Crowdsourcing experiment and fully convolutional neural networks for coastal remote sensing of seagrass and macro-algae

Brandon Hobley, Michal Mackiewicz, Julie Bremner, Tony Dolphin and Riccardo Arosio . (e-mail: b.hobley@uea.ac.uk).

***Abstract*—Recently, convolutional neural networks (CNNs) and fully convolutional neural networks (FCNs) have been successfully used for monitoring coastal marine ecosystems, in particular vegetation. However, even with recent advances in computational modelling and data acquisition, deep learning models require substantial amounts of good quality reference data to effectively self-learn internal representations of input imagery. The classical approach for coastal mapping requires experts to transcribe in-situ records and delineate polygons from high-resolution imagery such that FCNs can self-learn. However, labelling by a single individual limits the training data, whereas crowdsourcing labels can increase the volume of training data, but may compromise label quality and consistency. In this paper we assessed the reliability of crowdsourced labels on a complex multi-class problem domain for estuarine vegetation and unvegetated sediment. An inter-observer variability experiment was conducted in order to assess the statistical differences in crowdsourced annotations for plant species and sediment. The participants were grouped based on their discipline and level of expertise, and the statistical differences were evaluated using the Cochran's Q-test and the annotation accuracy of each group to determine for observation biases. Given the crowdsourced labels, FCNs were trained with majority-vote annotations from each group to check whether observation biases were propagated to FCN performance. Two scenarios were examined: first, a direct comparison of FCNs trained with transcribed in-situ labels and crowdsourced labels from each group was established. Then, transcribed in-situ labels were supplemented with crowdsourced labels to investigate the feasibility of training FCNs with crowdsourced labels in coastal mapping applications.**

We show that annotations sourced from discipline experts (ecologists and geomorphologists) familiar with the study site were more accurate than experts with no prior knowledge of the site and non-experts, with our results confirming that biases in participant annotation were propagated in FCN performance. Furthermore, FCNs trained with a combined dataset of in-situ and crowdsourced labels performed better than FCNs trained on the same imagery with in-situ labels.

***Index Terms*—Remote Sensing, Deep Learning, Convolutional Neural Network, Multispectral, Crowdsourcing**

I. Introduction

This work was supported by Cefas, EA and the Natural Environmental Research Council through Industrial CASE, grant number NE/R007888/1

B. Hobley and M.Mackiewicz are with School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, England, U.K.

Tony Dolphin and Julie Bremner are with Centre for Environment, Fisheries and Aquaculture Science, Lowestoft, NR33 0HT, England, U.K. and with School of Environmental Science, University of East Anglia, Norwich, NR4 7TJ, England, U.K.

Riccardo Arosio is with School of Biological, Earth and Environmental Sciences, University College Cork, Cork, Ireland

COASTAL ecosystems such as wetlands, estuaries and coral reefs represent dynamic and important nurturing habitats for a wide variety of plants, fish, shellfish and other wildlife [1]. With growing concerns over climate change, these coastal areas will be subject to changing atmospheric and ocean temperatures, sea levels, ocean chemistry, weather patterns and the increased demands of a growing global population. This emphasises the need to create and act on strategies that maintain a sustainable balance of coastal ecosystem health, while also effectively managing the use of resources that are derived from these ecosystems [2], [3].

In coastal monitoring, remote sensing has provided a major platform for ecologists to assess and monitor sites in many applications [4], [5]. Satellite imagery can provide global to regional observations at regular sampling intervals with successful applications for coastal management [2]. However, this avenue of data acquisition often struggles with cloud contamination, oblique views, costs for data acquisition and coarse resolution relative to the often narrow features of interest that stretch along the coast [6]. The shift to uncrewed aircraft systems (UASs) and commercially available cameras tackles the latter issues as it resolves coarse satellite resolution (typically 2 - 30m) by collecting several overlapping very high resolution (VHR) images and stitching sensor outputs together using Structure from Motion (SfM) techniques to create high-resolution orthomosaics [7], [8] (commonly less than 0.1 m).

Parallel to the advancements in data acquisition, computer vision (CV) has also improved in the last decade with deep learning (DL) [9] and the introduction of convolutional neural networks (CNNs) [10]. These methods have surpassed previous state-of-the-art results in a wide variety of computer vision applications [10]–[12]. Traditionally, supervised machine learning (ML) methods can be defined by two separate components: feature extraction and model training. Instead, CNNs learn hierarchical abstract representations of input imagery in a self-learning fashion, which in effect combines feature learning and supervised classifier training in one optimisation [9]. Fully convolutional neural networks (FCNs) are an adaptation of CNNs that perform per-pixel classifications, and enable contextual features to be extracted within a wide receptive field while also preserving the spatial origin of these features to produce a fine-grained and spatially explicit segmentation of the object [11], [13]. This is appropriate for remote sensing mapping applications where aerial imagery can be segmented to meaningful sets of classes in order to delineate objects or

species of interest [14]–[19].

This said, even with the advent of UASs and FCNs to map coastal environments, the quantity and quality of data labels is a pivotal concern in many real-world scenarios because DL models perform best with large, labelled, training datasets [9], [20]. In remote sensing, reference observations (FCN training data) are often acquired in-situ, which involves high logistic efforts, potential inaccuracies due to geo-location errors as well as sampling and observation bias [21], [22]. Moreover, the volume of data generated with UAS imagery may cover a substantial spatially-continuous area with respect to the real-world, yet the ratio between the area covered via in-situ surveying and the total area covered in imagery is often relatively small [16], [23]. Methods such as transfer learning [24], data-augmentation [25] and semi-supervision [26], [27] can provide tools for FCNs to self-learn if there are limited amounts of labelled data, as is often the case for environmental monitoring. However, an alternative for efficient in-situ data collection is visual identification and delineation of training data directly from orthomosaics [28]–[30] - possible in UAS imagery because the resolution is sufficiently high that even features as small as 10×10 cm can often be accurately identified and labelled. Further to this, crowdsourced labels can provide an even more cost-effective alternative to laborious labelling procedures from aerial imagery involving individual domain specific experts with studies showing that aggregated labels can provide better quality generalisation in machine learning modelling which draw parallels with field of expert frameworks and ensemble learning [31], [32].

Remote sensing applications have also leveraged the use of crowdsourced labels to supplement aerial imagery datasets in a variety of manners [33]. Commonly, web-based applications prompt participants to classify binary tasks with known GPS information for accurate geo-location. This has led to successful workflows that combine deep learning and crowdsourcing for several study sites: Guatemala, Laos and Malawi using MapSwipe [34]; the Missing Maps humanitarian project using OpenStreetMap [35]; settlements in Nigeria, Somalia, Pakistan and Afghanistan using Tomnod platform [36]; and for crop mapping in South East India using Plantix [37]. Furthermore, coastal surveying has also leveraged crowdsourced annotations for deep learning applications of litter mapping in the shores of Xabelia beach in Lesvos, Greece [38] and shoreline change mapping in two open-coast sandy beaches located within the Sydney metropolitan area [39].

These studies focus on combining crowdsourced labels with deep learning models on binary problem domains to avoid ambiguity for participants and erroneous labelling [33]. In contrast, coastal mapping requires the identification of multiple feature classes, some of which are superficially similar depending on the situation (e.g. sand and mud, seagrass and filamentous algae).

In this paper, we tackled the problem of deriving crowdsourced training data for estuarine vegetation and unvegetated sediment ecosystems at Budle Bay (Northumberland, UK). We performed an inter-observer experiment of crowdsourced annotations on a complex multi-class problem domain that includes intertidal coastal species, such as seagrass, saltmarsh and macro-algae. The experiment population consisted of 12 participants split into 3 groups based on their discipline and level of expertise in habitat mapping. The experiment was analogous to crowdsourcing labelled data in remote sensing applications as participants were prompted to classify pre-determined points. Our experimental setup comprised two sets of points: a set whereby the true semantic value of each human annotation was known according to an in-situ survey of the study site conducted by the UK Centre for Environment, Fisheries and Aquaculture Science (Cefas) and UK Environment Agency (EA), and an extra set of points created through expert photo-interpretation to balance class distribution (see Section II-D).

The analysis of our inter-observer variability experiment uses the Cochran's Q test to assess the statistical differences of crowdsourced annotations from each group. Furthermore, the annotation accuracy and a per-class analysis of crowdsourced annotations was used to assess for any potential observation biases.

Given the annotations from the inter-observer experiment, the feasibility of FCNs trained with crowdsourced annotations was investigated in two scenarios: first, four FCNs were trained with different versions of labelled data on the same imagery. Three FCNs were trained with labels based on majority-vote annotations from each participant group in the inter-observer experiment and the other FCN was trained with transcribed labels from the in-situ survey. This scenario allows for a direct performance comparison for FCNs trained with in-situ labels and crowdsourced labels, and evaluates whether biases in crowdsourced annotations were propagated in FCN performance. The second scenario investigates the feasibility of supplementing transcribed in-situ labels with crowdsourced labels using two FCNs. For this scenario, one FCN was trained with the set of points described in Section II-D, whereas the other FCN was trained with a combination of transcribed in-situ labels and crowdsourced labels on the same imagery. Consequently, we list the following contributions in the proposed manuscript.

- Discipline experts (ecologists and geomorphologists) familiar with the study site were more accurate than experts with no prior knowledge of the site and non-experts.
- FCNs trained with crowdsourced labels from discipline experts familiar with the site had comparable performance to FCNs trained with in-situ labels.
- FCNs trained with a combined labelled set of in-situ labels and crowdsourced labels were more accurate than FCNs trained with in-situ labels on the same imagery

Sections II-B and II-C detail the study site. Section II-E describes the experimental setup and Section II-F describes the FCN model and parameter training. Lastly, Sections III-A and III-B present the results of the inter-observer experiment and FCN experiments, and Sections IV-A, IV-B and IV-C the analysis and discussion of the inter-observer and FCN experiments.

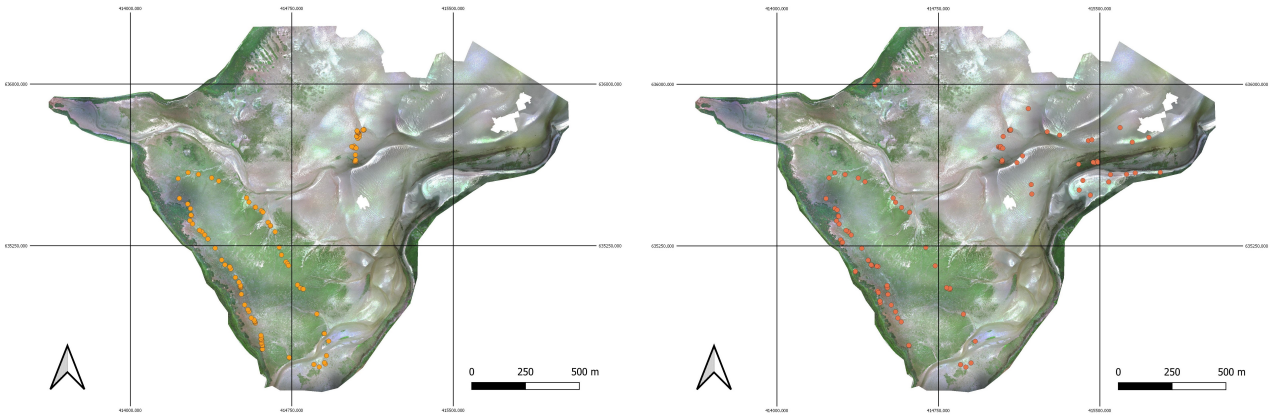

Fig. 1. Distribution of tags recorded during the in-situ survey (left) and the full set of points to be annotated, comprising the in-situ points plus those determined using expert photo-interpretation (right).

II. Methods

A. Study site

The research focused on Budle Bay, Northumberland, UK (55.625°N, 1.745°W). Budle Bay is a large (c. 300 ha) estuarine embayment with a single tidal inlet [40]–[42]. Sinuous and dendritic tidal channels are present within the estuary, and bordering the channels are areas of seagrass and various species of macroalgae. The tidal range varies between 1-4m for the majority of the year and the estuary is fully drained on low spring tides.

B. Image collection

Full details of the data collection can be found in [23]. Figure 1 displays a very high resolution orthomosaic of Budle Bay created from the Cefas and EA RPA survey in September 2017 using Agisoft's MetaShape [43] and SfM. SfM techniques rely on estimating intrinsic and extrinsic camera parameters from overlapping imagery [44]. A combination of appropriate flight planning in terms of altitude and aircraft speed, and the camera's field of view are important factors for producing good quality orthomosaics. For this work, a MicaSense RedEdge3 multispectral camera was used to capture the site. The camera consisted of 5 narrow band filters for red (655–680 nm), green (540–580 nm), blue (459–490 nm), red edge (705–730 nm) and near infra-red (800–880 nm) channels at a ground sampling distance of approximately 8 cm.

The resulting VHR orthomosaic was orthorectified using GPS logs of camera positions and ground control markers spread out across the site. This process ensured that the mosaic was well aligned with respect to the real-world and ecological features present within the coastal site. The orthomosaic had 32,647×26,534 pixels in 5 image bands. For ease of processing, the orthomosaic was split into 24 non-overlapping tiles of 6000×6000 pixel images with each image containing geographic information for further processing.

C. In-situ survey and class domain

The accompanying ground survey identified 13 ecological classes grouped into background sediment, algae, seagrass and saltmarsh.

Classes defining background sediment were rock, gravel, mud and sand. In-situ measurements of unvegetated sediment were predominately in the presence of water and moisture. However, as parts of the orthomosaic included dry sand, an extra sediment class was added through photo-interpretation (16 polygons). Two heuristics for delineating dry sand polygons were defined: first, the spectral reflectance of sand varies with presence of surface moisture and presents higher reflectance intensity for patches of dry sand [45]. Therefore, polygons were delineated by examining bright unvegetated areas in Figure 1. Second, each generated polygon was cross-checked with the topographic Digital Surface Model (DSM) to ensure that patches of dry sand only occur if the surface level was raised.

Algal classes include microphytobenthos, *Enteromorpha* sp. and other macroalgae (inc. *Fucus* sp.). Lastly, the coastal vegetation classes were seagrass and saltmarsh. Thus, a total of seven classes were listed as follows:

- Background sediment: dry sand and other bareground
- Algae: microphytobenthos, *Enteromorpha* and other macroalgae (including *Fucus*)
- Seagrass: *Zostera noltii* and *Zostera angustifolia* merged to a single class
- Other plants: saltmarsh

The in-situ survey recorded 108 geographically referenced tags with the percentage cover of all listed ecological features within a 300 mm radius. The percentage cover was estimated in quadrat sampling fashion [46], [47]. For each in-situ measurement, the class value with maximum percentage cover was chosen as the label.

D. Class distribution

The class distribution of in-situ measurements was not balanced, which may add cognitive bias and consequently skew results in human annotations for the experiment [48]. Recognising biases during crowdsourced data collection efforts is an important step to countering the effect these may impose on model training and is an enabling factor for algorithmic

fairness [49]. Therefore, a set of points from the in-situ survey were combined with extra points added through expert photo-interpretation (by the lead author) in order to balance the class distribution for the experiment setup. From the original set of 108 in-situ points, a balanced set of 53 points was chosen (the remaining 55 in-situ points were used for FCN performance evaluation see Section III-B). Then, added points through photo-interpretation were based on class dependant heuristics.

First, no extra points for dry sand were added as the set of photo-interpreted polygons covered a substantial area to generate enough points for both the experiment and FCN testing. Other bareground was a sediment class that comprised wet sediment features such as wet sand and mud. Selected points presented dark brown or gray color, rugged texture and low elevation values relative to the rest of the site. Generally, added points were sampled within a close vicinity of known in-situ records. But, this was not considered as an important factor for other bareground points as long as color, texture and elevation within a 300mm (6×6 image patch) radius was consistent.

Vegetation classes were split into three sets: algae, seagrass and saltmarsh. The geo-location of extra points for vegetation classes was always within the vicinity of known in-situ points to establish a baseline for comparing colour and texture.

Saltmarsh points were found to be easily identifiable due to slight elevation changes in the DSM but also because coastal saltmarsh occupy the interface between land and sea [50]. Therefore, saltmarsh points were most present on estuary borders. Identifying points for both species of intertidal seagrass was dependant on the following texture and colour features: both species occur in mixed beds of waterlogged depressions between free-draining hummocks dominated by *Zostera noltii* and presented sparse leaves with light yellow green or green colour. [51]–[53].

Microphytobenthos are microscopic organism that inhabit the upper millimetres of illuminated wet sediments, typically appearing only as a subtle greenish shading [54]. Identifying extra points for microphytobenthos was only possible within very close vicinity of known in-situ points, with colour (greenish shading) used as the identifier. Extra points for *Enteromorpha* sp. had to present bright green colour while other macroalgae (inc. *Fucus*), with similar texture to *Enteromorpha* sp., was presented in a dark brownish color [55], [56]. *Enteromorpha* sp. and other macroalgae were spatially continuous compared to seagrass, which were more likely to be sparse. This further aided distinguishing and picking extra points for these classes. While the vegetation species may be found in other circumstances (e.g. saltmarsh hummocks can grow amongst seagrass slightly away from estuary borders), our intent was to maximise our confidence that our selected points were classified correctly rather than to select across the range of possible appearances for each species. Overall, an extra 54 points were added through expert photo-interpretation to maintain the class distribution balance. Therefore, the set of points to be annotated for each participant comprised 119 points whereby 53 points were drawn from the in-situ survey and an extra 54 were created through photo-interpretation and the remaining 12 points were randomly selected from dry sand polygons.

E. Experiment setup

The goal of the experiment was to examine the variability in annotations from multiple participants with differing backgrounds in research and expertise with marine habitat mapping. Each participant was presented with a unique and random order of points to be annotated and a small set of labelled sample images representative of the vegetation classes, to assist with identification. Figures 2 and 3 respectively display the set of labelled sample images presented to each participant and the user interface available to participants during the experiment. Participants used ArcMap 10.6.1 to visualise and annotate samples.

Each participant generated 119 annotations with each cell containing a semantic value corresponding to the class domain in Section II-D. The participant population was split into three groups based on their level of expertise, to explore whether prior knowledge of the study site, research background and/or previous experience with marine annotation could influence experimental results. The criteria separating each group were as follows:

- Group A: expert ecologist or geomorphologist, present at the in-situ survey and/or had previous experience with annotating marine biology for the study site.
- Group B: expert ecologist or geomorphologist, but was not present at the in-situ survey and/or did not have experience with annotating marine biology for the study site.
- Group C: not an expert ecologist or geomorphologist, nor had experience with annotating marine biology from aerial imagery.

Therefore, annotations were grouped into 3 sets based on the stated groupings.

To evaluate the inter-observer variability within each group the Cochran's Q test was used to investigate the statistical significance of differences between K observations on the same n elements with binomial distribution [57], [58]. For this work, K series of observations corresponded to participants within a group and elements for each observation were individual annotations of participants. Therefore, the null hypothesis was that annotations for participants within a group were drawn from one common dichotomous distribution, which would imply low variability in annotations. However, the Cochran's Q test states that each annotation must be dichotomous and represented as 0 or 1. Since the experimental annotation setup was a complex multi-class problem, each annotation was compared with the assigned label (either in-situ or photo-interpreted) and represented as 1 if correct, otherwise the annotation was represented as 0.

The Cochran's Q test statistic with $K-1$ degrees of freedom follows a χ^2 distribution and is given in Equation 1.

$$Q = \frac{K(K-1)\sum_j (C_j - \bar{C})^2}{KS - \sum_i R_i^2} \tag{1}$$

Where, C_j is a column total, R_i is a row total, $\bar{C}$ is the average column total and S is the total score, i.e. $S = \sum_i R_i =$

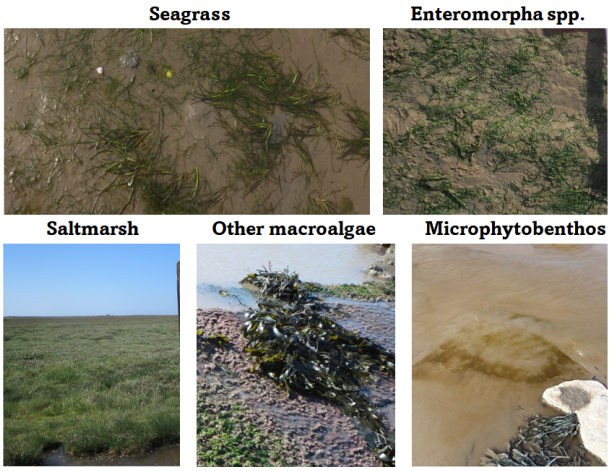

Fig. 2. Sample images representative of vegetation classes used in the analyses.

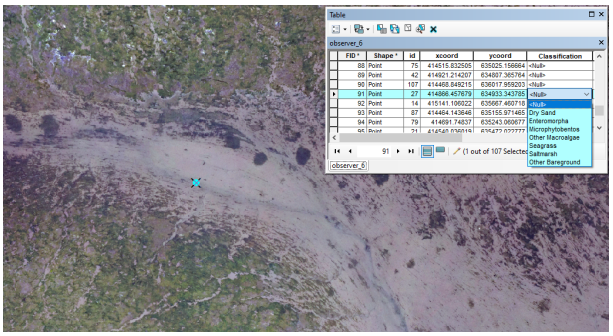

Fig. 3. User interface for providing participant annotations during the experiment.

$\sum_j C_j$. In this context, a column total is the sum of correct annotations for a single participant, and a row total is the sum of correct annotations for a single point across all participants.

F. Fully Convolutional Neural Networks

CNNs have proven to surpass prior-art techniques in a large number of different CV applications since the introduction of AlexNet [10]. The shift from supervised traditional machine learning algorithms, whereby tailored feature extraction methods and classifier tuning are replaced with a joint optimisation of both procedures, is an enabling factor for CNN success. The feature extraction process consists of repeated convolution and pooling operations that transform the input image into hierarchical abstract representations of data. The joint optimisation is achieved by adjusting convolutional kernel weights and biases through the derivative chain-rule that minimises the error between network outputs and annotated labels [9].

FCNs [11], [13] are an adaptation of CNNs for semantic segmentation. The architecture of FCNs can be broken down into three parts: an encoder, a decoder and a classification layer. The encoder network is a CNN without the final fully connected layer, the decoder network applies repeating upsample and convolution operations on feature maps created by the encoder network and the classification layer consists of 1×1 convolution kernels and a softmax transfer function to produce per pixel class probabilities. Figure 4 displays the architecture used for this work. The overall architecture was a U-Net [11] and the encoder network is a VGG-13 [60] pre-trained on ImageNet. However, the weights in the input layer were randomly initialised and changed to handle a 5-channel input image.

1) Data pre-processing and training parameters: FCNs were trained with segmentation maps that contain a one-to-one mapping of pixels encoded with a semantic value, with the goal to optimise this mapping [13]. Segmentation maps were generated using the geographic coordinates stored in each point and converting real-world coordinates to image-coordinates. If a point or multiple points resided within an image tile, then the candidate image was sampled into 256×256 image blocks centered on labelled parts of the image. For each point, a bounding box consistent to 300 mm was placed. Figure 5 shows a gallery of sample imagery used for training FCNs.

The loss was computed by processing a mini-batch of images with the FCN which result in per-pixel probabilities $P \in \mathbb{R}^{B \times K \times H \times W}$ and comparing network outputs with the corresponding annotated maps $Y \in \mathbb{Z}^{B \times H \times W}$; where B, K, H and W are respectively, batch size, number of target classes, height and width of image. Then, the negative log-likelihood loss was calculated between segmentation maps and network probabilities.

$$L_s(P, Y) = \begin{cases} 0, & \text{if } Y(\mathbf{x}) = -1 \\ -\sum_{k=1}^{K} Y_k(\mathbf{x}) \log(P_k(\mathbf{x})), & \\ & \text{if } Y(\mathbf{x}) \neq -1. \end{cases} \quad (2)$$

Where, $\mathbf{x} \in \Omega$; $\Omega \subseteq \mathbb{Z}^2$ is a pixel location and $P_k(\mathbf{x})$ is the probability for the k^{th} channel at pixel location $\mathbf{x}$, with $\sum_{k'=1}^{K} P_{k'}(\mathbf{x}) = 1$. For each image, the loss was the sum of all individual pixel losses using Equation (2) and averaged according to the number of labelled pixels within Y. Previous work on the same study site uses semi-supervision methods to improve the generalisation and performance of FCNs [23]. However, the use of an unsupervised loss term would influence the analysis of our experimental setup by allowing networks to adjust weights based on non-labelled parts of the image, whereas our goal was to determine the effects of aggregated crowdsourced labels.

During training, each image was augmented with stochastic transformations that consist of rotations up to 25° and horizontal or vertical flips. Each network was trained for 200 epochs with a batch-size of 12 with Adam optimiser. The optimiser learning rate was constant and set to 0.001. All FCNNs were implemented and trained using Pytorch version 10.2.

III. Results

A. Inter-observer experiment results

Table I and Figure 6 give the results of our experiment. The significance level for each control group was set to 5% and the degrees of freedom were set according to the number of participants within a particular group. Therefore, the critical values according to a χ^2 distribution were 9.49, 12.59 and 7.81, for control groups A, B and C respectively.

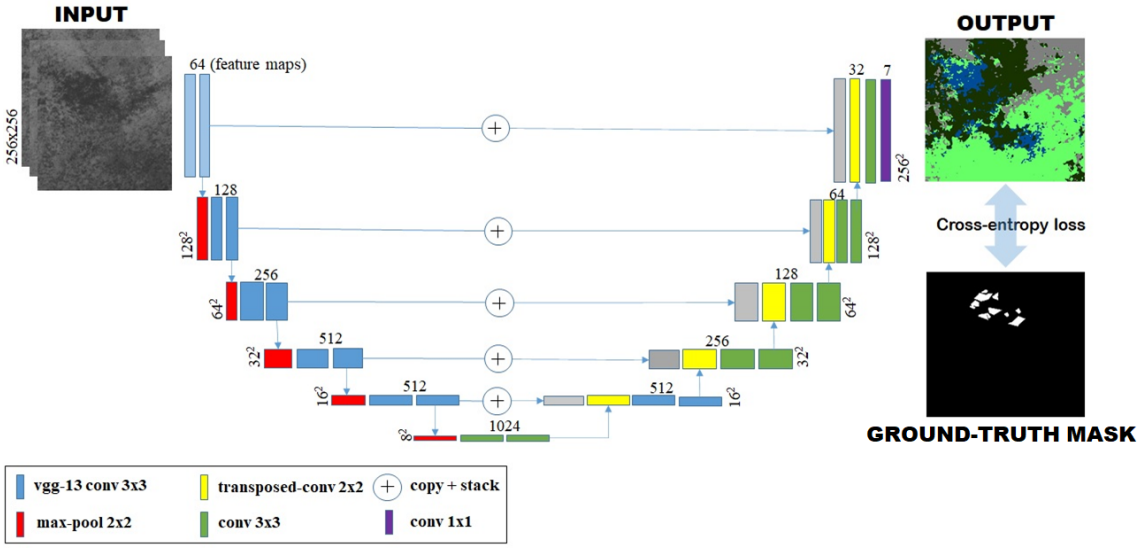

Fig. 4. U-Net architecture and loss calculation. The input channels were stacked and passed through the network. The encoder network applies repeated convolution and max pooling operations to extract feature maps, while the decoder network upsamples these and stacks features from the corresponding layer in the encoder path. The output is a segmented map, which was compared with the ground-truth mask using crossentropy loss. The computed loss was used to train the network, through gradient descent optimisation.

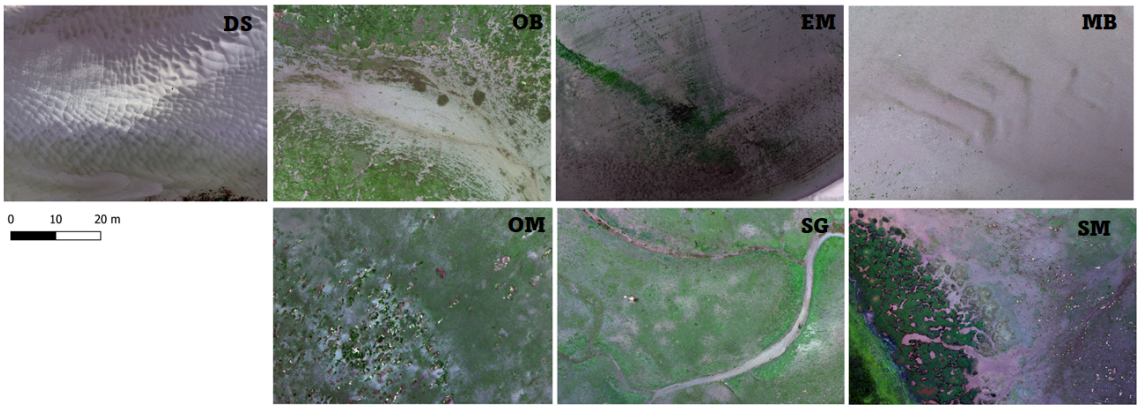

Fig. 5. Gallery of images and corresponding ecological target classes. OM—Other macroalgae inc. *Fucus*; MB—Microphytobenthos; EM—*Enteromorpha*; SM—Saltmarsh; SG—Seagrass; DS—Dry sand; OB—Other bareground.

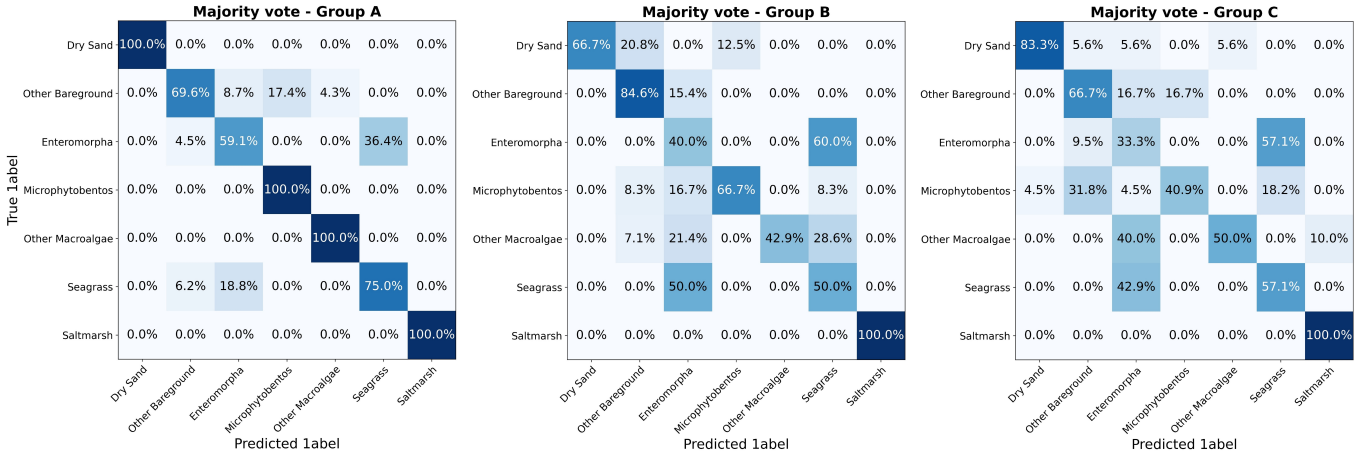

Fig. 6. Confusion matrices for the majority vote annotations for each control group.

Participants	Accuracy (%)	Group	Cochran Q	DoF /α	Outcome
1	70.09%	A	2.0842	4 / 0.05	Not reject
2	76.64%				
3	70.09%				
4	72.90%				
5	59.81%	B	78.8	6 / 0.05	Reject
6	27.10%				
7	63.55%				
8	61.68%				
9	75.70%				
10	63.55%	C	14.39	3 / 0.05	Reject
11	42.06%				
12	55.14%				

TABLE I
PARTICIPANT ANNOTATION ACCURACY AND COCHRAN Q TEST STATISTIC RESULTS. PARTICIPANTS WERE GROUPED BASED ON THEIR LEVEL OF EXPERTISE AND PRIOR KNOWLEDGE OF THE STUDY SITE AND/OR WHETHER PARTICIPANTS WERE FAMILIAR WITH AERIAL IMAGERY ANNOTATION.

	P	R	F1	P	R	F1
DS	0.982	0.956	0.968	0.982	0.997	**0.989**
OB	0.721	0.668	0.693	0.921	0.647	**0.76**
EM	0.433	0.769	0.554	0.517	0.738	**0.608**
MB	0.972	0.814	0.885	1.0	0.921	**0.959**
OM	0.99	1.0	**0.995**	0.982	0.809	0.887
SG	0.579	0.995	**0.73**	0.672	0.711	0.691
SM	0.928	0.944	**0.936**	0.918	0.915	0.917
	In-situ labels			**Majority-vote group A**		

TABLE II
PRECISION, RECALL AND F1 SCORES FOR MODELS TRAINED WITH IN-SITU LABELS AND FOR MODELS TRAINED WITH MAJORITY-VOTE ANNOTATIONS FROM GROUP A. DS—DRY SAND; OB—OTHER BAREGROUND; EM—ENTEROMORPHA; MB—MICROPHYTOBENTOS; OM—OTHER MACROALGAE; SG—SEAGRASS; SM—SALTMARSH

The test statistic described in Equation 1 objectively evaluates the statistical significance of differences between *K* observations on the same *n* elements with a binomial distribution. By comparing each annotation with the known in-situ label and representing correct annotations as 1 and incorrect as 0, the Cochran's Q test evaluates whether annotations, which can be correct or incorrect, were drawn from the same binomial distribution. Therefore, the test statistic for a group may not allow us to reject the null hypothesis, which would imply low inter-observer variability, but participants within that group could collectively annotate test points incorrectly. In fact, participants were more likely to be collectively incorrect than correct due to different incorrect annotations being represented as 0. For example, if the class label for a given point was dry sand but participants annotate the said point as other bareground and microphytobenthos, then both annotations were represented as 0 which would contribute to a smaller test statistic value. Hence, the test statistic was analysed along with the annotation accuracy metrics so that emphasis was placed on groups that were collectively correct and also yielded a test statistic that does not reject the null hypothesis.

B. FCNs results

The metrics to quantify FCN performance were pixel accuracy, precision, recall and F1-score. Pixel accuracy is the ratio between pixels that were classified correctly and the total number of labelled pixels in the test set for a given class. Equation 3 describes each metric, where TP, TN, FP and FN are respectively: True Positive, True Negative, False Positive and False Negative pixel classifications.

Our evaluation consisted of two different tests: the first test shows the effects of training several FCNs on different versions of labelled data based using majority-vote annotations from each group. This test evaluated whether errors in the annotation experiment were propagated to the FCN performance. For training the FCNs, we used the same points as in the inter-observer variability experiment - a set of 53 randomly selected points from the in-situ survey, an additional 54 points chosen through expert photo-interpretation and 12 points from dry sand polygons. The remaining 55 points recorded in-situ were used for model testing and a further 12 points from dry sand polygons. Therefore, FCNs were trained on the combined set of 119 points and the remaining 67 points comprised the test set. For our second test, the combined training set was reduced to the same initial set of 53 randomly selected in-situ points and the remaining 66 labels (54 from photo-interpretation plus 12 points from dry sand polygons) were replaced with majority-vote annotations from each group. The goal of the second experiment was to determine whether supplementing a reduced training set with majority-vote annotations still achieves comparable results to models trained with in-situ labels.

Figure 7 shows the results of our first experiment and Table II provides further insight into class specific performance on FCNs trained with in-situ data versus FCNs trained with majority-vote annotations from group A. Figure 8 shows the results of training FCNs on a reduced dataset of in-situ labels versus FCNs trained on a combined train set of in-situ labels and majority-vote annotations. The confusion matrices and tabled metrics contain the average results of 5 sequential train and test runs.

$$pixel\,accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = 2. \times \frac{recall \times precision}{recall + precision} \tag{6}$$

IV. DISCUSSION

A. nter-observer experiment analysis

From our results, the null hypothesis that participant annotations were drawn from the same distribution was not rejected only in group A. Moreover, group A also exhibited the highest mean and lowest variance in accuracy for annotations with $72.43^{+}_{-}3.106\%$, which showed that participants in group A were more likely to be correct than the other two groups. The pre-exposure of participants in group A to the target classes at the study site justified the lowest test statistic for participant annotations within this particular group. Furthermore, the latter statement can be also supported by examining the majority vote confusion matrix for group A (top-left matrix in in Figure 6), where the accuracy of the majority vote annotations was

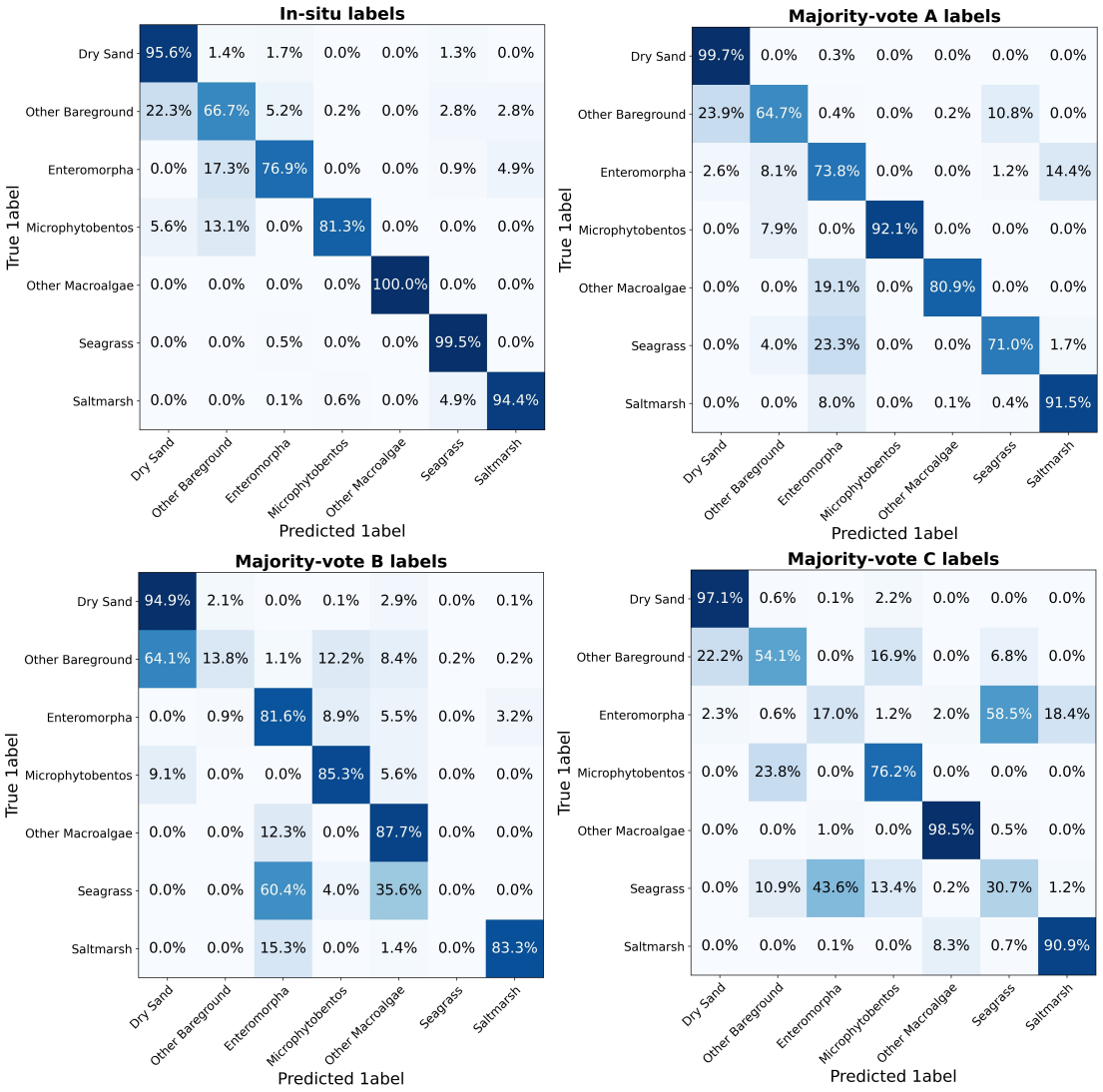

Fig. 7. Confusion matrices for FCNN models trained using different versions of labelled data. Results for models trained on in-situ labels (top-left) and majority-vote annotations for group A (top-right), group B (bottom-left) and group C (bottom-right)

81.31% for group A - higher than the highest accuracy of any participant in the experiment. This illustrates that annotations for participants in group A were better if performed collectively and as a whole group A were good candidates for crowdsourcing labels for this particular study site. Given the low variability in annotations for group A, examining Figure 6 also informed us about the problematic classes to annotate from aerial imagery. Other bareground was a sediment class composed of rock, mud and wet sand, and microphytobenthos typically appearing only as a subtle greenish shading on wet sediment [54], which could justify why both classes were mutually misannotated. The same reasoning can be applied to annotations for *Enteromorpha* sp. and seagrass, since both classes exhibit similar colour and texture from an aerial point of view.

The null hypothesis for participants in group B was rejected by a significant margin. This could be due to: (1) participants in this group were not familiar with annotating aerial imagery for this study site. In IR crowdsourcing, this is also known as the ambiguity effect whereby missing information makes annotations appear more difficult and consequently less attractive [59]. Alternatively, (2) the participant population contained experts from different disciplines who may have conflicting biases during annotation. If participants do not agree with each other, then the test statistic yields a high value based on whether annotations were correct or not. Specifically, the second highest overall annotation accuracy was from participant 9 while the lowest accuracy was from participant 6, both of whom belong to group B. In fact, participant 9 is a benthic ecologist with specific knowledge at identifying intertidal algae, while participant 6 is an expert in sedimentology. This contrast in discipline is reflected in annotation and subsequently in the test statistic due to correct or incorrect annotation on the same test points. The average accuracy was lower than in group A - $57.50^{+}_{-}18.16\%$ and the majority vote confusion matrix paints a similar picture - high variability and feature ambiguity lead to erroneous labelling, with an overall normalised majority-vote accuracy of 64.41%

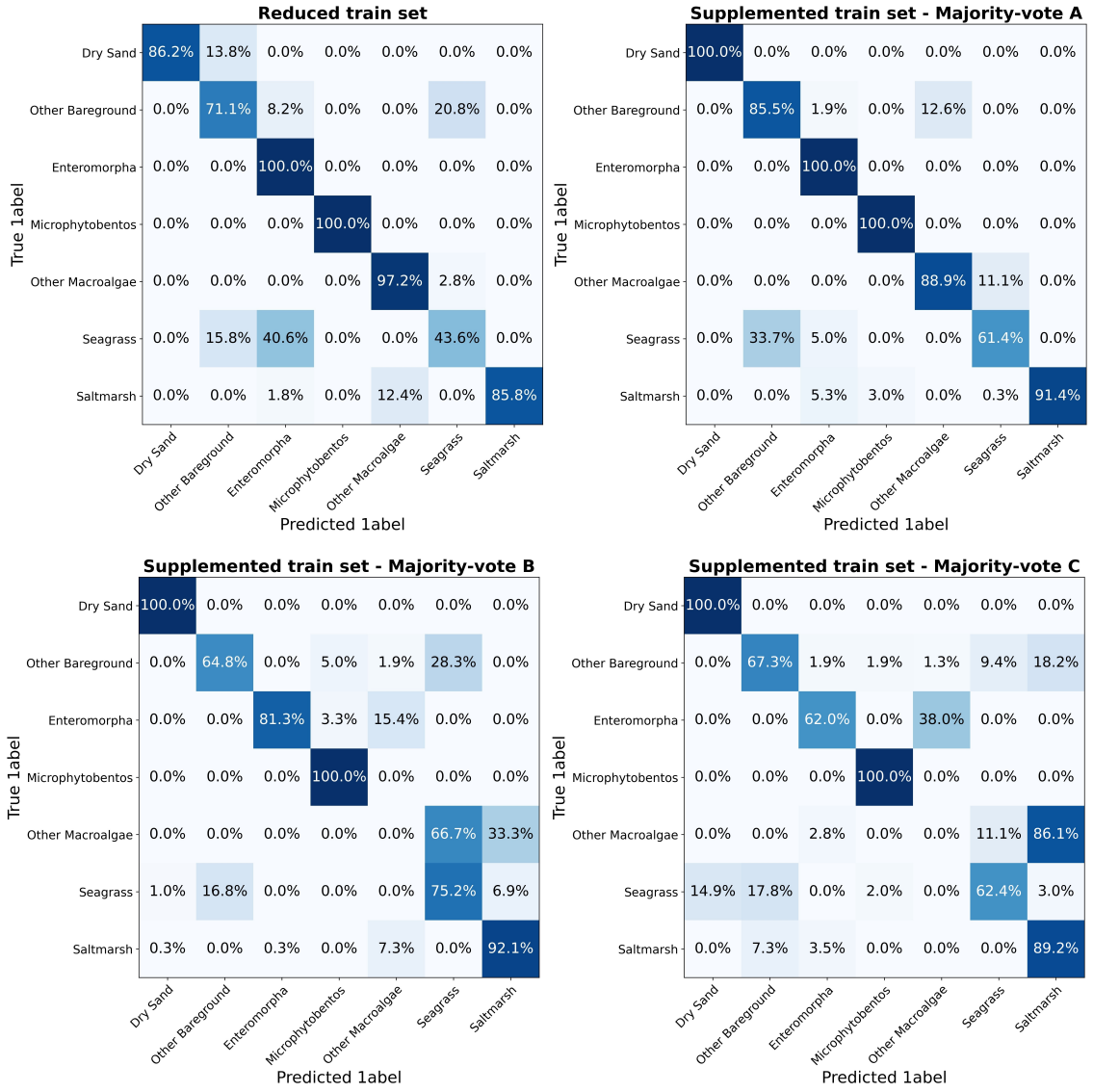

Fig. 8. Confusion matrices for FCNN models trained using set of in-situ labels (left), and using the same in-situ set supplemented with majority-vote annotations for groups A, B and C (top-right, bottom-left, bottom-right).

(middle-right matrix in Figure 6).

For participants in the final group C, the null hypothesis was also rejected, however by a smaller margin than group B. Again, this implies that participants within this group exhibit high inter-observer variability. Both the average accuracy and majority-vote accuracy were the lowest out of all groups, with $53.5^{+}_{-}10.82$ and 60.75% (bottom-left matrix in Figure 6), which also reflected low confidence in participant annotations. However, even with lower accuracy, participants within group C showed less variability in correct/incorrect annotations than the group B participants. This could be due to participants in group C not having any prior knowledge of the study site or with annotating aerial imagery and associating similar colour and texture based on the sample images in Figure 2 to the same class. The confusion matrix for group C provides insights into problematic target classes to annotate for subjects with the least experience. Algae classes, e.g. *Enteromorpha* sp. and other macroalgae, were often mutually mislabelled, while seagrass was often annotated as *Enteromorpha* sp.. This implies that vegetation classes were hard to discern from an aerial point of view with no prior knowledge. Furthermore and similarly to group A, other bareground, a sediment class that includes wet sand, was also incorrectly annotated as microphytobenthos, which again implies that these two classes are hard to discern from each other.

To sum up, this analysis covers three groups and assessed the inter-observer variability in participants with different backgrounds and expertise, while also assessing the accuracy of each participant, average group accuracy and majority vote accuracy. Participants in group A showed to have low inter-observer variability while also correctly annotating 81.31% of the points collectively. Participants in group B and C exhibited high inter-observer variability. Examining the criteria separating each group, having discipline expertise, prior

knowledge of the site and/or previous experience annotating marine biology play an important role in minimising inter-observer variability and ensure accurate annotation, and lack of exposure to these criteria leads to high variability and low confidence. While, our results also suggested that an expert ecologist or geomorphologist without in-situ exposure produced similar overall accuracy annotations as non-experts, this was influenced by the individual accuracy result of participant 6 since the majority of participants within group B yielded a higher accuracy in annotations than two of three participants in group C. Lastly, aggregating labels based on majority-vote annotations also draw parallels with field of expert frameworks in low-level image processing and ensemble learning [31], [32], [62], [63]. These frameworks model high-dimensional probability distributions by taking the product of several expert distributions, where each expert works on a low-dimensional subspace that is relatively easy to model. This is similar and accurate for annotations in all groups. In general, aggregating labels showed an increase in accuracy scores of 8.88%, 6.91% and 7.25%, respectively for groups A, B and C. This alludes to the specific and complementing nature of different research backgrounds aiding the accurate annotation.

B. FCNs with different versions of labelled data

The first test in our evaluation considered four FCNs trained with different versions of the labelled data.

First, FCNs trained with in-situ labels (top-left matrix in Figure 7) were viewed as the baseline for the remaining FCNs trained on majority-vote annotations from each group. The normalised accuracy with in-situ labels was 87.79% and models exhibited high confidence and accurate predictions for dry sand, other macroalgae, seagrass and saltmarsh. Other bareground proved to be a problematic class to model with a majority of predictions confused with microphytobenthos and *Enteromorpha* sp.. This paints a similar picture to majority-vote annotations for participants in group A (top-left matrix in Figure 6) whereby microphytobenthos was mislabelled as other bareground. However, FCNs do not mutually mislabel seagrass with *Enteromorpha spp.* which implies that FCNs were better at discerning these two specific vegetation classes than participants from group A.

The normalised accuracy for FCNs trained with majority-vote annotations from participants in group A was 81.99% (top-right matrix in Figure 7). This particular group exhibited low inter-observer variability and accurate annotations with the exception of microphytobenthos and other bareground; which may be due to both classes being present in wet sand. Furthermore, *Enteromorpha* sp. was mutually mislabelled with seagrass because both classes showed similar colour and texture from an aerial point of view. The latter bias in annotations from participants in group A was propagated to FCN performance - where 23.3% of seagrass labels were predicted as *Enteromorpha* sp. (top-right in Figure 7). However, examining *Enteromorpha* sp. predictions showed that this particular class was over represented due to erroneous predictions and confusion with other vegetation classes such as saltmarsh, seagrass and other macroalgae. Therefore, erroneous labels from participants in group A caused FCNs not only to mutually mislabel *Enteromorpha* sp. with seagrass, but also resulted in cascading errors for other vegetation classes due to overfitting for *Enteromorpha* sp.. Similarly to previous work using aerial imagery for annotation, this test also showed that empirical models can compensate certain degrees of erroneous human annotations [19], [28].

FCNs trained with majority-vote annotations from participants in group B yielded a normalised accuracy of 63.72% (bottom-left matrix in Figure 7). Annotations from participants in group B exhibited high inter-observer variability, resulting in low confidence in majority-vote annotations. This was due to conflicting biases between experts, i.e., ecologists, geomorphologists and sedimentologists, and the ambiguity effect through lack of exposure to the in-situ survey or aerial annotation of marine vegetation species from the study site. The main trends in human annotations from this group were other bareground mislabelled as dry sand, and a general confusion of vegetation classes between *Enteromorpha* sp., other macroalgae and seagrass. These errors were also propagated into FCN performance as 64.1% of other bareground predictions were mislabelled as dry sand and seagrass was severely misclassified and predicted as *Enteromorpha* sp. and other macroalgae, respectively 60.4% and 35.6% (bottom-left matrix Figure 7).

The final set of majority-vote labels from group C yielded a normalised accuracy of 66.36% (bottom-right matrix in Figure 7). Even though the average and majority-vote accuracy for annotations provided by group C were lower than results yielded by group B - FCNs trained with majority-vote annotations from subjects in group C yielded a higher test set accuracy than majority-vote annotations from group B. Our experiment showed that participants in group C presented high inter-observer variability but by less of a margin than group B (Table I in Section III-A). The analysis also showed that non-expert participants in group C exhibited low confidence predictions for other bareground with 31.8% of points labelled as microphytobenthos (bottom-left matrix in Figure 6). Similarly to participants in group B, they exhibited a general confusion in annotations for vegetation classes - in particular, seagrass and *Enteromorpha* sp. were often mutually misannotated. Again, these errors in human annotations were propagated to FCN errors, e.g. mutual misclassifications for seagrass and *Enteromorpha* sp. classes.

Our analysis supports the hypothesis that errors in crowd-sourced human annotation were propagated into the FCN performance. All groups had a similar trend whereby annotations for microphytobenthos were mislabelled with wet sediment classes. This bias was propagated into all models trained with majority-vote annotations where other bareground was either under represented (bottom-left matrix in Figure 7), over represented (bottom-right matrix in Figure 7) or confused with dry sand (top-right matrix in Figure 7). The mutual mislabelling of *Enteromorpha* sp. and seagrass points for participants in group A caused the FCN to misclassify all vegetation classes as *Enteromorpha* sp.. This showed that poor annotations not only propagated errors into the FCN performance, but could also cause cascading errors with classes that exhibit similar colour

and texture from an aerial point of view. This stresses the need for good quality labels as FCNs optimise their weights and biases based on a non-linear one-to-one mappings between image pixels and labelled maps [13]. However, our results also showed that FCNs trained with low inter-observer variability and high confidence annotations, as shown with subjects in group A, can demonstrate comparable performance to the FCNs trained with in-situ labels. Conversely, training with annotations from groups B or C, which manifested high inter-observer variability and higher rates of erroneous labelling, severely degraded FCN performance.

C. FCNs with balanced in-situ only versus crowdsourced supplemented labelled data

The second and final test in our evaluation considered two FCNs. One model was trained with only the balanced in-situ labels. Therefore, the training set was the initial balanced set of 53 randomly points with in-situ labels (see Section II-C) and the labels for the remaining 66 photo-interpreted points were replaced with the semantic value of majority-vote crowdsourced annotations.

For comparison, we considered a FCN trained with just the balanced set of 53 in-situ labels which yielded a normalised test set accuracy of 82.9% (top-left matrix in Figure 8). The accuracy was lower than FCNs trained with the combined full training set of 53 in-situ labels and 66 photo-interpreted labels (top-left Figure 7). This was expected as FCNs learn hierarchical representations of data through gradient descent [9], and if FCN kernel weight and bias adjustments were based on fewer image examples, then model performance and generalisation also degrades. The main affected and under represented class was seagrass where the accuracy dropped from 99.5% (top-left matrix in Figure 7) to 43.6% (top-left matrix in Figure 8).

The normalised accuracy for FCNs trained with the in-situ set supplemented with the labels from the participants in group A was 89.6% (top-right matrix in Figure 8) which was also the highest accuracy of all FCNs in our analysis. This setting improved the test set accuracy compared to the model trained with just in-situ labels. This was due to two reasons: first, supplementing the dataset allows for more unique samples to be incorporated into the training set, and second, the supplemented crowdsourced portion of the training set from group A exhibited low inter-observer variability and accurate annotations. Furthermore, this particular result provided an interesting comparison with the FCN trained on in-situ plus photo-interpreted labels (the top-left matrix in Figure 7). Both FCNs yielded satisfactory results which confirms that aggregated labels from multiple annotators within group A were as good as the efforts of a single expert annotator (lead author). This comparison also showed that in-situ efforts can be combined successfully with aerial imagery annotation, which could reduce costs and labour from in-situ surveys.

The accuracy for FCNs trained using in-situ labels supplemented with the labels from participants in groups B and C were respectively 73.34% and 68.7% (bottom-left and bottom-right matrices in Figure 8). Our analysis of both datasets was performed jointly as FCNs trained in both settings paint a similar picture. Both sets of models failed to achieve better results than models trained with just the balanced set of in-situ labels (top-left in Figure 8), which again stresses the need for good quality crowdsourced labels. FCNs trained with majority-vote annotations from participants in group B over represented seagrass and also misclassified all other macroalgae pixels, mostly as seagrass (bottom-left matrix in Figure 8). A similar outcome happened for models supplemented with the labels provided by group C - again all other macroalgae class instances are misclassified, this time mostly as saltmarsh (bottom-left matrix in Figure 8). In both settings this would be due to poor annotation performance from these two groups (see Fig. 6).

V. Conclusion

This work analysed the feasibility of using crowdsourced annotations on a complex multi-class problem domain that includes intertidal coastal species, such as seagrass, saltmarsh and macro-algae.

To assess the quality of crowdsourced annotations, an inter-observer variability experiment was performed with a population of 12 participants that were split into 3 sets of groups. The criterias for each group were based on discipline expertise and previous experience with either annotating aerial imagery for this study site or marine biology in general. The assessment was possible by analysing the statistical differences in crowdsourced annotations using the Cochran's Q test. Furthermore, the annotation accuracy and a per-class analysis was used to assess for any potential observation biases.

The results for our experiment show that discipline experts familiar with the study site were more accurate than experts with no prior knowledge of the site and non-experts. This confirms that discipline expertise, prior knowledge of the site and/or previous experience annotating marine biology play an important role in minimising inter-observer variability and ensuring accurate annotation, and that lack of exposure to either these criteria leads to high variability and low confidence. Furthermore, the results of our analysis also point to a small performance gain between annotators with expert discipline knowledge versus annotators with no previous experience in marine biology annotation or domain expertise. But, this may be skewed due to annotations from participant 6.

The experiment stressed the difficulty of labelling a complex multi-class marine biology problem and therefore, we conclude that pre-exposure to the study site is important for intertidal classification if good quality labels are to be guaranteed and that in-situ groundtruthing may be unavoidable to prevent confusion by site experts. For instance, the general confusion between microphytobenthos with other bareground and *Enteromorpha* sp. with seagrass (Sections III-A, IV-B and IV-C).

For the experiment with FCNs trained with crowdsourced annotations, two scenarios were considered: the first was a direct comparison of FCNs trained with majority-vote crowdsourced annotation from each participant group with FCNs trained with transcribed in-situ labels. This showed that annotations that exhibit low inter-observer variability and high

confidence annotations, as shown with subjects in group A, demonstrate comparable performance to the FCNs trained with in-situ labels. Conversely, training with annotations from groups B or C, which manifested high inter-observer variability and higher rates of erroneous labelling, severely degraded FCN performance. Therefore, we conclude that errors in crowdsourced human annotations were propagated into FCN performance. The second experiment considered two FCNs: one whereby the training set was the initial balanced set of 53 points with transcribed in-situ labels (see Section II-D), and the other where the training set was the initial set of 53 points with in-situ labels supplemented with majority-vote annotations from each participant group. In this scenario, FCNs supplemented with majority-vote annotations from participant group A reported a normalised accuracy of 89.6%, which was also the highest accuracy of all FCNs in our analysis. This showed that in-situ efforts can be combined successfully with crowdsourced aerial imagery annotation, which could reduce costs and labour from in-situ surveys, given that crowdsourced labels are consistent and accurate. Similarly to the previous scenario, FCNs supplemented with majority-vote annotations from participant groups B and C severely degraded FCN performance, which again stresses the need for good quality crowdsourced labels.

However, this work does not fully exclude in-situ surveying but merely affirms that a good quality labels can be found in-situ but a healthy quantity of labels can also be supplemented from aerial imagery which would reduce in-situ efforts and costs.

ACKNOWLEDGEMENTS

The authors thank Cefas and the EA for providing necessary imagery as well as data from the in-situ survey. The authors also thank the participant population comprised of expert ecologists and geomorphologists from Cefas and the EA for taking part in the experiment.

REFERENCES

[1] Klemas, Victor V, "Remote sensing techniques for studying coastal ecosystems: An overview", *Journal of coastal research*, vol. 27, no. 1, pp. 2–17, 2011

[2] McCarthy, Matthew J and Colna, Kaitlyn E and El-Mezayen, Mahmoud M and Laureano-Rosario, Abdiel E and Méndez-Lázaro, Pablo and Otis, Daniel B and Toro-Farmer, Gerardo and Vega-Rodriguez, Maria and Muller-Karger, Frank E, "Satellite remote sensing for coastal management: A review of successful applications", *Environmental management*, vol. 60, no. 2, pp. 232–339, 2017

[3] Pereira, Henrique M and Leadley, Paul W and Proença, Vânia and Alkemade, Rob and Scharlemann, Jörn PW and Fernandez-Manjarrés, Juan F and Araújo, Miguel B and Balvanera, Patricia and Biggs, Reinette and Cheung, William WL and others, "Scenarios for global biodiversity in the 21st century", *Science*, vol. 330, no. 6010, pp. 1260–1501, 2010

[4] Richards, John Alan, "Remote sensing digital image analysis", *Journal of coastal research*, vol. 3, 1999

[5] Klemas, Victor V, "Coastal and environmental remote sensing from unmanned aerial vehicles: An overview", *Journal of coastal research*, vol. 31, no. 5, pp. 1260–1267, 2015

[6] Anderson, Karen and Gaston, Kevin J, "Lightweight unmanned aerial vehicles will revolutionize spatial ecology", *Frontiers in Ecology and the Environment*, vol. 11, no. 3, pp. 138–146, 2013

[7] Duffy, James P and Pratt, Laura and Anderson, Karen and Land, Peter E and Shutler, Jamie D, "Spatial assessment of intertidal seagrass meadows using optical imaging systems and a lightweight drone", *Estuarine, Coastal and Shelf Science*, vol. 200, pp. 169–180, 2018

[8] Turner, Darren and Lucieer, Arko and Watson, Christopher, "An automated technique for generating georectified mosaics from ultra-high resolution unmanned aerial vehicle (UAV) imagery, based on structure from motion (SfM) point clouds", *Remote Sensing*, vol. 4, no. 5, pp. 1392–1410, 2012

[9] LeCun, Yann and Bengio, Yoshua and Hinton, Geoffrey, "Deep learning", *Nature*, vol. 521, no. 7553, pp. 436–444, 2015

[10] Krizhevsky, Alex and Sutskever, Ilya and Hinton, Geoffrey E, "Imagenet classification with deep convolutional neural networks", *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012

[11] Ronneberger, Olaf and Fischer, Philipp and Brox, Thomas, "U-net: Convolutional networks for biomedical image segmentation", in *International Conference on Medical image computing and computer-assisted intervention*, LOCATION, pp. 234–241, 2015

[12] He, Kaiming and Gkioxari, Georgia and Dollár, Piotr and Girshick, Ross, "Mask r-cnn", in *Proceedings of the IEEE international conference on computer vision*, LOCATION, pp. 2961–2969, 2017

[13] Long, Jonathan and Shelhamer, Evan and Darrell, Trevor, "Fully convolutional networks for semantic segmentation", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, LOCATION, pp. 3431–3440, 2015

[14] Dang, Thai-Viet, and Ngoc-Tam Bui "Multi-scale fully convolutional network-based semantic segmentation for mobile robot navigation", in *Electronics*, vol. 12, pp. 533, 2023

[15] Hamdi, Zayd Mahmoud and Brandmeier, Melanie and Straub, Christoph, "Forest damage assessment using deep learning on high resolution remote sensing data", *Remote Sensing*, vol. 11, no. 17, pp. 1976, 2019

[16] Bowler, Ellen and Fretwell, Peter T and French, Geoffrey and Mackiewicz, Michal, "Using deep learning to count albatrosses from space: Assessing results in light of ground truth uncertainty", *Remote Sensing*, vol. 12, no. 12, pp. 2026, 2020

[17] Li, Weijia and Fu, Haohuan and Yu, Le and Cracknell, Arthur, "Deep learning based oil palm tree detection and counting for high-resolution remote sensing images", *Remote Sensing*, vol. 9, no. 1, pp. 22, 2017

[18] Xu, Yongyang and Wu, Liang and Xie, Zhong and Chen, Zhanlong, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters", *Remote Sensing*, vol. 10, no. 1, pp. 144, 2018

[19] Kattenborn, Teja and Eichel, Jana and Fassnacht, Fabian Ewald, "Convolutional Neural Networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery", *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019

[20] Eickhoff, Carsten and de Vries, Arjen P, "Increasing cheat robustness of crowdsourcing tasks", *Information retrieval*, vol. 16, no. 2, pp. 121–137, 2013

[21] Congalton, Russell G, "Remote sensing and geographic information system data integration: error sources and research issues", *Photogrammetric Engineering & Remote Sensing*, vol. 57, no. 6, pp. 677–687, 1991

[22] Leitão, Pedro J and Schwieder, Marcel and Pötzschner, Florian and Pinto, José RR and Teixeira, Ana MC and Pedroni, Fernando and Sanchez, Maryland and Rogass, Christian and van der Linden, Sebastian and Bustamante, Mercedes MC and others, "From sample to pixel: multi-scale remote sensing data for upscaling aboveground carbon data in heterogeneous landscapes", *Ecosphere Engineering & Remote Sensing*, vol. 9, no. 8, pp. 2298, 2018

[23] Hobley, Brandon and Arosio, Riccardo and French, Geoffrey and Bremner, Julie and Dolphin, Tony and Mackiewicz, Michal, "Semi-supervised segmentation for coastal monitoring seagrass using RPA imagery", *Remote Sensing*, vol. 13, no. 9, pp. 1741, 2021

[24] Tan, Chuanqi and Sun, Fuchun and Kong, Tao and Zhang, Wenchang and Yang, Chao and Liu, Chunfang, "A survey on deep transfer learning" in *International conference on artificial neural networks*, LOCATION, pp. 270–279, 2018

[25] Shorten, Connor and Khoshgoftaar, Taghi M, "A survey on image data augmentation for deep learning", *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019

[26] Tarvainen, Antti and Valpola, Harri, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results", *Advances in neural information processing systems*, vol. 30, 2017

[27] French, Geoff and Laine, Samuli and Aila, Timo and Mackiewicz, Michal and Finlayson, Graham, "Semi-supervised semantic segmentation needs strong, varied perturbations", *arXiv preprint arXiv:1906.01916*, 2019

[28] Kattenborn, Teja and Lopatin, Javier and Förster, Michael and Braun, Andreas Christian and Fassnacht, Fabian Ewald, "UAV data as alternative to field sampling to map woody invasive species based on combined

Sentinel-1 and Sentinel-2 data", *Remote sensing of environment*, vol. 227, pp. 61–73, 2019

[29] Wagner, Fabien H and Sanchez, Alber and Tarabalka, Yuliya and Lotte, Rodolfo G and Ferreira, Matheus P and Aidar, Marcos PM and Gloor, Emanuel and Phillips, Oliver L and Aragao, Luiz EOC, "Using the U-net convolutional network to map forest types and disturbance in the Atlantic rainforest with very high resolution images", *Remote Sensing in Ecology and Conservation*, vol. 5, no. 4, pp. 360–375, 2019

[30] Lopatin, Javier and Dolos, Klara and Kattenborn, Teja and Fassnacht, Fabian E, "How canopy shadow affects invasive plant species classification in high spatial resolution remote sensing", *Remote Sensing in Ecology and Conservation*, vol. 5, no. 4, pp. 302–317, 2019

[31] Hinton, Geoffrey, "Training products of experts by minimizing contrastive divergence" *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002

[32] Polikar, Robi, "Ensemble learning" *Ensemble machine learning*, pp. 1–34, 2012

[33] Saralioglu, Ekrem and Gungor, Oguz, "Crowdsourcing in remote sensing: A review of applications and future directions", *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 4, pp. 89–110, 2020

[34] Herfort, Benjamin and Li, Hao and Fendrich, Sascha and Lautenbach, Sven and Zipf, Alexander, "Mapping human settlements with higher accuracy and less volunteer efforts by combining crowdsourcing and deep learning", *Remote Sensing*, vol. 11, no. 15, pp. 1799, 2019

[35] Albuquerque, João Porto de and Herfort, Benjamin and Eckle, Melanie, "The tasks of the crowd: A typology of tasks in geographic information crowdsourcing and a case study in humanitarian mapping", *Remote Sensing*, vol. 8, no. 10, pp. 859, 2016

[36] Gueguen, Lionel and Koenig, Jan and Reeder, Carl and Barksdale, Tim and Saints, Jon and Stamatiou, Kostas and Collins, Jeffery and Johnston, Carolyn, "Mapping human settlements and population at country scale from VHR images", *Remote Sensing*, vol. 10, no. 2, pp. 524–538, 2016

[37] Wang, Sherrie and Di Tommaso, Stefania and Faulkner, Joey and Friedel, Thomas and Kennepohl, Alexander and Strey, Rob and Lobell, David B, "Mapping crop types in southeast India with smartphone crowdsourcing and deep learning", *Remote Sensing*, vol. 12, no. 18, pp. 2957, 2019

[38] Papakonstantinou, Apostolos and Batsaris, Marios and Spondylidis, Spyros and Topouzelis, Konstantinos, "A Citizen Science Unmanned Aerial System Data Acquisition Protocol and Deep Learning Techniques for the Automatic Detection and Mapping of Marine Litter Concentrations in the Coastal Zone", *Drones*, vol. 5, no. 1, pp. 6, 2021

[39] Harley, Mitchell D and Kinsela, Michael A and Sánchez-García, Elena and Vos, Kilian, "Shoreline change mapping using crowd-sourced smartphone images", *Coastal Engineering*, vol. 150, pp. 175–189, 2019

[40] Ladle, M, "The Haustoriidae (Amphipoda) of Budle Bay, Northumberland", *Crustaceana*, vol. 28, no. 1, pp. 37–47, 1975

[41] Meyer, AN, "An investigation into certain aspects of the ecology of Fenham flats and budle bay, Northumberland", 1973

[42] Olive, PJW, "Management of the exploitation of the lugworm Arenicola marina and the ragworm Nereis virens (Polychaeta) in conservation areas", *Aquat. Conserv. : Mar. Freshw. Ecosyst.*, vol. 3, no. 1, pp. 1–24, 1993

[43] Agisoft, LLC, "Agisoft metashape user manual, Professional edition, Version 1.5", *Agisoft LLC St. Petersburg Russ. Https://Www. Agisoft. Com/Pdf/Metashape-Pro_1_5_En. Pdf Accessed June*, vol. 2, pp. 2019, 2018

[44] Cunliffe, Andrew M and Brazier, Richard E and Anderson, Karen, "Ultra-fine grain landscape-scale quantification of dryland vegetation structure with drone-acquired structure-from-motion photogrammetry", *Remote Sens. Environ.*, vol. 183, pp. 129–143, 2016

[45] Nolet, Corjan and Poortinga, Ate and Roosjen, Peter and Bartholomeus, Harm and Ruessink, Gerben "Measuring and modeling the effect of surface moisture on the spectral reflectance of coastal beach sand", *PLoS One*. vol. 9, no. 11, pp. 112–151, 2014

[46] Shuman, Craig S and Ambrose, Richard F, "A comparison of remote sensing and ground-based methods for monitoring wetland restoration success", *Restoration Ecology*, vol. 11, no. 3, pp. 325–333, 2003

[47] Mumby, PJ and Green, EP and Edwards, AJ and Clark, CD, "Measurement of seagrass standing crop using satellite and digital airborne remote sensing", *Marine ecology progress series*, vol. 159, pp. 51–60, 1997

[48] Eickhoff, Carsten, "Cognitive biases in crowdsourcing", in *Proceedings of the eleventh ACM international conference on web search and data mining*, LOCATION, pp. 162–170, 2018

[49] Hajian, Sara and Bonchi, Francesco and Castillo, Carlos, "Algorithmic bias: From discrimination discovery to fairness-aware data mining", in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, LOCATION, pp. 2125–2126, 2016

[50] Adam, Paul, "Saltmarsh ecology", 1993

[51] Hootsmans, MJM and Vermaat, JE and Van Vierssen, W, "Seed-bank development, germination and early seedling survival of two seagrass species from The Netherlands: Zostera marina L. and Zostera noltii Hornem", *Aquatic Botany*, vol. 29, no. 3, pp. 217–226, 1987

[52] Jiménez, Carlos and Niell, F Xavier and Algarra, Patricia, "Photosynthetic adaptation of Zostera noltii Hornem", *Aquatic Botany*, vol. 28, no. 3–4, pp. 275–285, 1987

[53] Hodges, J and Howe, M, "Milford Haven waterway monitoring of eelgrass, Zostera angustifolia, following the Sea Empress oils spill", *Report to Shoreline & Terrestrial Task Group. Sea Empress Environmental Evaluation Committee*, 1997

[54] MacIntyre, Hugh L and Geider, Richard J and Miller, Douglas C, "Microphytobenthos: the ecological role of the "secret garden" of unvegetated, shallow-water marine habitats. I. Distribution, abundance and primary production", *Estuaries*, vol. 19, no. 2, pp. 186–201, 1996

[55] Tillin, HM and Budd, Georgina, "Green seaweeds (Ulva spp. and Cladophora spp.) in shallow upper shore rockpools", *Marine Biological Association of the United Kingdom*, 2016

[56] Catarino, Marcelo D and Silva, Artur and Cardoso, Susana M, "Phycochemical constituents and biological activities of Fucus spp.", *Marine drugs*, vol. 16, no. 8, pp. 249, 2018

[57] Patil, Kashinath D, "Cochran's Q test: Exact distribution", *Journal of the American Statistical Association*, vol. 70, no. 349, pp. 186–189, 1975

[58] Kanji, Gopal K, "100 statistical tests", 2006

[59] Ellsberg, Daniel, "Risk, ambiguity, and the Savage axioms", *The quarterly journal of economics*, pp. 643–669, 1961

[60] Simonyan, Karen and Zisserman, Andrew, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.1556*, 2014

[61] Tinlin-Mackenzie, Ashleigh and Delany, Jane and Scott, Catherine L and Fitzsimmons, Clare, "Spatially modelling the suitability, sensitivity, and vulnerability of data poor fisheries with GIS: A case study of the Northumberland lugworm fishery", *Marine Policy*, vol. 109, 2019

[62] Roth, Stefan and Black, Michael J, "Fields of experts" *International Journal of Computer Vision*, vol. 82, no. 2, pp. 205–229, 2009

[63] Roth, Stefan and Black, Michael J, "Fields of experts: A framework for learning image priors" *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 860–867, 2005