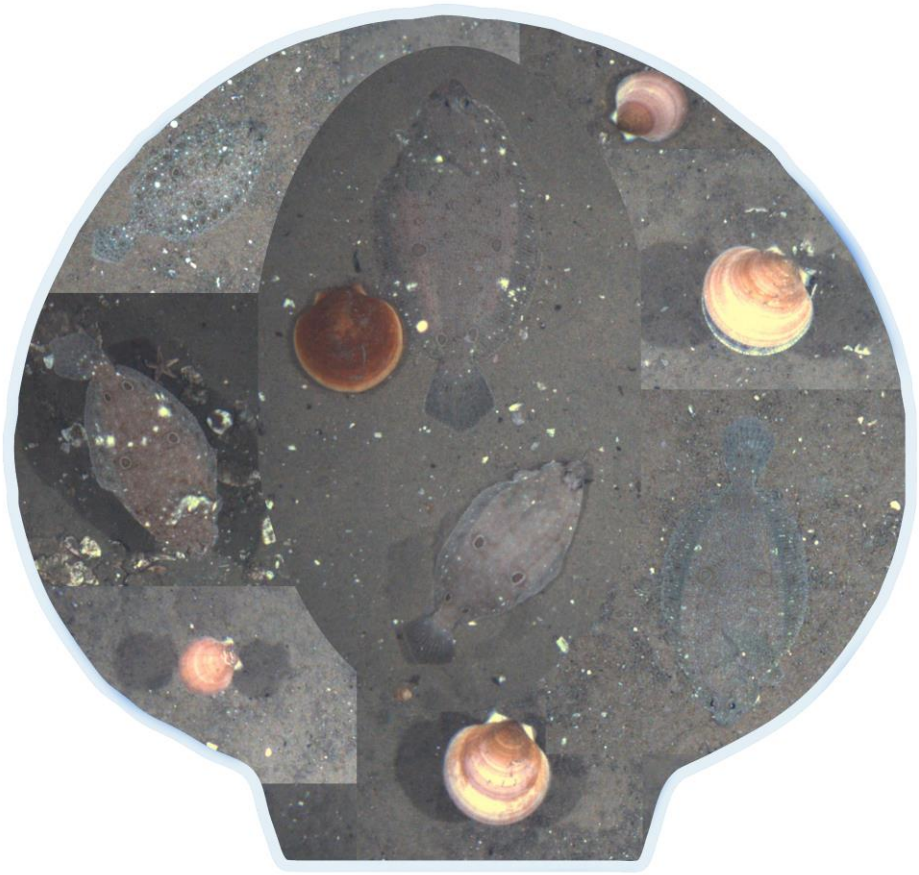
Improving automated detection of scallops and flounder in optical surveys with stereo detection methods

Final Report

Prepared for the 2019
NOAA Scallop Research Set-Aside
(Grant # NA19NMF4540017)
June 2021



Submitted By

Liese Siemann and Tasha O'Hara
Coonamessett Farm Foundation, Inc

In collaboration with
Matt Dawkins, Jon Crall, Matt Leotta, and Anthony Hoogs – Kitware, Inc.

Coonamessett Farm Foundation, Inc

277 Hatchville Road
East Falmouth, MA 02536

508-356-3601 FAX
508-356-3603
contact@cfarm.org

www.cfarm.org

Project Title: Improving automated detection of scallops and flounder in optical surveys with stereo detection methods

Principal Investigators: Liese Siemann and Tasha O'Hara (replaced Jason Clermont)

Organization: Coonamessett Farm Foundation, Inc. (CFF)

NOAA Grant Number: NA19NMF4540017

Report date: June 29, 2021

Executive Summary

Coonamessett Farm Foundation, Inc. (CFF) conducts one of the optical surveys of the sea scallop resource using a HabCam towed vehicle. The CFF HabCam v3 collects 5-7 images per second, providing a continuous track of imagery as it is towed over scallop grounds. Annotations from HabCam images are translated into biomass estimates through a multi-step process that relies on adequate image subsampling and accurate scallop counts and measurements. Annotating images is the most time-consuming part of generating these scallop biomass estimates. Reliably automating scallop image annotations could increase annotation rates and improve estimates of biomass, while also addressing questions about scallop patchiness and distribution over small-to-large scales on scallop grounds. Furthermore, optical survey images provide a wealth of raw data that is not consistently used at this time. One high priority goal for the use of HabCam images is providing data for improving flounder stock assessments because the incidental bycatch of non-target species, like yellowtail and windowpane flounder, could have negative impacts on the long-term sustainability of the scallop fishery.

Kitware, Inc. developed the Video and Image Analytics for Marine Environments (VIAME) system for analysis of underwater imagery with support from the NOAA Automated Image Analysis Strategic Initiative. CFF collaborated with Kitware to develop improved automated detectors for scallop classes (live scallops, swimming scallops, and clappers) and flounder by incorporating depth disparity information (i.e., depth maps) obtained from stereo image pairs taken during the 2017-2019 HabCam v3 surveys. The accuracy and precision of the new detectors with depth maps incorporated was compared to that of single-image detectors when both utilized upgraded deep learning networks.

We expected the addition of depth disparity information to improve object detection models for all of the target classes, resulting in (1) better discrimination between live scallops, swimming scallops, and clappers based on characteristics like distance off bottom and gape distance and (2) enhanced detection of flounders based on their vertical profiles on the bottom. The addition of depth information improved detection of live scallops, swimming scallops, and flounders, but did not improve detection of clappers. We hypothesize that detection of swimming scallops was improved only slightly, despite strong depth map signatures, because detection models using single images may incorporate the same visual cues used by human annotators (e.g., shadows). Flounder detector improvements may have been small because flounder were observed partially buried, resting on the surface, or settled in hollows, and flounders only occasionally created a large signature in computed depth maps, often overshadowed by differences in terrain elevation and background content. The lack of model improvement for clappers could be due to variability in their appearance and orientation in the sediment, combined with weak differences in the stereo disparity maps between live scallops and clappers.

The addition of new images with annotated flounder and less common scallop classes (swimming scallops and clappers) will improve models for these relatively rare target organisms, and CFF will continue to supply HabCam imagery to Kitware over the next few years.

Project timeline

Funding period: April 1, 2019 – March 31, 2021

Kick-off meeting: April 15, 2019. Attended by Liese Siemann, Jason Clermont, Matt Dawkins, and Anthony Hoogs

Algorithm development (Kitware): May 1, 2019 – March 31, 2021

HabCam v3 image annotation and consolidation: April 16, 2019 – January 13, 2020

Scallop autodetection meeting: May 7, 2020. Attended by representatives from CFF, Kitware, the Northeast Fisheries Science Center (NEFSC), the University of Massachusetts Dartmouth School for Marine Science & Technology (SMAST), C-Vision (subcontractor working with SMAST), the Woods Hole Oceanographic Institution (WHOI), Coastal Ocean Vision (working with WHOI), Rutgers University, the Greater Atlantic Regional Fisheries Office (GARFO) , and the New England Fishery Management Council (NEFMC)

Project management and participation

Project management and reporting: Liese Siemann

Image annotations and consolidation: Liese Siemann and Tasha O'Hara (replaced Jason Clermont)

Algorithm development: Kitware, Inc. (Matt Dawkins, Jon Crall, Matt Leotta, and Anthony Hoogs)

Contents

Background 1
Scallop stock assessment *1*
Use of HabCam images for flounder stock estimations *3*
VIAME (Video and Image Analytics for Marine Environments) *4*
Project objectives 5
Methods 5
Image Set Acquisition and Consolidation *5*
Improved camera calibration and stereo image-pair alignment *5*
Development of improved RGB and stereo auto-detection algorithms *5*
Results 7
Improved stereo calibration *7*
Scallop detectors *7*
Project Outreach 9
Evaluation 9
Accomplishments by objective *9*
Appendix A 13
Appendix B 16

Background

The Atlantic sea scallop (*Placopecten magellanicus*) is the focus of one of the most valuable fisheries on the east coast of the United States (NMFS 2020). The continued profitability of the fishery has depended on reliable surveys of the scallop stock. Biomass estimates are currently generated yearly using the results of federal and industry-funded surveys of the resource, with three out of four industry surveys conducted using optical survey equipment in recent years (NEFMC 2019, 2020). NEFSC conducts a combined optical/dredge survey. Three of the optical surveys use HabCam towed vehicles to collect imagery, with the vehicles operated by CFF (**Figure 1A**), WHOI, and NEFSC. These systems collect 5-7 images per second, providing a continuous track of imagery (**Figure 1B**) as they are towed along predetermined tracks across important scallop grounds (**Figure 1C**).

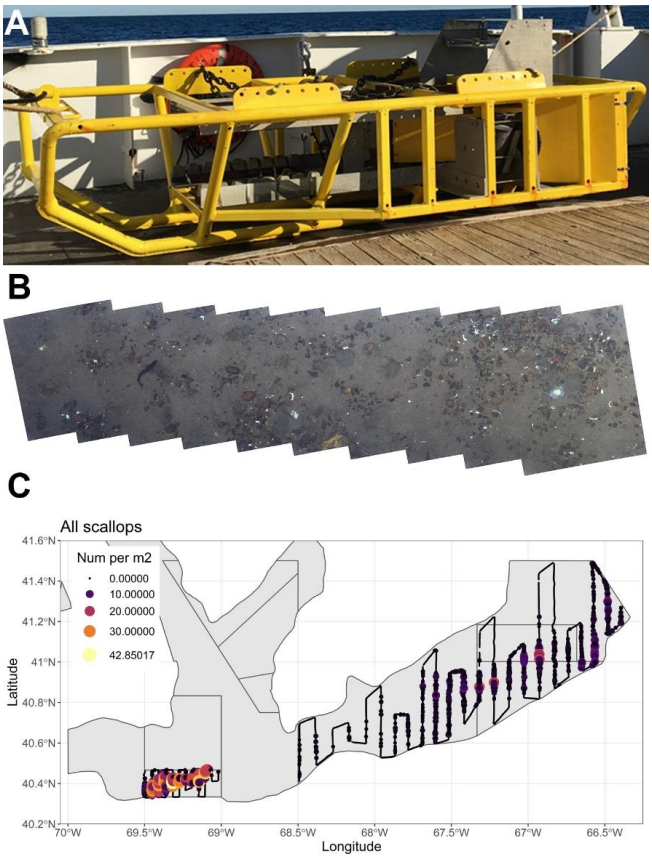


Figure 1. *(A) HabCam V3, operated by Coonamessett Farm Foundation. (B) Mosaic of images from the 2017 HabCam V3 survey. (C) Track and distribution of scallops from the 2020 industry HabCam survey on Georges Bank.*

Scallop stock assessment

Annotations from HabCam images are translated into biomass estimates through a multi-step process that relies on adequate image subsampling (i.e., the annotation rate) and accurate scallop counts and measurements. Based on information collected during HabCam camera calibration, scallop lengths in pixels are converted to shell heights in millimeters. The field of view (FOV) of each image is also calculated to estimate optical survey coverage. Each shell height (SH) measured from the HabCam images is converted to a meat weight (MW) in grams using published location-specific SHMW equations that include depth as a covariate (e.g., Hennen & Hart 2012). Biomass per meter^2 is calculated by summing all MWs in an image and dividing by the FOV of that image. Biomass in a defined stock area is estimated using a combination of a hurdle generalized additive model (GAM) and ordinary kriging (Chang et al. 2017). The hurdle GAM (quasi-binomial distribution for the presence/absence model and quasi-Poisson distribution for the positive model) is used to estimate the large scale trends in biomass with respect to latitude, longitude, and depth. Kriging on the model residuals improves estimates over smaller scales.

Annotating images is the most time consuming part of generating reliable scallop biomass estimates. NEFSC currently annotates 1:50 images collected during their surveys, while CFF annotates 1:200 to 1:400 images. During recent HabCam v3 surveys by CFF, 25-30% of annotated images included live sea scallops. The NEFSC scallop group examined scallop abundance along a

HabCam track line and estimated that a minimum of 50 images needed to be sampled over every 1000 meters to generate stable abundance estimates from the models used to generate biomass estimates in their surveys (presented at the March 2015 Peer Review Meeting of Sea Scallop Survey Methodologies summarized in Maguire 2015). The closer track lines in the more intensive industry surveys (**Figure 1C**) relax these annotation rate requirements, although improved annotation rates would ultimately lead to more confidence in the model inputs. The very short timeline between survey cruises and deadlines for submitting biomass estimates makes higher annotation rates difficult to achieve, but increasing image annotation rate continues to be a goal for CFF and other optical survey groups. Reliably automating scallop image annotations could increase annotation rates for all of the optical surveys, improving estimates of biomass and addressing questions about scallop patchiness and distribution over small-to-large scales on scallop grounds.

Annotation of scallops in HabCam images is complicated by the need to distinguish between live and dead scallops to generate accurate counts for biomass estimates. Human annotators differentiate between live and dead scallops based on a suite of characteristics (**Figure 2**), with trained annotators confident in their decisions ~88% of the time (Maguire 2015). Gaping, one characteristic of dead scallops, as well as the presence of depressions created by live scallops (Shumway & Parsons 2016), could be more easily detected using stereo images.

Distinguishing between scallops resting on the sea floor and those swimming in the water column is important for accurate sizing of scallops in survey images. Scallop lengths in pixels are overestimated when measured in two versus three dimensions, with errors increasing as scallops swim off the sea floor (**Figure 3**). Because the scallop MWs used in stock biomass estimations are derived from SHMW equations, overestimation of scallop lengths, and therefore SHs, translates directly into overestimations of scallop biomass. Consequently, increasing the use of stereo images for scallop length estimations would have a direct impact on stock assessments.

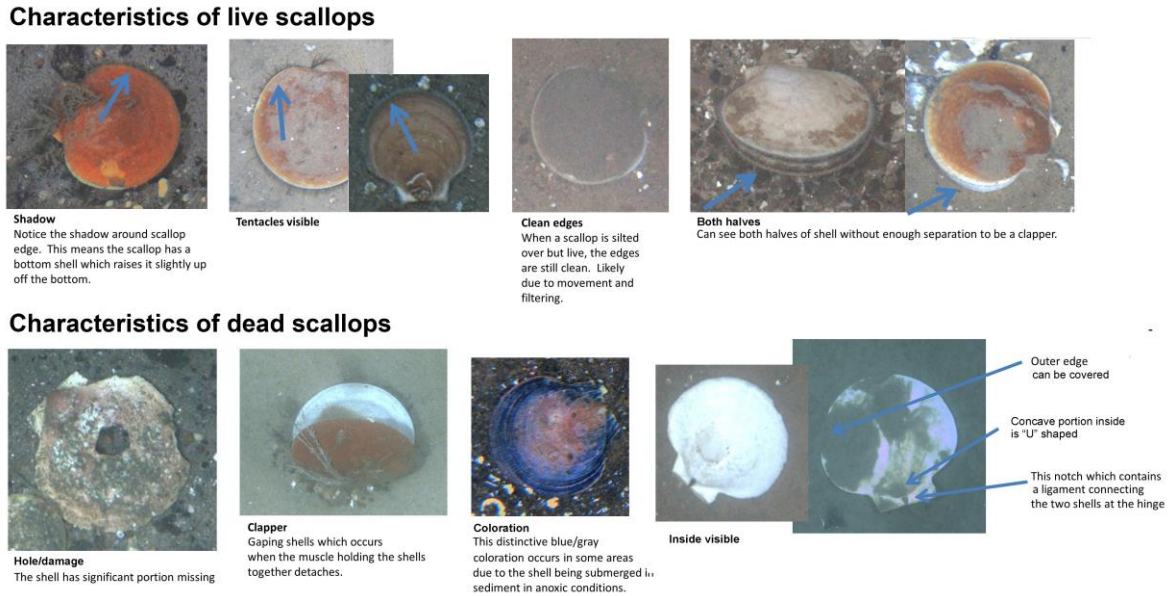


Figure 2. *Characteristics of live and dead scallops in survey images. Adapted from figures presented at the March 2015 Peer Review Meeting of Sea Scallop Survey Methodologies, summarized in Maguire 2015.*

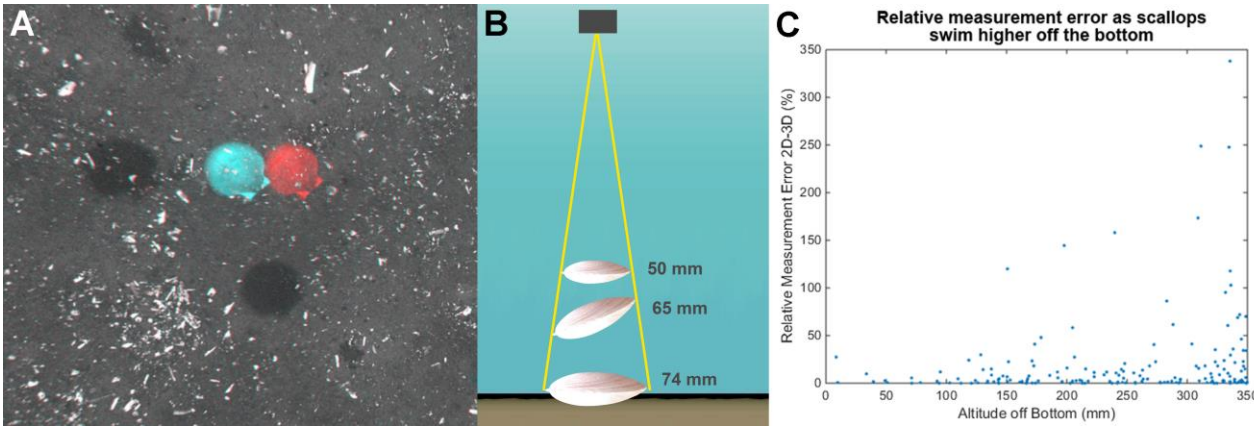


Figure 3. *Impact of swimming scallops locations in the water column on the estimation of scallop lengths. (A) Stereo image of swimming scallops, colored for 3D viewing, showing the characteristic shadows used for identification. (B) Schematic highlighting changes in real (shown) vs. estimated size as scallops move off the bottom or change their angle relative to horizontal. All three scallops in the example would have estimated 74-mm shell heights based on pixel coverage and camera altitude using a single 2D image. (C) Plot showing increasing measurement error as scallops move higher off the bottom. Figures A and C are adapted from*

Use of HabCam images for flounder stock estimations

Optical survey images provide a wealth of raw data that is not consistently used at this time. One high priority goal for these images is to build capacity to provide data for improving flounder stock assessments because the incidental bycatch of non-target species could have negative impacts on the long-term sustainability of the scallop fishery. Bycatch of yellowtail flounder (*Limanda ferruginea*) and windowpane flounder (*Scopthalmus aquosus*) continues to impact catch levels and quota, thereby putting the continued sustainability of the scallop fishery at risk (O'Keefe & DeCelles 2013).

Flounders are relatively rare in scallop optical survey images, with fewer than 1% of annotated images from recent HabCam v3 surveys including any flounder species. Moreover, identifying flounder in survey images can be difficult. Flounder can change their body patterns to camouflage on different substrates (**Figure 4**), and they often bury themselves in finer substrates like sand. Consequently, there is a need for increasing image annotation rates with a focus on flounder and a more reliable method for detecting flounder in survey images. Stereo disparities have been used to improve flounder detection algorithms, taking advantage of changes in bottom topography caused by flounder presence (**Figure 5**).

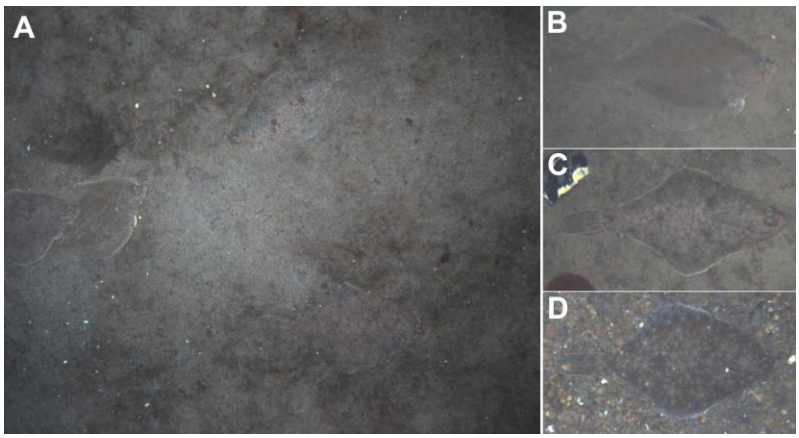


Figure 4. *Images of yellowtail flounder taken during the 2015 HabCam v3 survey. (A) A single image with at least five yellowtail flounder. Flounders with (B) uniform, (C) light mottle, and (D) dark mottle body patterns.*

VIAME (Video and Image Analytics for Marine Environments)

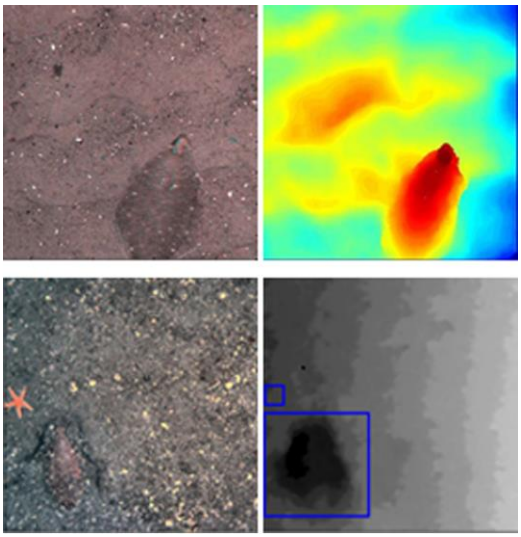


Figure 5. *Image segmentation of flounder using stereo disparities. (Top pair) Flounder resting on the surface is highlighted in red (red=close to blue=far). Adapted from presentations summarized in Maguire 2015. (Bottom pair) Flounder in a depression is highlighted in black (white=close to black=far).*

VIAME is an open-source system for analysis of underwater imagery developed by Kitware (Clifton Park, NY) with support from the NOAA Automated Image Analysis Strategic Initiative. It was originally designed as an integration platform for several different video and image processing algorithms (Dawkins et al. 2017), but it has since evolved into an end-to-end toolkit for producing analytics on an archive of imagery and/or video. At the core of VIAME is an image processing system which can link C/C++, Python, and MATLAB processing nodes together into a graph-like pipeline architecture. This can be used to easily build sequences of image processing algorithms without recompiling any source code or writing software, enabling rapid experimentation and re-use of different functional modules. For example, a common HabCam processing sequence consists of an image color correction module followed by a scallop detector. Using VIAME, different color correction modules can be tested with the same scallop detector simply by changing the pipeline configuration. Alongside the pipelined image processing system are a number of standalone utilities for object detector model training, output detection visualization, groundtruth annotation, detector evaluation (a.k.a. scoring), and image archive search.

Although the pre-defined VIAME detection and classification modules provide a substantial amount of functionality, there are many marine science image and video annotation problems that they do not address. Consequently, VIAME includes the capability for end-users to apply VIAME algorithms to new datasets and problems without any programming or knowledge of machine learning. Leveraging annotation user interfaces, there are two deep-learning capabilities in VIAME that enable users to create object detectors, classifiers, and other analytics. Through image search and interactive query refinement, users can quickly build a complete detection and classification capability for a novel problem, then run it on any amount of imagery or video. For more challenging analytical problems, users can manually annotate images, then train a deep learning detection and classification capability specific to their problem. Both methods were successfully used by marine scientists during VIAME training sessions at all of the NOAA Fisheries Science Centers (FSCs) to develop analytics on datasets that VIAME had not seen previously.

However, the model generation methods in VIAME did not make use of stereo imagery, a limitation that this project was designed to address so that any user with stereo data would be able to create a detector which exploits depth information. Stereo data collection is very common across the NOAA FSCs and the marine science community in general, as stereo enables 3D measurements of scallops, fish, flora, and even habitat. Stereo also significantly improves automated detection and classification under difficult conditions as indicated above for HabCam data, and also for many other underwater video collections such as small fish against cluttered backgrounds. Adding a

complete stereo capability to VIAME will have broad impact across a wide spectrum of marine scientists using VIAME now and in the future.

Project objectives

This project addressed ***Scallop RSA High Priority #1*** – Survey Related Research – by improving the accuracy of automated detectors for scallop categories (live scallops, swimming scallops, and clappers) and expanding the useful data generated from optical surveys with the goal of generating more accurate and precise biomass estimates. It also addressed ***Scallop RSA General Research Priority #5*** – Bycatch Research – by improving the detection of and annotation rates for key bycatch species, specifically flounder. Project efforts led to development of new algorithms and modules that incorporated stereo image pairs taken during HabCam surveys to:

(1) improve detection of scallops and differentiate between live, dead, and swimming scallops and

(2) detect and count flounder, fish that are difficult to reliably distinguish due to their camouflage abilities, changing body patterns, and tendency to bury themselves in sandy substrates.

Methods

Image Set Acquisition and Consolidation

CFF consolidated all of the jpg images, raw tif stereo image pairs, and annotations from the 2017 – 2019 HabCam v3 surveys and supplied them to Kitware for developing the new detector algorithms. Because swimming scallops and flounder are rare and dead scallops other than clappers are not normally annotated for the scallop assessment surveys, we examined over 260,000 additional images from the 2019 survey for the presence of swimming or dead scallops, clappers, and flounder. Images with these organisms were aggregated and annotated. In total, CFF supplied Kitware with 8,048 jpgs and 8,048 paired stereo tifs and over 400,000 annotations. The most recent intrinsic and extrinsic calibration matrices for the HabCam v3 cameras were also supplied to Kitware for creating depth disparity maps using the stereo image pairs.

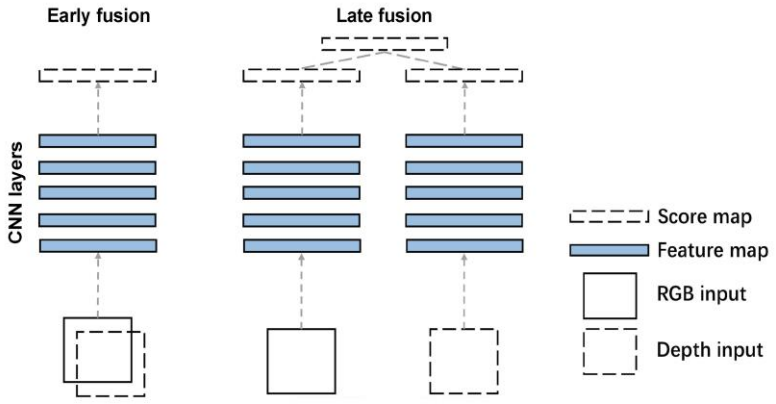


Figure 6. *Early vs. late fusion of stereo depth information. For early fusion, RGB images were joined with depth maps at the start to create a 4-channel input. For late fusion, RGB images and depth maps were processed in separate streams, and the resulting feature maps were summed to create final scoring maps. Figure modified from Li et al. 2017.*

Improved camera calibration and stereo image-pair alignment

Using the checkerboard images collected during the last calibration sessions for the HabCam v3 and v4 systems, Kitware re-calibrated the raw stereo-pair imagery provided by CFF and NEFSC using different frame selection and calibration techniques available in OpenCV. The improved calibrations were used to generate 3D sea floor models and a 2D stereo depth disparity maps.

Development of improved RGB and stereo auto-detection algorithms

Kitware tested a range of object detectors for scallops and flounder using both three-channel Red

Green Blue (RGB) images and images with depth information incorporated (RGBD images) through use of depth disparities derived from stereo pairs. Depth maps, generated using depth disparity information, were fused early or late during training of the detectors to determine which option resulted in better detections of scallop classes and flounder (**Figure 6**). Results from detectors developed using stereo information were compared to alternative methods using secondary chip classifiers (**Figure 7**) or secondary fine-tuning of models. The main upgraded deep learning networks used during the project included Cascade Faster-Region Convolutional Neural Networks (Cascade FRCNN, Cai & Vasconcelos 2018) and High-Resolution Networks (HRNet, Wang et al. 2020). The methods used included the following:

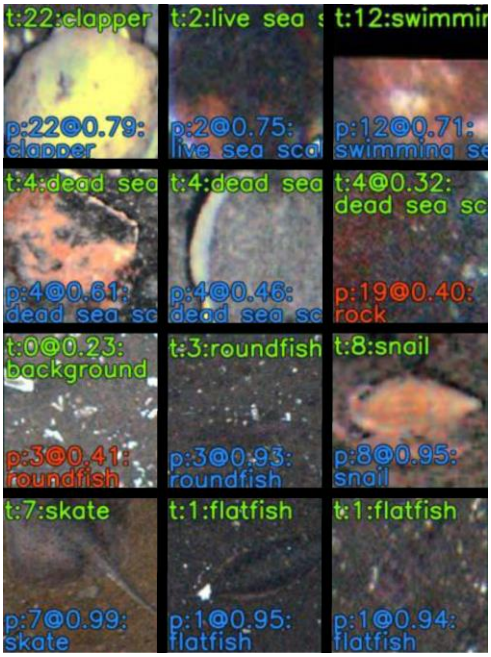


Figure 7. *Examples of secondary chips used for object classification.*

- Detection and classification (Cascade FCRNN) with and without ***early depth fusion***, with RGBD images treated as 4-channel inputs,
- Detection and classification (Cascade FCRNN and HRNet) with and without ***late depth fusion***, with RGB images and depth maps in separate streams to produce feature maps that were summed into a single scoring map,
- Detection and classification (Cascade FCRNN) based on ***secondary chips*** created from regions that were proposed using masks, and
- Detection and classification (Cascade FCRNN and HRNet) based on ***fine-tuning*** of models generated using multiple categories of target organisms or specific categories (e.g., flounders).

Appendix A provides a brief overview of image representation as layered matrices of numbers and the methods used in this report for quantifying model skill (precision-recall curves and confusion matrices).

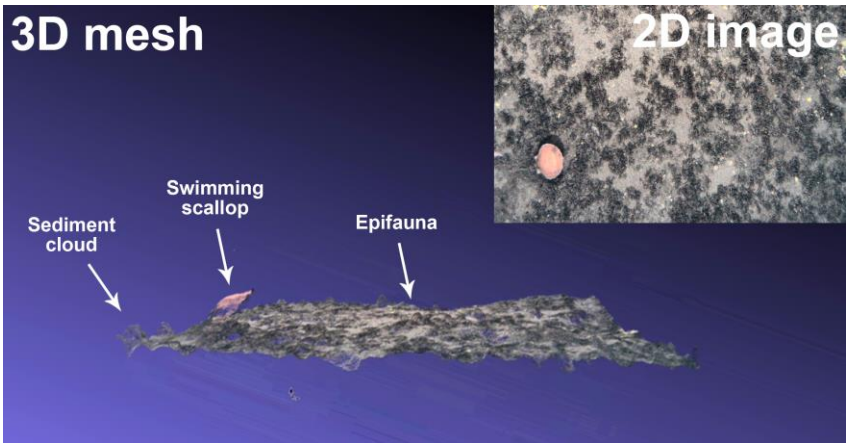


Figure 8. *Example of a 3D model of the sea floor viewed in MeshLab showing a swimming scallop and the sediment cloud it generated when jetting off the sea floor.*

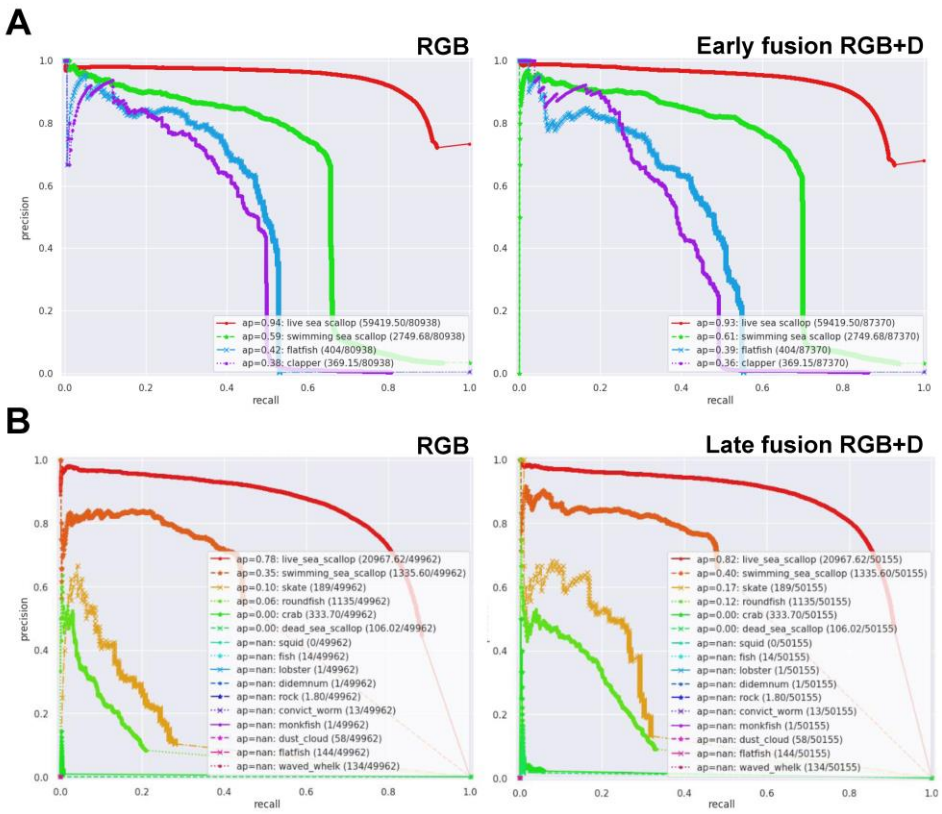


Figure 9. *Precision-Recall curves for detectors using only RGB images as inputs vs. (A) RGB images with a depth layer added as a fourth channel using early fusion and (B) RGB images and depth maps with the two joined using late fusion.*

Results

Improved stereo calibration

Recalibration improved the depth disparity maps generated from the CFF imagery. It also provided the data needed to generate 3D sea floor models that could be viewed in MeshLab, open-source software for processing, editing, and viewing 3D triangular meshes (http://www.meshlab.net/). The 3D models provided striking visual representations of the seafloor and the locations of scallops and flounder in the water column (**Figure 8**).

Scallop detectors

The inclusion of depth information via early fusion improved the mean average precision (mAP) values for swimming scallop detection by 0.04 (**Table 1** and **Figure 9A**). No improvements were seen for other scallop classes using early fusion, with mAP values decreasing slightly for live scallops by -0.004 and clappers by -0.04 (**Table 1** and **Figure 9A**). Late fusion of depth information improved detection average precision for swimming scallops by 0.05 and live scallops by 0.04 (**Table 1** and **Figure 9B**). Discrimination between swimming scallops and clappers was slightly improved with the addition of depth information through early fusion (**Figure 10**).

The results from object detectors using secondary chip classifiers were better than those incorporating depth information, with mAP values increasing for swimming scallops by 0.10 and clappers by 0.20 (**Table 1**). Discrimination between all three scallop classes was also improved (**Figure 11**).

Table 1. *Shifts in mean average precision (mAP) values from best model runs with the addition of depth disparity information (RGBD – RGB), a secondary chip classifier (RGB+chip – RGB), or secondary training for fine tuning (RGB fine – RGB) for the study target organisms using each object detection model.*

Name/type	Live scallops	Swimming scallops	Clappers	Flounder
Early Fusion	-0.004	0.04	-0.02	-0.04
Late Fusion	0.04	0.05	NA	0.03
Secondary chip classifiers	-0.05	0.10	0.20	0.04
Secondary fine-tuning	-0.002	NA	NA	0.09

Flounder detectors

The inclusion of depth information via early fusion did not improve detection of flounder, with average precision decreasing by 0.04 (**Table 1** and **Figure 9A**). The opposite trend was seen when using secondary chip classifiers, with average precision increasing by 0.04, and when using secondary fine tuning, with average precision increasing by 0.09 (**Table 1**). Discrimination between flounder and live scallops or background regions also improved more with secondary chip classifiers (**Figure 11**) than early fusion of depth maps (**Figure 10**).

It should be noted that in the above experiments, models were trained jointly across all annotated categories, including all scallop categories and flounder categories together; this biased the models toward scallops due to large class imbalances in the training sets, with orders-of-magnitude more scallops present in the training set annotations. When models were fine-tuned only for flounders (i.e., only training the detectors over the flounder category), mAP values for models with depth

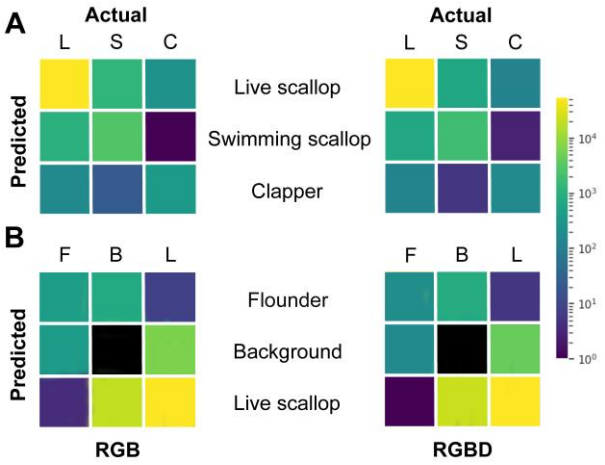


Figure 10. *Confusion matrices showing the performance of the RGB vs. RGBD early fusion detectors for classifying (A) live scallops, swimming scallops, and clappers and (B) flounder, background regions, and live scallops.*

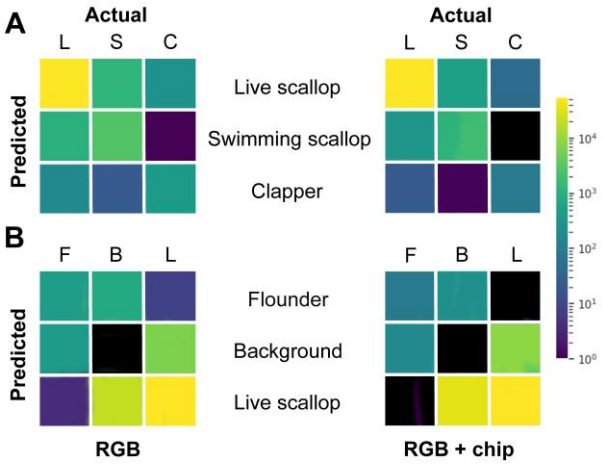
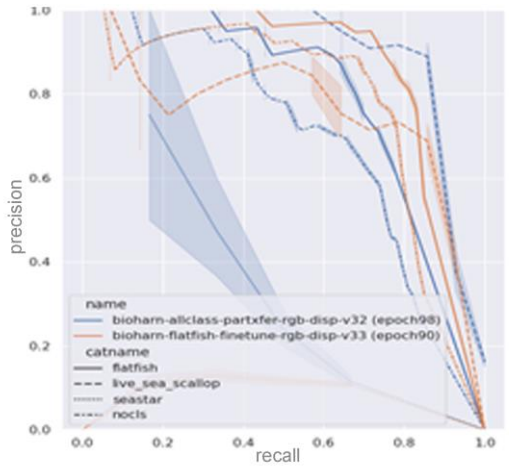


Figure 11. *Confusion matrices showing the performance of the RGB vs. RGB+chip detectors for classifying (A) live scallops, swimming scallops, and clappers and (B) flounder, background regions, and live scallops.*

information improved by a couple of percentage points (**Figure 12**). The baseline flounder-only RGB detector was also significantly improved over the detector for the flounder category in the multi-class experiments (**Figure 12**).

Examples of scallop and flounder detections are included in **Appendix B**.



Model/training	Flounder mAP
RGB all classes	0.520
RGB flounder only	0.790
RGBD flounder only	0.814

Figure 12. *Precision-recall curves and flounder mAP values for models that were fine-tuned for all categories vs. for flounder only. The flounder curve from the all-category model is shown in solid blue, while the flounder curve from the flounder-only model is shown in solid orange, shifted to the right indicating improved model performance.*

Project Outreach

Results from the project were presented at the following meetings:

- May 7, 2020 – meeting of scientific groups involved in automated detection of sea scallops, particularly those funded by an RSA grants. This meeting included representatives from CFF, Kitware, NEFSC, SMAST, C-Vision (subcontractor working with SMAST), WHOI, Coastal Ocean Vision (working with WHOI), Rutgers University, GARFO, and NEFMC.
- May 19, 2020 – NEFMC Scallop RSA Share Day. The audience for this meeting included scientists, fisheries managers, and members of the scallop fishing industry.
- September 22, 2020 – 2nd NOAA Workshop on Leveraging Artificial Intelligence in Environmental Sciences. The audience was primarily scientists from NOAA and other federal agencies.
- October 5, 2020 – Tenth Annual Meeting of the Marine Alliance for Science and Technology for Scotland. This presentation led to 50+ European scientists registering for VIAME accounts.
- May 6, 2021 – NEFMC Scallop Research Share Day. The audience for this meeting included scientists, fisheries managers, and members of the scallop fishing industry.

Scallop and flounder detectors developed with project funding are available on github (https://github.com/VIAME/VIAME#installations) as the optional patch *HabCam Models (Scallop, Skate, Flatfish), All OS*.

Evaluation

Accomplishments by objective

Accomplishments by objective are described below.

(1) Improve detection of scallops and differentiate between live, dead, and swimming scallops.

The addition of depth information from stereo image pairs through early and late fusion improved classification of swimming scallops but not clappers. Classification of live scallops was improved with late fusion but not early fusion. These result were unexpected. We hypothesized that stereo information would improve detection of and discrimination between all classes of scallops (live

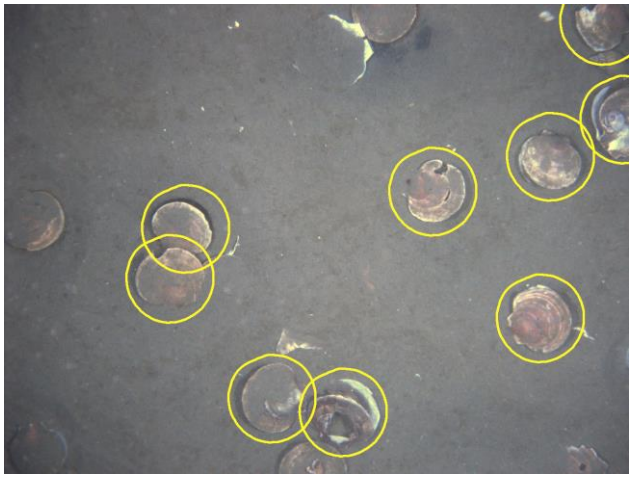


Figure 13. *Examples of the variability in the appearance of clappers (yellow circles).*

scallops, swimming scallops, and clappers) based on characteristics like distance off bottom for live vs. swimming scallops and gape distance for live scallops vs. clappers. While stereo information improved discrimination of swimming scallops slightly, it was less than expected since swimming scallops have strong signatures in depth maps. In hindsight, this result may not be surprising because human annotators do not have trouble discriminating swimming scallops in RGB images when shadows are present (**Figure 3**), and the object detector models could use the same visual cue. The lack of model improvement for detecting clappers could be due to variability in the orientation and therefore appearance of clappers (**Figure 13**). This variability also makes it difficult for human annotators to distinguish between live scallops, dead scallops, and clappers.

(2) Detect and count flounder.

The addition of depth information from stereo image pairs did not improve detection of flounder as much as expected, although there was an improvement with late fusion of depth information by a few percentage points. Flounders only occasionally created a large signature in computed depth maps, often overshadowed by differences in terrain elevation and background content. Training of flounder detectors using RGB and RGBD imagery was also problematic because flounders are sparse targets in the HabCam image archive. Even with all the HabCam survey images provided by CFF and NEFSC data from 2015 (~10,000 images for training and ~10,000 images for testing), there are only ~2100 annotations of flounder. While this number of annotations is good enough for training initial models, deep learning is traditionally performed on orders-of-magnitude more annotations.

Model average precision may have decreased with the use of RGBD images and early fusion because the depth map characteristics for a flounder can differ significantly, with some flounder being closer to the camera lens than the sediment when resting on the surface but further from the camera lens when they have buried themselves or dug a hollow in the sediment (**Figure 5**).

Discussion and future research

CFF will continue to supply annotated HabCam imagery to Kitware over the next few years. The addition of new images with annotated flounder and less common scallop classes (swimming scallops and clappers) will improve models for these relatively rare target organisms in VIAME, leading to more reliable automated detectors for these organisms of key importance to the scallop

One of the surprising results from this work is how often background regions are mistaken for both scallops and flounder. Live scallops are frequently scored as background regions and vice versa using RGB images with and without secondary chip classification and RGBD images (**Figures 10** and **11**). This may occur because human annotators mistakenly label rocks or other bivalves that are not annotated as scallops (**Figure 14**). Because object detector models are based on human annotations, which are used to train the models and quantify their skill, inaccurate human annotations would have impacts at all stages.

CFF, in collaboration with Kitware, has just received a new grant from the Scallop RSA program to develop new algorithms to improve automated classification of benthic habitats in HabCam imagery and incorporate automated measurement of target organisms, including scallop shell height measurements, into the VIAME system. Because this effort will expand automated detection beyond organism detection to habitat classification, it could reduce the number of mistaken classifications of background regions as target organisms and lead to better use of image databases

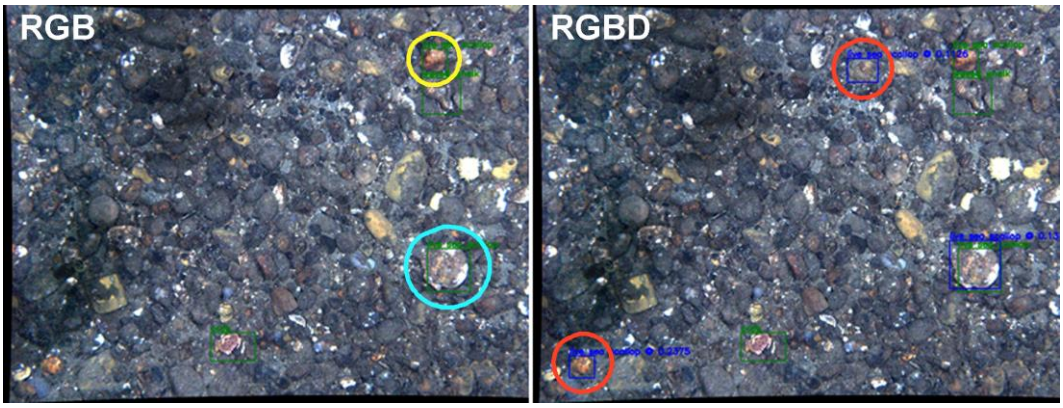


Figure 14. *Annotation and object detection model issues. The annotated RGB image included an incorrectly labeled live scallop that was actually a rock (yellow circle). The model trained on RGB images did not detect the rock and label it as a scallop, but it did miss a correctly labeled scallop (blue circle). The model trained on RGBD images did detect this scallop, but it also incorrectly labeled two rocks as scallops (red circles).*

for delving into questions about ecology and habitat changes. As part of the newly funded project, CFF plans to investigate habitat preferences of pre-recruit scallops using HabCam imagery collected over multiple years in Closed Area II and the southern flank of Georges Bank and in the Elephant Trunk and Hudson Canyon access areas in the Mid-Atlantic.

Literature Cited

Cai Z, Vasconcelos N. 2018. Cascade r-cnn: Delving into high quality object detection. Proceedings of the IEEE conference on computer vision and pattern recognition: 6154-6162.

Chang JH, Shank BV, and Hart DR. 2017. A comparison of methods to estimate abundance and biomass from belt transect surveys. Limnology and Oceanography: Methods 15: 480-494.

Dawkins M, Sherrill L, Fieldhouse K, Hoogs A, Richards B, et al. 2017. An open-source platform for underwater image and video analytics. In IEEE Winter Conference on Applications of Computer Vision 2017. pp 898-906.

Hennen DR and Hart DR. 2012. Shell height-to-weight relationships for Atlantic sea scallops (*Placopecten magellanicus*) in offshore US waters. Journal of Shellfish Research 31: 1133-1144.

Li Y, Zhang J, Cheng Y, Huang K, Tan T. 2017. Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation. 2017 IEEE International Conference on Image Processing: 1262-1266.

Maguire JJ. 2015. Summary Report of the Review of Sea Scallop Survey Methodologies and Their Integration for Stock Assessment and Fishery Management.

https://www.nefsc.noaa.gov/saw/scallop-2015/pdfs/scallop-surveys-review-summary-report-april-9.pdf.

National Marine Fisheries Service (NMFS). 2020. Fisheries of the United States 2018. Current Fishery Statistics No. 2018. 167 pp.

New England Fishery Management Council (NEFMC). 2019. Atlantic Sea Scallop Fishery Management Plan Framework Adjustment 30. https://s3.amazonaws.com/nefmc.org/190307-FW30-Final-Submission.pdf.

NEFMC. 2020. Scallop Fishery Management Plan Framework Adjustment 32. https://s3.amazonaws.com/nefmc.org/Framework-32-Final-Submission_signed-FONSI.pdf.

O'Keefe, C and DeCelles, G. 2013. Forming a partnership to avoid bycatch. Fisheries 38: 434-444.

Shumway SE and Parsons GJ, eds. 2016. Scallops: Biology, Ecology, Aquaculture, and Fisheries. Elsevier Publishing. Amsterdam, Netherlands. 1196 pp.

Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, Liu D, Mu Y, Tan M, Wang X, Liu W. 2020. Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence.

Appendix A

RGB and image representation in computer vision programs

When loaded into computer programs, images are normally represented as three layers, or channels, of two dimensional matrices of numbers, with each pixel in each layer represented as a number (**Figure A1**). That number can range from 0 (black or no color) to 2^{bit}-1 (white or maximum color); the maximum value per pixel is 255 in 8-bit images, 4095 in 12-bit images, and 65,535 in 16-bit images. Mathematical operations can be performed on the stack of matrices to quantify and adjust brightness, contrast, and color balance. They can also be used to quantify pattern scale and orientation, shapes, and edges. Taken together, these image characteristics are used for object detection, which combines image localization (defining location of objects that share certain characteristics), and classification (labeling the objects by class).

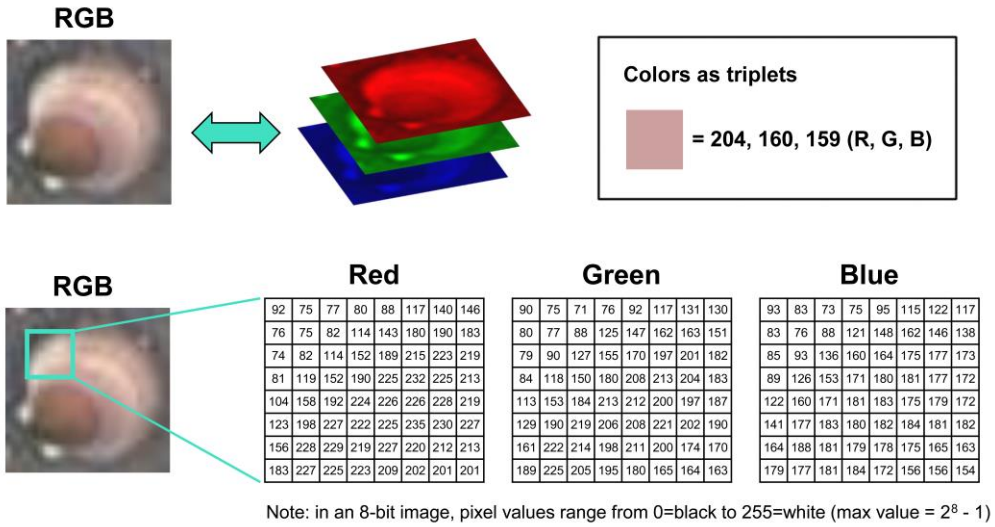


Figure A1. *RGB images as three layers or channels of matrices.*

Stereo images can be used to compute a depth map by looking at the differences in the location of key points in a pair of images taken by camera lenses with a defined spatial relationship. That depth map can be added as a fourth layer or channel to the RGB image, or it can be treated as its own one-channel image map (**Figure A2**).

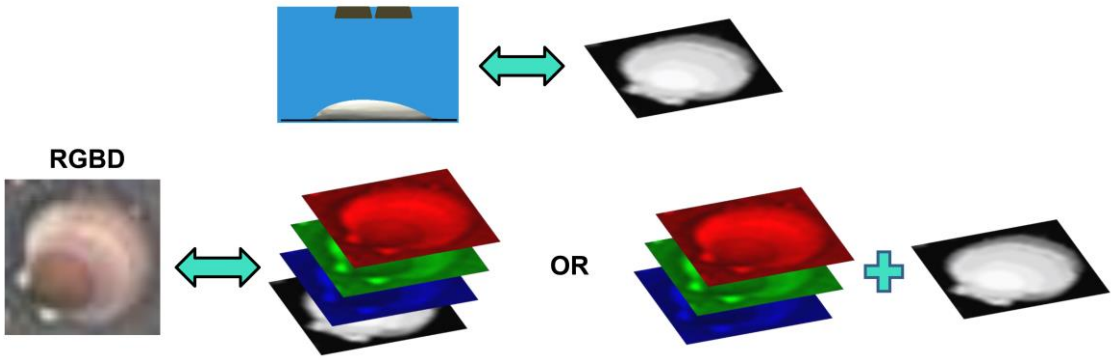


Figure A2. *RGBD images as four layers, with depth disparities based on point distances from the camera lenses.*

Quantifying object detection model skill

The skill of an object detection model can be quantified based on the number of correct detections and the number of errors. The groundtruth for determining errors comes from the human annotations. Types of errors include:

- False Positives – detect an object when none is there (i.e., predict a scallop when the boxed location is a background region) or the object does not belong to that class (i.e., predict a scallop when it is really a flounder)
- False Negatives – predict no object when one is there or label the object incorrectly (i.e., label a true scallop as a flounder)

Models can include threshold values to balance these two types of errors.

Model skill can be summarized using these error rates in a variety of ways including Precision-Recall (PR) curves (**Figure A3**). PR curves plot Recall on the x-axis vs. Precision on the y-axis with these quantities defines as:

- Precision = True Positives / (True Positives + False Positives)
- Recall = True Positives / (True Positives + False Negatives)

The Average Precision (AP) can be calculated as the area under the resulting curve.

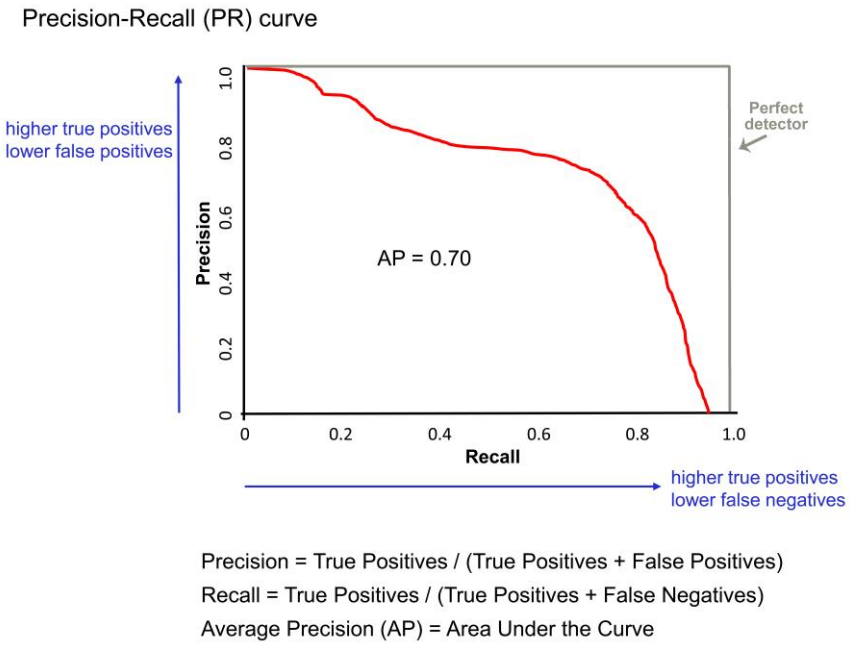


Figure A3. *A Precision-recall curve.*

Model skill can also be visualized using a confusion matrix. (**Figure A4**). For a multiclass model, each entry in the matrix denotes the number of predictions in each class, parsed out by each actual class. Performance measures can be calculated from a confusion matrix by summing the true positives, true negatives, false positives, and false negatives in the matrix.

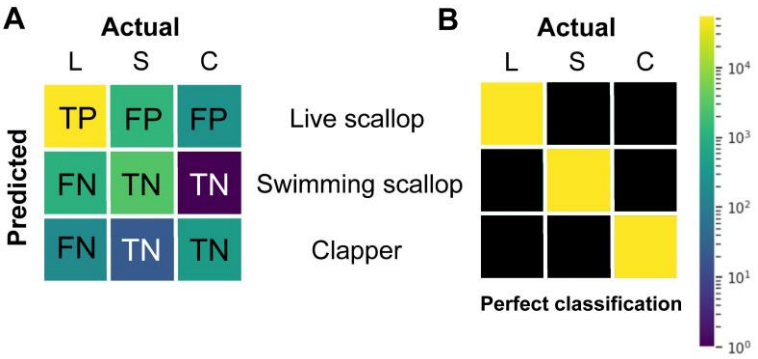
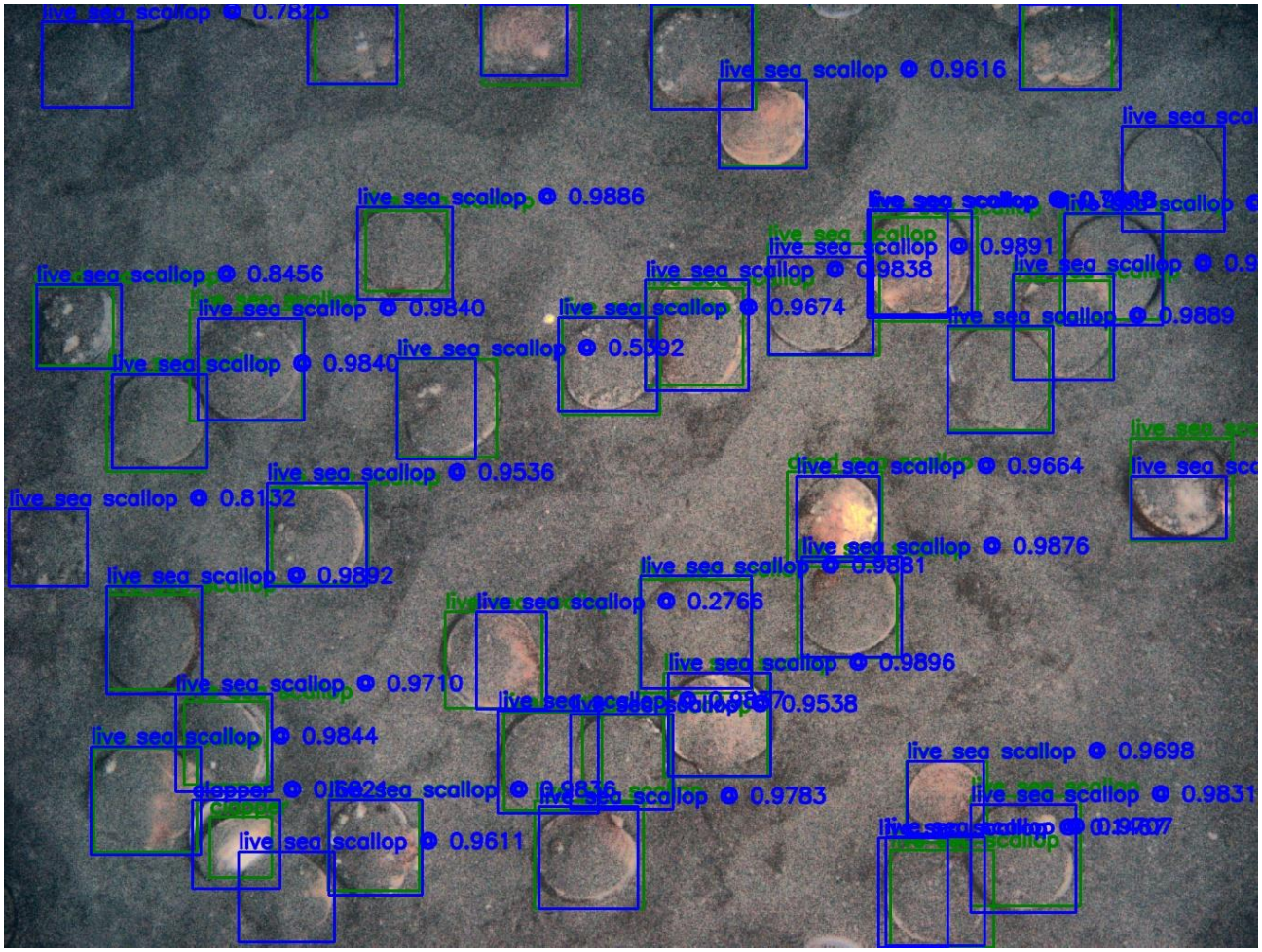


Figure A4. *Example of a confusion matrix showing (A) the locations of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) boxes for live scallops. Actual object classes are in columns and predicted object classes are in rows. Based on the color scale used, live scallops are incorrectly classified as swimming scallops more often than clappers. (B) A confusion matrix for a perfect set of detectors using the same color scale.*

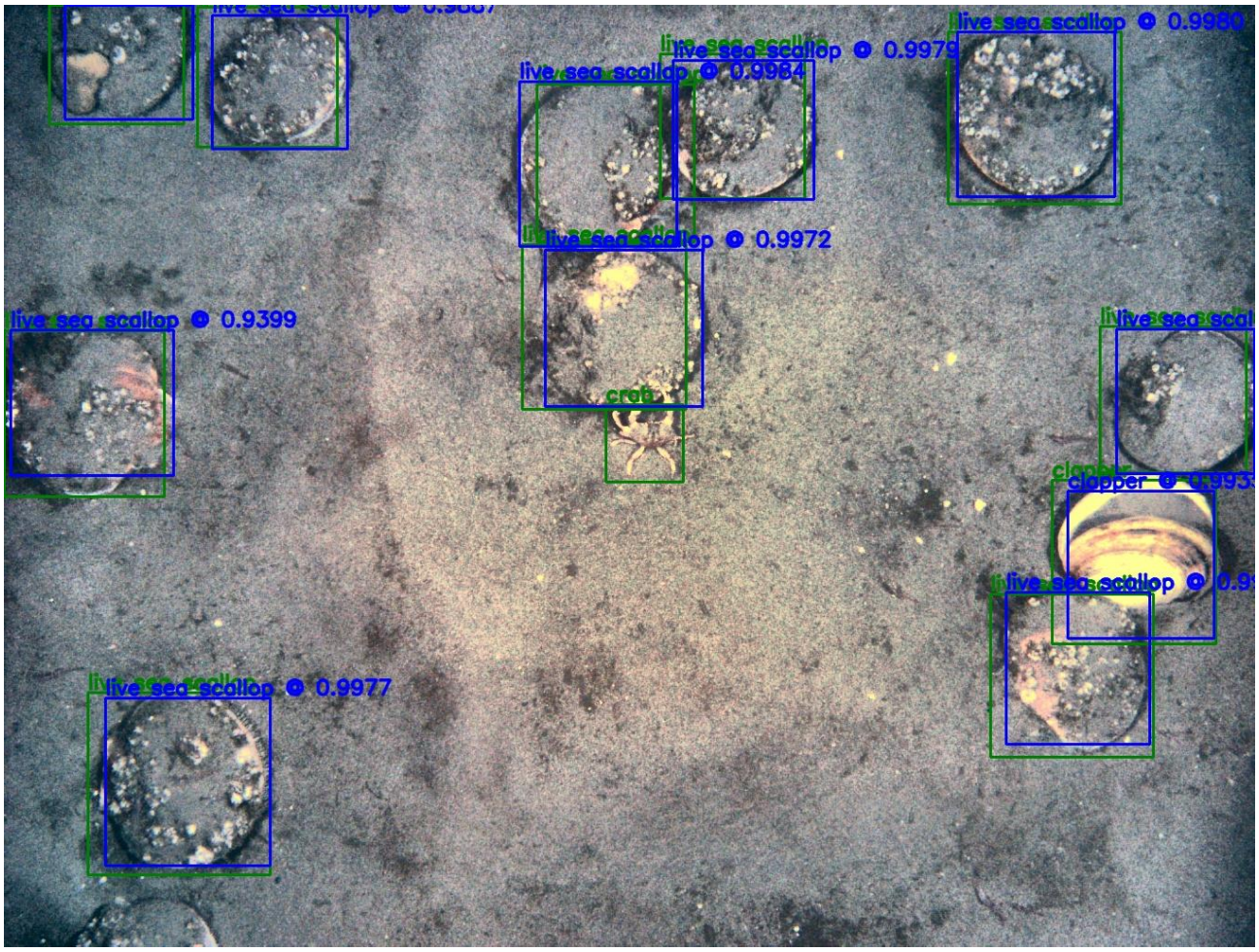
Appendix B

Examples of model detections. Unless otherwise indicated, groundtruth (i.e., human) annotations are shown in green and detector-model predictions are shown in blue. The numbers by the predicted boxes indicate how much the predicted box overlaps with the groundtruth box (Intersection over Union).

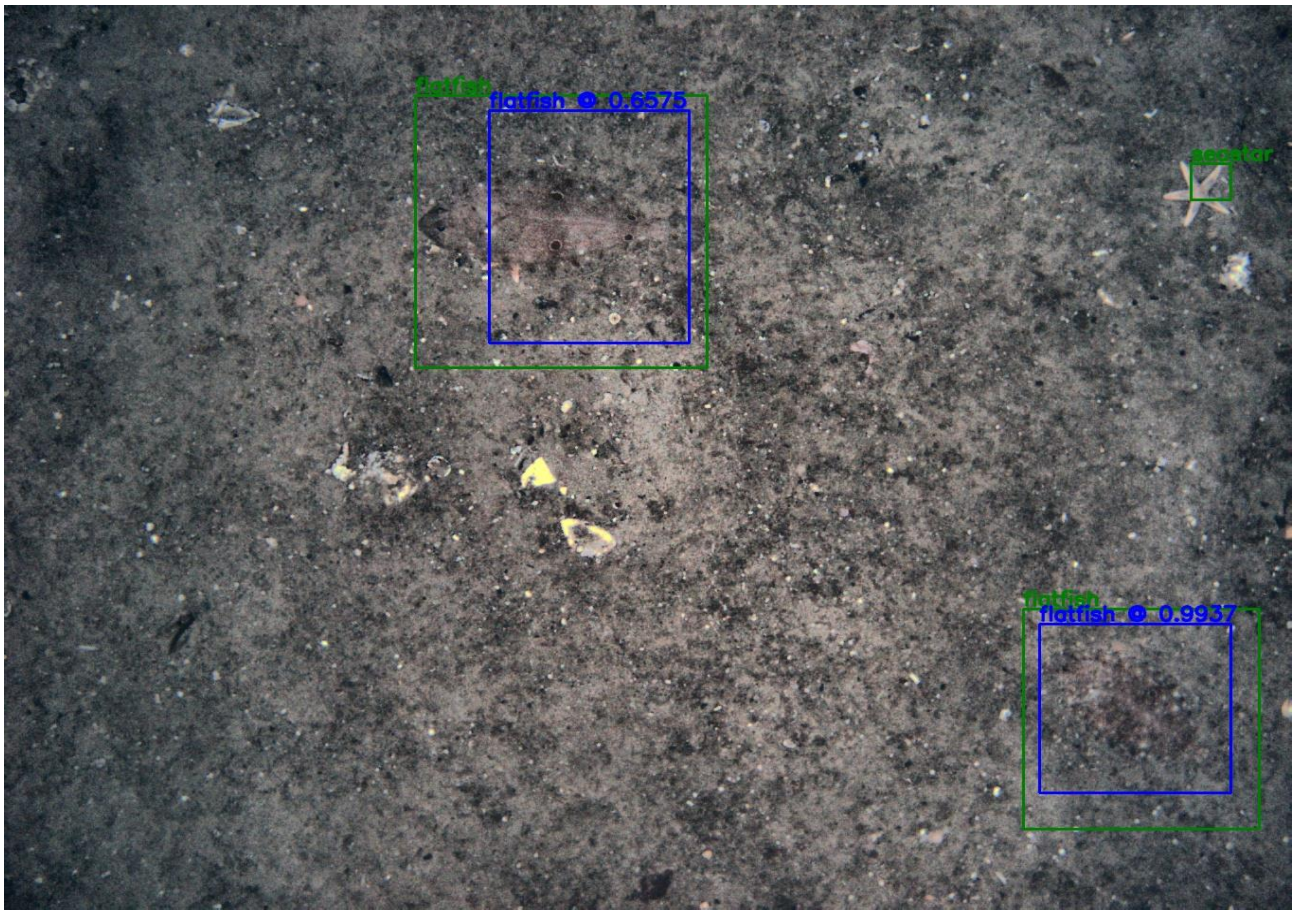
RGB – example of scallop annotations in a dense bed.



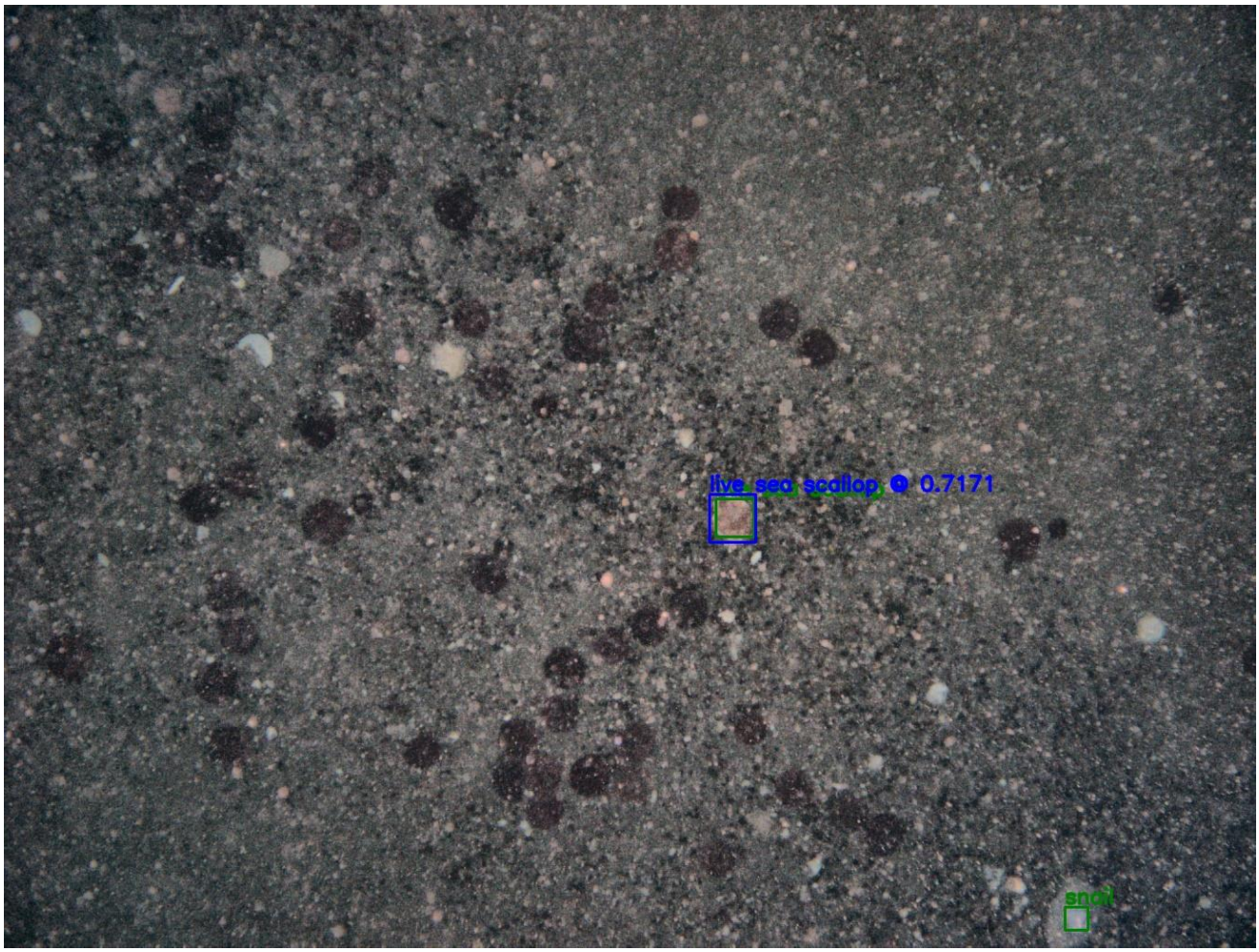
RGB – live scallops and a clapper.



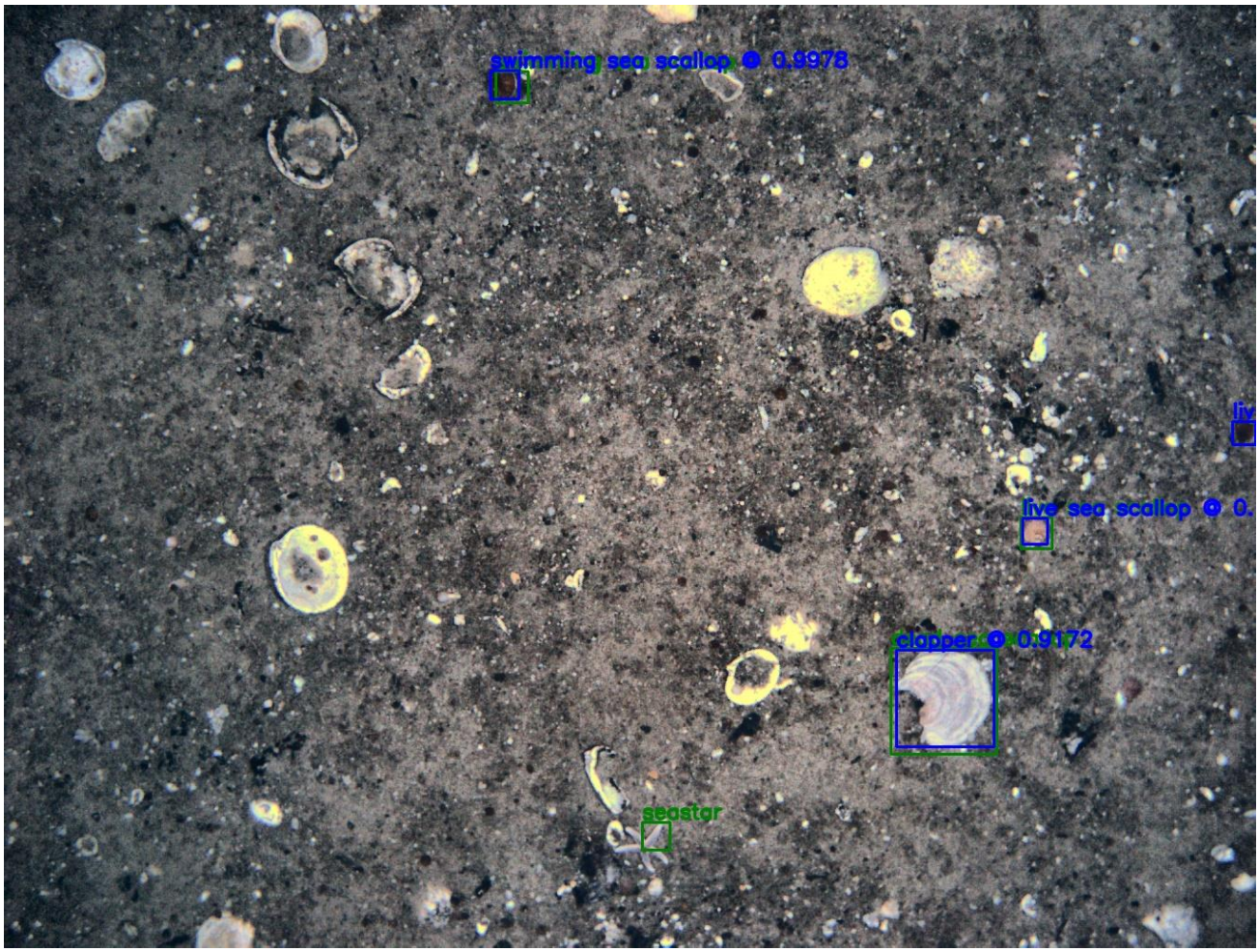
RGB – flounder annotations.



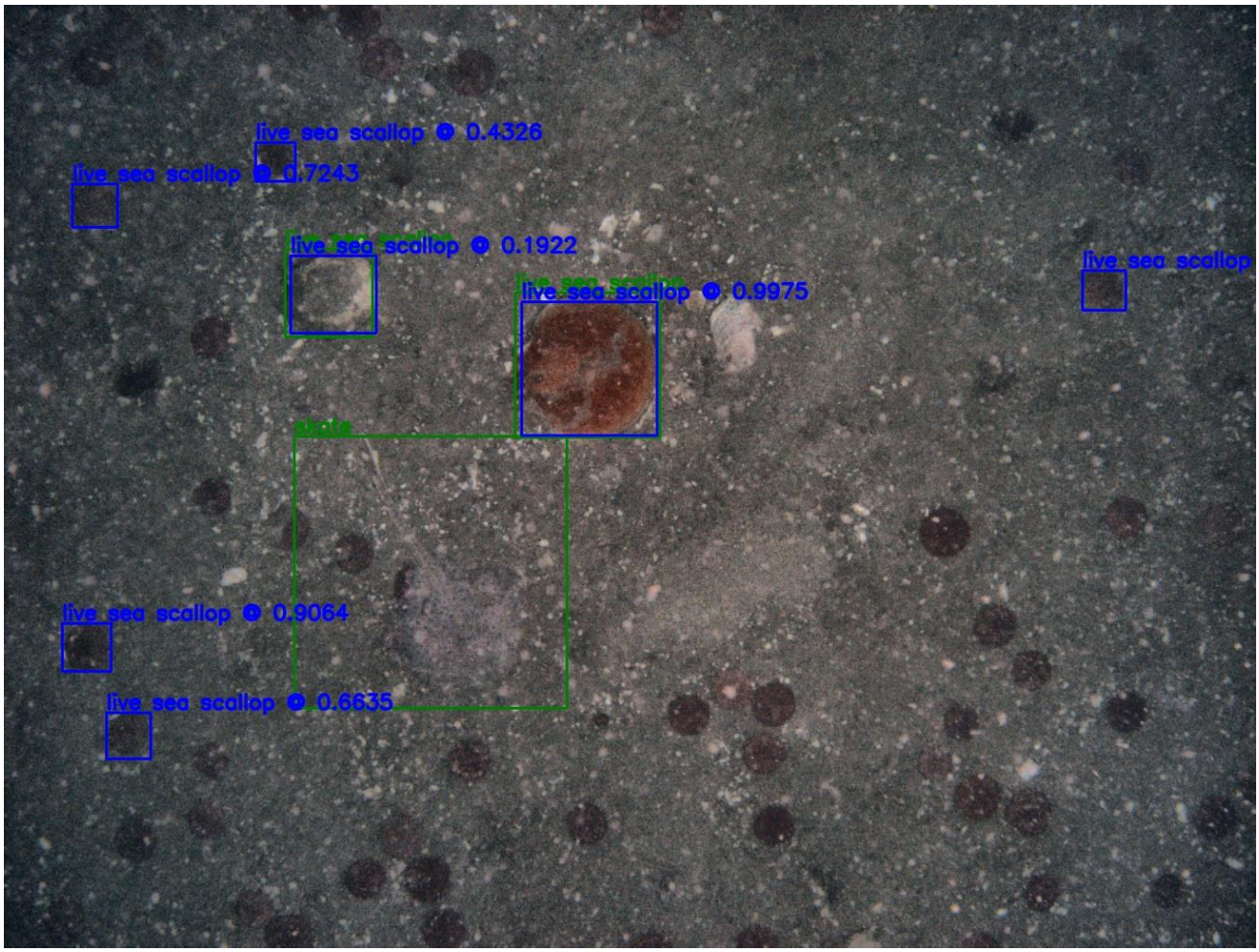
RGB- scallop picked out in a bed of sand dollars.



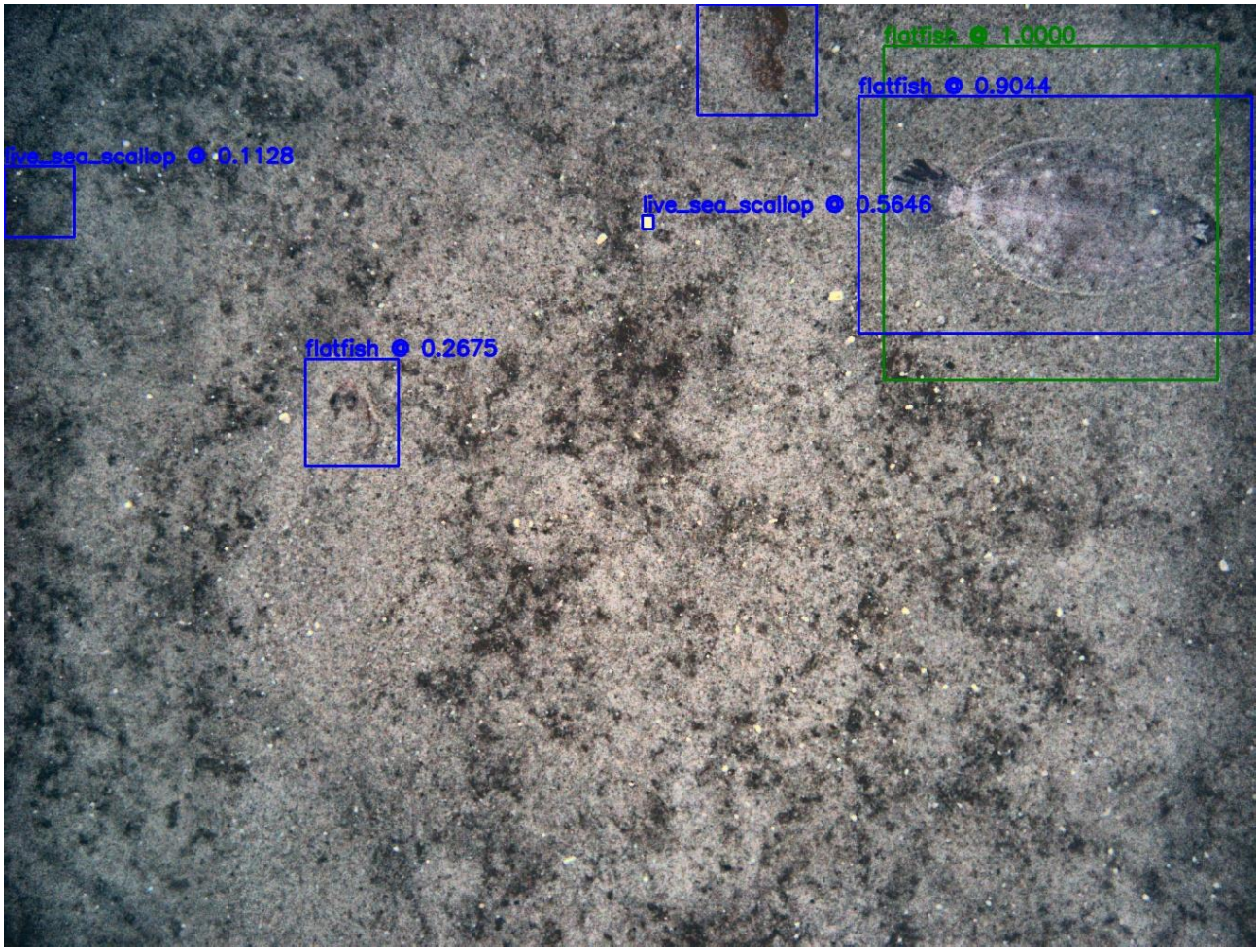
RGB- live scallop, swimming scallop, and clapper.



RGB – false positives with sand dollars labeled as scallops.



Secondary training – model detects flounder missed by human.



RGB (magenta) vs. RGBD with late fusion (orange) – addition of depth map through late fusion avoids false positive flounder detections.