

Title: Automating the analysis of fish abundance using object detection: optimising animal ecology with deep learning

Author list: Ellen M. Ditria<sup>1\*</sup>, Sebastian Lopez-Marcano<sup>1</sup>, Michael K. Sievers<sup>1</sup>, Eric L. Jinks<sup>1</sup>, Christopher J. Brown<sup>2</sup> Rod M. Connolly<sup>1</sup>

\*Corresponding author: Ellen Ditria: [ellen.ditria@griffithuni.edu.au](mailto:ellen.ditria@griffithuni.edu.au)

<sup>1</sup>Australian Rivers Institute – Coast & Estuaries, and School of Environment and Science, Griffith University, Gold Coast, QLD 4222, Australia

<sup>2</sup>Australian Rivers Institute – Coast & Estuaries, and School of Environment and Science Griffith University, Nathan, QLD 4111, Australia

All authors have seen and approved this manuscript. This manuscript has not been accepted or published elsewhere.

1 Abstract

2 Aquatic ecologists routinely count animals to provide critical information for conservation  
3 and management. Increased accessibility to underwater recording equipment such as cameras  
4 and unmanned underwater devices have allowed footage to be captured efficiently and safely.  
5 It has, however, led to immense volumes of data being collected that require manual  
6 processing, and thus significant time, labour and money. The use of deep learning to  
7 automate image processing has substantial benefits, but has rarely been adopted within the  
8 field of aquatic ecology. To test its efficacy and utility, we compared the accuracy and speed  
9 of deep learning techniques against human counterparts for quantifying fish abundance in  
10 underwater images and video footage. We collected footage of fish assemblages in seagrass  
11 meadows in Queensland, Australia. We produced three models using a MaskR-CNN object  
12 detection framework to detect the target species, an ecologically important fish, luderick  
13 (*Girella tricuspidata*). Our models were trained on three randomised 80:20 ratios of  
14 training:validation data-sets from a total of 6,080 annotations. The computer accurately  
15 determined abundance from videos with high performance using unseen footage from the  
16 same estuary as the training data (F1 = 92.4%, mAP50 = 92.5%), and from novel footage  
17 collected from a different estuary (F1 = 92.3%, mAP50 = 93.4%). The computer's  
18 performance in determining MaxN was 7.1% better than human marine experts, and 13.4%  
19 better than citizen scientists in single image test data-sets, and 1.5% and 7.8% higher in video  
20 data-sets, respectively. We show that deep learning is a more accurate tool than humans at  
21 determining abundance, and that results are consistent and transferable across survey  
22 locations. Deep learning methods provide a faster, cheaper and more accurate alternative to  
23 manual data analysis methods currently used to monitor and assess animal abundance. Deep  
24 learning techniques have much to offer the field of aquatic ecology.

25

26 Keywords: automation, deep learning, object detection, computer vision, fish abundance,  
27 monitoring tools

28

29

30

31

## 32 1. Introduction

33 The foundation for all key questions in animal ecology revolves around the abundance,  
34 distribution and behaviour of animals. Collecting robust, accurate and unbiased information  
35 is therefore vital to understanding ecological theories and applications. Many of the invasive  
36 data collection methods traditionally used to collect this information in animal ecology, such  
37 as tagging, netting and trawling, are now largely unnecessary due to remote data collection  
38 using cameras. The development and availability of these devices have facilitated more  
39 accurate and cheaper methods of data collection, with reduced risk to the operator (Hodgson  
40 et al. 2013). Most importantly from a scientific perspective, they have increased sampling  
41 accuracy as well as replicability and reproducibility (Weinstein 2017), which form the basis  
42 of a sound scientific study (Leek & Peng 2015). However, the amount of data now being  
43 generated can be overwhelming. The solution has become the new problem.

44

45 Much like the physical collection of data, manual processing of data is often labour-intensive,  
46 time-consuming and extremely costly (Weinstein 2017). This has led to invaluable data  
47 collected over large temporal and spatial scales laying unused in storage libraries. In  
48 Australia, for example, the Integrated Marine Observing System (IMOS) collects millions of  
49 images of coral reefs every year, yet despite affiliations and partnerships with a range of  
50 universities and management agencies, less than 5% of these are analysed by experts  
51 (Moniruzzaman et al. 2017). This apparently never-ending stream of data brings a new  
52 challenge for ecologists; to find or develop the analytical tools needed to extract information  
53 from the immense volumes of incoming images and video content (Valletta et al. 2017).

54

55 Fortunately, recent advances in machine learning technologies have provided one such tool to  
56 help combat this problem; deep learning. Deep learning is a subset of machine learning  
57 consisting of a number of computational layers within an architectural framework designed to  
58 process data that is difficult to model analytically, such as raw images and video footage  
59 (LeCun et al. 2015). Although neural networks are not a new technology (Rawat & Wang  
60 2017), the relatively recent advances in graphics processing units (GPUs) have spurred an  
61 increase in their application for computer vision data. In the CNN, data are fed into an input  
62 layer, while an output layer is sorted into categories pre-determined by manual training, in a  
63 process known as supervised learning (Rawat & Wang 2017).

64

65 Although deep learning techniques are being implemented enthusiastically in terrestrial  
66 ecology, it is currently an under-exploited tool in aquatic environments (Moniruzzaman et al.  
67 2017, Xu et al. 2019). As the global challenges in marine science and management increase  
68 (Halpern et al. 2015), it is critical for marine science to realise the potential automated  
69 analysis offers (Malde et al. 2019). Relative to terrestrial environments, however, obtaining  
70 useable footage in marine environments to achieve acceptable computational performance  
71 presents a unique set of challenges. For example, there are often high levels of environmental  
72 complexities in marine environments which can interfere with clear footage, including  
73 variable water clarity, complex background structures, decreased light at depth, and  
74 obstruction due to schooling fish (Mandal et al. 2018, Salman et al. 2019). Although these  
75 factors may affect the quality of images and videos, deep learning methods have proven  
76 successful in a range of marine applications (Galloway et al. 2017, Arellano-Verdejo et al.  
77 2019).

78

79 Efforts to use deep learning methods in marine environments currently revolve around the  
80 automated *classification* of specific species. Attempts to classify tropical reef fish have  
81 achieved high levels of performance and have also outperformed humans in species  
82 recognition (Villon et al. 2018). There have also been suggestions from classification studies  
83 on freshwater fish to incorporate other strategies for increasing performance, such as  
84 including taxonomic family and order (dos Santos & Gonçalves 2019). Although all marine  
85 environments have challenging conditions, the tropical reef studies by Villon et al. (2018)  
86 and Salman et al. (2019) typically operate with high visibility, high fish abundance, and  
87 highly variable inter-specific morphology, which makes distinguishing different species  
88 easier (Xu & Matzner 2018). Conversely, coastal and estuarine systems often suffer poor  
89 visibility due to complex topography, anthropogenic eutrophication, and sediment induced  
90 turbidity (Lehtiniemi et al. 2005, Baker & Sheaves 2006, Lowe et al. 2015).

91

92 Although classification enables the determination of species, its usefulness for answering  
93 broad ecological questions is rather limited. Object detection allows us to classify both *what*  
94 is in a frame and *where* it is and therefore enables us to determine both the species in an area  
95 and their abundance (eg. Maire et al. 2015, Salberg 2015, Gray et al. 2019b).

96

97 Here, we use fish inhabiting subtropical seagrass meadows as a case study to explore the  
98 viability of computer vision and deep learning as a suitable, non-invasive technique using  
99 remotely collected data in a variable marine environment. Seagrass meadows provide critical  
100 ecosystem services such as carbon sequestration, nutrient cycling, shoreline stabilisation and  
101 enhanced biodiversity (Waycott et al. 2009, Sievers et al. 2019). However, many seagrass  
102 meadows are being lost and degraded due to a range of anthropogenic stressors, such as  
103 overfishing, eutrophication and physical disturbances (Orth et al. 2006). Due to their  
104 background complexity, constant movement, and ability to obscure fish, seagrass may prove  
105 to be a difficult habitat to implement a deep learning solution. Luderick (*Girella tricuspidata*)  
106 is a common herbivorous fish found along the east coast of Australia and is abundant in  
107 coastal and estuarine systems, including seagrass meadows (Ferguson et al. 2013). Unlike  
108 most herbivorous fish in seagrass meadows, this species grazes on both the epiphytic algae  
109 that grows on seagrass and the seagrass itself, making it of interest ecologically (Gollan &  
110 Wright 2006). Using this ecologically important ecosystem, we specifically aim to deduce  
111 whether deep learning techniques can be used to determine: (1) the accurate object detection  
112 of a target species, (2) the flexibility of algorithms in analysing data across locations, and (3)  
113 the comparative performance between computers and humans in determining abundance  
114 from images and video footage. As far as we are aware, this is the first time that humans and  
115 deep learning algorithms have been compared in their ability to quantify abundance from  
116 underwater video footage, or that object detection and computer vision methods have been  
117 used in estuarine systems.

118

## 119 2. Methods

### 120 2.1 Training data-set

121 We used submerged action cameras (Haldex Sports Action Cam HD 1080p) to collect video  
122 footage of luderick in the Tweed River estuary in southeast Queensland (-28.169438,  
123 153.547594), between February and July 2019. Each sampling day, six cameras were  
124 deployed for 1 h over a variety of seagrass patches; the angle and placement of cameras was  
125 varied among deployment to ensure a variety of backgrounds and fish angles. Videos were  
126 trimmed for training to contain only footage of luderick and split into 5 frames per second.

127

128 *2.2 Convolutional Neural Network*

129 The object detection framework we used is an implementation of Mask R-CNN developed by  
130 Massa & Girshick (2018). Mask R-CNN works by classifying and localising the region of  
131 interest (RoI). It extends previous frameworks in that it can predict a segmentation mask on  
132 the RoI, and currently has the highest performance output for deep learning models (He et al.  
133 2017, Dai et al. 2019). To develop our model, we used a ResNet50 configuration, pre-trained  
134 on the ImageNet-1k data-set. This configuration provides an acceptable balance between  
135 training time and performance (Massa & Girshick 2018). We conducted the model training,  
136 testing and prediction tasks on a Microsoft Azure Data Science Virtual Machine powered by  
137 an NVIDIA V100 GPU. Data preparation and annotation tasks were carried out using  
138 software developed at Griffith University. While deep learning has begun to be adopted for  
139 ecological data analysis in the last two years, its use in the environmental sciences requires  
140 substantial software engineering knowledge, as unfortunately there is not yet an accessible  
141 software package for ecologists (Piechaud et al. 2019). The development of this interface for  
142 manual annotation, that can be retrained for different species, takes strides towards an end-to-  
143 end, user-friendly application tailored for ecologists. A trained team in fish identification  
144 manually drew segmentation masks around luderick (i.e. our RoI, Fig. 2.1) and annotated  
145 6,080 fish for the training data-set. Luderick were annotated if they could be positively  
146 identified at any time within the video the image came from.

147



148

149 Fig. 1. Training data-set image demonstrating manual segmentation mask (white dashed line  
150 around fish) denoting the region of interest (RoI).

151

152 The utility of the model depends on how accurately the computer identifies the presence of  
153 luderick, which we quantified in two ways based on the interactions between precision (P)  
154 and recall (R). Precision is how rigorous the model is at identifying the presence of luderick,  
155 and recall is the number of the total positives the model captured (Everingham et al. 2010).  
156 Generally, an increase in recall results in decreased precision and vice versa and were  
157 calculated as follows:

$$158 \quad \text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

159

$$160 \quad \text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

161

162 Firstly, the computer's ability to fit a segmentation mask around the RoI was determined by  
163 the mean average precision value (mAP) (Everingham et al. 2010).

$$164 \quad mAP = \int_0^1 P(R) dR$$

165 We used the mAP50 value in this study, which equates to how well the model overlapped a  
166 segmentation mask around at least 50% of the ground truth outline of the fish. The higher this  
167 value, the more accurate the model was at overlapping the segmentation mask. Secondly, the  
168 success of our model in answering ecological questions on abundance was determined by an  
169 F1 score:

$$170 \quad F1 = 2 \times \frac{P \times R}{P + R}$$

171 We used the F1 score and mAP50 values to assess the performance of the computer model.  
172 All predictions were made with a confidence threshold of 90%, that is, the algorithm was at  
173 least 90% sure that it was identifying a luderick to minimise the occurrence of false  
174 negatives. This threshold was chosen as it typically maximised F1 performance by filtering  
175 out false positives.

176

### 177 *2.3 Model Validation and Performance Curve*

178 Models were trained using a random 80% sample of the annotated dataset, with the remaining  
179 20% used to form a validation dataset (Alexandropoulos et al. 2019). Training performance  
180 was then measured against the validation set to monitor for overfitting. Overfitting is a  
181 phenomenon when the computer becomes dependent on, and memorises the training data,

182 failing to perform well when tested on data it has not encountered previously (Chicco 2017).  
183 We minimised overfitting by using the early-stopping technique (Prechelt 1998). In our case,  
184 this was achieved by assessing the mAP50 on the validation set at intervals of 2,500  
185 iterations and determined where the performance began to drop (Chicco 2017).  
186 The same computer algorithm was used to train three different models on three different  
187 randomised 80/20 subsets of the whole training data set to account for variation in the  
188 training and validation split. These models were subsequently used to compare the unseen  
189 and novel test data-set, and in the human vs computer test.

190

191 We generated a performance curve to confirm that variation among models was sufficiently  
192 low to ensure consistency in in performance across the three models. Random subsets of still  
193 images were selected from the training data-set. These subsets of data increased in volume to  
194 determine the performance of the model as training data increase. As the volume of training  
195 data increased, the risk of overfitting decreased so the number of training iterations were  
196 adjusted to maintain optimum performance.

197

198 Manual annotation cost can be a significant factor to consider when training CNN networks  
199 and can also be monitored by using the performance curve. Time stamps were added to the  
200 training software to record the speed at which training data was annotated to infer total  
201 annotation time of the training data by humans. We used this data to determined how much  
202 training is required by this model to produce high accuracy, and thus also the effort needed to  
203 produce a consistent and reliable ecological tool.

204

#### 205 *2.4 Model performance*

206 The 80/20 validation test is an established method in machine learning to assess the expected  
207 performance of the final model (Alexandropoulos et al. 2019). However, using deep learning  
208 to answer ecological questions requires another testing procedure to accurately reflect the  
209 usability of the model when analysing new data. We therefore also tested the model against  
210 annotations from two types of new footage not used for the training data-set. We used unseen  
211 footage from the same location in the Tweed River estuary ('Unseen'), as well as from a  
212 novel location ('Novel'), being seagrass meadows in a separate estuary system in  
213 Tallebudgera Creek (-28.109721, 153.448975). A t-test was used to compare the performance



214 of the three models between the unseen test-set from Tweed estuary, and the novel test-set  
215 from Tallebudgera.

216

### 217 *2.5 Human vs Computer*

218 Creating an automated data analysis system aims to lessen the manual workload of humans  
219 by creating a faster, yet accurate, alternative. Therefore, it is crucial to not only know how  
220 well the model performs, but also to assess its capabilities in speed and accuracy, compared  
221 to current human methods. This “human vs computer” method analysis compared Citizen  
222 Scientists and Experts against the computer: 1) Citizen Scientists were undergraduate marine  
223 science students and interested members of the public (n = 20) 2) Experts were fish scientists  
224 with a PhD or currently studying for one (n = 7),, and 3) the computer models (n = 3). We  
225 compared these groups using both video footage (n=31) and images (n=50), and analysed  
226 differences in test speed and performance. Both the image set and videos were run through  
227 the three deep learning models to account for variation in performance in the 80% of training  
228 data used to train the models. The number of false negatives, false positives, proportion of  
229 accurate answers (observed answers divided by ground truth) as well as the overall F1 score  
230 were recorded. Citizen Scientist and Experts were provided with a package that contained a  
231 link to the video test uploaded to YouTube, the image set sent as a zip file, instruction sheet,  
232 example images of the target species and datasheets. This process was set up to minimise bias  
233 in training the human subjects that may have occurred if the test was explained verbally.  
234 Humans were instructed to only record the target species if they could visually identify the  
235 Luderick with confidence. Participants were required to estimate the maximum number of  
236 Luderick in any single frame per video and per still image (MaxN), simulating the most  
237 popular manual method currently used in analysing videos (e.g. Gilby et al. 2017). Start and  
238 end time of each test was also recorded to compare how quickly the participants completed  
239 the task, compared to the deep learning algorithm. The still image data-set was randomly  
240 selected from the “unseen” test video footage and used as the ground truth for images. The  
241 video footage was expertly annotated at five frames per second and used as the ground truth  
242 for videos. Luderick were only annotated if they could be positively identified at least at one  
243 instance in the video. This enabled us to quantitatively compare the human and computer  
244 accuracy in determining MaxN, assessed using the overall F1 score for each test.

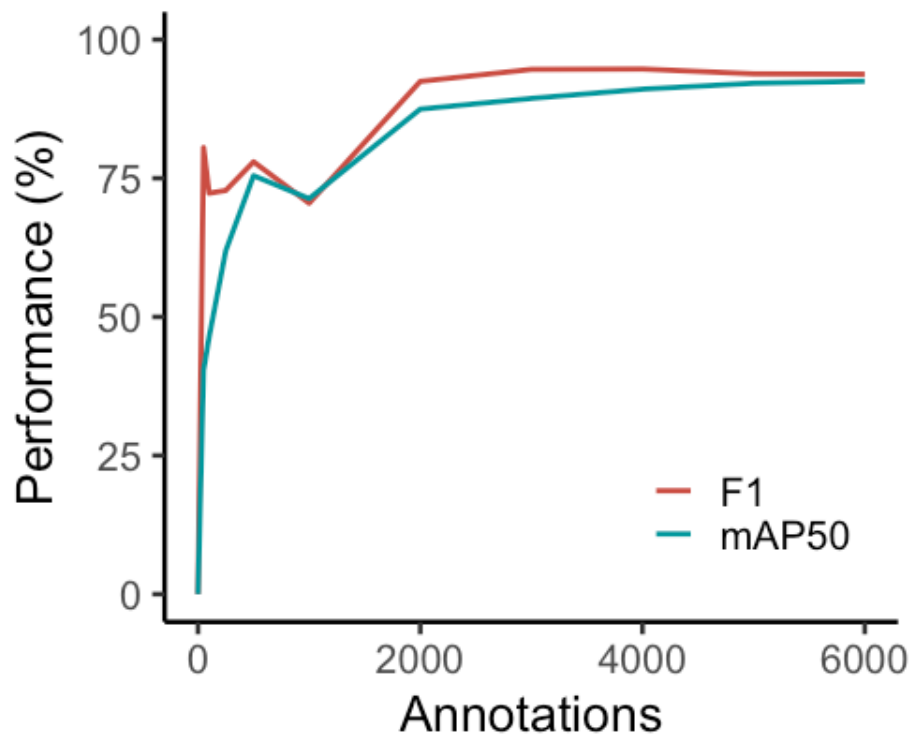
245

246 3. Results

247 3.1 Performance curve

248 Based on the computer algorithm curve, F1 performance began to plateau earlier than mAP50  
249 (Fig. 2.). F1 varied only 0.9% from 2,000 annotations to 6,000 annotations compared to an  
250 increase of 3.1% by mAP50 at the same annotations. At lower volumes of training  
251 annotations (between 0 and 1,000), the performance of both mAP50 and F1 fluctuated. Even  
252 with our streamlined process for annotation, the average time for an operator to annotate one  
253 fish was 36 seconds, and the total time to annotate all 6,080 images was in the order of 60  
254 hours.

255



256

257 Fig. 2. Performance curve showing the computer's ability to fit a segmentation mask around  
258 the luderick (performance scored by mAP50) and in identifying abundance (performance  
259 scored by F1).

260

261

262 *Model performance*

263 Performance was high for both the Unseen and Novel test sets (mAP and F1 both >92%).

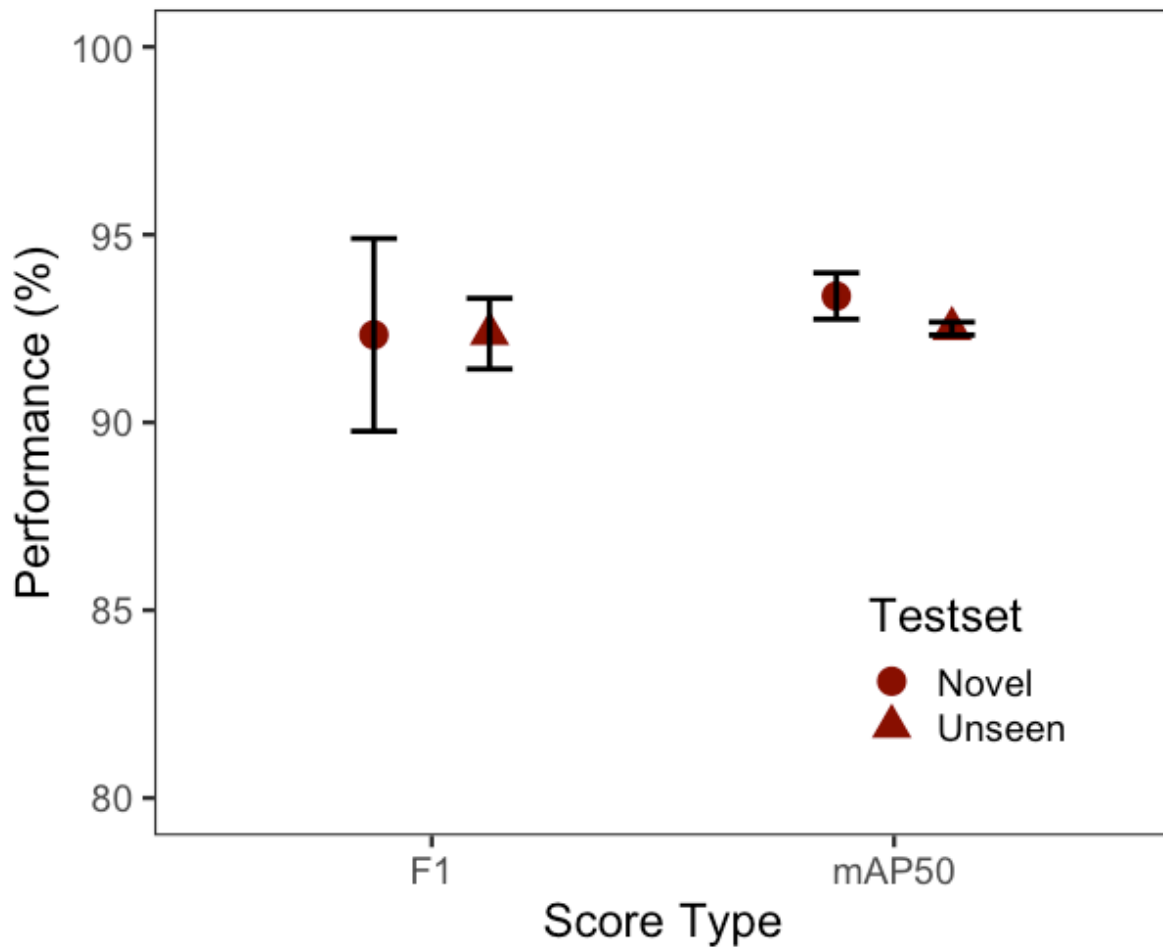
264 Based on F1 scores, the computer performed equally well (t-test;  $t = -0.01$ ,  $p = 0.99$ ) on the

265 Unseen (92.4%) and novel (92.3%; Fig 3). Similarly, the difference in performance for

266 mAP50 was non-significant ( $t = 1.4$ ,  $p = 0.29$ ) on the Unseen (92.5%) and Novel (93.4%)

267 test-sets.

268



269

270

271 Fig. 3. The performance of the three model's F1 and mAP50 scores (mean, SE) for the

272 unseen test footage from the same location and novel footage (Unseen; 32 videos, Novel; 32

273 videos).

274

275 *Human vs Machine*

276 The computer algorithm achieved the highest mean F1 score in both the image (95.4%) and  
 277 the video-based tests (86.8%), when compared with the experts and citizen scientists. The  
 278 computer also had fewer false positives (incorrectly identifying another species as luderick)  
 279 and false negatives (incorrectly ignoring a luderick) in the image test. The computer models  
 280 also had the lowest rate of false positives in the video-based test when compared to both  
 281 human groups, but had the highest rate of false negatives. The computer performed the task  
 282 far faster than both human groups. Experts on average performed better (F1) than the citizen  
 283 scientists in both tests, and had higher accuracy scores (Table 1).

284

285 Table 1. Summary of performance measures comparing averaged scores from computer vs  
 286 humans (citizen scientists and experts). Accuracy is displayed as the observer answer divided  
 287 by the ground truth. Speed is measured as seconds per image, and minutes per minute of  
 288 video. Images N = 50, Videos N = 31.

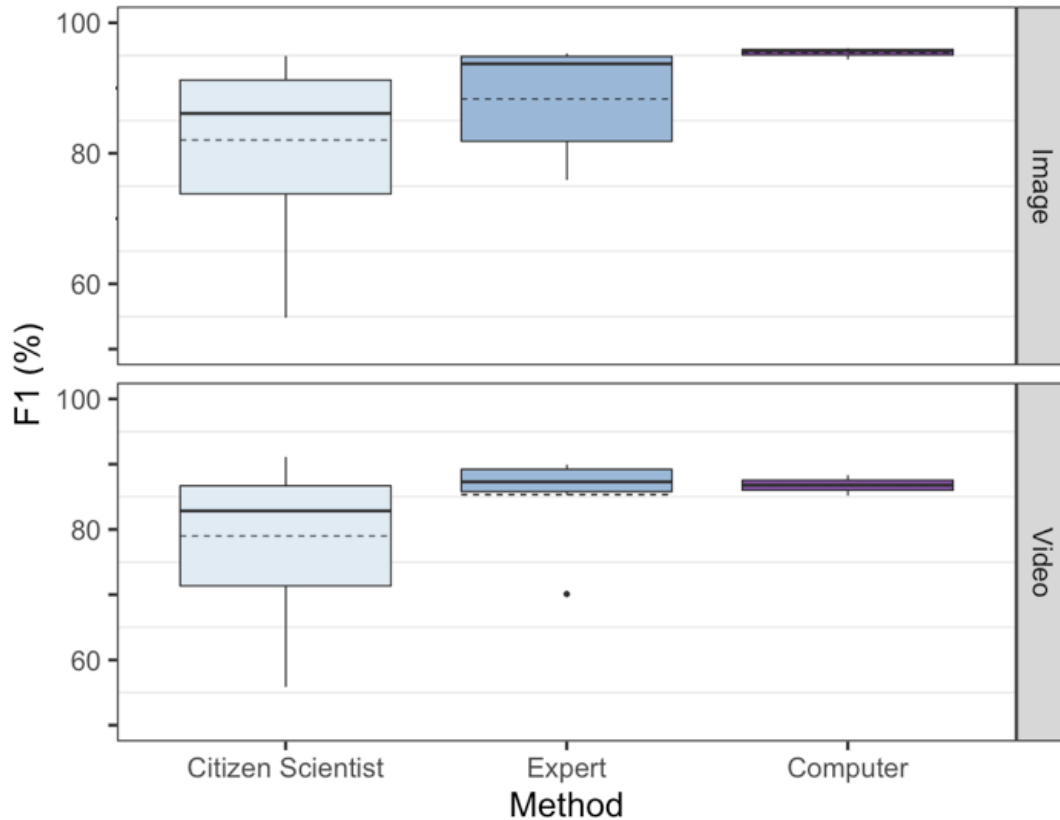
289

<b>Analysis Method</b>	<b>False Negatives</b>	<b>False Positives</b>	<b>Accuracy (prop. +/-)</b>	<b>F1 (%) (SE)</b>	<b>Speed (mins) (SE)</b>
<b>Images</b>					
Citizen Scientist	28.6	7.2	-0.14	82.0 (2.8)	12.6 (1.4)
Expert	18.1	5.6	-0.08	88.3 (8.4)	14.3 (4.0)
Computer	11.7	4.7	-0.12	95.4 (0.9)	0.4 (0.0)
<b>Videos</b>					
Citizen Scientist	20.9	12.6	-0.10	79.0 (2.4)	2.4 (2.4)
Expert	12.1	11.9	+0.06	85.3 (6.9)	2.8 (4.4)
Computer	24.3	2.7	-0.10	86.8 (1.6)	1.2 (0.3)

290

291 F1 scores were most variable for the citizen scientist group, with the difference between the  
292 lowest and the highest score for the image and video tests being 40.1% and 35.1%,  
293 respectively. The computer achieved the lowest variance, with these values only 3.1% for the  
294 video test and 1.7% for the image test (Fig. 4).

295



296

297

298 Fig. 4. Overall test performance in determining abundance (F1) by Computer vs Humans  
299 (Citizen Scientists and Experts) based on identical tests using 50 images and 31 videos.

300 Variance was highest and performance lowest in the citizen scientist group while the  
301 computer had the lowest variance and highest performance. Solid line denotes median,  
302 dashed line the mean.

303

304 *5. Discussion*

305 Our object detection models achieved high performance on a previously unseen data set, and  
306 maintained this performance on footage collected in a novel location. It outperformed both  
307 classes of humans (citizen scientists and experts) in speed and performance, with high  
308 consistency (i.e. low variability).

309

310 We clearly show that our model is fully capable of accurately performing the same on novel  
311 footage from locations beyond the data used for training. Few previous demonstrations of the  
312 utility of deep learning have tested algorithms under these novel conditions, but is one which  
313 consider important for determining how transferable the model is to practising environmental  
314 scientists. For our example, our intention was to test how robust and flexible the algorithm  
315 was in identifying luderick under different environmental conditions which can vary with  
316 tides, water clarity, ambient light, differences in non-target fish species and backgrounds. In a  
317 study conducted by Xia et al. (2018) on sea cucumbers, a novel test data set comprised of  
318 internet images demonstrated an accuracy of 76.3%. This performance was significantly  
319 lower than the test data set the model was trained on which achieved an accuracy of 97.6%.  
320 Similarly, Xu and Matzner (2018) attempted to monitor the effects of water turbines on local  
321 fish species at three different sites, but their model only generated a 53.9% accuracy. All  
322 three sites exhibited their own unique challenges to underwater data collection, including  
323 occlusion due to bubbles from fast-flowing water and debris, that made fish detection  
324 difficult even for a human observer. Their study demonstrates the aforementioned  
325 environmental challenges marine scientist face in using computer vision. Despite the  
326 performance limitations of deep learning when provided with limited training data, one  
327 reason that our models produced high-performance results from the novel location is the  
328 broad variation in environmental conditions and camera angles in the training data. Future  
329 work on this topic could extend the novel test to include an even wider array of novel  
330 locations to further assess the robustness of the model.

331

332 The computer's high performance, speed and low variance compared to humans suggests that  
333 it is a suitable model to replace manual efforts to determine MaxN in marine environments.  
334 Deep learning may be the solution for researchers to avoid analytical bottlenecks (Gray et al.  
335 2019a) as the computer performed the image-based test considerably faster on average than  
336 humans. The image test results are consistent with other deep learning related models

337 comparing human and computer performance. Villon et al. (2018) trained a classification  
338 model which outperformed humans by approximately 5% in classifying still images of nine  
339 coral reef fish species. Similar results were found by Torney et al. (2019) using object  
340 detection to accurately survey wildebeest abundance in Tanzania at a rate of approximately  
341 500 images per hour. Torney et al. (2019) calculated that computer analysis could reduce  
342 analysis of surveys from around three to six weeks done manually by up to four wildlife  
343 experts, down to just 24 hours using a deep learning algorithm. Additionally, they found  
344 accuracy was not compromised, with the abundance estimate from deep learning within 1%  
345 of that from expert manual analysis. Like humans, the computer is reliant on the quality of  
346 the image it receives. Deep learning methods tend to decrease in performance when the  
347 picture quality is blurred or subject to excessive noise (Salman et al. 2016). In low light or  
348 high turbidity situations, image processing to improve the quality of the picture (such as  
349 cancelling noise and improving contrast) can improve the performance of the model (Salman  
350 et al. 2016).

351

352 Previous studies comparing humans versus computers have predominantly used images  
353 rather than videos. When analysing video footage, there is an assumption that humans have  
354 the comparative advantage when addressing uncertainty and ambiguity (Jarrahi 2018). Fish  
355 that could not be positively identified early in the video may be identifiable later and vice  
356 versa. Humans can move back and forward within the video to correctly identify each fish  
357 when calculating MaxN, an ability our deep learning model lacks. The results show that even  
358 when humans seem to have the spatio-temporal advantage, the computer model still  
359 outperforms both the experts and citizen scientists. In our set-up, inference time for video  
360 footage by the computer was about half that of humans. Analytical time could be further  
361 reduced by using multiple GPUs or by implementing parallel processing using multiple  
362 virtual machines. Consistency in estimating populations is important in ecology, as  
363 quantifying population trends is critical to understanding ecosystem health. The computers  
364 low variation indicates that it may prove an advantage for monitoring, when data relies on  
365 consistency to determine fluctuations in species abundance. Errors that can occur when using  
366 humans in data analysis can include individual observer bias and even bias estimation of  
367 trends (Yoccoz et al. 2001). This variance is inter-personal and could be standardised by  
368 having a single observer across all data sets. This is unrealistic, however, given the large  
369 volumes of data often generated by video monitoring (Weinstein 2017). Deep learning

370 methods standardise observer affects not only within data-sets, but also between data-sets  
371 from different periods, without personal bias.

372

373 The performance curves for our models suggest that they may be just as useful in determining  
374 fish abundance with fewer annotations than our full training set of 6,080 annotations.

375 Therefore, less time was needed for training the algorithm as the accuracy of the model's  
376 ability to predict the whole fish (mAP50) is not needed to determine abundance. As our  
377 model took approximately 60 hours to train, running a performance curve while training we  
378 can see that the time to reach optimum performance could be two-thirds quicker at 20 hours.

379 Creating a performance curve is a useful step when calculating the cost-benefits of  
380 implementing a high performing model as well as monitoring algorithm issues such as  
381 overfitting. However, this does not take into account the time for human to be trained on  
382 which species to annotate. Fish identification experts may not need additional training while  
383 citizen scientists may. However studies have shown that citizen scientist annotated data for  
384 deep learning can be as reliable as expertly annotated data (Snow et al. 2008) providing an  
385 additional low-cost solution for model training.

386

387 Although recent advances in deep learning can make image analysis for animal ecology more  
388 efficient, there are still some ecological and environmental limitations. Ecological limitations  
389 include the difficulty in detection of small, rare or elusive species and therefore abundance  
390 may not be able to be estimated in-situ. Nevertheless, even plankton classification using deep  
391 learning has been attempted (Li & Cui 2016, Py et al. 2016). This approach may be used to  
392 calculate the relative abundance of these microscopic organisms and therefore estimate a wild  
393 population density. This may be particularly useful in predicting and monitoring outbreaks of  
394 nuisance species such as crown-of-thorns sea stars (Hock et al. 2014) or stinging sea jellies  
395 (Llewellyn et al. 2016). Another key ecological issue when using computer vision is low  
396 sampling resolution due to the limited field of view from cameras, limiting the accuracy of  
397 determining abundance. Campbell et al. (2018) discovered that using cameras with a 360-  
398 degree field-of-view improved the accuracy of fish counts compared with single-camera  
399 MaxN counts. Improvements for future studies could include combining deep learning with a  
400 360-degree camera aspect when assessing abundance. The current limitations in computer  
401 vision imply that this technology is not suitable for all facets of animal ecology.

402 Environmental conditions such as water clarity and light availability currently dictate the



403 useability of footage in marine environments which subsequently affects the performance of  
404 the model (Salman et al. 2019). However, these limitations are also experienced by human  
405 observers in manual data analysis.

406

407 Deep learning methodologies provide a useful tool for consistent monitoring and estimations  
408 of abundance in marine environments, surpassing the overall performance of manual, human  
409 efforts in a fraction of the time. As this field advances, future ecological applications can  
410 include automation in estimating fish size (Costa et al. 2006), estimating abundance for  
411 multiple species simultaneously (Mandal et al. 2018), studying animal behaviour (Valletta et  
412 al. 2017, Norouzzadeh et al. 2018), and monitoring pest species populations (Clement et al.  
413 2005). Future technological advances in the application of the “internet of things” may also  
414 provide ecologists with fully automated management systems via remote sensors connected  
415 to machine learning algorithms to achieve continuous environmental information at high  
416 temporal resolution (Allan et al. 2018). Given the significant advantages that these algorithms  
417 can provide, deep learning can indeed be a highly successful and complementary tool for  
418 marine animal ecology.

419

420 Acknowledgements

421 We thank the many fish experts and citizen scientists who participated in the study. This  
422 work benefitted from the support of the Global Wetlands Project.

423

424 Funding

425 CJB was supported by a Discovery Early Career Researcher Award (DE160101207) from the  
426 Australian Research Council. CJB and RMC were supported by a Discovery Project from the  
427 Australian Research Council (DP180103124).

428

429 References

- 430 Alexandropoulos S-AN, Aridas CK, Kotsiantis SB, Vrahatis MN (2019) Multi-Objective  
431 Evolutionary Optimization Algorithms for Machine Learning: A Recent Survey. In:  
432 Demetriou IC, Pardalos PM (eds) Approximation and Optimization. Springer, Cham  
433 Allan BM, Nimmo DG, Ierodiaconou D, VanDerWal J, Koh LP, Ritchie EG (2018)  
434 Futurecasting ecological research: the rise of technoecology. *Ecosphere* 9:e02163  
435 Arellano-Verdejo J, Lazcano-Hernandez HE, Cabanillas-Terán N (2019) ERISNet: deep  
436 neural network for Sargassum detection along the coastline of the Mexican  
437 Caribbean. *PeerJ* 7:e6842
- 438 Baker R, Sheaves M (2006) Visual surveys reveal high densities of large piscivores in  
439 shallow estuarine nurseries. *Mar Ecol Prog Ser* 323:75-82
- 440 Campbell MD, Salisbury J, Caillouet R, Driggers WB, Kilfoil J (2018) Camera field-of-view  
441 and fish abundance estimation: A comparison of individual-based model output and  
442 empirical data. *J Exp Mar Biol Ecol* 501:46-53
- 443 Chicco D (2017) Ten quick tips for machine learning in computational biology. *BioData*  
444 mining 10:35
- 445 Clement R, Dunbabin M, Wyeth G (2005) Toward robust image detection of crown-of-thorns  
446 starfish for autonomous population monitoring. *Proc Australasian Conference on*  
447 *Robotics and Automation 2005*. Australian Robotics and Automation Association Inc
- 448 Costa C, Loy A, Cataudella S, Davis D, Scardi M (2006) Extracting fish size using dual  
449 underwater cameras. *Aquacult Eng* 35:218-227
- 450 Dai Z, Carver E, Liu C, Lee J, Feldman A, Zong W, Pantelic M, Elshaikh M, Wen N (2019)  
451 Segmentation of the Prostatic Gland and the Intraprostatic Lesions on Multiparametric  
452 MRI Using Mask-RCNN. *arXiv preprint arXiv:190402575*
- 453 dos Santos AA, Gonçalves WN (2019) Improving Pantanal fish species recognition through  
454 taxonomic ranks in convolutional neural networks. *Ecol Inform*:100977
- 455 Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual  
456 object classes (voc) challenge. *Int J Comput Vis* 88:303-338
- 457 Ferguson AM, Harvey ES, Taylor MD, Knott NA (2013) A herbivore knows its patch:  
458 luderick, *Girella tricuspidata*, exhibit strong site fidelity on shallow subtidal reefs in a  
459 temperate marine park. *PLoS One* 8:e65838

- 460 Galloway A, Taylor GW, Ramsay A, Moussa M (2017) The Ciona17 Dataset for Semantic  
461 Segmentation of Invasive Species in a Marine Aquaculture Environment. 14th  
462 Conference on Computer and Robot Vision (CRV). IEEE
- 463 Gilby BL, Olds AD, Connolly RM, Yabsley NA, Maxwell PS, Tibbetts IR, Schoeman DS,  
464 Schlacher TA (2017) Umbrellas can work under water: Using threatened species as  
465 indicator and management surrogates can improve coastal conservation. *Estuar Coast*  
466 *Shelf Sci* 199:132-140
- 467 Gollan JR, Wright JT (2006) Limited grazing pressure by native herbivores on the invasive  
468 seaweed *Caulerpa taxifolia* in a temperate Australian estuary. *Mar Freshwat Res*  
469 57:685-694
- 470 Gray PC, Bierlich KC, Mantell SA, Friedlaender AS, Goldbogen JA, Johnston DW (2019a)  
471 Drones and convolutional neural networks facilitate automated and accurate cetacean  
472 species identification and photogrammetry. *Methods Ecol Evol* 10:1490-1500
- 473 Gray PC, Fleishman AB, Klein DJ, McKown MW, Bézy VS, Lohmann KJ, Johnston DW  
474 (2019b) A convolutional neural network for detecting sea turtles in drone imagery.  
475 *Methods Ecol Evol* 10:345-355
- 476 Halpern BS, Frazier M, Potapenko J, Casey KS, Koenig K, Longo C, Lowndes JS,  
477 Rockwood RC, Selig ER, Selkoe KA (2015) Spatial and temporal changes in  
478 cumulative human impacts on the world's ocean. *Nat Commun* 6:7615
- 479 He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. *Proceedings of the IEEE*  
480 *international conference on computer vision*:2961-2969
- 481 Hock K, Wolff NH, Condie SA, Anthony KR, Mumby PJ (2014) Connectivity networks  
482 reveal the risks of crown-of-thorns starfish outbreaks on the Great Barrier Reef. *J*  
483 *Appl Ecol* 51:1188-1196
- 484 Hodgson A, Kelly N, Peel D (2013) Unmanned aerial vehicles (UAVs) for surveying marine  
485 fauna: a dugong case study. *PloS one* 8:e79556
- 486 Jarrahi MH (2018) Artificial intelligence and the future of work: human-AI symbiosis in  
487 organizational decision making. *Bus Horiz* 61:577-586
- 488 LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436
- 489 Leek JT, Peng RD (2015) Opinion: Reproducible research can still be wrong: Adopting a  
490 prevention approach. *Proc Natl Acad Sci USA* 112:1645-1646
- 491 Lehtiniemi M, Engström-Öst J, Viitasalo M (2005) Turbidity decreases anti-predator  
492 behaviour in pike larvae, *Esox lucius*. *Environ Biol Fishes* 73:1-8

- 493 Li X, Cui Z Deep residual networks for plankton classification. Proc OCEANS 2016  
494 MTS/IEEE Monterey. IEEE
- 495 Llewellyn L, Bainbridge S, Page G, O'Callaghan M, Kingsford M (2016) StingerCam: A tool  
496 for ecologists and stakeholders to detect the presence of venomous tropical jellyfish.  
497 Limnol Oceanogr Methods 14:649-657
- 498 Lowe M, Morrison M, Taylor R (2015) Harmful effects of sediment-induced turbidity on  
499 juvenile fish in estuaries. Mar Ecol Prog Ser 539:241-254
- 500 Maire F, Alvarez LM, Hodgson A (2015) Automating marine mammal detection in aerial  
501 images captured during wildlife surveys: a deep learning approach. Australasian Joint  
502 Conference on Artificial Intelligence:379-385
- 503 Malde K, Handegard NO, Eikvil L, Salberg A-B (2019) Machine intelligence and the data-  
504 driven future of marine science. ICES J Mar Sci
- 505 Mandal R, Connolly RM, Schlacher TA, Stantic B (2018) Assessing fish abundance from  
506 underwater video using deep neural networks. 2018 International Joint Conference on  
507 Neural Networks (IJCNN):1-6
- 508 Massa F, Girshick R (2018) maskrcnn-benchmark: Fast, modular reference implementation  
509 of Instance Segmentation and Object Detection algorithms in PyTorch. Accessed  
510 03/06. <https://github.com/facebookresearch/maskrcnn-benchmark>
- 511 Moniruzzaman M, Islam SMS, Bennamoun M, Lavery P (2017) Deep learning on underwater  
512 marine object detection: a survey. International Conference on Advanced Concepts  
513 for Intelligent Vision Systems:150-160
- 514 Norouzzadeh MS, Nguyen A, Kosmala M, Swanson A, Palmer MS, Packer C, Clune J (2018)  
515 Automatically identifying, counting, and describing wild animals in camera-trap  
516 images with deep learning. Proc Natl Acad Sci 115:E5716-E5725
- 517 Orth RJ, Carruthers TJ, Dennison WC, Duarte CM, Fourqurean JW, Heck KL, Hughes AR,  
518 Kendrick GA, Kenworthy WJ, Olyarnik S (2006) A global crisis for seagrass  
519 ecosystems. Bioscience 56:987-996
- 520 Piechaud N, Hunt C, Culverhouse PF, Foster NL, Howell KL (2019) Automated  
521 identification of benthic epifauna with computer vision. Mar Ecol Prog Ser 615:15-30
- 522 Prechelt L (1998) Early stopping-but when? In: Müller K-R, Orr G (eds) Neural Networks:  
523 Tricks of the trade. Springer, Berlin

- 524 Py O, Hong H, Zhongzhi S (2016) Plankton classification with deep convolutional neural  
525 networks. Proc 2016 IEEE Information Technology, Networking, Electronic and  
526 Automation Control Conference. IEEE
- 527 Rawat W, Wang Z (2017) Deep convolutional neural networks for image classification: A  
528 comprehensive review. *Neural Comput* 29:2352-2449
- 529 Salberg A-B (2015) Detection of seals in remote sensing images using features extracted  
530 from deep convolutional neural networks. 2015 IEEE International Geoscience and  
531 Remote Sensing Symposium (IGARSS):1893-1896
- 532 Salman A, Jalal A, Shafait F, Mian A, Shortis M, Seager J, Harvey E (2016) Fish species  
533 classification in unconstrained underwater environments based on deep learning.  
534 *Limnol Oceanogr Methods* 14:570-585
- 535 Salman A, Siddiqui SA, Shafait F, Mian A, Shortis MR, Khurshid K, Ulges A, Schwanecke  
536 U (2019) Automatic fish detection in underwater videos by a deep neural network-  
537 based hybrid motion learning system. *ICES J Mar Sci*
- 538 Sievers M, Brown CJ, Tulloch VJ, Pearson RM, Haig JA, Turschwell MP, Connolly RM  
539 (2019) The role of vegetated coastal wetlands for marine megafauna conservation.  
540 *Trends Ecol Evol*
- 541 Snow R, O'Connor B, Jurafsky D, Ng AY (2008) Cheap and fast---but is it good?: evaluating  
542 non-expert annotations for natural language tasks. Proc Proceedings of the conference  
543 on empirical methods in natural language processing. Association for Computational  
544 Linguistics
- 545 Torney CJ, Lloyd-Jones DJ, Chevallier M, Moyer DC, Maliti HT, Mwita M, Kohi EM,  
546 Hopcraft GC (2019) A comparison of deep learning and citizen science techniques for  
547 counting wildlife in aerial survey images. *Methods Ecol Evol*
- 548 Valletta JJ, Torney C, Kings M, Thornton A, Madden J (2017) Applications of machine  
549 learning in animal behaviour studies. *Anim Behav* 124:203-220
- 550 Villon S, Mouillot D, Chaumont M, Darling ES, Subsol G, Claverie T, Villéger S (2018) A  
551 Deep learning method for accurate and fast identification of coral reef fishes in  
552 underwater images. *Ecol Inform* 48:238-244
- 553 Waycott M, Duarte CM, Carruthers TJ, Orth RJ, Dennison WC, Olyarnik S, Calladine A,  
554 Fourqurean JW, Heck KL, Hughes AR (2009) Accelerating loss of seagrasses across  
555 the globe threatens coastal ecosystems. *Proc Natl Acad Sci* 106:12377-12381
- 556 Weinstein BG (2017) A computer vision for animal ecology. *J Anim Ecol* 87:533-545

- 557 Xia C, Fu L, Liu H, Chen L (2018) In Situ Sea Cucumber Detection Based on Deep Learning  
558 Approach. 2018 OCEANS-MTS/IEEE Kobe Techno-Oceans (OTO):1-4
- 559 Xu L, Bennamoun M, An S, Sohel F, Boussaid F (2019) Deep learning for marine species  
560 recognition. In: Balas V, Roy S, Sharma D, Samui P (eds) Advances in  
561 Computational Intelligence. Springer
- 562 Xu W, Matzner S (2018) Underwater Fish Detection using Deep Learning for Water Power  
563 Applications. arXiv preprint arXiv:181101494
- 564 Yoccoz NG, Nichols JD, Boulinier T (2001) Monitoring of biological diversity in space and  
565 time. Trends Ecol Evol 16:446-453
- 566