Martech 2024.
Marine Technology Workshop
06, 07th June. Palma, Balearic Islnads, Spain

Automated Mask Generation for Efficient YOLO-based Instance Segmentation in Marine Environments for Fish Detection

Xènia Rovira Coll[1], Antoni Burguera Burguera[2]

[1] *Universitat Oberta de Catalunya (Spain). E-mail: xroviraco@uoc.edu*
[2] *Universitat de les Illes Balears (Spain). E-mail: antoni.burguera@uib.es*

***Abstract* – This paper addresses the laborious, time-consuming and error-prone process of generating ground truth data to perform instance segmentation of fish in their natural habitat. Our proposal is to use the *Segment Anything Model* (SAM), which allows zero-shot inference, to automatically build the segmentation masks, significantly reducing the dataset creation time and enhancing scalability to larger datasets. Experimental results using *You Only Look Once* (YOLO) demonstrate only marginal performance differences between our approach –with segmentation masks created with no human intervention– and a standard training using a fully human-labeled dataset. The results underscore the effectiveness of the automated workflow discussed herein, showcasing substantial reduction in dataset creation time, particularly in demanding underwater scenarios.**

***Keywords* -** *Instance segmentation;deep learning;SAM;YOLO*

I. MOTIVATION AND TECHNICAL BACKGROUND

Object detection and instance segmentation are related computer vision tasks focused on identifying and localizing objects within images. Object detection involves recognizing and labeling objects while providing bounding boxes around them. Instance segmentation takes this a step further by not only classifying objects but also precisely detecting the boundaries of each instance at the pixel level.

The first successful object detector was Viola-Jones [1], which classified a sliding window over the entire image. The rise of deep learning lead to approaches relying on *Convolutional Neural Networks* (CNN) [2], which resulted in large processing times. Even though further approaches, such as *Faster R-CNN* [3], concentrated on the inference speed, they were soon outperformed by *You Only Look Once* (YOLO) [4]. Only models with added functionalities, like *Mask R-CNN* [5], capable of instance segmentation, managed to stay competitive. Notably, YOLO has integrated support for instance segmentation as well since September 2023.

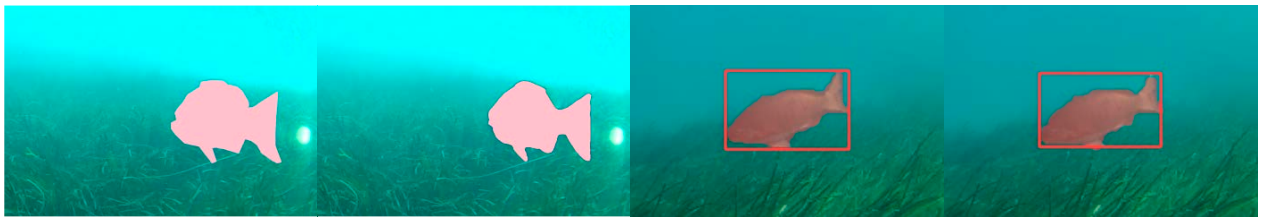
Labeling images for instance segmentation training is time-consuming and error-prone, since pixel-level annotations are required for each instance. This is particularly challenging in underwater environments where poor visibility and unfamiliar visual structures add up to the general problems of subjective interpretation, visual ambiguities or occlusions. Thus, automating this labeling process is crucial.

The *Segment Anything Model* (SAM) [6], introduced in April 2023, is a notable pre-trained model with zero-shot inference capabilities, allowing segmentation of objects without explicit training. SAM's performance improves with additional information such as bounding boxes. Since this additional information is available when creating instance segmentation datasets from object detection ones, SAM seems to be the perfect choice to automate the segmentation mask creation.

Our proposal is to use SAM to enrich an object detection dataset with segmentation masks in order to be useable to train an instance segmentation model. In this way, the human intervention in limited to bounding box labeling, which is much faster, and the remaining data is automatically created from the images and the bounding boxes. Also, existing object detection datasets can be automatically transformed into instance segmentation ones with no human intervention. More specifically, we will focus on the special case of fish detection and segmentation, which poses additional difficulties due to the particulariries of underwater environments, using YOLOv8.

II. EXPERIMENTAL RESULTS

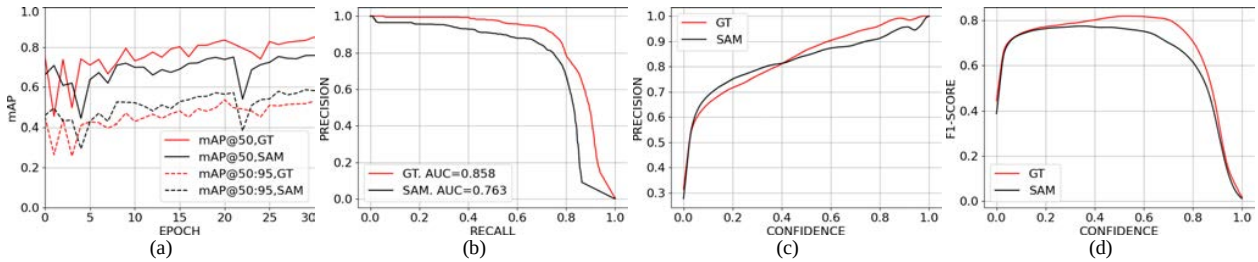
Our proposal has been tested using the Luderick-Seagrass dataset [7], which comprises annotated footage of fish in two estuary systems in South East Queensland, Australia. The dataset contains 9429 fish annotations distributed in 4280 images. The annotations include both the bounding box and the segmentation mask. The former have been used to feed SAM in order to automatically build the new segmentation masks. These new segmentation masks have been used to train YOLOv8 for instance segmentation during 30 epochs. The original, hand-labeled, segmentation mask has been solely used to evaluate the quality of the SAM generated ones.



(a) (b) (c) (d)

Fig 1. Segmentation masks. (a) Hand-labeled, (b) SAM, (c) YOLO trained with hand-labeled masks and (d) YOLO trained with SAM-generated masks.

Figure 1 exemplifies hand-labeled segmentation masks (Figure 1-a) and SAM-generated masks (Figure 1-b), being nearly identical with SAM actually capturing the fish contours more accurately. Instance segmentation with YOLOv8's trained with hand-labeled data (Figure 1-c) versus trained with SAM-generated masks (Figure 1-d) shows minimal differences, making it difficult to tell which one is better.



(a) (b) (c) (d)

Fig 2. Quality metrics. (a) Evolution of mAP during training. (b) Recall-Precision curve. (c) Confidence-Precision curve. (d) Confidence-F1-Score curve.

Figure 2 displays instance segmentation quality metrics. In Figure 2-a, validation mAP@0.5 and mAP@0.5:0.95 trends across training epochs are compared using hand-labeled masks (ground truth or GT) and SAM-generated ones. While GT performs slightly better at mAP@0.5, SAM based training excels in capturing fine details (mAP@0.5:0.95). In Figure 2-b, the Recall-Precision curve after 30 epochs indicates a small decrease in mAP@0.5 with SAM with respect to GT. Figures 2-c and 2-d depict Precision and F1-Score evolution with respect to the confidence score threshold, showing the use of hand-labeled and SAM-based masks yield similar results, though SAM exhibits slightly worse overall metrics.

Overall, the proposed approach enables fully automated segmentation mask creation, demonstrating comparable quality in YOLOv8 training with hand-labeled images. While quantitative results show a slight dip when using automated masks, visual inspection suggests comparable, even superior, performance. Discrepancies between visual inspection and quantitative data may stem from biases in the evaluation ground truth, which was hand-labeled. In essence, SAM facilitates automated mask creation, enhancing YOLOv8 training without compromising quality.

ACKNOWLEDGMENTS

This work is partially supported by Grant PID2020-115332RB-C33 funded by MCIN/AEI/10.13039/501100011033 by "ERDF A way of making Europe" and by Grant PLEC2021-007525/AEI/10.13039/501100011033 funded by the Agencia Estatal de Investigacion, under NextGeneration EU/PRTR

REFERENCES

[1] Viola, P. and Jones, M. (2001). *Rapid object detection using a boosted cascade of simple features*. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 1.

[2] Sermanet, P. and Eigen, D. and Zhang, X. and Mathieu, M. and Fergus, R. and LeCun, Y. *Overfeat: Integrated recognition, localization and detection using convolutional networks.* arXiv:1312.6229 [cs.CV], 2013.

[3] Ren, S. and He, K. and Girshick, R. and Sun, J. *Faster R-CNN: Towards real-time object detection with region proposal networks.* Advances in Neural Information Processing Systems, 2015.

[4] Redmon. J. and Divvala, S. and Girshick, R. and Farhadi, Ali. *You Only Look Once: Unified, Real-Time Object Detection.* Proceedings of the 2016 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2016.

[5] He, K. and Gkioxari, G. and Dollár, P. and Girshick, R. *Mask R-CNN* Proceedings of IEEE International Conference on Computer Vision, ICCV 2017

[6] Kirillov, A. and Mintun, E. and Ravi, N. and Mao, H. and Rolland, C. and Gustafson, L. and Xiao, T. and Whitehead, S. and Berg, A.C. and Lo, W. and Dollár, P. and Girshick, R. *Segment Anything*. arXiv:2304.02643 [cs.CV], 2023

[7] Ditria, E.M. and Connolly, R.M. and Jinks, E.L. and López-Marcano, S., *Annotated video footage for automated identification and counting of fish in unconstrained seagrass habitats*, Frontiers in Marine Science, vol. 8, ISSN 2296-7745, 2021.