Aquaculture 626 (2027) 744459

Contents lists available at ScienceDirect

Aquaculture

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

Automatic visual detection of fish in Recirculated Aquaculture Systems using the Segment Anything Model

Hilla Fred [a,*], Mogens Agerbo Krogh [b], Britt Bang Jensen [c,b], Laura Ruotsalainen [d], Jouni Vielma [a], Matti Pastell [a]

[a] Natural Resources Institute Finland, Helsinki, 00790, Finland
[b] Department of Animal and Veterinary Sciences, Aarhus University, Tjele, DK-8830, Denmark
[c] Norwegian Veterinary Institute, Ås, 1404, Norway
[d] Department of Computer Science, University of Helsinki, Helsinki, 00014, Finland

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.15528511

Keywords:
Segmentation
Computer vision
RAS
Aquaculture
Fish welfare

ABSTRACT

Detecting individual animals is an important step for building monitoring tools for animal based welfare indicators. Latest advances in foundation models (FM) have shown promising performance in zero-shot segmentation over various domains. Here, we show that an FM Segment Anything Model (SAM) can be used to find rainbow trout (*Oncorhynchus mykiss*) and Atlantic salmon (*Salmo salar*) in a real-life recirculated aquaculture system (RAS) commercial production environment with a stocking density of $32-40\,\mathrm{kg/m^3}$, without the need for costly training data. We evaluate two prompting methods over a benchmark containing 88 manually verified images and thousands of masks, comparing these hand-annotated prompts with fully automatic segmentation. We also propose a fast annotation workflow utilizing a lightweight classifier to refine the set of output masks, improving the proportion of good-quality whole fish or head masks for automatic detection from 60.5% to 77.7%, and for hand-annotated prompts from 75.6% to 85.2%.

1. Introduction

The global need for fish protein for human consumption continues to grow, and because fishing is already at the maximum capacity, aquaculture is expected to cover the increased demand in the future (EC, 2020). Recirculating aquaculture systems (RAS) are a scalable solution to increase fish production. As the whole environment in RAS is controlled by humans, strict monitoring systems are necessary to ensure health and welfare of the fish. Inferior health and welfare decreases profitability and leads to a suboptimal use of resources. Therefore, improvements in fish health and welfare practices are key for achieving optimal and sustainable fish farming (Macaulay et al., 2021). During the last decade, computer vision and advances in deep learning have provided ample opportunities for intensive monitoring in animal production, with most of the focus being on terrestrial animals (Borges Oliveira et al., 2021), but showing great potential also for aquaculture (Yang et al., 2021; Fitzgerald et al., 2025).

Animal based measures (ABM) are the best indicators of animal welfare (EFSA AHAW Panel, 2012). For two species commonly farmed in the Nordic countries, rainbow trout (*Oncorhynchus mykiss*) and Atlantic salmon (*Salmo salar*), such ABM include external morphological damage such as skin lesions, eye, gill and fin damage, abnormalities in body shape, and swimming-related indicators such as activity level, schooling behaviour, aberrant swimming and loss of balance, aggression, apathy and stereotypic movement (Pettersen et al., 2014; Noble et al., 2020). Detecting individual fish from video data and deriving these indicators automatically in a commercial RAS environment would be a significant building block for fish welfare monitoring, as expert observations of these indicators are very costly and time-consuming to do in production facilities. In a review, Fitzgerald et al. found that the second most common use of computer vision applications in aquaculture was behavioural monitoring (Fitzgerald et al., 2025).

Segmentation of fish in tanks is a special computer vision task, because of multiple similar-sized objects of interest densely scattered against the background. The fish are of the same species, of roughly similar size and the area of interest is restricted. This makes the task a well-suited candidate for experiments with general-purpose models showing good zero-shot performance. Segment Anything Model (SAM) (Kirillov et al., 2023) is the first foundation model proposed for image segmentation, showing promising results over a wide array of tasks without additional training (Wang et al., 2024). Previous work on fish instance segmentation is largely based on either background–foreground subtraction (Zhao et al., 2016; Zhou et al., 2018; Yu et al.,

* Corresponding author.
E-mail address: hilla.fred@luke.fi (H. Fred).

https://doi.org/10.1016/j.aquaculture.2026.744459
Received 3 June 2025; Received in revised form 3 June 2026; Accepted 2 August 2026
Available online 6 August 2026

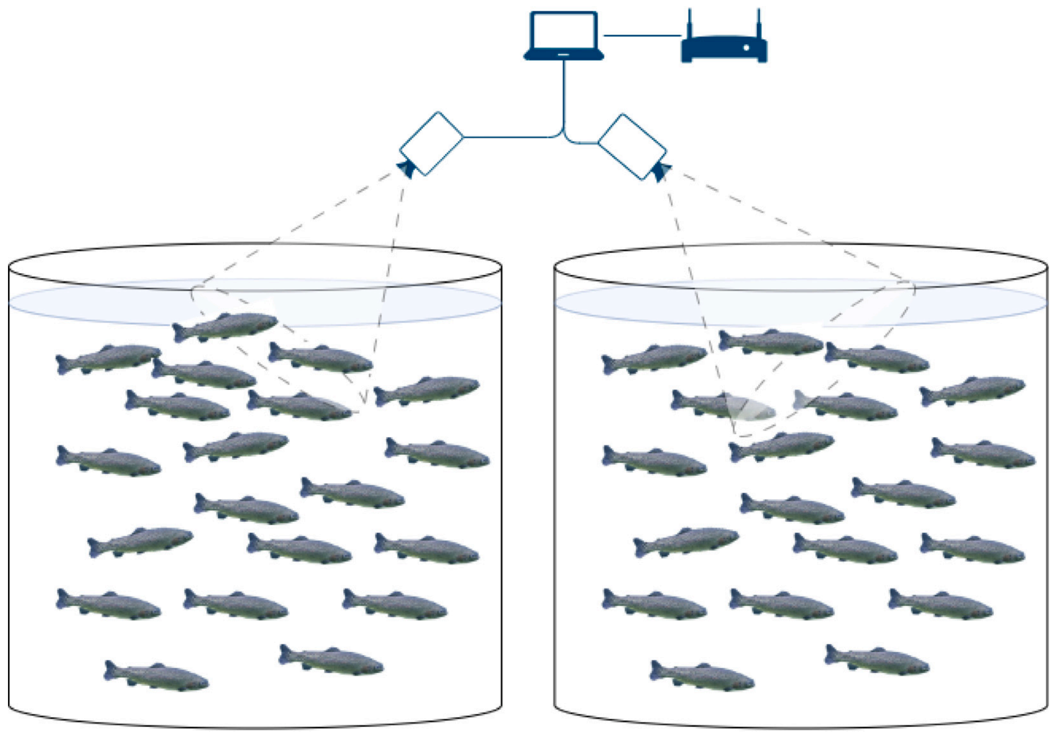

Fig. 1. In farm 1, the cameras were connected to Jetson Nanos, from which the video data was streamed to a mini-pc located close to the tanks. The video data was further transferred to a cloud storage for offline analysis and model training.

2021), a traditional computer vision method that requires a low stocking density and a static background, or a supervised approach based on neural networks, using manually annotated data to train a model (L. Zhang et al., 2024). Manual annotation is costly, which has somewhat been mitigated by using weakly supervised learning and various annotation tools (Laradji et al., 2021). The strong zero-shot performance of SAM is based on its architecture utilizing recent advances in deep learning and the large scale of its training dataset (11M images, 1B masks). The model architecture is based on Transformers, best known as the building blocks behind modern Large Language Models (LLMs) (Vaswani et al., 2017). Vision Transformers (ViTs), Transformers adapted for visual tasks, are a higher capacity alternative to previous state-of-the-art architectural paradigm Convolutional Neural Networks (CNNs) (Dosovitskiy et al., 2021).

Therefore, we experiment with SAM and compare its different prompting approaches using both annotated data and a fully automatic segmentation mode. As part of the publication, we release a manually verified segmentation dataset consisting of 88 images. To our knowledge, datasets from large-scale commercial RAS environment have not been published for fish instance segmentation before. Previous experiments using computer vision in a RAS environment, including detecting (Wang et al., 2022), segmenting (L. Zhang et al., 2024) and tracking fish (Liu et al., 2024; Li et al., 2024; Ciani et al., 2024), classifying appetite (Razman et al., 2019) and detecting dead fish (Zhou et al., 2025), have typically been carried out at small aquariums and tanks, 100 – 1000 times smaller in scale and intensity.

The aim of the paper was to study if fish detection using SAM-based automatic instance segmentation is applicable in a large-scale commercial production environment. We evaluated SAM on the fish object detection task and conducted a categorical assessment of the mask quality. In addition, we developed a lightweight post-processing workflow leveraging SAM and a mask shape-based classifier, showing a clear precision-increasing trade-off on both on automatic and hand-annotated prompts.

2. Materials and methods

Image data was collected from two commercial Nordic RAS farms, one raising rainbow trout in farm 1, and one raising Atlantic salmon in farm 2. From both farms, RGB cameras were installed over the tank surface to record over several days' time period, and the resulting image dataset was manually annotated.

2.1. Data collection

2.1.1. Image data

In farm 1, Alvium 1800 U-240c cameras were set up 2 m above the surface of two 540 m^3 tanks with a height of 5.5 m and 12 m in diameter containing rainbow trout (Fig. 1). During the observation period, the estimated mean weight of the stock was 300 – 350 g with a stocking density of 32 – 38 kg/m^3. Cameras were placed and oriented to avoid occlusion caused by foam and reflections from the light sources, and to ensure view of the constantly lit feed landing area from the centralized pipe feeding system. Camera settings were adjusted for better image quality, deciding on exposure time of 33 ms, gain of 18 dB, continuous white balance adjustment and a frame rate of 27 FPS. In total, 470 h of semi-continuous video data was collected in June 2022, recording two minutes at a time at varying intervals over a 12-day period. Images were extracted from a selection of the videos, using a minimum interval of 100 ms to avoid strong temporal correlation between the frames, resulting in 1501 RGB images with a size of 1600 × 999 from Camera 1 and 369 images of size 1450 × 1100 from Camera 2.

At farm 1, recirculating water goes through drum filters to remove solids, fixed-bed biofilters to transfer ammonia to nitrate, and a trickling filter to remove CO_2. The water is treated with UV before reaching the fish tanks. NaOH and liquid oxygen are added to maintain pH and oxygen at appropriate levels. Temperature is typically maintained at 16 – 17 °C. The centralized pipe feeding system provides feed 10 – 14 times per day.

In farm 2, one Ubiquiti Unifi G4 Pro camera was set up 4 m above the surface of a circular concrete tank with a depth of 3 m and a

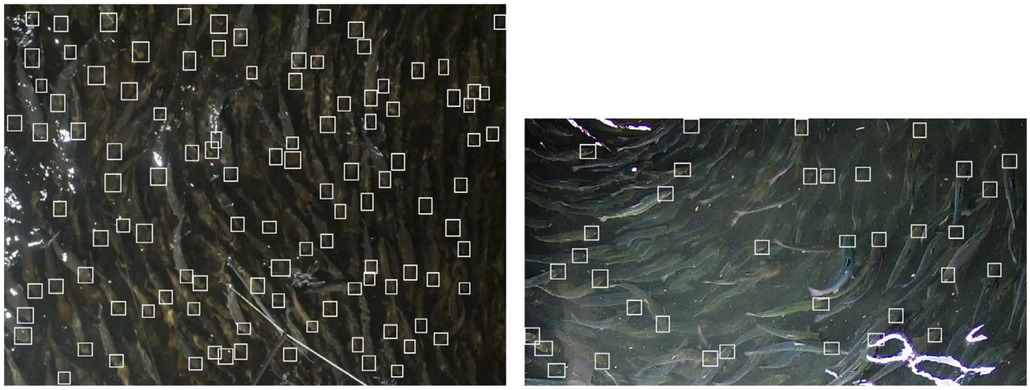

Fig. 2. Annotated fish heads from Farm 1 (left) and from Farm 2 (right).

diameter of 7 m containing Atlantic salmon. The salmon had an average size of 475 g and a density of 40 kg of biomass per cubic metre. In total, 36 h of semi-continuous video data was collected in the spring of 2022, recording five minutes at a time at varying intervals over a 19-day period. From the 24 FPS high-resolution video, two interesting non-overlapping sublocations with less noise due to reflections and foam were identified and cropped. Images were extracted from the videos, using a minimum interval of 100 ms to avoid strong temporal correlation between the frames, resulting in 370 extracted RGB images of size 2000×1159 and 414 images of size 1600×850.

At farm 2, intake water is sourced directly from the sea and undergoes no further treatment than filtering and UV-treatment before entering the system. The water has a salinity of around 30 g/L. Within the recirculation system, the water is treated in fixed-bed biofilters and reoxygenated to 85 – 100 percent saturation (around 10 mg/L) with a tank diffuser and high pressure cones. Temperature is typically maintained at 12 °C. Sprinkle feeders provided feed continuously during a 16 time window each day and artificial light were on for 24 h/day. Ammonium and carbon-dioxide concentrations and pH are monitored daily with classical chemical methods. Alkalinity were controlled using sodium hydroxide (NaOH) to keep an average pH of 7.15. Carbodioxide concentrations were on average 12 mg/L.

2.1.2. Manual annotation

We used a manually annotated fish object detection dataset in our experiments. The detection annotation task involved drawing rectangular bounding boxes around the heads of the fish to mark all fish in the images (Fig. 2). Three annotators conducted their work independently over several months, using a concise guidebook that included visual examples, authored by one of the annotators. All utilized the open-source MakeSense (Skalski, 2019) online annotation tool for their tasks.

By focusing on fish heads rather than entire bodies, the annotators aimed to minimize variations in bounding box size resulting from differences in fish orientation and camera angles. Identifying all fish consistently posed a challenge due to variable lighting and high densities of fish within the tanks. The resulting annotated dataset consisted of 1684 images from both farms. A subset of the data was used as model input in our experiments. No additional manual annotation was performed; in particular, no segmentation masks were manually created.

Inter-annotator agreement was assessed on a separate subset of 76 images that was sampled to reflect the camera distribution of the evaluation data. The three annotators independently labelled fish head bounding boxes on the same images. Then, inter-annotator agreement was evaluated using precision, recall, F_1 and mIoU:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{Area of Overlap}_i}{\text{Area of Union}_i} \tag{4}$$

In these object detection metrics, predicted box is considered a True Positive (TP) if it has an Intersection over Union (IoU) above a given threshold with a ground truth (GT) box. False Positives (FP) are predicted boxes that do not have a matching GT box based on this criterion, and False Negatives (FN) are GT boxes that do not have a matching predicted box.

When evaluating model performance, these object detection metrics are computed by comparing predicted bounding boxes to manually annotated GT bounding boxes. We calculated inter-annotator agreement by doing a pairwise comparison over the annotator pairs, treating one as the "predicted" set and the other as the "ground truth" set. Precision, recall, F_1 and mIoU between the pairs were then computed using one-to-one bounding-box matching at different IoU thresholds. All bounding boxes were matched only once, and the matching was performed in descending IoU order.

Overall, the curve suggested highest object-level agreement at IoU $\leq$ 0.30, with localization differences becoming more influential at higher IoU thresholds (Fig. 3). Overall agreement was highest between Annotators A and B (F1 = 0.809), while comparisons involving Annotator C showed lower recall due to fewer annotated objects (Table 1).

2.2. Computational experiments

2.2.1. Automatic segmentation

The encoder–decoder architecture of SAM consists of a heavy image encoder and a more lightweight prompt encoder and mask decoder (Kirillov et al., 2023). The authors leverage this architectural design when training the model, as the image only needs to be encoded once, and any subsequent prompting of that image uses less than 1% of the image encoder's compute time. This enables iterative predictions for mask refinement during training, improving model performance (Kirillov et al., 2023). The image encoder used is a ViT model Masked Autoencoder (ViTMAE) (He et al., 2022), whose training is based on self-supervision by masking random patches of the input image and reconstructing the missing parts. This self-supervision serves as a starting point for multi-stage model training, which together with the architectural choices allow for the large-scale training needed for a foundation model. In addition, the authors of SAM define prompt in segmentation context to mean any information that specifies what to segment in the image (Kirillov et al., 2023), e.g. a point or a bounding box, which makes SAM flexible in terms of model input.

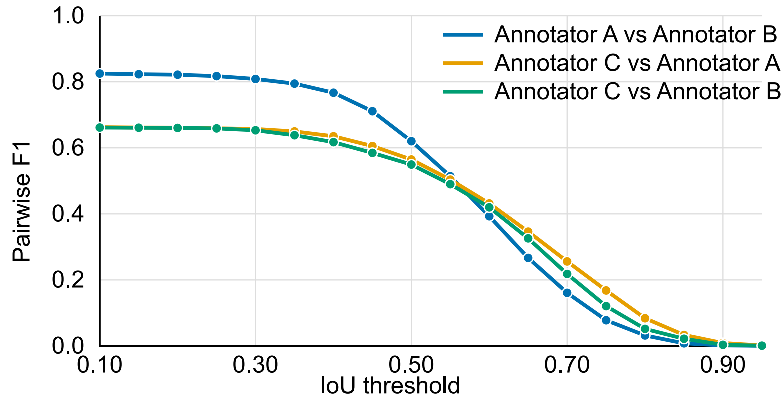

Fig. 3. Inter-annotator agreement across IoU thresholds.

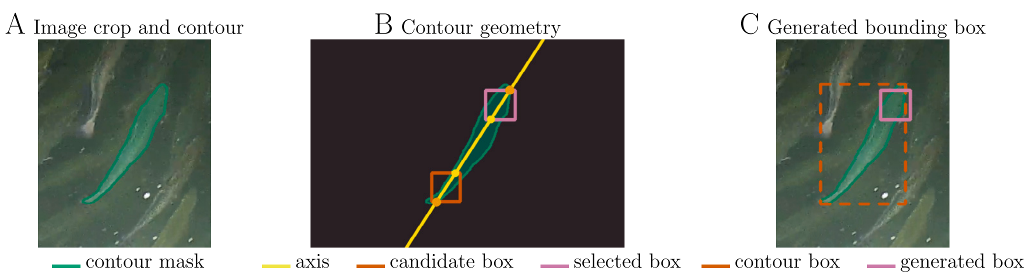

Fig. 4. Bounding-box generation workflow. We estimate the fish long axis from the mask contour, generate two candidate square boxes at opposite ends of the mask using an anatomy-informed inward fraction of 0.25, and select the candidate containing the greater mask pixel mass.

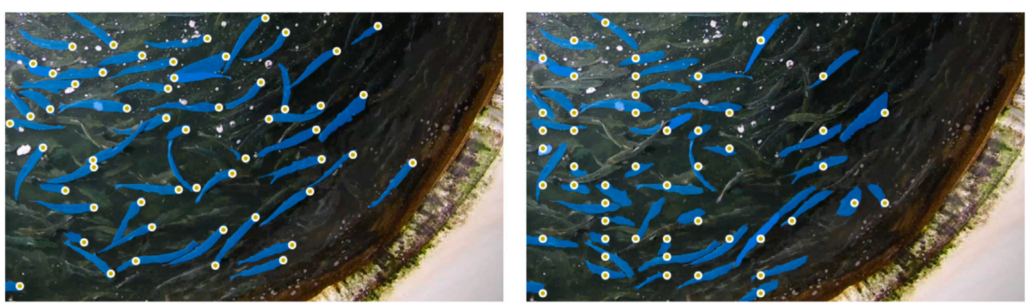

Fig. 5. Prompts (dots) and resulting segmentation masks (blue). Very small masks with an area smaller than 1000 pixels and very large masks with an area larger than 25000 pixels and the corresponding prompts have been removed for clarity, as well as grid points that did not produce a mask in automatic segmentation due to low model confidence. Data from Farm 2. Left: Hand annotated prompts (dots) and resulting segmentation masks. Right: Grid of points as prompts (dots) and resulting segmentation masks.

Table 1
Pairwise inter-annotator agreement on a sampled 76-image subset. Object-level matches were computed using same-class bounding boxes with IoU $\geq$ 0.30.

Annotators	Images	N_A	N_B	Precision	Recall	F1	Mean IoU
A vs B	76	4702	4520	0.793	0.825	0.809	0.594
C vs A	76	2436	4702	0.963	0.499	0.657	0.652
C vs B	76	2436	4520	0.933	0.503	0.653	0.635

Utilizing the property of the model to be prompted by any set of points in the image, we compared the performance of SAM when it is prompted using hand-annotated points locating the fish, and when it is prompted using "automatic segmentation", an evenly sampled grid of points, a method proposed by Kirillov et al. (2023), without any prior knowledge of the locations of the fish.

Using 88 randomly split images from the dataset described in Section 2.1.2, SAM was prompted with the centre points of the N=5,828 annotated bounding boxes. Another set of masks for the same images was produced using automatic segmentation (Fig. 5). Suitable parameters for the automatic segmentation were chosen manually, by visually inspecting the outputs of a coarse grid search described in Appendix A. The largest image encoder variant ViT-H was used in all the experiments.

The data used in this study had originally been annotated for a fully supervised fish head object detection task. As a result, there was a systematic scale mismatch between the detection ground truth, which consisted of head bounding boxes, and the SAM predictions, which typically segmented the whole fish. To report standard object detection metrics, we derived head bounding boxes automatically from the predicted segmentation masks using a contour-guided head box generation procedure (Fig. 4). As a parameter, we use a biologically informed

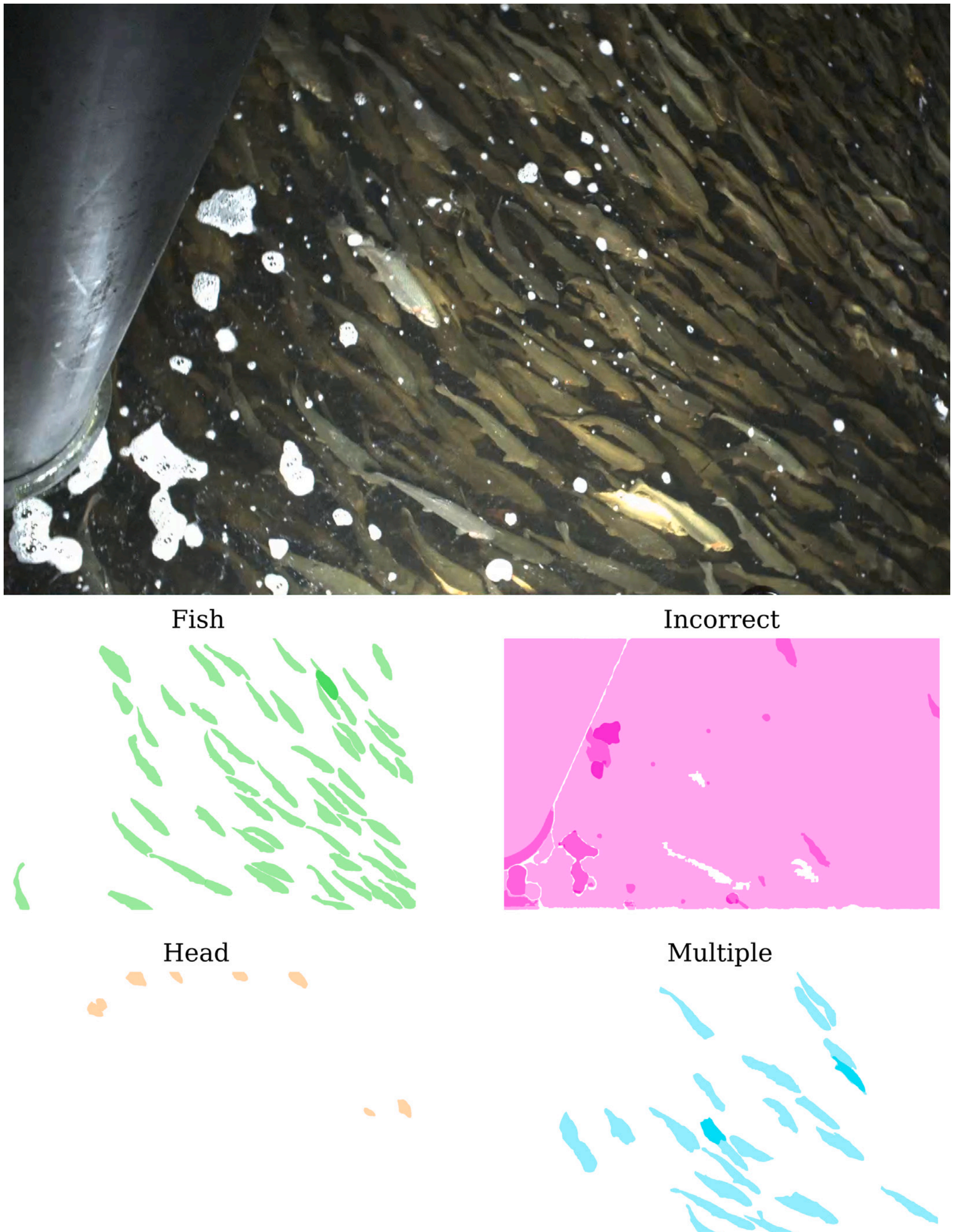

Fig. 6. Examples of the mask classes before any post-processing.

inward fraction of 0.25: Atlantic Salmon head length is approximately 20% of fork length (Garber et al., 2019), and we use a slightly larger value to account for the fact that human-annotated detection boxes are not necessarily tight around the head. This pre-processing step was designed for the present imaging setup to compute standard detection metrics, and may be less reliable for fish with substantially different proportions, for example due to age, body condition, or camera viewpoint. Full details and a sensitivity analysis of the fraction parameter are provided in Appendix A.

Additionally, to evaluate the performance of the two different prompting strategies independent of annotator recall and bounding box localization, and provide qualitative insight mask quality and consistency, both segmentation outputs were manually reviewed, and the masks were assigned a class based on their quality (Fig. 6). The criteria for the classes were as follows:

- ‘incorrect’: Non-fish, or a rugged (non-smooth) or multi-part mask, with no real occlusion to justify its roughness or patchiness, or a very small mask covering only a portion of the fish or its head
- ‘multiple’: A smooth mask covering multiple fish (e.g. a head and another whole fish, or two whole fish)
- ’head: A smooth mask covering the head of the fish, but less than 50% of the body of the fish
- ‘fish’: A smooth mask covering exactly one fish, or more than 50% of a fish

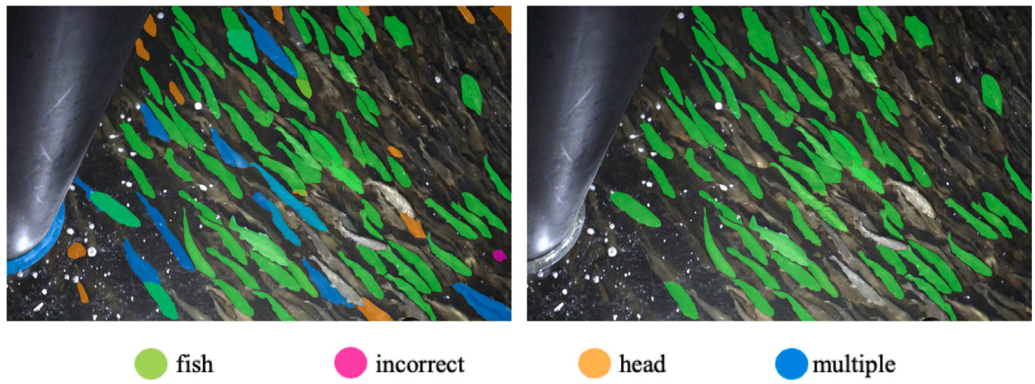

Fig. 7. Filtering undesired masks using a classifier, data from Farm 1. Left: Output of automatic segmentation and class predictions produced by our post-processing method. Right: The proposed post-processing applied, removing all masks that are not predicted to be of good quality 'fish' masks by the model. Overlapping masks are coloured in brighter green.

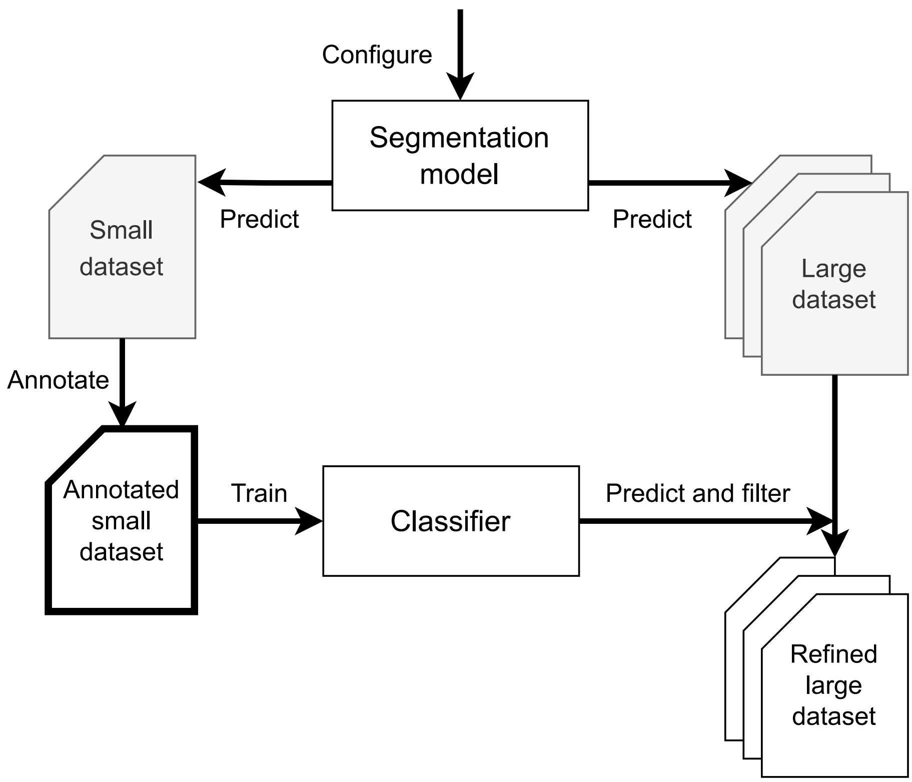

Fig. 8. Our proposed post-processing workflow to refine the automatic segmentation output.

2.2.2. Post-processing

With either prompting method, some of the resulting masks were low quality. They contained ragged edges, multi-part masks, multiple fish or no fish at all, and were difficult or impossible to filter using heuristic rules based on mask size and shape. To reduce the number of incorrect masks from the dataset, we propose to use a classifier to filter out the low quality masks using extracted mask features (Algorithm 1, Appendix A).

Assigning the possibly undesirable masks to different classes provides more information on the underlying data and enables more granular post-processing, as filtering requirements can depend on the application at hand (Fig. 7). For example, in tracking tasks it might be beneficial to include heads as well as whole fish masks, as there is substantial occlusion in commercial fish farming tanks due to stocking density. Similarly, 'multiple' masks could be used to help infer tracks with overlapping detections.

The proposed workflow is described in Fig. 8. First, reasonable parameters for automatic segmentation are chosen in the configuration phase. Then, a small set of segmentation masks are produced using automatic segmentation, annotated by hand and used to train a classifier. The trained classifier is then applied to a larger dataset, and the masks that are predicted to be of undesired class are filtered out. The proposed workflow can be plugged in with any classifier. In our experiments we compared three different non-linear classifiers, which we trained on a combination of simple global descriptors, convexity defects and invariant moments.

The models were first evaluated comparing their one-vs-all binary classification accuracies for the different classes. Binary accuracy, fraction of correct predictions, gives insight into the class-specific performance of the model. In addition, the proposed post-processing was applied by using a multi-class classifier trained to distinguish all of the classes from each other. Mask precision %, i.e. the true proportion of

Table 2
Detection results at IoU 0.30 without post-processing. Bounding boxes are derived from the segmentation mask with the contour-guided generation.

Prompt	N predictions	TP	FP	FN	Precision	Recall	F1	Mean IoU
Manual	5307	3183	2124	2641	0.600	0.547	0.572	0.485
Automatic	6376	2103	4273	3721	0.330	0.361	0.345	0.488

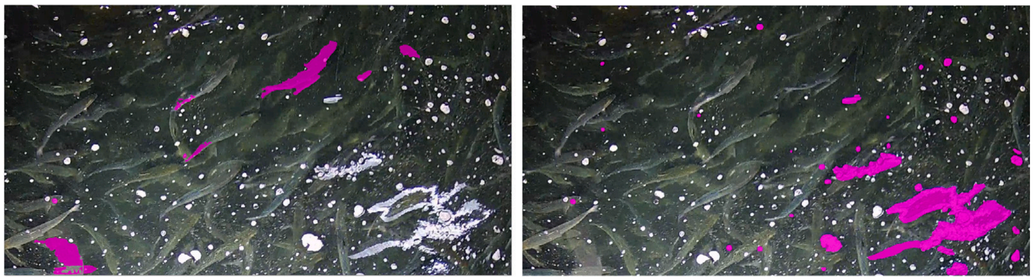

Fig. 9. Example of differences in 'incorrect' masks between prompting methods from Farm 2. Left: Manual prompts. Right: Automatic segmentation. Large amount of surface disturbance combined with the grid prompt caused the reflections and bits of foam to be segmented.

the desired class in the data before and after filtering, was calculated by comparing the manually annotated class label and predicted labels for a given mask on the left-out test set:

$$\text{Mask Precision } \% = \frac{TP}{TP + FP} \cdot 100\%, \quad (5)$$

where True Positives (TP) are predictions correctly assigned to the target class, and False Positives (FP) are predictions incorrectly assigned to that class.

The focus on precision in the evaluation metrics stems from the observation that, when monitoring fish in RAS tanks, the group-level metrics from the detected fish may serve as an indicator of the group. The fraction of fish visible at the surface compared to the total number of fish in the tank is very small, also in human visual observation, making image-level recall less relevant when considering the scale of the fish monitoring problem in RAS. Mask precision % does not require bounding box localization with a matching human-produced bounding box annotation, and the precision depends only on the manually verified class labels. In addition, precision of the multiclass classifier reflects the method's applicability in real-life scenarios in cases where multiple classes need to be considered simultaneously. Because we evaluated the mask quality based on coarse categorical assessment, segmentation metrics that depend on pixel accuracy were not applicable to this data.

3. Results

3.1. Automatic segmentation

Manual prompts produced significantly better results than automatic, with an F_1 score of 0.57 (Table 2). The manual annotation prompts are a strong prior for object localization, and the resulting difference of manual to the original bounding boxes can be attributed to SAM and the contour-based bounding box generation. Automatic mode does not have any training for the specific task or prior information on fish locations, and is relying solely on pre-training and the visual structure of the image, resulting in significantly lower precision and recall, with an F_1 score of 0.35

Automatic segmentation produced more masks overall, with 72 masks on average per image, while having 49.9% of the masks with a confirmed label of 'fish' (Table 3). A notable difference between the prompting methods was the proportion of 'incorrect' masks produced. Automatic segmentation produced more masks on surface disturbance, as the evenly distributed point grid sampling is likely to hit a location where these are present (Fig. 9). Using automatic segmentation, 20.0% of masks were of 'incorrect' class.

Table 3
Distribution of manually confirmed mask classes produced using two different mask generation methods with SAM on 88 images without post-processing.

Prompt	N	Fish	Incorrect	Head	Multiple
Manual	5307	62.0%	6.8%	13.5%	17.7%
Automatic	6376	49.9%	20.0%	10.6%	19.5%

With manually annotated prompts, the percentage of 'fish' masks was 62.0% (Table 3), with an average number of 60 masks per image. There was a higher proportion of 'head' masks, which was to be expected due to the prompts being located directly on the fish heads. Human annotators were also less likely to mark a bounding box on an area with low visibility, resulting in a smaller quantity of 'incorrect' masks. For one image with low visibility due to surface disturbance, both prompting methods produced zero masks.

3.2. Post-processing

We evaluated the proposed method using both sets of the manually classified masks from the 88 images. Features were extracted from the mask shapes: area, major and minor axis lengths, number of submasks, number of convexity defects and Hu moment invariants (Hu, 1962) (Fig. 10). The extracted features were used to train three different non-linear classifiers: K-Nearest Neighbours (KNN), Multilayer perceptron (MLP) and Random Forest (RF). The classifiers were configured as follows: KNN($n_{neighbours}$ = 3), RF(max_depth = 10, $n_{estimators}$ = 20, max_features = 3), MLP(α = 0.05, max_iter = 5000).

For the dataset with hand-annotated prompts, mean test accuracy over the three evaluated classifiers for distinguishing 'fish' masks from other classes with a one-vs-all binary classification scheme was 82.7% (Table 4). On the automatically produced masks, the mean test accuracy for one-vs-all binary classification was slightly lower, especially on the accuracy on 'incorrect' and 'head' masks (Table 4). The accuracy of 'fish' masks over the evaluated classifiers on automatic prompts was 82.4%. On both prompting modes, correctly classifying 'multiple' masks was a difficult task due to the close similarity of 'fish' and 'multiple' classes in cases where the fish are swimming closely parallel, leading to more confusion between these classes.

Overall, there was little difference in the performance between the different evaluated classification models on either prompting methods, with KNN having a somewhat lower accuracy compared to the other two models. For clarity, we have reported only RF in Table 5, as it achieves the highest accuracy on automatic 'fish' binary classification,

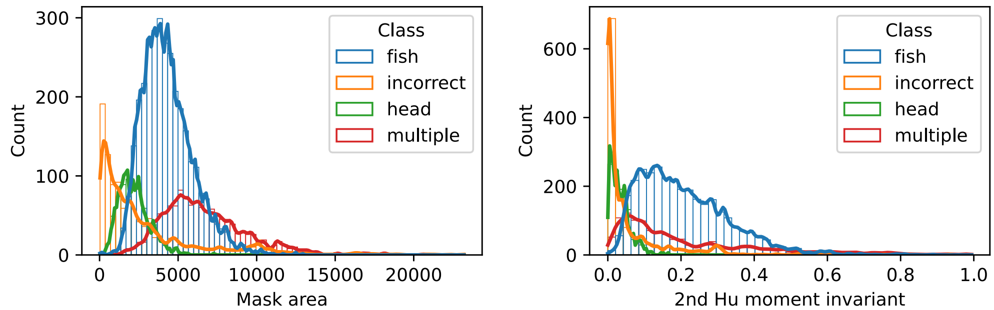

Fig. 10. Kernel density estimation (KDE) over histogram of mask area sizes (left) and second Hu moment invariant (right) for the different manually confirmed mask classes in automatic segmentation. Bandwidth adjustment factor of 0.25 is used to retain the characteristics of the distributions. Outliers (i.e. masks with pixel area greater than 25000 and moment invariants greater than 1.0) are omitted for clarity.

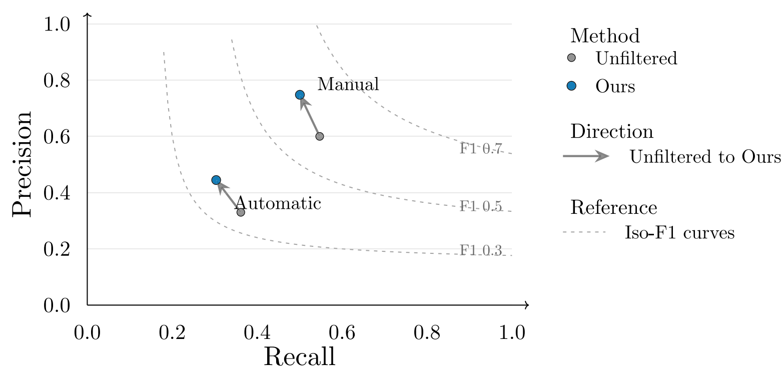

Fig. 11. Precision–recall at IoU 0.30. Our post-processing consistently improves precision but lowers recall.

Table 4
One-vs-all binary classification mean test accuracy and its standard deviation for the different models and classes over a stratified 5-fold cross-validation, compared to manually confirmed mask labels.

Prompt	Model	Fish %	Incorrect %	Head %	Multiple %
Manual	**KNN**	80.3 ± 1.0	94.7 ± 0.5	95.2 ± 0.9	83.7 ± 1.0
	MLP	84.2 ± 0.4	94.9 ± 0.4	96.1 ± 0.4	86.5 ± 0.8
	RF	83.6 ± 1.1	94.8 ± 0.5	95.9 ± 0.3	86.0 ± 1.4
Automatic	**KNN**	79.1 ± 1.3	88.1 ± 1.0	90.8 ± 0.6	82.4 ± 0.9
	MLP	83.6 ± 1.1	89.8 ± 1.4	91.5 ± 0.9	86.5 ± 1.0
	RF	84.6 ± 1.3	90.3 ± 0.8	92.1 ± 0.6	85.9 ± 0.5

and the conclusions are nearly identical for MLP for post-filtering precision.

For the detection evaluation, bounding boxes were produced as follows: for masks predicted 'fish', the contour-guided bounding box generation procedure with head fraction 0.25 was used; for masks predicted 'head', the SAM-produced bounding box containing the mask was used. We have included the analysis on the SAM-produced head boxes in Appendix A. Box selection was based only on the classifier-predicted mask class. Ground-truth mask labels were not used during detection evaluation.

Filtering increased precision for both prompting strategies at the cost of recall (Fig. 11). For manually prompted SAM masks, this tradeoff slightly reduced mean F_1, whereas for automatically prompted masks the precision gain outweighed the recall loss, producing a modest increase in mean F_1 (Table 5). Although filtering changed the mean precision–recall balance and the mean F_1 slightly improved on both modes, the fold-wise F_1 distributions overlapped with the unfiltered setting, indicating that the post-processing primarily traded recall for precision rather than producing a clear overall F_1 gain or reduction (Table 5).

When evaluating multiclass model precision by using the true mask class quantities left after the post-processing on the automatic mode, we see that it effectively filtered out undesired masks: the mask precision % on 'fish' and 'head' masks increased from 60.5% to 77.7% (Table 5). By applying the filtering on manually prompted masks, the mask precision % on 'fish' and 'head' in the test set increased from 75.6% to 85.2% (Table 5). The resulting number of masks after filtering was 31%–33% lower than ground truth on the two modes. In practical terms, on average, the post-processed automatic prompt produced 45 masks per image, of which 35 confirmed relevant masks, whereas human annotators marked 66 fish per image.

4. Discussion

This study demonstrated that SAM could be used to automatically detect fish in two large-scale commercial RAS environments, in two different fish species with a production level stocking density and standard production fish size with reasonable precision. We also showed that detecting fish in RAS setting is a challenging task also for a human annotator, resulting in low inter-annotator agreement.

The video equipment used were RGB surface cameras, which makes the fish instance segmentation a challenging and time-consuming task, because of the low contrast, surface disturbances and high level of occlusion caused by a commercial level stocking density (Fig. 12). The performance of SAM was better with manually annotated prompts than when using a fully automatic mode. This is to be expected, as the manually annotated prompts effectively incorporate prior information, and guide the localization for the model.

Both sets of mask outputs were effectively refined using a lightweight post-processing procedure, showing that simple shape descriptors are enough to improve the detection precision and mask precision of SAM. The three different one-vs-all binary classifiers trained to distinguish good quality, whole 'fish' masks from other types reached a mean accuracy of 82.7% on manually annotated data and 82.4% on automatic prompts. Using a multiclass classifier trained to distinguish all four of the presented classes, the proportion of 'fish' masks increased

Table 5
Final post-processing results for the filtering target using RF classifier. Values are stratified 5-fold cross-validation fold means and standard deviations, shown on separate rows. For the proposed filtering method, detection metrics were computed from out-of-fold classifier predictions. Detection was evaluated at IoU 0.30 using contour-guided generated head boxes with head fraction 0.25 for 'fish' predictions, and SAM-produced bounding boxes for 'head' predictions. Mask P % is the categorical mask precision: the fraction of TP 'fish' and 'head' masks present after filtering, regardless of human annotator bounding box locations.

Prompt/	Mask count			Mask	Detection at IoU 0.30						
filter	GT	Before filter	After filter	P (%)	TP	FP	FN	P	R	F1	Mean IoU
M/None	1165	1061	1061	75.6	637	425	528	0.600	0.547	0.572	0.485
	± 50	± 26	± 26	± 0.7	± 17	± 25	± 44	± 0.017	± 0.021	± 0.019	± 0.004
M/Ours	1165	1061	784	85.2	586	198	579	0.747	0.504	0.602	0.486
	± 50	± 26	± 19	± 1.4	± 20	± 12	± 67	± 0.015	± 0.037	± 0.030	± 0.003
A/None	1165	1275	1275	60.5	421	855	744	0.330	0.361	0.345	0.487
	± 42	± 14	± 14	± 1.0	± 19	± 31	± 43	± 0.018	± 0.019	± 0.017	± 0.007
A/Ours	1165	1275	801	77.7	355	446	810	0.443	0.305	0.361	0.489
	± 42	± 14	± 26	± 2.2	± 29	± 26	± 44	± 0.031	± 0.024	± 0.026	± 0.005

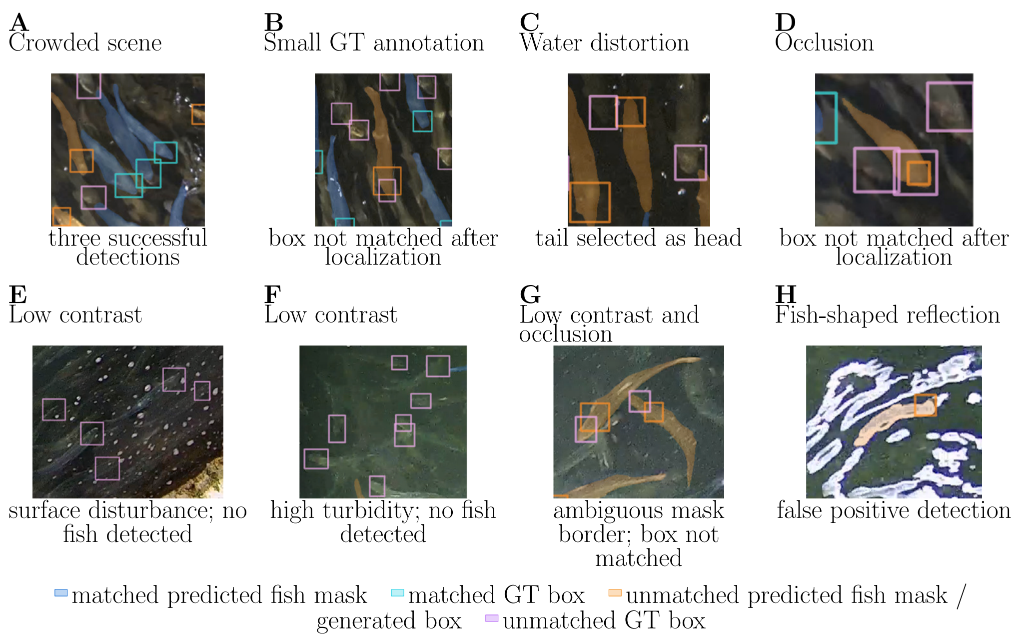

Fig. 12. Examples of failure cases of SAM and the bounding box generation procedure, successful and unsuccessful bounding box localization and matching.

from 75.6% to 85.2% on the manually annotated data and from 60.5% to 77.7% on the automatic prompts. The improvements are at the cost of recall, which is clearly reduced for both prompting methods, showing a decrease on the object detection recall from 0.36 to 0.31 on automatic. This precision–recall tradeoff is also seen on the object detection precision, which increases from 0.33 to 0.44 on automatic. However, F_1 stays relatively stable on both prompting methods, suggesting a controlled tradeoff.

Monitoring fish on a commercial scale is considered difficult (Macaulay et al., 2021). Among animal production systems, some of the highest stocking densities outside aquaculture are typically found in broiler chicken production (Buijs et al., 2009). However, broiler chickens do not exhibit the vertical stacking and underwater imaging distortions present in RAS, and their colouration generally provides better contrast against the background. The small fraction of fish visible at the surface compared to the total number of fish in the tank makes the broad visual monitoring of the whole tank a very challenging problem.

In previous work on detecting swimming-related ABM, studies have relied on classical computer vision techniques in small-scale tanks. Using background–foreground subtraction measures such as dispersion, velocity and turning angle of the shoal (Zhao et al., 2016) or different feeding-related quantitative measures (Zhou et al., 2018; Li et al., 2020) can be calculated. Optical flow has been suggested to avoid some of the limitations of background–foreground subtraction, and used to find behaviours related to stress and feeding based on outlier detection (Yu et al., 2021). However, modern deep learning methods have been shown to surpass classical computer vision methods in accuracy (Fitzgerald et al., 2025). Supervised deep learning methods has been used in both detecting individual fish and classifying group behaviour (Zhao et al., 2021). Supervised detection combined with stereo cameras to measure swimming speed, pattern and dispersion has been applied to study the effects of dissolved hydrogen sulphide in Atlantic salmon (Ciani et al., 2024). The downside to supervised models is that they need a lot of training data specific to each environment, and thus require costly manual annotation for good performance. This is often not feasible in real world applications, highlighting the usefulness of models such as SAM that demonstrate viable zero-shot performance in fish instance segmentation without additional training.

In underwater images, object detectors have been used to find fish skin surface damage (Liu et al., 2025), and to detect sea lice in Atlantic Salmon (Gupta et al., 2022; C. Zhang et al., 2024). Similarly, object detectors have been used to detect skin peeling and whitening in groupers (Chen et al., 2022). In surface cameras, the image quality is greatly affected by reflections, waves, foam and water turbidity (Li et al., 2020). On the other hand, underwater cameras are more expensive and require more maintenance due to biofilm accumulation. In our setting, underwater cameras would have likely given a very limited view on the fish due to water turbidity and high occlusion from the stocking density.

Pre-trained foundation models have sparked enthusiasm, due to their good zero-shot performance on a variety of tasks and the possibility for quick adaptation for real-world applications (Zhou et al., 2024). A general-purpose model that is quick to fine-tune to specific tasks saves training resources and time spent on annotation work. Within a short time of the release of SAM, the numerous adaptations (Awais et al., 2025), many fine-tuning experiments (Song et al., 2024; Huang et al., 2024; Archit et al., 2025), and its already prevalent use as a tool to assist manual segmentation show the significance of FM in real-world image segmentation tasks. Using a point-level weak supervision scheme requiring a single click per fish (Laradji et al., 2021) with SAM is much more efficient than drawing the masks by hand. Our proposed workflow, once trained on a small number of images, can be used to obtain a larger dataset using automatic segmentation. Depending on the number of objects in the image and the performance of the automatic model, our method can be efficient when leveraging the trained classifier together with automatic segmentation to create a large number of annotations with sufficient quality for practical applications. However, the proposed workflow is not likely suitable for applications where a very high, consistent mask quality or high recall is needed.

When comparing the prompting techniques of SAM, similar results to ours have been found, for example, in medical imaging: using manual annotations seem to outperform automatic segmentation (Huang et al., 2024), but the performance can be improved by fine-tuning the model (Huang et al., 2024) and post-processing the segmentation output (Williams et al., 2024). Using different post-processing techniques and incorporating prior information to refine the set of output masks to improve model performance is a common technique used in segmentation tasks (Chen et al., 2018; Sauer et al., 2024; Furtado, 2021), including using a classifier to assess segmentation mask quality (Lin et al., 2020). When training classifiers, simple hand-crafted features can outperform CNN-based feature extraction when the training data sample size is small (Lin et al., 2020).

The limitations of our study includes lack of pixel-level ground truth and consequently lack of pixel-level mask evaluation. In our paper, we evaluate the mask quality based on coarse categorical assessment. Pixel-level ground truth of the segmentation masks would be necessary to comprehensively validate mask quality and border consistency of SAM in RAS with standard instance segmentation metrics. Our low inter-annotator agreement on object detection task indicates that obtaining gold-standard pixel-level ground truth is not a trivial task in commercial RAS setting.

The bounding box generation used in our experiment was designed as a preprocessing step for the present imaging setup and metric reporting, and it may be less reliable for fish with substantially different proportions, for example due to age, body condition, or camera viewpoint. The lower but very stable mean IoU across cross-validation folds compared to human annotators suggests stable but coarse object localization, reflecting limitations of the bounding-box generation procedure.

The detection recall metric for our proposed method is very low. Our recall metric suffers from two sources of noise unrelated to the model itself: low ground truth quality due to low inter-annotator agreement and localization inaccuracies in bounding box generation. Therefore, as recall is confounded by noise factors independent of the model, the measured performance is likely underestimated and may increase once these factors are mitigated. This interpretation is further supported by the large difference in detection precision and categorical mask precision (0.44 vs 0.78), where the difference of the metrics is that detection precision requires bounding box localization with a matching human-produced annotation (IoU$\geq$ 0.3). Higher quality, task-specific ground truth data and a more thorough analysis would be needed to comprehensively assess and distinguish these effects, to get a better understanding of the recall when using SAM and our proposed post-processing classifier.

The proposed classifier in our post-processing method is a supervised machine learning model and thus dependent on the data it learns on. Having a relatively small number of images and fish farms presented, as we have in our experiments, may limit the generalizability of the model. On the other hand, the annotation and re-training effort associated with out method is very moderate, possibly making it more suitable for a case-by-case fit than traditional supervised object detection models.

In our paper we use a simplistic grid based prompt for SAM. As demonstrated by our results in the differences of manual mode vs automatic, better localization priors as prompts are likely to further increase the usefulness of SAM. Further development of SAM itself is active, showing continuing improvement of its zero-shot performance (Ke et al., 2023). Multi-modal implementations combining prompts in text format (Zou et al., 2023), a feature already supported by SAM, are of particular interest: could a simple ‘fish’ prompt compete with annotations made by humans in our task in the future? Human-in-the-loop SAM-based annotation workflow utilizing this multi-modality has already been proposed for livestock, showing considerable reduction in the amount of manual labour required (Wu et al., 2024).

It has been shown that improved localization accuracy through object segmentation improves tracking and scene understanding performance in various settings compared locating the objects using bounding boxes (Voigtlaender et al., 2019; Behl et al., 2017). Many recent adaptations of SAM focus on tracking (Zhang et al., 2025; Bai et al., 2024; Cheng et al., 2023). On fish, a robust zero-shot tracking model based on instance segmentation FMs has not yet been established. We see the approach as promising and hope that our experiments contribute towards this goal.

Challenges for model development for fish detection and instance segmentation include the lack of data from large-scale fish farms and benchmarks for model development, especially when taking into account the differences in ABM when assessing welfare in different fish species and the vast range of different rearing environments. We hope that our contribution in this regard encourages other researchers to share their data in the future to tackle these challenges.

5. Conclusion

We have demonstrated the feasibility of SAM in detecting fish in a large-scale commercial RAS environment. We compared the performance of SAM on manually annotated bounding box prompts to a fully automatic segmentation mode, finding that manual annotations produce clearly better results when no additional post-processing is applied. Our proposed post-processing method improved precision on both prompting methods by classifying the mask outputs and filtering out undesirable masks. While providing a clear precision-oriented trade-off, our method had stable F_1 overall performance.

CRediT authorship contribution statement

Hilla Fred: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation. **Mogens Agerbo Krogh:** Writing – review & editing, Investigation, Funding acquisition, Data curation, Conceptualization. **Britt Bang Jensen:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization. **Laura Ruotsalainen:** Writing – review & editing, Supervision, Methodology. **Jouni Vielma:** Writing – review & editing, Investigation, Conceptualization. **Matti Pastell:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used OpenAI Codex to assist with restructuring and refactoring analysis and visualization code, improving readability, modularity, and reproducibility. All code logic, analyses, and outputs were reviewed and validated by the authors, who take full responsibility for the results.

Funding

This study was funded by the Nordic Research and Innovation Programme for Sustainable Aquaculture under the NordForsk umbrella, project 104876 "IntelliRAS - Intelligent farming and health control in landbased recirculating aquaculture systems".

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.

Acknowledgements

The authors would like to thank Lea Ala-Ilomäki, Mille Stenholt Jensen and Sofie Bælum Bugge for their contributions in conducting the data annotation. In addition, the authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.aquaculture.2026.744459.

Data availability

Data and code available at https://doi.org/10.5281/zenodo.15528511.

References

Archit, A., Freckmann, L., Nair, S., Khalid, N., Hilt, P., Rajashekar, V., Freitag, M., Teuber, C., Buckley, G., von Haaren, S., Gupta, S., Dengel, A., Ahmed, S., Pape, C., 2025. Segment anything for microscopy. Nature Methods 1–13. http://dx.doi.org/10.1038/s41592-024-02580-4, Publisher: Nature Publishing Group. URL https://www.nature.com/articles/s41592-024-02580-4.

Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.-H., Khan, F.S., 2025. Foundation models defining a new era in vision: a survey and outlook. IEEE Trans. Pattern Anal. Mach. Intell. 1–20. http://dx.doi.org/10.1109/TPAMI.2024.3506283, Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. URL https://ieeexplore.ieee.org/abstract/document/10834497.

Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Liu, L., Zhang, Z., Shou, M.Z., 2024. One token to seg them all: Language instructed reasoning segmentation in videos. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (Eds.), Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/0cf3e7eefb9d643e93e16ff1d94090a7-Abstract-Conference.html.

Behl, A., Jafari, O.H., Mustikovela, S.K., Alhaija, H.A., Rother, C., Geiger, A., 2017. Bounding boxes, segmentations and object coordinates: How important is recognition for 3D scene flow estimation in autonomous driving scenarios? In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2593–2602. http://dx.doi.org/10.1109/ICCV.2017.281, ISSN: 2380-7504. URL https://ieeexplore.ieee.org/document/8237543.

Borges Oliveira, D.A., Ribeiro Pereira, L.G., Bresolin, T., Pontes Ferreira, R.E., Reboucas Dorea, J.R., 2021. A review of deep learning algorithms for computer vision systems in livestock. Livest. Sci. 253, 104700. http://dx.doi.org/10.1016/j.livsci.2021.104700, URL https://www.sciencedirect.com/science/article/pii/S1871141321003085.

Buijs, S., Keeling, L., Rettenbacher, S., Van Poucke, E., Tuyttens, F.A.M., 2009. Stocking density effects on broiler welfare: Identifying sensitive ranges for different indicators. Poult. Sci. 88 (8), 1536–1543. http://dx.doi.org/10.3382/ps.2009-00007, URL https://www.sciencedirect.com/science/article/pii/S0032579119389564.

Chen, J.C., Chen, T.-L., Wang, H.-L., Chang, P.-C., 2022. Underwater abnormal classification system based on deep learning: A case study on aquaculture fish farm in Taiwan. Aquac. Eng. 99, 102290. http://dx.doi.org/10.1016/j.aquaeng.2022.102290, URL https://www.sciencedirect.com/science/article/pii/S0144860922000668.

Chen, L., 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. http://dx.doi.org/10.1109/TPAMI.2017.2699184.

Cheng, H.K., Oh, S.W., Price, B.L., Schwing, A.G., Lee, J., 2023. Tracking anything with decoupled video segmentation. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. IEEE, pp. 1316–1326. http://dx.doi.org/10.1109/ICCV51070.2023.00127.

Ciani, E., Kvæstad, B., Stormoen, M., Mayer, I., Gupta, S., Ribičić, D., Netzer, R., 2024. Early warning through video monitoring: Dissolved hydrogen sulphide (H2S) affects atlantic salmon swimming behavior in recirculating aquaculture systems. Aquaculture 581, 740201. http://dx.doi.org/10.1016/j.aquaculture.2023.740201, URL https://www.sciencedirect.com/science/article/pii/S0044848623009754.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, URL https://openreview.net/forum?id=YicbFdNTTy.

EFSA Panel on Animal Health and Welfare (AHAW), 2012. Statement on the use of animal-based measures to assess the welfare of animals. EFSA J. 10 (6), 2767. http://dx.doi.org/10.2903/j.efsa.2012.2767, URL https://onlinelibrary.wiley.com/doi/abs/10.2903/j.efsa.2012.2767. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2903/j.efsa.2012.2767.

European Commission: Directorate-General for Maritime and Fisheries Affairs, EU-MOFA, 2020. Recirculation Aquaculture Systems. Publications Office of the European Union, http://dx.doi.org/10.2771/66025.

Fitzgerald, A., Ioannou, C.C., Consuegra, S., Dowsey, A., Garcia de Leaniz, C., 2025. Machine vision applications for welfare monitoring in aquaculture: Challenges and opportunities. Aquac. Fish Fish. 5 (1), e70036. http://dx.doi.org/10.1002/aff2.70036, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aff2.70036._eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aff2.70036.

Furtado, P., 2021. Loss, post-processing and standard architecture improvements of liver deep learning segmentation from computed tomography and magnetic resonance. Informatics Med. Unlocked 24, 100585. http://dx.doi.org/10.1016/j.imu.2021.100585, URL https://www.sciencedirect.com/science/article/pii/S2352914821000757.

Garber, A.F., Amini, F., Gezan, S.A., Swift, B.D., Hodkinson, S.E., Nickerson, J., Bridger, C.J., 2019. Genetic and phenotypic evaluation of harvest traits from a comprehensive commercial atlantic salmon, *salmo salar* l., broodstock program. Aquaculture 503, 242–253. http://dx.doi.org/10.1016/j.aquaculture.2019.01.001, URL https://www.sciencedirect.com/science/article/pii/S0044848618312869.

Gupta, A., Bringsdal, E., Knausg˚ard, K.M., Goodwin, M., 2022. Accurate wound and lice detection in atlantic salmon fish using a convolutional neural network. Fishes 7 (6), 345. http://dx.doi.org/10.3390/fishes7060345, Number: 6 Publisher: Multidisciplinary Digital Publishing Institute. URL https://www.mdpi.com/2410-3888/7/6/345.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.B., 2022. Masked autoencoders are scalable vision learners. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, la, USA, June 18-24, 2022. IEEE, pp. 15979–15988. http://dx.doi.org/10.1109/CVPR52688.2022.01553.

Hu, M., 1962. Visual pattern recognition by moment invariants. IRE Trans. Inf. Theory 8 (2), 179–187. http://dx.doi.org/10.1109/TIT.1962.1057692.

Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., Liu, S., Chi, H., Hu, X., Yue, K., Li, L., Grau, V., Fan, D.-P., Dong, F., Ni, D., 2024. Segment anything model for medical images? Med. Image Anal. 92, 103061. http://dx.doi.org/10.1016/j.media.2023.103061, URL http://arxiv.org/abs/2304.14660. arXiv:2304.14660 [eess].

Ke, L., Ye, M., Danelljan, M., Liu, Y., Tai, Y., Tang, C., Yu, F., 2023. Segment anything in high quality. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (Eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, la, USA, December 10 - 16, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/5f828e38160f31935cfe9f67503ad17c-Abstract-Conference.html.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B., 2023. Segment anything. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. IEEE, pp. 3992–4003. http://dx.doi.org/10.1109/ICCV51070.2023.00371.

Laradji, I.H., Saleh, A., Rodriguez, P., Nowrouzezahrai, D., Azghadi, M.R., Vazquez, D., 2021. Weakly supervised underwater fish segmentation using affinity LCFCN. Sci.

Rep. 11 (1), 17379. http://dx.doi.org/10.1038/s41598-021-96610-2, Publisher: Nature Publishing Group. URL https://www.nature.com/articles/s41598-021-96610-2.

Li, W., Liu, Y., Wang, W., Li, Z., Yue, J., 2024. TFMFT: Transformer-based multiple fish tracking. Comput. Electron. Agric. 217, 108600. http://dx.doi.org/10.1016/j.compag.2023.108600, URL https://www.sciencedirect.com/science/article/pii/S0168169923009882.

Li, D., Wang, Z., Wu, S., Miao, Z., Du, L., Duan, Y., 2020. Automatic recognition methods of fish feeding behavior in aquaculture: A review. Aquaculture 528, 735508. http://dx.doi.org/10.1016/j.aquaculture.2020.735508, URL https://www.sciencedirect.com/science/article/pii/S0044848620315258.

Lin, W., Hasenstab, K., Moura Cunha, G., Schwartzman, A., 2020. Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment. Sci. Rep. 10 (1), 20336. http://dx.doi.org/10.1038/s41598-020-77264-y, Publisher: Nature Publishing Group. URL https://www.nature.com/articles/s41598-020-77264-y.

Liu, H., Li, H., Ma, X., Jiang, W., Man, X., Zheng, R., 2025. Fish surface damage detection with improved YOLOv7. In: Intelligent Robotics and Applications. Springer, Singapore, pp. 157–168. http://dx.doi.org/10.1007/978-981-96-0792-1_13, ISSN: 1611-3349. URL https://link-springer-com.libproxy.helsinki.fi/chapter/10.1007/978-981-96-0792-1_13.

Liu, Y., Li, B., Zhou, X., Li, D., Duan, Q., 2024. FishTrack: Multi-object tracking method for fish using spatiotemporal information fusion. Expert Syst. Appl. 238, 122194. http://dx.doi.org/10.1016/j.eswa.2023.122194, URL https://www.sciencedirect.com/science/article/pii/S0957417423026969.

Macaulay, G., Bui, S., Oppedal, F., Dempster, T., 2021. Challenges and benefits of applying fish behaviour to improve production and welfare in industrial aquaculture. Rev. Aquac. 13 (2), 934–948. http://dx.doi.org/10.1111/raq.12505, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/raq.12505. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/raq.12505.

Noble, C., Gismervik, K., Iversen, M.H., Kolarevic, J., Nilsson, J., Stien, L.H., Turnbull, J.F., 2020. Welfare indicators for farmed rainbow trout: tools for assessing fish welfare.

Pettersen, J.M., Bracke, M.B., Midtlyng, P.J., Folkedal, O., Stien, L.H., Steffenak, H.a., Kristiansen, T.S., 2014. Salmon welfare index model 2.0: an extended model for overall welfare assessment of caged atlantic salmon, based on a review of selected welfare indicators and intended for fish health professionals. Rev. Aquac. 6 (3), 162–179. http://dx.doi.org/10.1111/raq.12039, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/raq.12039. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/raq.12039.

Razman, M.A.M., Susto, G.A., Cenedese, A., Abdul Majeed, A.P.P., Musa, R.M., Abdul Ghani, A.S., Adnan, F.A., Ismail, K.M., Taha, Z., Mukai, Y., 2019. Hunger classification of *lates calcarifer* by means of an automated feeder and image processing. Comput. Electron. Agric. 163, 104883. http://dx.doi.org/10.1016/j.compag.2019.104883, URL https://www.sciencedirect.com/science/article/pii/S0168169919305332.

Sauer, A., Tian, Y., Bewersdorf, J., Rittscher, J., 2024. Refining biologically inconsistent segmentation masks with masked autoencoders. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024. IEEE, pp. 6904–6912. http://dx.doi.org/10.1109/CVPRW63382.2024.00684.

Skalski, P., 2019. Make sense. https://github.com/SkalskiP/make-sense/.

Song, B., Yang, H., Wu, Y., Zhang, P., Wang, B., Han, G., 2024. A multispectral remote sensing crop segmentation method based on segment anything model using multistage adaptation fine-tuning. IEEE Trans. Geosci. Remote Sens. 62, 1–18. http://dx.doi.org/10.1109/TGRS.2024.3411398, Conference Name: IEEE Transactions on Geoscience and Remote Sensing. URL https://ieeexplore.ieee.org/abstract/document/10551868.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B., 2019. MOTS: multi-object tracking and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, pp. 7942–7951. http://dx.doi.org/10.1109/CVPR.2019.00813, URL http://openaccess.thecvf.com/content_CVPR_2019/html/Voigtlaender_MOTS_Multi-Object_Tracking_and_Segmentation_CVPR_2019_paper.html.

Wang, H., Zhang, S., Zhao, S., Wang, Q., Li, D., Zhao, R., 2022. Real-time detection and tracking of fish abnormal behavior based on improved YOLOV5 and siamrpn++. Comput. Electron. Agric. 192, 106512. http://dx.doi.org/10.1016/j.compag.2021.106512, URL https://www.sciencedirect.com/science/article/pii/S0168169921005299.

Wang, Y., Zhao, Y., Petzold, L., 2024. An empirical study on the robustness of the segment anything model (SAM). Pattern Recognit. 155, 110685. http://dx.doi.org/10.1016/j.patcog.2024.110685, URL https://www.sciencedirect.com/science/article/pii/S0031320324004369.

Williams, D., Macfarlane, F., Britten, A., 2024. Leaf only SAM: A segment anything pipeline for zero-shot automated leaf segmentation. Smart Agric. Technol. 8, 100515. http://dx.doi.org/10.1016/j.atech.2024.100515, URL https://www.sciencedirect.com/science/article/pii/S2772375524001205.

Wu, Y., Zhou, S., Wu, Z., Chen, Z., Hu, X., Li, J., 2024. Accelerated data engine: A faster dataset construction workflow for computer vision applications in commercial livestock farms. Comput. Electron. Agric. 226, 109452. http://dx.doi.org/10.1016/j.compag.2024.109452, URL https://www.sciencedirect.com/science/article/pii/S0168169924008433.

Yang, L., Liu, Y., Yu, H., Fang, X., Song, L., Li, D., Chen, Y., 2021. Computer vision models in intelligent aquaculture with emphasis on fish detection and behavior analysis: A review. Arch. Comput. Methods Eng. 28 (4), 2785–2816. http://dx.doi.org/10.1007/s11831-020-09486-2, Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 4 Publisher: Springer Netherlands. URL https://link-springer-com.libproxy.helsinki.fi/article/10.1007/s11831-020-09486-2.

Yu, X., Wang, Y., An, D., Wei, Y., 2021. Identification methodology of special behaviors for fish school based on spatial behavior characteristics. Comput. Electron. Agric. 185, 106169. http://dx.doi.org/10.1016/j.compag.2021.106169, URL https://www.sciencedirect.com/science/article/pii/S0168169921001861.

Zhang, C., Bracke, M., da Silva Torres, R., Gansel, L.C., 2024. Rapid detection of salmon louse larvae in seawater based on machine learning. Aquaculture 592, 741252. http://dx.doi.org/10.1016/j.aquaculture.2024.741252, URL https://www.sciencedirect.com/science/article/pii/S0044848624007130.

Zhang, L., Qiu, Y., Fan, J., Li, S., Hu, Q., Xing, B., Xu, J., 2024. Underwater fish detection and counting using image segmentation. Aquac. Int. 32 (4), 4799–4817. http://dx.doi.org/10.1007/s10499-024-01402-w.

Zhang, Z., Zhang, S., Dai, Z., Dong, Z., Zhu, S., 2025. MWVOS: Mask-free weakly supervised video object segmentation via promptable foundation model. Pattern Recognit. 159, 111100. http://dx.doi.org/10.1016/j.patcog.2024.111100, URL https://www.sciencedirect.com/science/article/pii/S0031320324008513.

Zhao, J., Gu, Z., Shi, M., Lu, H., Li, J., Shen, M., Ye, Z., Zhu, S., 2016. Spatial behavioral characteristics and statistics-based kinetic energy modeling in special behaviors detection of a shoal of fish in a recirculating aquaculture system. Comput. Electron. Agric. 127, 271–280. http://dx.doi.org/10.1016/j.compag.2016.06.025, URL https://www.sciencedirect.com/science/article/pii/S0168169916304252.

Zhao, S., Zhang, S., Liu, J., Wang, H., Zhu, J., Li, D., Zhao, R., 2021. Application of machine learning in intelligent fish aquaculture: A review. Aquaculture 540, 736724. http://dx.doi.org/10.1016/j.aquaculture.2021.736724, URL https://www.sciencedirect.com/science/article/pii/S0044848621003860.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P.S., Sun, L., 2024. A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. Int. J. Mach. Learn. Cybern. http://dx.doi.org/10.1007/s13042-024-02443-6.

Zhou, C., Lin, K., Xu, D., Chen, L., Guo, Q., Sun, C., Yang, X., 2018. Near infrared computer vision and neuro-fuzzy model-based feeding decision system for fish in aquaculture. Comput. Electron. Agric. 146, 114–124. http://dx.doi.org/10.1016/j.compag.2018.02.006, URL https://www.sciencedirect.com/science/article/pii/S0168169917312917.

Zhou, C., Wang, C., Sun, D., Hu, J., Ye, H., 2025. An automated lightweight approach for detecting dead fish in a recirculating aquaculture system. Aquaculture 594, 741433. http://dx.doi.org/10.1016/j.aquaculture.2024.741433, URL https://www.sciencedirect.com/science/article/pii/S0044848624008949.

Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J., 2023. Segment everything everywhere all at once. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (Eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, la, USA, December 10 - 16, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/3ef61f7e4afacf9a2c5b71c726172b86-Abstract-Conference.html.