doi: 10.1109/JBHI.2020.3024262.

[54] M. S. Ibrahim, A. Vahdat, M. Ranjbar, and W. G. Macready, "Semi-supervised semantic image segmentation with self-correcting networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12712–12722.

[55] G. K. Mahani, R. Li, N. Evangelou, S. Sotiropolous, P. S. Morgan, A. P. French, and X. Chen, "Bounding box based weakly supervised deep convolutional neural network for medical image segmentation using an uncertainty guided and spatially constrained loss," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–5.

[56] J. Wicaksana, Z. Yan, D. Zhang, X. Huang, H. Wu, X. Yang, and K.-T. Cheng, "FedMix: Mixed supervised federated learning for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 7, pp. 1955–1968, Jul. 2023, doi: 10.1109/TMI.2022.3233405.

[57] Y. Xie, Z. Zhang, S. Chen, and C. Qiu, "Detect, Grow, Seg: A weakly supervision method for medical image segmentation based on bounding box," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105158, doi: 10.1016/j.bspc.2023.105158.

[58] A. Saleh, M. Sheaves, D. Jerry, and M. R. Azghadi, "Unsupervised fish trajectory tracking and segmentation," 2022, *arXiv:2208.10662*.

[59] S. Mittal, J. Niemeijer, J. P. Schäfer, and T. Brox, "Best practices in active learning for semantic segmentation," in *Proc. German Conf. Pattern Recognit. (GCPR)*, 2023, pp. 1–14.

[60] S. A. Golestaneh and K. M. Kitani, "Importance of self-consistency in active learning for semantic segmentation," 2020, *arXiv:2008.01860*.

[61] R. Mackowiak, P. Lenz, O. Ghori, F. Diego, O. Lange, and C. Rother, "CEREALS—Cost-Effective REgion-based Active Learning for Semantic segmentation," 2018, *arXiv:1810.09726*.

[62] K. Sofiiuk, I. A. Petrov, and A. Konushin, "Reviving iterative training with mask guidance for interactive segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 3141–3145.
[63] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023, *arXiv:2304.02643*.
[64] G. Lenczner, A. Chan-Hon-Tong, B. Le Saux, N. Luminari, and G. Le Besnerais, "DIAL: Deep interactive and active learning for semantic segmentation in remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3376–3389, 2022.
[65] J. McGlinchy, B. Johnson, B. Müller, M. Joseph, and J. Diaz, "Application of UNet fully convolutional neural network to impervious surface segmentation in urban environment from high resolution satellite imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 3915–3918.
[66] S. Kartal, "Comparison of semantic segmentation algorithms for the estimation of botanical composition of clover-grass pastures from RGB images," *Ecol. Informat.*, vol. 66, Dec. 2021, Art. no. 101467.
[67] A. Sapkal, Arti, D. Pawar, and P. Singh, "Lane detection techniques for self-driving vehicle: Comprehensive review," *Multimedia Tools Appl.*, vol. 82, no. 22, pp. 33983–34004, Sep. 2023.
[68] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. 4th Int. Workshop Deep Learn. Med. Image Anal. (DLMIA), 8th Int. Workshop Multimodal Learn. Clin. Decis. Support (ML-CDS)*. Granada, Spain: Springer, Sep. 2018, pp. 3–11.
[69] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.
[70] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
[71] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[72] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[73] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, "Conformal risk control," 2022, *arXiv:2208.02814*.
[74] P. Soille, *Morphological Image Analysis: Principles and Applications*, 2nd ed. Berlin, Germany: Springer-Verlag, 2003.
[75] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT++ better real-time instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 1108–1121, Feb. 2022, doi: 10.1109/TPAMI.2020.3014297.
[76] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1280–1289.
[77] M. Francescangeli, S. Marini, E. Martínez, J. Del Río, D. M. Toma, M. Nogueras, and J. Aguzzi, "Image dataset for benchmarking automated fish detection and classification algorithms," *Sci. Data*, vol. 10, no. 1, p. 5, Jan. 2023, doi: 10.1038/s41597-022-01906-1.
[78] M. Zurowietz, 2020, "Annotated marine image datasets S083, S155, S171 and S233," *IEEE Dataset*, doi: 10.21227/vm46-vd05.
[79] E. M. Ditria, C. A. Buelow, M. Gonzalez-Rivero, and R. M. Connolly, "Artificial intelligence and automated monitoring for assisting conservation of marine ecosystems: A perspective," *Frontiers Mar. Sci.*, vol. 9, Jul. 2022, Art. no. 918104, doi: 10.3389/fmars.2022.918104.

JOSEP S. SÁNCHEZ received the M.S. degree in computer engineering and the M.S. degree in artificial intelligence and computer vision from the University of the Balearic Islands, Spain, in 2021 and 2023, respectively. He is currently a Research Staff Member with the University of the Balearic Islands and the Mediterranean Institute of Advanced Studies. His research interests include video object recognition and tools for automatic annotation generation.

JOSE-LUIS LISANI (Member, IEEE) received the Ph.D. degree in computer science and applied mathematics from Universitat de les Illes Balears, Spain, and Paris Dauphine University, France, in 2001. He is currently an Associate Professor with the University of the Balearic Islands, Spain. He is also the co-inventor of five U.S. patents, in collaboration with Cognitech Inc., USA. His research interests include the analysis and processing of color images and video sequences. He has recently focused his research on the combination of image-processing techniques with deep learning methods. He has coauthored more than 15 articles indexed in *JCR* and more than 25 conference papers, one book, and one book chapter. He is an Associate Editor of *Image Processing on Line* (IPOL) and IEEE Geoscience and Remote Sensing Letters.

IGNACIO A. CATALÁN received the M.Sc. degree in fisheries and shellfish culture from the School of Ocean Sciences, Bangor, U.K., in 1998, and the Ph.D. degree in biology from the University of Barcelona, in 2003. He conducted several postdoctoral contracts and stages with UiB, Bergen, Norway, and CSIC, Spain. He was the Vice-Director of the Mediterranean Institute for Advanced Studies (IMEDEA), where he is currently the Head of the Department of Marine Ecology. Since 2009, he has been a tenured Scientist with IMEDEA, a joint center between the Spanish National Research Council (CSIC) and the University of the Balearic Islands (UIB). He has authored more than 75 SCI articles on marine ecology, particularly fisheries oceanography, and has led 15 (three EU) research projects. He has supervised five Ph.D. and several postdoctoral students (including two Marie Curie grantees).

AMAYA ÁLVAREZ-ELLACURÍA received the Ph.D. degree in marine science from the University of the Balearic Islands, in 2010. She was hired as a Technician with the Balearic Islands Coastal Observing and Forecasting System (SOCIB), Spain, from 2010 to 2017. Since 2018, she has been a Technician with the Mediterranean Institute for Advanced Studies (IMEDEA), a joint center between the Spanish National Research Council (CSIC) and the University of the Balearic Islands. She has authored more than 15 SCI articles on beach morphodynamics and fish ecology, focusing the last years on the use of deep learning in fish ecology.