Uncertainty as a Proxy of the Generalization Error for Marine Species Identification

David Serrano Lozano

Abstract

Marine Protected Areas monitoring is a must to understand ecological processes and assess whether management aims are fulfilled. One of the best ways of doing it is by using a Remotely Operated Underwater Vehicle to collect images. However, the main drawback is the large amount of data that has to be annotated by specialists. In this thesis, we propose to go one step further and use a deep learning system to maximize the system's performance while reducing the human workload. The algorithm reports, in addition to the deterministic decision, uncertainty estimations to identify potential misclassifications. However, evaluating the model doubtfulness is not trivial and, therefore, we test several well-known and a novel metric which evaluates the quality of the estimations regarding its ranking. Furthermore, we propose a systematic method to reduce the workload from non-annotated datasets, using uncertainty as a proxy of the generalization error and automatically labelling with the model the most certain samples.

https://github.com/davidserra9/UncertaintyProxy

Index Terms

Uncertainty, Uncertainty Evaluation, Generalization Error, Deep Learning, Marine Species Classification

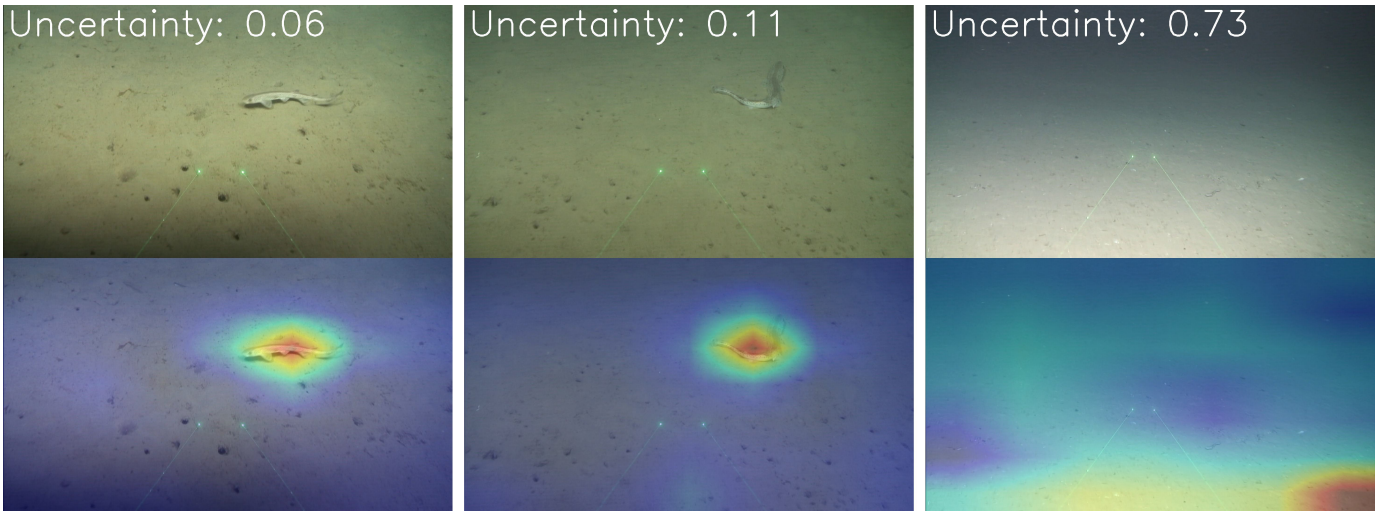


Fig. 1. Three different images of a *Scyliorhinus canicula* with the corresponding uncertainty estimations (predictive entropy) and CAMs predicted by our system. The first two images are predicted correctly with low uncertainty as the creatures are relatively easy to see. However, the third sample is misclassified as other marine species since it is challenging to identify the *S. canicula* even for a human being. We can observe how the uncertainty value is higher and, thus, could be used as proxy of the generalization error.

I. Introduction

THE oceans have it all: from microscopic life to the largest animal that has ever lived on Earth, from the colourless to the shimmering, from the frozen to the boiling and from the sunlit to the mysterious dark of the deepest part of the planet. Oceans are an essential component of the Earth's ecosystems and jointly with marine life, they make the planet habitable for humankind. The huge body of salt water that covers 71% of Earth's surface is the main life support and regulation tool to keep the global climate system in check [1]. It is the world's largest ecosystem, home to nearly a million known species and containing the vast untapped potential for scientific discovery [2]. Despite the critical importance of conserving oceans, decades of irresponsible exploitation and human footprint have led to an alarming level of degradation. Overfishing, toxic pollution, invasive species, nutrient over-enrichment, habitat degradation and destruction, biodiversity loss, the dependence of a growing

Author: David Serrano Lozano, 99d.serrano@gmail.com
Advisor 1: David Masip Rodó, Scene Understanding and Artificial Intelligence, Universitat Oberta de Catalunya
Advisor 2: José Antonio García del Arco, Institut de Ciències del Mar
Thesis dissertation submitted: September 2022

global population on its goods and services, coastal development and acidification, all threaten the sustainability of the ocean ecosystems [3, 4]. This is why in 2015, the United Nations introduced the SDG (Sustainable Development Goal) 14 about "Life below water" which aims to conserve and sustainably use the oceans, seas and marine resources focusing on 10 different targets [5].

Marine ecosystems contain a diverse array of living organisms and abiotic processes. That is why governments have placed Marine Protected Areas (MPAs) worldwide, spatially-delimited areas of the marine environment that are managed, at least in part, for the conservation of biodiversity and the recovery of threatened species [6, 7]. In order to assess whether management aims are fulfilled, and to understand ecological processes that confine MPA success, it is necessary to implement a program of research and monitoring of the systems in question. However, as of September of 2022, there are more than 17 thousand MPAs worldwide with a total area of 29 million square kilometres, representing 8.13% of the ocean [8] which monitoring cannot be achieved by sporadic observation.

Remotely Operated Underwater Vehicles (ROVs) are a powerful tool used in a lot of expeditions to survey and monitor aquatic habitats and species for large areas [9–11]. Their image quality and their manoeuvrability allow researchers to capture videos of large regions without any need to use invasive techniques or put both people and animals at risk. Nevertheless, the main drawback is the fact that large amounts of multimedia content have to be annotated with great precision. In addition, labelling underwater images is even harder as water attenuates light sharply, causes high noise and introduces haziness when light is absorbed or scattered many times by floating matter [12]. This is why, to help biologists with the tedious and repetitive task of annotating hundreds of hours of underwater videos, we propose a deep learning system capable of identifying marine species using uncertainty as a proxy of the generalization error to both reduce the quantity of image-level labels and recognise potential misclassifications.

As there exist around a million different marine species [13] and most of them only live in small areas we propose a system that is easily reproducible for any MPA or location in the world reducing the workload involved in labelling tonnes of images efficiently. Furthermore, instead of requiring the exact bounding box of the animals in order to train the model, our solution only uses image-level labels to save time when annotating with the possibility of obtaining an approximate position using Class Activations Maps (CAMs) [14].

In recent years, deep neural networks (DNNs) [15] have revolutionized computer vision and gained considerable traction in challenging scientific data analysis problems [16]. Surprisingly, however, most DNN-based solutions for automatically identifying animals have so far produced deterministic outputs and do not quantify or handle uncertainty in the prediction. Especially in underwater imagery, risk management plays a crucial role due to environmental characteristics and differences between animals of the same species as the appearance of creatures changes during growth, such as size, shape and colour. Notwithstanding the above, information about the reliability of automated decisions should be a key requirement to avoid misclassifications, allowing a specialist to correct the doubtful examples. Thus, DNNs should report, in addition to the decision, an associated estimate of uncertainty to identify potential misclassifications, in particular since some images may be more challenging to analyze and classify than others. In Figure 1 we show three different images of a *Scyliorhinus canicula*, the first two predicted correctly and the third one incorrectly. It can be seen how by the use of CAMs, the system is capable of localizing the creature correctly, however, depending on its difficulty to be seen, the uncertainty, in this case, predictive entropy, is higher or lower. Furthermore, we can observe how the uncertainty is higher in the wrongly predicted image and, therefore, we hypothesize that uncertainty can be used as an indicator of the error.

The main disadvantage of estimating uncertainty is its evaluation as there is no conceivable way of obtaining ground truth labels and the lack of a unified protocol. In DNNs, the gold standard is to use the softmax operator to convert the continuous activations of the output layer to class probabilities. However, the estimates are unreliable, as the distance of the predicted label of a newly seen observation is not useful for the conclusion besides its comparative value against other classes. Furthermore, the probabilistic interpretation of the cross-entropy loss is mere Maximum Likelihood Estimation (MLE) which is not capable of inferring the predictive distribution variance [17]. However, despite all of the above, in classification models, the probability vector obtained at the softmax output is often erroneously interpreted as model confidence due to the fact that imperceptible perturbations to a real image or out-of-distribution samples could change a DNNs softmax output to arbitrary values [18–20].

Due to the lack of a unified definition of the uncertainty, we test several well-known and new metrics, as well as, visualization tools to demonstrate that uncertainty can be used as a proxy of the generalization error and, in addition, the model doubtfulness can be used for identifying error-prone samples. Furthermore, we also propose a systematic protocol to reduce the human workload involved on annotating datasets by training a DNN to automatically label the most certain samples, leaving for the expert the most doubtful images.

Our contributions to this thesis are as follows. (1) Create a new dataset from videos acquired using a ROV in Mediterranean locations. (2) Test the viability of the benchmark with several CNN architectures. (3) Test well-known evaluation metrics and propose new ones for evaluating uncertainty for both individual images and the entire system as one. (4) Besides the scalar statistics, propose new graphical plots to ease the illustration of the uncertainty estimations. (5) Demonstrate that uncertainty can be used as a proxy of the generalization error when using the right metrics and hyperparameters proposing the Uncertainty Driven Classification protocol which reduces the workload involved in labelling images using uncertainty estimates.

Regarding the structure of this thesis, it is organized as follows: Section II describes the previous marine species classification

works, reviews the main concepts about uncertainty and briefly exposes the main techniques for both estimation and evaluation. Section III is divided into four main parts where it describes the DNNs used, the uncertainty estimation and evaluation and the proposed protocol. Section IV summarizes the implemented experiments and Section V its results. Finally, Section VI exposes the achievements of this study, the conclusions and the future lines of research.

II. State of the art

In this section, some of the most important work and material about fish identification will be reviewed, as well as sources and types of uncertainty and both its relationship with deep learning and different approaches to estimating it in DNNs.

A. Fish Detection

In recent years, the analysis of marine ecology has been on the rise. The main challenges remain to differentiate the species from the background and associate distinct creatures with the same species although having different colours, shapes or sizes. Regarding the data, the major drawback is the need for qualified staff and the amount of time that the annotator has to be fully focused as an animal may only pass through the scene for a few frames. Furthermore, the available datasets are limited to only a few annotated marine species. In this subsection, we first review the different animal detection methods, and then we survey the most useful datasets created worldwide to the best of our knowledge.

1) Appearance-Based Techniques: Although the appearance of marine species changes drastically through their growth, appearance-based techniques can identify static visual appearance features for all living creatures with only one image per species. Several works deal with the features obtained from colour [21–24], texture [25, 26] and both combined [27–30]. The task is very challenging because the high variability of classes and the light attenuation through the water cause loss of colour information. Trying to manage the almost colourless environment, other works use shape descriptors which solely consider the boundaries of the creatures and neglect the information contained in the interior [31–33]. Local feature descriptors such as SIFT, SURF [34, 35], Webber [24], HOG [21, 36] and LBP [37] are also used to extract features and later on classify the detected species using SVM [28, 34, 35, 38–40], KNN [41–43] or AdaBoost [24, 44].

2) Motion-Based Techniques: Appearance-based techniques are generally based on static images, which ignores motion information of creatures in 2D and 3D space. For marine species behaviour and detection, tracking is crucial, thus motion-based techniques are used to establish dynamic background models. Generally, motion-based techniques fall into two categories: background subtraction and optical flow.

Background subtraction is a common method which is utilized to segment the moving region and allows detection and distinction of dynamic objects obtained by static cameras. To avoid the influence of illumination changes, the background image needs to be updated in real-time based on the current image frame. A majority of studies have utilized GMM [38, 45, 46] which is reliable enough to eliminate false positives but is incredibly challenging to extract pure background frames since slow-moving objects may fail. At the same time, GMM execution speed is slow and cannot be performed well in short video sequences [47]. To solve that, Hsiao and Chen [48] proposed a faster background modelling method using non-parametric histograms.

Optical flow is used to represent motion information defined as the pattern of apparent motion of objects and surfaces caused by the relative velocity between an observer and a scene. Ye *et al.* [49] used it to track entire fish groups instead of using sophisticated methods to track them individually. Zhao *et al.* [50] used a particle advection scheme and optical flow (PAOF) to detect global fish gathering and scattering behaviours. Experimental results show that the optical flow method is susceptible to slight fluctuation in the hyperparameters and like background subtraction methods, works poorly when the camera is constantly moving and with species that hardly move.

3) Deep Learning Techniques: Hand-crafted features and traditional machine learning require a large effort and can only deal with small datasets. However, deep learning can automatically extract low-level and high-level features from big data being robust to changes in illumination, translation, and rotation, thus making them suitable for computer vision modelling. Among the most popular deep learning techniques, convolutional neural networks (CNNs) have demonstrated the ability to achieve high performance. The vast majority of studies focus on purely detecting and locating multiple species requiring a huge amount of labelled bounding boxes. Hence, designing an effective detection framework is essential for reducing both computational costs and the required amount of data.

Region-based two-stage detection methods have been widely used in underwater videos. The first stage consists in generating region proposals while the second one is extracting features and classifying the possible object. In [51–53], authors used Faster-RCNN, outperforming DPM, R-CNN and Fast R-CNN both in Mean Average Precision and speed. Nevertheless, algorithms based on the R-CNN series are slow and cannot achieve real-time detection. To further improve detection speed, the end-to-end You Only Look Once (YOLO) algorithm adopted the idea of regression performing detection and classification on a predefined grid [54–57].

However, when implementing an object detection framework not only the type of species has to be annotated but also its position. Other studies have focused on identifying and classifying marine species without interest in the exact position of

TABLE I
AVAILABLE UNDERWATER DATASETS

	Number of videos/images	Resolution (pix)	Total number of labels	Number of species
Fish4-Knowledge	700,000 videos of 10 min	320 x 240	-	3,000
LCF-14	1,000 videos of 10 min	640 x 480	19,868	10
LCF-15	93 videos of 10 min	640 x 480	9,000	15
DeepFish	40,000 images	1920 x 1080	280,000	-
NOAA	4,000 images	720 x 480	4,119	-
ICM-20.1 (ours)	24,145 images	1920 x 1080	24,145	6

the creatures. Chhabra *et al.* [58] used a pre-trained VGG16 model to classify 8 different fish, Mathur and Goel [59] used a ResNet-50 in a small dataset and Agarwal *et al.* [60] used a variety of CNNs to identify and classify 9 different types of seafood.

4) Datasets: The datasets play an important role in fish detection, not only for measuring the performance of the algorithm but also for driving more researchers to solve more complex and challenging problems. In this section, we summarized the generic public datasets with their corresponding number of videos or images, resolution, number of labelled creatures and number of species, which can be seen in Table I.

Fish4-Knowledge [61] is a research project which investigated information abstraction and storage methods for reducing the massive amount of video data for describing fish. The dataset contains 700 thousand clips of ten minutes of underwater clips monitoring the Nanwan, Lanyu and Houbi Lakes of Taiwan. The public dataset was taken within 2 years showing several phenomena from sunrise to sunset *e.g.*, murky water, algae, etc.

LifeClef2014 (LCF-14) [62] and LifeClef2015 (LCF-15) [63] are two datasets derived from the Fish4-Knowledge repository, which consists of only 1000 and 93 videos of 10 and 15 fish species, correspondingly.

DeepFish [64] is an underwater fish species image dataset with around 40 thousand images that capture high variability fish habitats. The videos were collected for 20 habitats from remote coastal marine environments of tropical Australia. The dataset is suitable for classification, detection and segmentation since it has image-level, point-level and per-pixel annotations.

To evaluate the performance of the recognition, detection and tracking algorithms, a novel public dataset was developed on natural habits by National Oceanic and Atmospheric Administration (NOAA) [37]. The dataset was taken by using a camera on a ROV in the Southern California Bight from 2000 to 2012 consisting of more than three thousand images labelled with bounding boxes.

B. Uncertainty

1) Uncertainty in Deep Neural Networks: Understanding what a model does not know is essential in many deep learning systems. Nowadays, DNNs are able to learn powerful representations which can map high-dimensional data to an array of outputs. Regardless, these mappings are often taken blindly and assumed to be accurate, which is not always the case. Due to enormous advances, deep learning has reached almost every field of science becoming crucial to achieving state-of-the-art performance. However, the vast majority of the models are not able to represent uncertainty since, for example, classification models, often give normalized score vectors, which do not necessarily capture model doubtfulness.

While there are many sources of uncertainty, they are generally characterized as epistemic or aleatoric [65]. Epistemic or model uncertainty captures the uncertainty in the model parameters. It arises from the lack of sufficient data to train the model to infer the underlying data-generating function correctly - uncertainty which captures our ignorance about which model generated our collected data. Therefore, epistemic uncertainty is inversely proportional to the density of training examples and could be reduced by collecting and training on more data. On the other hand, aleatoric uncertainty is described by the noise in the observations. This type of uncertainty arises due to hidden variables or measurement errors and cannot be explained away by capturing more data [66].

However, especially in practical applications employing real-world data, the training data is only a subset of all possible input data, which means that a miss-match between the DNN domain and the unknown actual data domain is often unavoidable. An exact representation of the uncertainty of a DNN prediction is not possible to compute, since the different uncertainties can in general not be modelled accurately and are most often even unknown. Therefore, methods for estimating uncertainty in a DNN are a popular and vital field of research.

2) Uncertainty Estimation: In [67] split the methods for estimating uncertainty into four different types based on the number (single or multiple) and the nature (deterministic or stochastic) of the used DNNs, as can be seen in Figure 2.

- *Single deterministic methods* give the prediction based on one single forward pass within a deterministic network and each repetition of a forward pass delivers the same result. They can be categorized into approaches where one single network is explicitly modelled and trained in order to quantify uncertainties [17, 68] and approaches that use additional

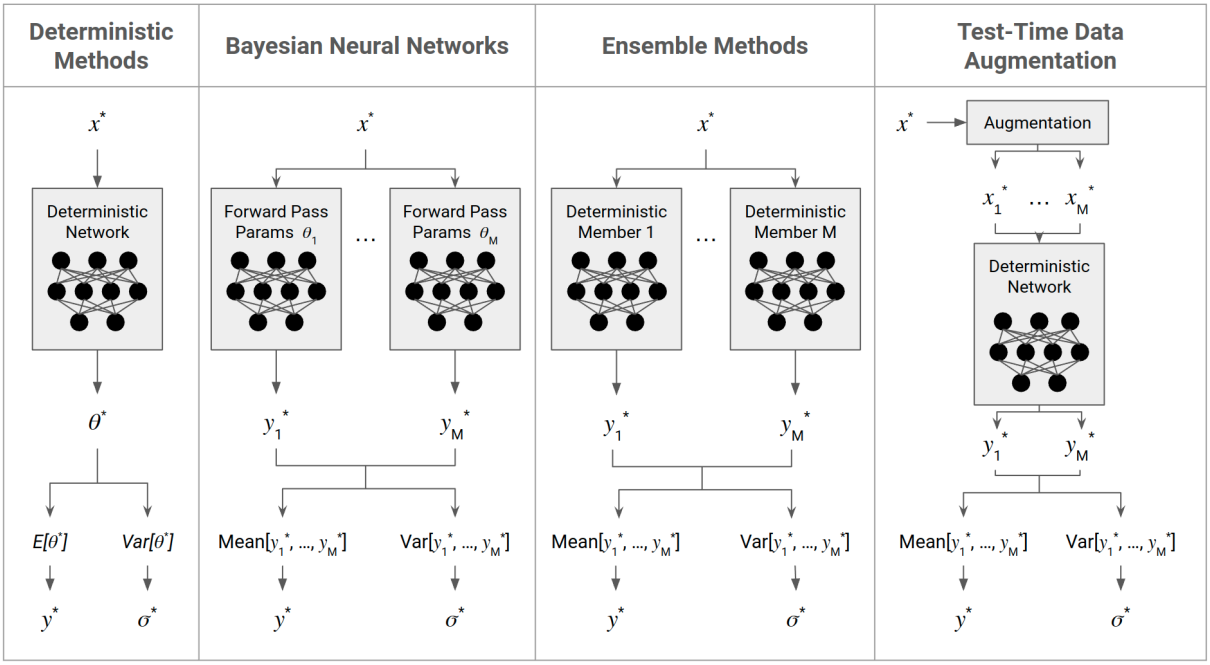


Fig. 2. A visualization of the basic principles of uncertainty modelling of the four general types of uncertainty estimation using neural networks. For a given input sample x^* each approach delivers a prediction y^*, and a representation of the uncertainty σ. The mean and the standard deviation are only used to keep the visualization simple. In practice, other methods could be utilized. Image adapted from [67].

components to give an uncertainty estimate on the prediction of a network [69, 70]. Compared to many other principles, single deterministic methods are computationally efficient in training and evaluation, however, its main disadvantage is the fact that they rely on a single opinion and can therefore become very sensitive to the underlying network architecture, training procedure, and the training data.

- *Bayesian methods* cover all kinds of stochastic DNNs, where several forward passes of the same sample generally lead to different results. Bayesian Neural Networks (BNNs) have the ability to combine the scalability, expressiveness and predictive performance of neural networks. This is achieved by inferring the probability distribution over the network parameters. However, bayesian inference techniques have been proven to be demanding as the size of the data and the number of parameters are too large for the use-cases of DNNs, and thus, approximation techniques are therefore typically applied. One of the successes in bayesian methods has been accomplished by approximating the posterior distribution by optimizing over a family of tractable distributions - by casting existing stochastic elements of deep learning as variational inference.
- *Ensemble methods* combine predictions of several dissimilar deterministic networks at inference. They target a better generalization by making use of synergy effects among the different models, arguing that a group of decision-makers tend to make better decisions than a single decision maker [71, 72]. Besides the improvement in the accuracy, ensembles give an intuitive way of representing the model uncertainty on a prediction by evaluating the variety among the member's predictions.
- *Test-time augmentation methods* give the prediction based on a single deterministic network but augment the input data at test-time in order to generate several predictions that are used to evaluate the certainty of the prediction. The basic method is to create multiple test samples from each test sample by applying data augmentation techniques on it and then test all those samples to compute a predictive distribution in order to measure uncertainty. Mostly, this technique has been used in medical image processing [73, 74] since that field already makes heavy use of data augmentations while using deep learning [75].

3) Uncertainty Measures in Classification Tasks: In order to evaluate the approaches to estimating uncertainty, measures have to be applied to the derived values since the correctness and trustworthiness of these estimations are not automatically given. As exposed in [67] by Gawlikowski *et al.*, there are several reasons why evaluating the quality of uncertainty estimates is a challenging task. First, the quality of the estimation depends on the underlying method as demonstrated by Yao *et al.* [76] which shows that different approximates of bayesian inference result in different qualities. Second, there is a lack of ground truth uncertainty estimates. Its measurement using human annotators would become a very subjective task. For instance, if we defined ground truth estimates across several human subjects we would not know the number of samples required or how to

choose the subjects in order to create a truthful set. Third, there is a lack of a unified quantitative evaluation metric as the uncertainty is defined differently in different machine learning tasks.

When assessing uncertainty, in this paper, we classify the metrics into two different types. The individual metrics evaluate the uncertainty separately for each sample and the system metrics which evaluate the performance of a model, pipeline, method or the usefulness of an individual metric as one. With these types of metrics, the user can choose the most suitable individual metric, architecture and estimation method for a specific task.

Some of the most used individual metrics are the predictive variance which evaluates the variance on the softmax outputs [74]; the predictive entropy, which measures the average level of information or uncertainty inherent in the possible outcomes; the Bhattacharyya Coefficient [77] which assesses the distance between the top-2 classes; the mutual information which evaluates the dependence between the softmax output and the expected information in the softmax output [78] and the Kullback-Leiber divergence which measures the deviation among the possible softmax outputs.

Regarding the system metrics, Milanés-Hermosilla *et al.* [79] proposed a novelty metric based on the Bhattacharyya distance to compare the ability of individual metrics for discriminating between correct and incorrect classified predictions. Asgharnezhad *et al.* [80] proposed a sensitivity (USen), specificity (USpec), precision (UPrec) and accuracy (UAcc) using an uncertainty threshold to measure the model's ability to estimate the doubtfulness analogously to the ordinary machine learning metrics. Brando *et al.* [81], although it was proposed to measure uncertainty models in temporal data series, proposed a metric to analogise different methods by comparing the ordering of the uncertainty values.

III. Method

A. Image Classification

CNNs are widely used to classify images between different classes and they are one of the best methods when the amount of data is limited. In this thesis, we use a wide range of architectures to classify marine species:

- *VGG* [82] is a very deep CNN proposed by Simonyan and Zisserman achieving top performance in 2014.
- *ResNet* proposed by He *et al.* [83], is a set of CNNs with novel identity connections between layers being able to create deeper models without vanishing gradients.
- *EfficinetNet* [84] is a set of architectures that not only provided better accuracy but also improved the efficiency of the models by reducing the parameters with a compound scaling method.
- *EfficientNetV2* [85] improves its previous version by using progressive learning and a more dynamic scaling approach.
- *ConvNext* [86] is a family of pure CNNs which compete favourably with Transformers in terms of accuracy and scalability while maintaining the simplicity and efficiency of standard CNNs.

The reason for using such a diverse number of CNN is to evaluate the performance of estimating uncertainty in different types of architectures and choose better hyperparameters.

B. Uncertainty Estimation

The dropout technique is commonly used to reduce the model complexity and also avoid overfitting [87]. A dropout layer multiplies the output of each neuron by a binary mask that is drawn following a Bernoulli distribution, randomly setting some neurons to zero in the neural network, during the training time. However, Gal and Ghahramani [88] demonstrated that dropout could be used at test time as an approximation of probabilistic Bayesian models in deep Gaussian processes. Monte Carlo dropout (MC dropout) quantifies the uncertainty of network outputs from its predictive distribution by sampling T new dropout masks for each forward pass. As a result, instead of one output model, T model outputs for each input sample are obtained. Then, the set of different predictions $p_t(x)$ can be interpreted as samples from the predictive distribution, which is useful to extract information regarding the prediction's dispersion. The main drawback of MC dropout is its computational complexity at test time, which can be proportional to the number of forward passes T.

C. Uncertainty Evaluation

Due to the lack of a unified protocol and metrics for evaluating uncertainty estimations, it is difficult to compare the various available methods. To solve that, we propose some techniques to evaluate uncertainty estimation methods. The main problems prevail in that the definition of uncertainty is not trivial as it can be defined in many ways and its performances are sometimes data-dependent. For instance, for measuring uncertainty in a single image in this thesis, defined as individual metrics, we use the predictive standard deviation, the predictive entropy and the Bhattacharyya Coefficient. However, if we want to know the method that fits best both the model and dataset, the need for new evaluation methods arises. An individual metric is considered to be good if it is correlated with the model behaviour *i.e.* the incorrect images have generally more uncertainty than the correct ones. To measure that, in this Section we present the metrics used.

For an image x we obtain T different predictions $p_t(x)$ by the MC dropout method, where each prediction is a vector of softmax scores for the C classes. We compute the average prediction score for the T samples as:

$$\bar{p}_T(x) = \frac{1}{T}\sum_{t=1}^{T} p_t(x) \tag{1}$$

Regarding the individual metrics we use the predictive standard deviation, the predictive entropy and the Bhattacharyya Coefficient.

- *Predictive standard deviation* measures the amount of variation or dispersion of the T predictions for each class [88]. A low standard deviation indicates that the values tend to be close to the mean and the prediction is certain, while a high standard deviation indicates that the values are spread out and the model is doubtful about the outcome. However, the main drawback is that both the predictive variance and standard deviation typically have a small value range, which is hard to interpret. The standard deviation is the square root of the average of the squared deviations from the mean:

$$\sigma(x) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(p_t(x) - \bar{p}_T(x))^2} \tag{2}$$

- *Predictive entropy* is a measure interpreted as the average level of information inherent in the possible outcomes of a random variable [74]. The lower the entropy the more certain the model is about the prediction. It is computed as follows:

$$H(x) = -\sum_{c=1}^{C} \bar{p}_T(x)[c]\log(\bar{p}_T(x)[c]) \tag{3}$$

- *Bhattacharyya Coefficient (BC)*, proposed by Molle *et al.* [77], is a measure of the amount of overlap between two statistical samples or populations, the two softmax distributions with the highest mean. In the first step, only those distributions d_1 and d_2 for the top-2 classes are selected, having the highest and the second highest mean. In a second step, construct histograms h_1 and h_2 for d_1 and d_2, respectively, both with N bins. Finally, the BC is obtained by measuring the overlap between these distributions with the normalized Bhattacharyya distance, given by:

$$BC(h_1, h_2) = \frac{1}{N}\sum_{i=1}^{N}\sqrt{h_1(i)h_2(i)} \tag{4}$$

Concerning the system metrics, we use: correct/incorrect intersection (C∩I) and the Uncertainty Ordering Curve (UOC) and its corresponding area under the curve (AUOC) and normalized area under the curve (NAUOC).

- *Correct/Incorrect intersection (C∩I)* is a measure inspired by the Bhattacharyya distance proposed in [79]. The prediction's uncertainty can be intuitively expected to be correlated with the classification performance *i.e.* miss-classified samples are expected to have more uncertainty than correctly predicted samples. Taking the expected behaviour into consideration and to ease the visualization we can plot boxplots and histograms of all the samples' individual metrics, spilt into predictions classified by correct and incorrect samples. Ideally, the histogram of correct samples should have lower uncertainty than the incorrect ones and they should not intersect *i.e.* the two histograms should be separated from each other with the lowest uncertainty being the histogram of correct predictions. In line with this idea, the C∩I measures the implementation of an individual metric, method or model on correlating uncertainty with the overall performance. The intersection of two histograms of N bins is defined as follows:

$$C{\cap}I(h_c, h_i) = \frac{1}{N}\sum_{i=1}^{N} min(h_c(i), h_i(i)) \tag{5}$$

- *Uncertainty Ordering Curve (UOC)* is a novelty approach to comparing the ordering quality of different individual metrics, methods and architectures. UOC is a monotone increasing curve created by plotting the accuracy against the percentage of supervised samples sorted by uncertainty in descending order. It follows a similar idea of C∩I since, ideally, the misclassified samples should have more uncertainty than the correct predictions and, thus, if we ranked them, the incorrect images should all be grouped at the beginning. The starting point of UOC is the model's accuracy, and the curve starts to increase as the number of samples are being supervised and corrected if they are incorrect. If the current prediction is incorrect, the curve will rise, while if it is correct, the accuracy will remain the same. For that reason, a good method has a large rise in the first steps of the curve, but then once the set has reached the top accuracy, the curve remains flat. We propose this new approach as a feature to compare different individual metrics, methods and models. However, while the UOC is a useful visual tool, it is more convenient to have a scalar summary statistic. That is why we

propose measuring the area under the UOC both unnormalized and normalized concerning the ideal curve to be able to compare the more easily the outcomes.

- *Areas Under the Uncertainty Ordering Curve (AUOC and NAUOC)* is a scalar performance metric that measures the correctness of the uncertainty ranking *i.e.* the correctness of the UOC. As the starting point of the UOC is the model accuracy, the simple area under the UOC (AUOC) measures a mixture between the model performance and the ability to estimate uncertainty which is suitable when the goal is to create the best possible system for a fixed dataset. However, when the objective is only to analyse the ability to estimate uncertainty or the individual metrics regardless of the architecture performance, the Normalized Area Under the UOC (NAUOC) is more suitable. The NAUOC is computed as AUOC but normalized concerning the ideal AUOC, where the ordering is perfect *i.e.* steadily increasing curve until it reaches the maximum precision.

D. Uncertainty Driven Classification Protocol

The vast majority of challenging scientific data analysis problems require a huge amount of labelled data which can be very expensive and time-consuming. To assist with the tedious task of annotating tones of images, we propose a protocol to ease the creation of a classification system that automatically annotates the most certain samples and allows the agent only having to supervise the most uncertain ones. The goal of the proposed Uncertainty Driven Classification protocol is to annotate large sets of data minimizing the human labelling while still achieving a good dataset accuracy.

For a fair evaluation of the system, we propose dividing the dataset in three disjoint splits, as in classic machine learning problems. The validation and test sets require to have all the images manually annotated which we propose to be around 10% selected randomly. The validation set is used for hyperparameter search, model selection and ablative studies, while the test set is used to evaluate the operation of those parameters. The training set only contains a reduced number of the rest of images which are annotated to train the model.

First, we manually annotate and use a reduced set of samples N_{RT} to train an initial deep learning model capable of estimating uncertainty. Then, the model is used to automatically obtain the labels and uncertainty scores for the validation data. We assess the accuracy, and we find an uncertainty threshold t_{un} by computing the number of samples that must be corrected N_C to reach a certain desired accuracy A_{un} (95% in our experiments). We conjecture that ranking the samples according to their degree of uncertainty, and starting the manual corrections on the most uncertain ones, will reach the A_{un} with less annotation effort. The number of initial training samples N_{RT} also influences the quality of the labels and uncertainty scores, so we experimentally explore different values for N_{RT}. Then, once the threshold t_{un} has been found, the model is used to automatically obtain the labels of the non-annotated images, allowing the agent to only supervise the samples with a higher uncertainty value than the threshold t_{un}.

As exposed in Section I, evaluating the quality of uncertainty estimations is not trivial and, we propose using the UOC for finding the best threshold t_{un}. The UOC allows us to measure the optimal threshold on the number of samples that need to be supervised to reach a certain accuracy A_{un}. We propose that a good A_{un} is 95% as some of the most widely used databases can have up to 5% miss-labellings and still achieve good results. In Figure 3 we present a flowchart of the Uncertainty Driven Classification protocol.

Uncertainty is mainly used for ranking the dataset samples by doubtfulness and, therefore, the agent has only to supervise the ones with higher uncertainty. The advantage of the Uncertainty Driven protocol is that it ensures that at the end the labels of the dataset will have the stipulated accuracy A_{un}. While, if for instance, a model was trained with a reduced number of samples and used for annotating the remaining images without using the protocol, the accuracy of the labels would be approximately the same as the model accuracy, which would be low due to the number of images used. Furthermore, if instead of using uncertainty to rank the predictions, it was done randomly, we would need to annotate many more images to obtain the desired accuracy A_{un}. That is why uncertainty is a good proxy for detecting potential misclassifications and the protocol helps on reducing the workload involved in labelling large datasets.

IV. Experiments

A. Dataset

The Institut de Ciències del Mar (ICM) is the fourth largest research institute of the Spanish National Research Council (CSIC) and the largest dedicated to marine research. The ICM conducts frontier research with one of the scopes being underwater biodiversity. In some Spanish regions, the aquatic animals and vegetation continue to disappear due to the human impact and the ICM tries to monitor and track how the species and their population evolve. However, as exposed in Section I, monitoring and tracking underwater species is much harder than doing it with terrestrial species. For that reason, the ICM used the Liropus 2000 (Super Mohawk II), a ROV capable of operating at depths of more than two thousand meters equipped with a high-definition compact colour inspection camera (Kongsberg OE14-502A).

During the expeditions, in September 2020, the ROV collected videos from three different Mediterranean regions (Figure 5). These videos were collected by lowering the ROV to the seabed at different depths and moving at a continuous speed. Video

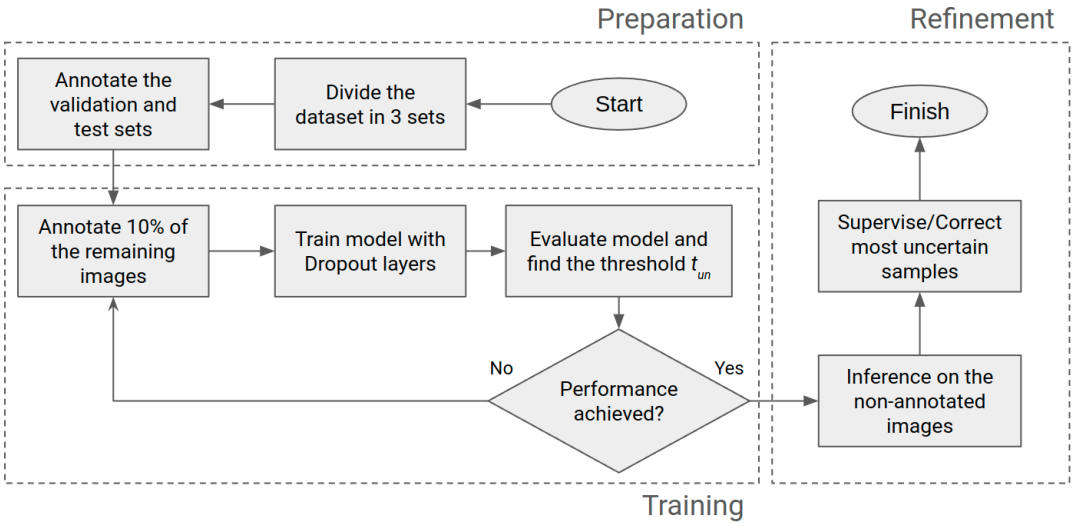


Fig. 3. Flowchart of the Uncertainty Driven Classification protocol. It is based on three different steps, preparation, training and refinement. The preparation step consists mainly to ready the protocol and annotating the validation and test sets. The following stage is about training and evaluating the model for several periods until the desired performance is achieved. Finally, the last step consists in annotating the required number of images sorted by uncertainty.

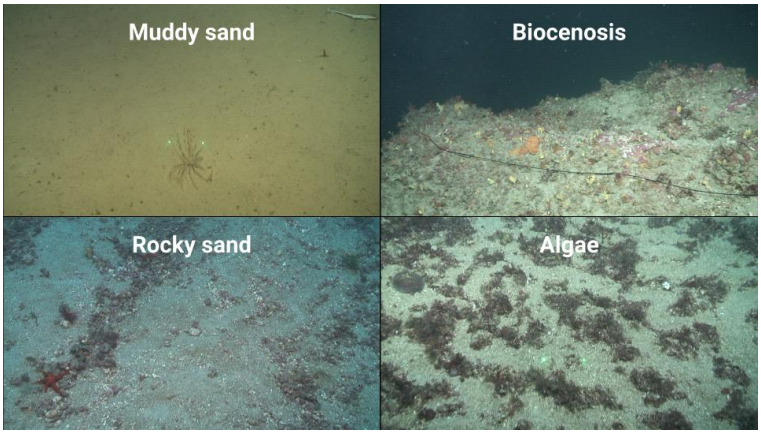


Fig. 4. Image samples of 4 different seabeds and habitats with different marine species. In the muddy sand seabed, a *scyliorhinus canicula* in the upper-right can be seen. In the biocenosis, a *bonellia viridis* in the lower part of the image can be seen. In the rocky sand habitat, a *echinaster sepositus* in the lower left can be seen. In the algae seabed, a *spatangus purpureus* on the left-hand side can be seen.

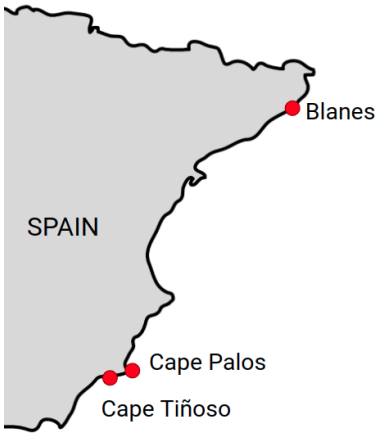


Fig. 5. Mediterranean locations in which the ROV collected the videos during September of 2020 expeditions.

recordings were carried out during daylight hours and in relatively low turbidity periods. The video clips were captured in full HD resolution. In total, the contents exceed 67 hours, the 13 thousand annotations and the 2 hundred species. Examples of these video frames and the different habitats can be seen in Figure 4.

Yet the characteristics of the dataset make it suitable as a machine learning benchmark, its original purpose was to evaluate the environmental status of the most emblematic Spanish habitats and at the same time the most impacted by fisheries. That is why, instead of annotating the images precisely or with the exact position of the creatures, the specialists decided to only annotate an approximate timestamp of every event. This entails the need to create a robust pipeline against inaccurate annotations over time and miss-labellings.

As can be seen in the Annex A, a huge number of different species were sighted during the expeditions. However, some of them are extremely difficult to see such as shrimps, crabs or gobies due to their size, colour, transparencies, occlusions and speed. For this reason and in addition to the fact that some species only appear in a small number of videos, a first version of the dataset is created. The ICM-20.1 contains 6 different marine species. Nevertheless, as the labels are not precise enough, 5 images are taken for each annotation - 0,5 and 1 second before and after the set timestamp and the actual frame. This allows more flexibility when evaluating the system and less probability of miss-classification since for each annotated species we have a small window of video compressed into 5 images. In Table I the characteristics of the dataset can be seen.

To create a suitable machine learning benchmark, we have divided the annotations into three different sets as can be seen in Table II. In addition, to identify when there is no creature on screen, we added an extra class in which it does not appear

TABLE II
ICM-20.1 DATASET DISTRIBUTION

Species Name	Image	Train Set	Validation Set	Test Set
Spatangus purpureus		858	237	213
Echinaster sepositus		688	243	262
Cerianthus membranaceus		235	92	104
Bonellia viridis		159	62	63
Scyliorhinus canicula		103	43	38
Ophiura ophiura		94	31	27
Background	-	678	256	256
Total	-	2905	962	962

any animal. The split has been done in a specific way to ensure that annotations very close in time are not used in the same set. If two creatures appear within 30 seconds, they are assigned to the same set, if not, they can be assigned to different ones. In doing so, we avoid training and evaluating the model with images which could be very similar. As far as the *Background* images are concerned, we took one sample between annotations which are more than 30 seconds apart in time, ensuring that there is no creature.

B. Marine Species Classification

Experiments were designed to evaluate the effectiveness of the proposed networks over the ICM-20.1 dataset following the protocol proposed in this thesis in Section III-D. The data splits used are stated in Section IV-A and can be seen in Table II. However, the ICM-20.1 contains 5 images per annotation, which is treated differently for training and evaluation. When training the CNN architectures, each image, regardless of whether they belong to the same annotation or not, is treated isolately *i.e.* as a sort of data augmentation over time. While in evaluation, only one prediction is taken for each annotation or set of 5 images. The label with the maximum mean value over the 5 images is taken as the prediction.

We trained the networks with the pre-trained weights of Imagenet, cross-entropy loss, Adam optimizer [89] and a learning rate found by a using the validation set. The best weights were chosen by the best F1-Score over all the epochs. The training data was augmented with simple rotations, translations and colour transformations using Albumentations [90].

Furthermore, we used CAMs [14] to obtain qualitative results and to inspect which parts of the image have contributed more to the final output. This is done mainly to ensure that the model is not using any shortcut to classify the species since some of them could only live in specific seabeds or habitats and we want to create a robust system able to classify the creatures in any environment.

In Table III the test accuracies and F1-Scores of all CNN architectures are presented, while in Figure 8 the confusion matrix of the best performing model, ConvNeXt-L is shown.

TABLE III
MARINE SPECIES CLASSIFICATION RESULTS

	Accuracy	F1-Score	Parameters (M)
VGG16	84.1	83.4	138
ResNet18	81.5	80.4	11
ResNet101	81.5	80.4	42
EfficientNet-B0	83.7	82.7	5.3
EfficientNet-B4	84.2	83.8	19
EfficientNet-B7	84.4	83.8	66
EfficientNetV2-S	84.4	84.8	22
EfficientNetV2-L	84.7	84.3	120
ConvNeXt-T	86.2	86.3	29
ConvNeXt-L	**87.0**	**86.7**	198

TABLE IV
UNCERTAINTY METRICS RESULTS

	STD			Entropy			BC		
	C∩I	AUOC	NAUOC	C∩I	AUOC	NAUOC	C∩I	AUOC	NAUOC
VGG16	0.7847	0.9454	0.9595	0.4533	0.9561	0.9703	**0.9932**	**0.9360**	**0.9500**
ResNet18	**0.6029**	0.9368	0.9502	0.5141	0.9394	0.9607	0.9972	0.9180	0.9389
ResNet101	0.7773	0.9352	0.9533	0.4569	0.9475	0.9658	0.9987	0.9137	0.9313
EfficientNet-B0	0.8237	0.9467	0.9622	0.4391	0.9552	0.9708	1.0	0.9265	0.9417
EfficientNet-B4	0.8802	**0.9548**	**0.9689**	0.4422	0.9563	0.9704	1.0	0.9264	0.9401
EfficientNet-B7	0.7872	0.9449	0.9593	0.4329	0.9561	0.9708	1.0	0.9263	0.9404
EfficientNetV2-S	1.0	0.9243	0.9381	0.3927	0.9591	0.9734	1.0	0.9253	0.9391
EfficientNetV2-L	1.0	0.9215	0.9349	0.4250	0.9572	0.9711	1.0	0.9214	0.9348
ConvNeXt-T	0.9872	0.9278	0.9386	0.3847	0.9660	0.9775	1.0	0.9306	0.9440
ConvNeXt-L	0.9814	0.9286	0.9394	**0.3366**	**0.9700**	**0.9812**	1.0	0.9342	0.9450
Average	0.8656	0.9373	0.9512	**0.4280**	**0.9563**	**0.9712**	0.9989	0.9258	0.9405

C. Uncertainty Evaluation

The experiments taken in this Section aim to demonstrate that the metrics proposed in Section III-C are correlated with the classification performance.

According to [88], to estimate the uncertainties using MC dropout, $T = 50$ is considered a safe choice. However, among all the architectures used, the ResNets do not have any dropout layer and, thus, at inference time a dropout layer is added in the classifier block following the structure of the other models. For all other CNNs, the dropout layers used are the ones by default. All the dropout rates used are $p = 0.2$.

We compute all the system metrics, C∩I, AUOC and NAUOC, with the different individual metrics, predictive standard deviation, predictive entropy and BC. The number of bins N of the BC histograms has been found by grid search on the validation set. The results are presented in Table IV.

Furthermore, to demonstrate that uncertainty is a good proxy of the generalization error because miss-classified samples are expected to have higher uncertainty than correct samples, we plot the boxplots and histograms classified by incorrect and correct images. As it is more convenient to have a scalar statistic, we also measure the C∩I. The charts and the metrics corresponding to the best performing model both in accuracy and uncertainty estimation, ConvNeXt-L, can be seen in Figure 6.

We also plotted the UOC curve of the ConvNeXt-L, with both predictive entropy and predictive standard deviation, which is shown in Figure 7.

D. Uncertainty Driven Classification Protocol

To demonstrate the usability of the Uncertainty Driven Classification protocol, we test an implementation with the best performing network, ConvNeXt-L. The goal of the protocol is to achieve the desired accuracy A_{un} with the less amount of manually annotated data N_{RT} and supervision N_C. As exposed in Section III-D and Figure 3, once the dataset has been divided, the first step is to annotate the validation and test sets for both hyperparameter tuning and evaluation. Then, the training set is created, in the first iteration with the 10% of the remaining images.

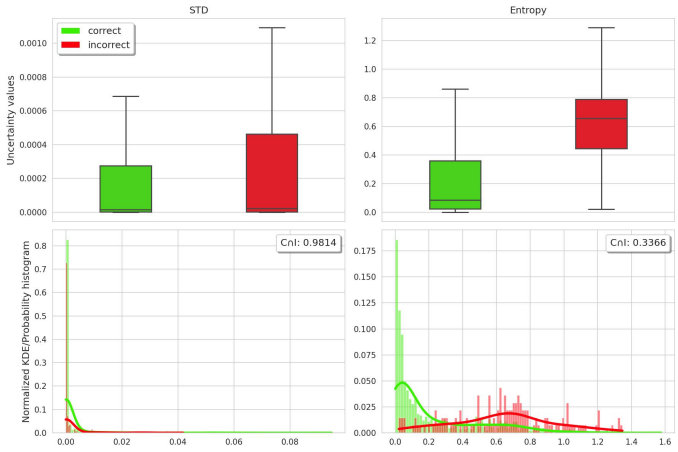


Fig. 6. ConvNeXt-L predictive standard deviation and predictive entropy box plots and histograms classified by correct and incorrect samples. In the first row we plot the boxplots showing the interquartile range and ignoring the outliers. In the second row we plot the histograms, as well as, the kernel distribution estimation. The C∩I is also computed.

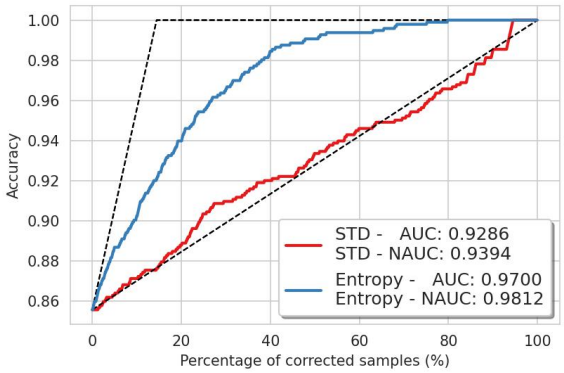


Fig. 7. ConvNeXt-L predictive standard deviation and predictive entropy UOC. The dashed lines correspond to the ideal and random ordering.

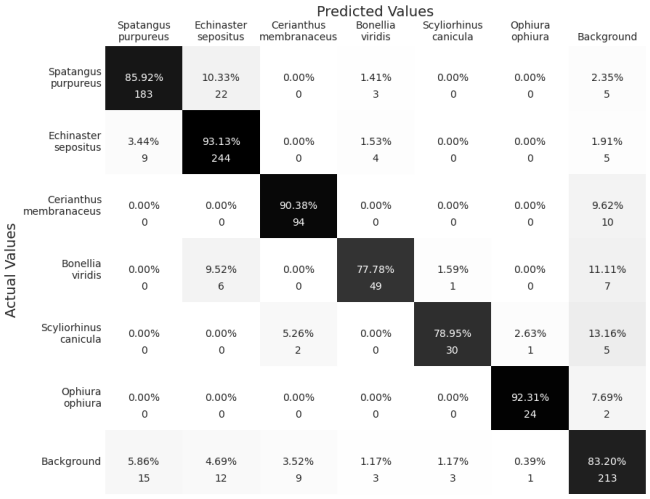


Fig. 8. ConvNeXt-L confusion matrix.

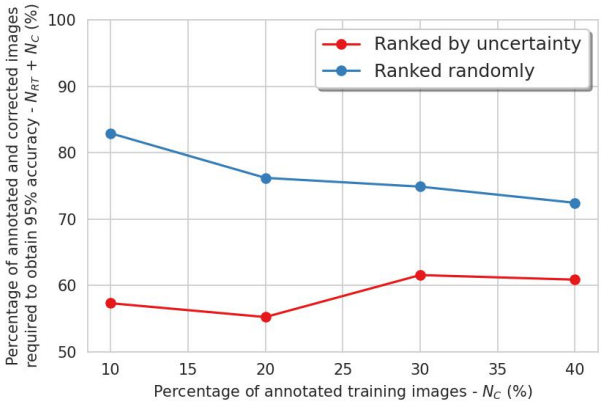


Fig. 9. Percentage of the amount of images that have to be manually annotated or supervised (N_{RT} + N_C) against the percentage of training images N_{RT} to obtain a desired accuracy A_{un} of 95%. The red line is by using the proposed protocol, using uncertainty to rank the predictions, while the blue line is by selecting the samples to supervise randomly.

After the creation of the training set, we train the model finding the best model hyperparameters with the validation set. Then, the threshold t_{un} is found to know the number of samples that have to be corrected N_C to obtain the desired accuracy A_{un}. In addition, the model is used to predict the labels of the test set and by the use of UOC we can observe if the threshold t_{un} has the appropriate behaviour.

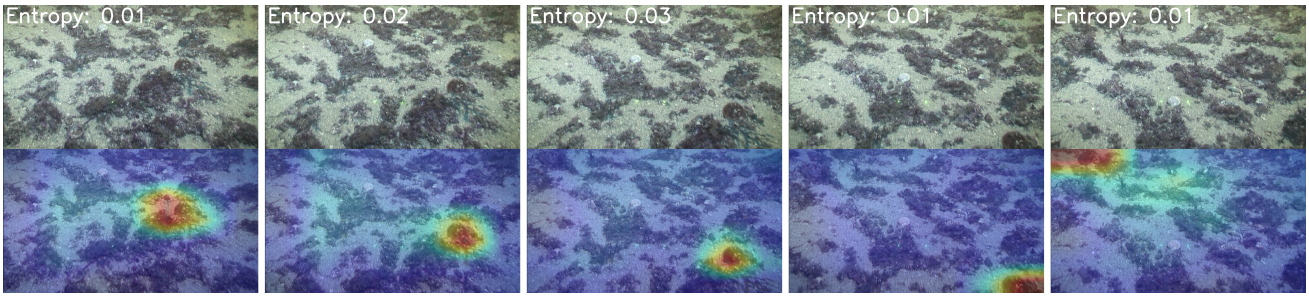
In the event that the system performance is not as expected, we can increase the training images N_{RT} by annotating more samples. That is why, in this experiment, we test different values of training images N_{RT}. In Figure 9 we plot the sum of manually annotated images N_{RT} and number of images to correct N_C against different values of N_{RT} in order to obtain a desired accuracy Aun of 95%. In Figure IV-D we show both using the proposed protocol and choosing the images to be supervised randomly.

V. Results

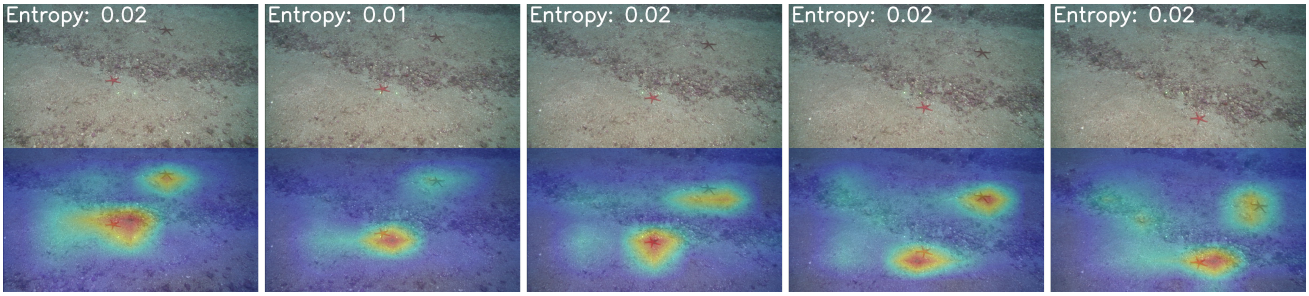
In this Section, we will explain the final results obtained from the proposed set of experiments.

A. Classification Results

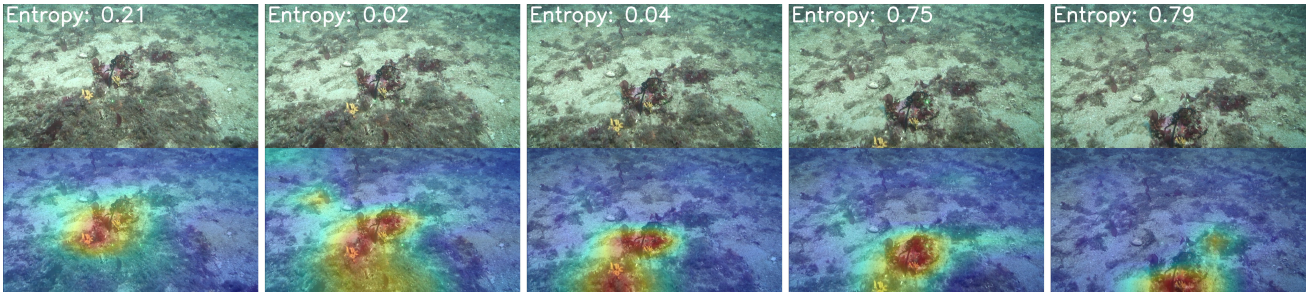
In Table III, we present the final results regarding the marine species classification of the ICM-20.1 dataset. We measured both the accuracy and F1-Score to evaluate the architectures although having an imbalanced dataset. We can observe that,



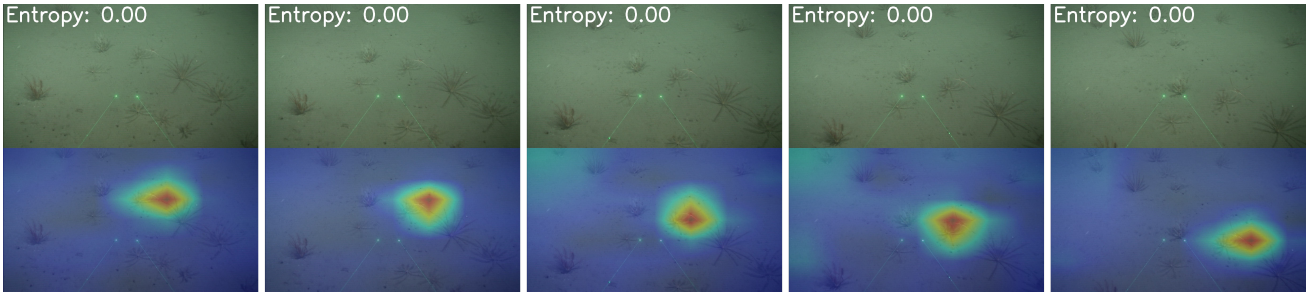
(a) *Spatangus purpureus*



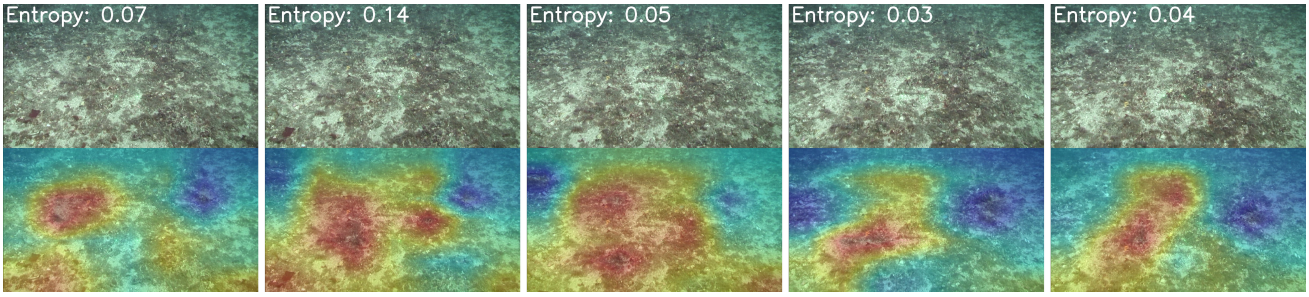
(b) *Echinaster sepositus*



(c) *Bonellia viridis*



(d) *Ophiura ophiura*



(e) Background

Fig. 10. Five examples of five different classes of the operation of the system. Each image has its corresponding CAM underneath and the estimated predictive entropy. This time the uncertainty is measured for each image individually, although for evaluation purposes, only one scalar is taken for each set of 5 images.

as expected, the best-performing model is the ConvNeXt-L with an 87.0% accuracy and 86.7% F1-Score. The ConvNeXt family is the state-of-the-art of neural network since they achieved the best ImageNet top-1 accuracy of all the CNNs so far in 2022. However, the large model has almost 200 million parameters, while the tiny model obtains a similar performance with 6 times fewer parameters which highlights the excellent efficiency of the entire family. The rest of the architectures remain in an acceptable range between 80% and 85% of accuracy whereas the EfficientNet-B0 also stands out for its low number of parameters.

In Figure 8, we show the confusion matrix of the ConvNeXt-L model. As it is known from the goodness of the F1-Score, all the marine species are classified correctly by a good percentage. However, we can observe that the vast majority of misclassifications are predicted as background. This is because some creatures may appear for a small time or be very small in size, causing the network to predict that nothing occurred. But in general, it can be seen that the models can identify and classify all species excellently.

B. Uncertainty Results

In Table IV, we present the results regarding the uncertainty estimations. We computed all the proposed system scalar metrics (C∩I, AUOC and NAUOC) with all the individual metrics (predictive standard deviation, predictive and BC) for all the CNN architectures. We can observe from the average of all the outcomes that the individual metric that correlates best with the classification performance is the predictive entropy. Predictive entropy is best metric to measure the uncertainty from a single image softmax vector since it has by far the lowest C∩I, with an average over all the models of 0.4280. Not only that, but ConvNeXt-L achieves a standing C∩I of 0.3366. Regarding the predictive standard deviation, the main problem is that its values are extremely small due to the similarity between softmax vectors over all the MC samples. From Table IV, we can also observe that the BC is by far the worst performing uncertainty metric. This confirms that, at least in the ICM-20.1 dataset, the distance between the top-2 softmax classes' histograms is not correlated with whether the sample has been correctly predicted or not.

Also in Table IV, we can observe how well the different individual metrics and architectures sort the predictions by uncertainty. The highest AUOC is obtained by the mixture between ConvNeXt-L and predictive entropy, as was expected since this metric measures a combination of model performance and uncertainty estimations. However, it is the same combination that achieves the best results using the NAUOC, too. We can conclude that ConvNeXt-L is the best-performing model and the best architecture for estimating uncertainty.

In Figure 6, we present both the box plots and histograms of the predictive standard deviation and predictive entropy estimated with the ConvNeXt-L. As we already discussed from the scalar metrics, in both plots we can corroborate that the entropy is a better metric than the standard deviation since the correct and incorrect box plots are more distant having a greater average difference. Furthermore, in the histogram plots, we also computed the Kernel Density Estimation (KDE) which allows us to see that the correct distribution has lower variance than the incorrect distribution which implies that the correct samples' are grouped around a low uncertainty value.

In Figure 7, we present the novel Uncertainty Ordering Curve of both the predictive standard deviation and predictive entropy estimated using ConvNeXt-L. In the plot, we display two black dashed lines which correspond to the ideal and random ordering. We can observe that ordering the predictions by the predictive standard deviation is similar to random ordering. While sorting the predictions by the predictive entropy entails a better outcome. This means that measuring the doubtfulness of a model with the predictive entropy allows using the uncertainty as a proxy of the generalization error since the miss-classified samples have greater value, as a general rule. From the entropy UOC, we can also observe that if the agent annotated the 20% most uncertain images and let the model classify the rest, the accuracy would rise to 95% approximately. However, if 20% of images were selected randomly, the outcome accuracy would only be around 88%, as shown by the lower dashed line. This means that thanks to estimating the uncertainty of a model and using the UOC curve to evaluate it, if the agent would annotate 20% of images, the accuracy would be 7% higher than without using uncertainty.

In figure 10 we present five output examples of five different classes. At first sight, it can be seen that the regions of the images which most contribute to the final decision are the ones which contain a creature. In the case of the background image, the activation map is scattered since there is no specific region in which indicates the class to which it belongs. In both the *Spatangus purpureus* and *Ophiura ophiura* examples, it can be seen that although the difficulty of finding the species the model can accomplish its work perfectly, even with low uncertainty. In the *Echinaster sepositus* example, we show that, although appearing two starfish on the frame, the model can both predict them correctly and localize them. However, in the *Bonellia viridis* example, we can observe that model is more uncertain about its prediction as the creature is very thin and is surrounded by biocenosis.

C. Uncertainty Driven Classification Results

In Figure 9, we present the Uncertainty Driven Classification protocol results using different amount of training images N_{RT}. The plot shows percentage of the amount of images that have to be manually annotated or supervised (N_{RT} + N_C) against the percentage of training images N_{RT} to obtain a desired accuracy A_{un} of 95%. We demonstrate that ranking the

predictions using uncertainty reduces the total amount of images that have to be labelled or supervised with respect to choosing the samples to supervise randomly. Uncertainty takes a vital point in the Uncertainty Driven Classification protocol as it reduces 20% approximately the data that has to be annotated manually. Ranking the predictions by uncertainty is most effective when the model is trained on a reduced number of training data N_{RT}, as the minimum is obtained at 20%.

This demonstrates the usefulness of estimating uncertainty as the system has the ability of distinguishing the most error-prone samples and by the use of the UOC and the Uncertainty Driven Classification protocol, the arduous task of annotating a dataset is mitigated by reducing the number of manually annotated images required. Furthermore, estimating uncertainty and ranking the predictions accordingly could be crucial in a lot of applications such as finding out-of-distribution samples, recognizing potential misclassifications in dangerous applications or even the Uncertainty Driven Classification protocol.

VI. Conclusions

During this thesis we have shown that we have achieved the objectives defined in Section I. We have presented a new useful dataset containing six different marine species taken in three different Mediterranean locations using a ROV. The seabed and species variability enable the ICM-20.1 dataset to work as an underwater machine learning benchmark. Since it is a novel dataset, we have used a wide variety of CNNs to verify the proper functioning of the benchmark. Although the main objective of this thesis was not to obtain the highest performance on the dataset, we have produced fairly good results. The ConvNeXt family has been the best-performing architecture obtaining an accuracy of 87.0% and an F1-Score of 86.7%. Furthermore, by using CAMs we have verified which parts of the image have contributed more to the final output demonstrating that the model can localize the distinct species correctly.

Regarding uncertainty, we have shown that it can be used as a proxy of the generalization error since when choosing the most suitable measures and architectures, the uncertainty is highly correlated with the model performance. However, the main challenge remained in finding the best uncertainty definition or metrics as it is not trivial, as well as, its evaluation. To solve that problem, in this thesis, we have tested a well-known set of individual metrics which measure the dispersion of a distribution, the predictive standard deviation, the predictive entropy and the BC. In order to evaluate the performances of the individual metrics, we proposed two metrics the C$\cap$I and the novel UOC. We have demonstrated that these metrics are useful to measure the correlation between uncertainty and the accuracy of the predictions and their ordering. Furthermore, the areas under the UOC allow for the comparison of several methods at once by a mixture between both model and uncertainty estimation performances which is useful when what is desired is to obtain the best performance possible in a certain dataset using uncertainty estimations. Nevertheless, the NAUOC allows measuring solely the uncertainty estimation performance as it is normalized concerning the ideal ordering.

In the introduction of this thesis, we introduced one of the main problems of monitoring not only underwater ecosystems but also other challenging scientific data analysis problems: data annotation. To solve that, we have proposed a protocol to reduce the workload of labelling huge amounts of images by the use of uncertainty and its evaluation. The Uncertainty Driven Classification protocol uses a reduced portion of labelled samples to train a CNN and, then, uses the trained model with MC dropout to report, in addition to the decision, an associated estimate of the uncertainty. The validation set is used for hyperparameter tuning, model selection, ablative studies and finding the threshold which delivers the desired model performance, while the test set is used to evaluate the uncertainty estimations and both the hyperparameters and threshold reliability. Once the uncertainty threshold found on the validation set has been verified on the test set, we know the exact amount of annotated data required to obtain the desired accuracy. In Figure 9, we showed that labelling the images with the highest uncertainty reduces the workload with respect to doing it randomly. Furthermore, the UOC plot enables us to visualise the quality of the uncertainty estimations to rank and minimize the human correction effort.

Finally, besides the good results we have obtained, much remains to be done in future work. First of all, ICM-20.1 is a very good underwater benchmark, but the dataset could have more species which should be tested as there are a lot of them that only appear in just a few video sequences. Secondly, all the experiments have been tested in only one benchmark which would be more reliable if they had been tested not only in underwater datasets but also in other types of data analysis problems. Another enhancement of the presented work could be in evaluating the usefulness of the uncertainty to detect out-of-distributions samples. In the specific scenario of the ICM-20.1 dataset, we plan to evaluate how marine species that are not labelled behave when estimating their uncertainty. Regarding the uncertainty metrics, we suspect that the low performance of the predictive standard deviation method in uncertainty estimation is due to the fact that all the estimations have similar and small values. We plan to explore calibration methods to improve these results. Finally, although it was not a requirement for the animal counting application, we plan to extend the system to obtain the exact bounding box of the animals using the same annotations from the Class Activation Maps. This can be done by finding a threshold over all the heatmaps.

Acknowledgment

I would like to thank my thesis advisors David Masip Rodó and José Antonio Garcia del Arco for allowing me to develop this project, trusting me from day one and for their help in everything I have needed. Also our acknowledgments to CRIMA project (RTI2018-095770-B-100), which is funded by MCIU/AEI/FEDER, EU for all videos and biological data provided. Finally, I would like to thank my parents, my friends and my girlfriend for their support during this long and arduous journey.

APPENDIX A
MARINE SPECIES ANNOTATIONS

TABLE V
TOP 62 SPECIES WITH THE HIGHEST OCCURRENCE

Marine Species	Total	ROV04	ROV05	ROV06	ROV07	ROV08	ROV18	ROV19	ROV20	ROV22	ROV23	ROV24
Spatangus purpureus	1308	45	27	2	1	0	946	287	0	0	0	0
Echinaster sepositus	1191	297	159	40	0	0	325	262	108	0	0	0
Crab	465	1	2	0	13	3	13	5	0	60	341	27
Cerianthus membranaceus	431	2	3	1	1	0	5	0	4	194	57	164
Polychaeta	400	4	9	0	16	2	20	7	1	110	146	85
Callionymus sp	394	0	0	0	174	18	4	0	3	82	102	11
Gobiidae	377	5	0	0	77	1	51	40	0	81	71	51
Bonellia viridis	284	55	7	0	3	4	121	70	15	7	2	0
Plesionika sp	277	0	0	0	0	0	0	0	0	70	0	207
Hermit	267	16	14	2	3	1	66	34	13	28	13	77
Sponge	253	31	3	0	0	0	103	99	15	1	0	1
Flatfish	247	3	0	0	55	5	1	1	0	100	57	25
Sphaerechinus granularis	207	0	0	0	0	0	139	64	4	0	0	0
Serranus hepatus	184	2	15	0	8	2	6	17	2	92	40	0
Scyliorhinus canicula	184	0	0	0	14	4	0	0	0	86	6	74
Ophiura ophiura	151	0	0	0	64	74	0	0	0	13	0	0
Serranus cabrilla	137	8	1	0	4	1	43	61	7	10	2	0
Chaetaster longipes	112	9	4	2	0	0	67	11	19	0	0	0
Halocynthia papillosa	108	34	12	5	0	0	27	28	2	0	0	0
Helicolenus dactylopterus	97	0	0	0	4	0	0	0	0	40	0	53
Capros aper	74	0	0	0	37	5	0	0	0	17	0	15
Nemertesia antennina	71	0	0	0	0	0	30	37	4	0	0	0
Cidaris cidaris	62	1	2	0	0	0	48	0	11	0	0	0
Holothuria tubulosa	59	28	24	1	0	0	6	0	0	0	0	0
Posidonia oceanica	59	2	8	0	0	0	1	48	0	0	0	0
Coris julis	52	11	1	0	0	0	26	12	2	0	0	0
Paguridae	50	3	0	0	0	1	0	0	0	12	12	22
Goneplax rhomboides	50	2	0	0	0	0	0	0	0	31	17	0
Chelidonichthys lastoviza	47	7	8	1	0	0	19	9	3	0	0	0
Cereus pedunculatus	43	0	0	0	0	0	34	4	5	0	0	0
Triglidae	42	2	3	1	0	1	6	0	14	2	13	0
Ebalia sp	40	0	0	0	17	23	0	0	0	0	0	0
Serranidae	39	1	0	0	0	0	1	24	0	13	0	0
Hacelia attenuata	37	2	0	0	0	0	3	32	0	0	0	0
Myxicola infundibulum	37	0	0	0	0	0	0	0	0	37	0	0
Ulva sp	36	0	0	0	0	0	19	17	0	0	0	0
Echinus melo	36	2	0	0	14	7	3	3	5	2	0	0
Holothuria forskali	36	20	10	0	0	0	3	0	3	0	0	0
Alcyonium palmatum	36	2	0	0	0	1	0	1	0	4	28	0
Parastichopus regalis	35	7	0	0	8	2	0	1	0	5	0	12
Funiculina quadrangularis	34	0	0	0	16	1	0	1	0	15	1	0
Trachinus draco	34	19	0	0	0	0	0	2	0	12	1	0
Merluccius merluccius	34	0	0	0	1	4	0	0	0	16	4	9
Gastropoda	33	2	0	0	3	0	3	0	0	14	4	7
Anthias anthias	33	5	0	0	0	0	14	7	7	0	0	0
Holothuria sp	29	3	17	0	0	0	2	0	0	6	1	0
Paramuricea clavata	27	15	0	0	0	0	6	3	3	0	0	0
Pennatula rubra	23	0	1	0	0	0	0	0	0	5	17	0
Eutrigla gurnardus	21	0	0	0	0	0	0	0	0	7	2	12
Axinella polypoides	21	9	0	0	0	0	11	0	0	1	0	0
Holosepia	21	0	0	0	0	0	0	0	0	19	0	2
Sphyraena sphyraena	21	0	0	0	21	0	0	0	0	0	0	0
Aplidium sp	20	0	0	0	0	0	16	4	0	0	0	0
Citharus linguatula	19	0	0	0	4	0	0	0	0	4	11	0
Astropecten irregularis	19	1	0	0	9	2	1	0	0	2	4	0
Luidia ciliaris	17	0	0	0	0	0	10	3	4	0	0	0
Alcyonium acaule	16	0	10	0	0	0	0	3	0	3	0	0
Pennatula sp	15	0	0	0	0	0	0	2	0	4	9	0
Eunicella verrucosa	15	0	0	0	0	0	6	9	0	0	0	0
Myriapora truncata	15	1	0	0	0	0	14	0	0	0	0	0
Octopus vulgaris	14	2	0	0	3	0	2	1	0	5	1	0
Phallusia mammillata	14	2	2	0	0	0	5	3	2	0	0	0

REFERENCES

[1] P. C. Reid, A. C. Fischer, E. Lewis-Brown, M. P. Meredith, M. Sparrow, A. J. Andersson, A. Antia, N. R. Bates, U. Bathmann, G. Beaugrand *et al.*, "Impacts of the oceans on climate change," *Advances in marine biology*, vol. 56, pp. 1–150, 2009.

[2] N. P. M. Sevilla, M. Lopez, and A. Hanhausen, "The ocean: Life below water and why it matters," 2020.

[3] K. Sherman and G. McGovern, "Toward recovery and sustainability of the world's large marine ecosystems during cli-mate change," *IUCN, Gland, Switzerland:(i+ 19 pages)*, 2011.

[4] D. Roberts, "Ocean acidification: Connecting science, industry, policy and public," 2011.

[5] R. Cormier and M. Elliott, "Smart marine goals, targets and management–is sdg 14 operational or aspirational, is 'life below water'sinking or swimming?" *Marine pollution bulletin*, vol. 123, no. 1-2, pp. 28–33, 2017.

[6] G. J. Edgar, G. R. Russ, and R. C. Babcock, "Marine protected areas," *Marine ecology*, vol. 27, pp. 533–555, 2007.

[7] D. Laffoley, "Towards networks of marine protected areas: the mpa plan of action for iucn's world commission on protected areas," 2008.

[8] "Protected Planet Marine Protected Areas," https://www.protectedplanet.net/en/thematic-areas/marine-protected-areas, accessed: 2022-09-01.

[9] D. Rosen and A. Lauermann, "It's all about your network: Using rovs to assess marine protected area effectiveness," in *Oceans 2016 Mts/Ieee Monterey*. IEEE, 2016, pp. 1–6.

[10] C. Domínguez Carrió, "Rov-based ecological study and management proposals for the offshore marine protected area of cap de creus (nw mediterranean)," 2018.

[11] L. G. Costanzo, G. Marletta, and G. Alongi, "Assessment of marine litter in the coralligenous habitat of a marine protected area along the ionian coast of sicily (central mediterranean)," *Journal of Marine Science and Engineering*, vol. 8, no. 9, p. 656, 2020.

[12] H. R. Gordon, "Can the lambert-beer law be applied to the diffuse attenuation coefficient of ocean water?" *Limnology and Oceanography*, vol. 34, no. 8, pp. 1389–1409, 1989.

[13] W. Appeltans, S. T. Ahyong, G. Anderson, M. V. Angel, T. Artois, N. Bailly, R. Bamber, A. Barber, I. Bartsch, A. Berta *et al.*, "The magnitude of global marine species diversity," *Current biology*, vol. 22, no. 23, pp. 2189–2202, 2012.

[14] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[15] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[17] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*, vol. 31, 2018.

[18] Y. Gal *et al.*, "Uncertainty in deep learning," 2016.

[19] T. Pearce, A. Brintrup, and J. Zhu, "Understanding softmax confidence and uncertainty," *arXiv preprint arXiv:2106.04972*, 2021.

[20] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[21] R. S. Dhawal and L. Chen, "A copula based method for fish species classification," in *2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. IEEE, 2016, pp. 471–478.

[22] D. Joo, Y.-s. Kwan, J. Song, C. Pinho, J. Hey, and Y.-J. Won, "Identification of cichlid fishes from lake malawi using computer vision," *PloS one*, vol. 8, no. 10, p. e77686, 2013.

[23] D. S. Y. Kartika and D. Herumurti, "Koi fish classification based on hsv color space," in *2016 International Conference on Information & Communication Technology and Systems (ICTS)*. IEEE, 2016, pp. 96–100.

[24] A. Tharwat, A. A. Hemedan, A. E. Hassanien, and T. Gabel, "A biometric-based model for fish species classification," *Fisheries research*, vol. 204, pp. 324–336, 2018.

[25] M.-C. Chuang, J.-N. Hwang, F.-F. Kuo, M.-K. Shan, and K. Williams, "Recognizing live fish species by hierarchical partial classification based on the exponential benefit," in *2014 IEEE international conference on Image Processing (ICIP)*. IEEE, 2014, pp. 5232–5236.

[26] A. Hernández-Serna and L. F. Jiménez-Segura, "Automatic identification of species with neural networks," *PeerJ*, vol. 2, p. e563, 2014.

[27] M. K. Alsmadi, K. B. Omar, S. A. Noah, and I. Almarashdeh, "Fish recognition based on robust features extraction from color texture measurements using back-propagation classifier," *Journal of Theoritical and Applied Information Technology*, vol. 18, no. 1, pp. 11–18, 2010.

[28] J. Hu, D. Li, Q. Duan, Y. Han, G. Chen, and X. Si, "Fish species classification by color, texture and multi-class support vector machine using computer vision," *Computers and electronics in agriculture*, vol. 88, pp. 133–140, 2012.

[29] P. X. Huang, "Hierarchical classification system with reject option for live fish recognition," in *Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data*. Springer, 2016, pp. 141–159.

[30] S. Palazzo, I. Kavasidis, and C. Spampinato, "Covariance based modeling of underwater scenes for fish detection," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 1481–1485.
[31] M. Ravanbakhsh, M. R. Shortis, F. Shafait, A. Mian, E. S. Harvey, and J. W. Seager, "Automated fish detection in underwater images using shape-based level sets," *The Photogrammetric Record*, vol. 30, no. 149, pp. 46–62, 2015.
[32] M.-C. Chuang, J.-N. Hwang, K. Williams, and R. Towler, "Multiple fish tracking via viterbi data association for low-frame-rate underwater camera systems," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2013, pp. 2400–2403.
[33] ——, "Tracking live fish from low-contrast and low-frame-rate stereo videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 1, pp. 167–179, 2014.
[34] M. M. M. Fouad, H. M. Zawbaa, N. El-Bendary, and A. E. Hassanien, "Automatic nile tilapia fish classification approach using machine learning techniques," in *13th international conference on hybrid intelligent systems (HIS 2013)*. IEEE, 2013, pp. 173–178.
[35] C. Spampinato, S. Palazzo, P.-H. Joalland, S. Paris, H. Glotin, K. Blanc, D. Lingrand, and F. Precioso, "Fine-grained object recognition in underwater visual data," *Multimedia Tools and Applications*, vol. 75, no. 3, pp. 1701–1720, 2016.
[36] M.-C. Chuang, J.-N. Hwang, J.-H. Ye, S.-C. Huang, and K. Williams, "Underwater fish tracking for moving cameras based on deformable multiple kernels," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 9, pp. 2467–2477, 2016.
[37] G. Cutter, K. Stierhoff, and J. Zeng, "Automated detection of rockfish in unconstrained underwater videos using haar cascades and a new image dataset: labeled fishes in the wild," in *2015 IEEE Winter Applications and Computer Vision Workshops*. IEEE, 2015, pp. 57–62.
[38] E. Hossain, S. S. Alam, A. A. Ali, and M. A. Amin, "Fish activity tracking and species identification in underwater video," in *2016 5th International conference on informatics, electronics and vision (ICIEV)*. IEEE, 2016, pp. 62–66.
[39] S. A. Siddiqui, A. Salman, M. I. Malik, F. Shafait, A. Mian, M. R. Shortis, and E. S. Harvey, "Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data," *ICES Journal of Marine Science*, vol. 75, no. 1, pp. 374–389, 2018.
[40] X. Sun, J. Shi, J. Dong, and X. Wang, "Fish recognition from low-resolution underwater images," in *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2016, pp. 471–476.
[41] B. Iscimen, Y. Kutlu, A. Uyan, and C. Turan, "Classification of fish species with two dorsal fins using centroid-contour distance," in *2015 23nd Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2015, pp. 1981–1984.
[42] Y. Kutlu, B. Iscimen, and C. Turan, "Multi-stage fish classification system using morphometry," *Fresenius Environmental Bulletin*, vol. 26, no. 3, pp. 1911–1917, 2017.
[43] B. V. Deep and R. Dash, "Underwater fish species recognition using deep learning techniques," in *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 2019, pp. 665–669.
[44] D. Zhang, D.-J. Lee, M. Zhang, B. J. Tippetts, and K. D. Lillywhite, "Object recognition algorithm for the automatic identification and removal of invasive fish," *Biosystems Engineering*, vol. 145, pp. 65–75, 2016.
[45] M. H. Sharif, F. Galip, A. Guler, and S. Uyaver, "A simple approach to count and track underwater fishes from videos," in *2015 18th international conference on computer and information technology (ICCIT)*. Ieee, 2015, pp. 347–352.
[46] E. Lantsova, T. Voitiuk, T. Zudilova, and A. Kaarna, "Using low-quality video sequences for fish detection and tracking," in *2016 SAI Computing Conference (SAI)*. IEEE, 2016, pp. 426–433.
[47] Y.-H. Shiau, C.-C. Chen, and S.-I. Lin, "Using bounding-surrounding boxes method for fish tracking in real world underwater observation," *International Journal of Advanced Robotic Systems*, vol. 10, no. 7, p. 298, 2013.
[48] Y.-H. Hsiao and C.-C. Chen, "A sparse sample collection and representation method using re-weighting and dynamically updating omp for fish tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 3494–3497.
[49] Z. Ye, J. Zhao, Z. Han, S. Zhu, J. Li, H. Lu, and Y. Ruan, "Behavioral characteristics and statistics-based imaging techniques in the assessment and optimization of tilapia feeding in a recirculating aquaculture system," *Transactions of the ASABE*, vol. 59, no. 1, pp. 345–355, 2016.
[50] J. Zhao, Z. Gu, M. Shi, H. Lu, J. Li, M. Shen, Z. Ye, and S. Zhu, "Spatial behavioral characteristics and statistics-based kinetic energy modeling in special behaviors detection of a shoal of fish in a recirculating aquaculture system," *Computers and Electronics in Agriculture*, vol. 127, pp. 271–280, 2016.
[51] X. Li, M. Shang, J. Hao, and Z. Yang, "Accelerating fish detection and recognition by sharing cnns with objectness learning," in *OCEANS 2016-Shanghai*. IEEE, 2016, pp. 1–5.
[52] X. Li, Y. Tang, and T. Gao, "Deep but lightweight neural networks for fish detection," in *OCEANS 2017-Aberdeen*. IEEE, 2017, pp. 1–5.
[53] R. Mandal, R. M. Connolly, T. A. Schlacher, and B. Stantic, "Assessing fish abundance from underwater video using deep neural networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–6.

[54] W. Xu and S. Matzner, "Underwater fish detection using deep learning for water power applications," in *2018 International conference on computational science and computational intelligence (CSCI)*. IEEE, 2018, pp. 313–318.
[55] K. Raza and H. Song, "Fast and accurate fish detection design with improved yolo-v3 model and transfer learning," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, 2020.
[56] Y. Wageeh, H. E.-D. Mohamed, A. Fadl, O. Anas, N. ElMasry, A. Nabil, and A. Atia, "Yolo fish detection with euclidean tracking in fish farms," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 1, pp. 5–12, 2021.
[57] X. Hu, Y. Liu, Z. Zhao, J. Liu, X. Yang, C. Sun, S. Chen, B. Li, and C. Zhou, "Real-time detection of uneaten feed pellets in underwater images for aquaculture using an improved yolo-v4 network," *Computers and Electronics in Agriculture*, vol. 185, p. 106135, 2021.
[58] H. S. Chhabra, A. K. Srivastava, and R. Nijhawan, "A hybrid deep learning approach for automatic fish classification," in *Proceedings of ICETIT 2019*. Springer, 2020, pp. 427–436.
[59] M. Mathur and N. Goel, "Fishresnet: Automatic fish classification approach in underwater scenario," *SN Computer Science*, vol. 2, no. 4, pp. 1–12, 2021.
[60] A. K. Agarwal, R. G. Tiwari, V. Khullar, and R. K. Kaushal, "Transfer learning inspired fish species classification," in *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 2021, pp. 1154–1159.
[61] R. B. Fisher, Y.-H. Chen-Burger, D. Giordano, L. Hardman, F.-P. Lin *et al.*, *Fish4Knowledge: collecting and analyzing massive coral reef fish video data*. Springer, 2016, vol. 104.
[62] A. Joly, H. Goëau, H. Glotin, C. Spampinato, P. Bonnet, W.-P. Vellinga, R. Planque, A. Rauber, R. Fisher, and H. Müller, "Lifeclef 2014: multimedia life species identification challenges," in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2014, pp. 229–249.
[63] A. Joly, H. Goëau, H. Glotin, C. Spampinato, P. Bonnet, W.-P. Vellinga, R. Planqué, A. Rauber, S. Palazzo, B. Fisher *et al.*, "Lifeclef 2015: multimedia life species identification challenges," in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2015, pp. 462–483.
[64] A. Saleh, I. H. Laradji, D. A. Konovalov, M. Bradley, D. Vazquez, and M. Sheaves, "A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis," *Scientific Reports*, vol. 10, no. 1, pp. 1–10, 2020.
[65] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?" *Structural safety*, vol. 31, no. 2, pp. 105–112, 2009.
[66] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.
[67] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher *et al.*, "A survey of uncertainty in deep neural networks," *arXiv preprint arXiv:2107.03342*, 2021.
[68] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," *Advances in neural information processing systems*, vol. 31, 2018.
[69] M. Raghu, K. Blumer, R. Sayres, Z. Obermeyer, B. Kleinberg, S. Mullainathan, and J. Kleinberg, "Direct uncertainty prediction for medical second opinions," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5281–5290.
[70] P. Oberdiek, M. Rottmann, and H. Gottschalk, "Classification uncertainty of deep neural networks based on gradient information," in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer, 2018, pp. 113–125.
[71] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.
[72] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE transactions on pattern analysis and machine intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.
[73] N. Moshkov, B. Mathe, A. Kertesz-Farkas, R. Hollandi, and P. Horvath, "Test-time augmentation for deep learning-based cell segmentation on microscopy images," *Scientific reports*, vol. 10, no. 1, pp. 1–7, 2020.
[74] M. Combalia, F. Hueto, S. Puig, J. Malvehy, and V. Vilaplana, "Uncertainty estimation in deep neural networks for dermoscopic image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 744–745.
[75] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[76] J. Yao, W. Pan, S. Ghosh, and F. Doshi-Velez, "Quality of uncertainty quantification for bayesian neural network inference," *arXiv preprint arXiv:1906.09686*, 2019.
[77] P. V. Molle, T. Verbelen, C. D. Boom, B. Vankeirsbilck, J. D. Vylder, B. Diricx, T. Kimpe, P. Simoens, and B. Dhoedt, "Quantifying uncertainty of deep neural networks in skin lesion classification," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures*. Springer, 2019, pp. 52–61.
[78] L. Smith and Y. Gal, "Understanding measures of uncertainty for adversarial example detection," *arXiv preprint arXiv:1803.08533*, 2018.
[79] D. Milanés-Hermosilla, R. Trujillo Codorniú, R. López-Baracaldo, R. Sagaró-Zamora, D. Delisle-Rodriguez, J. J. Villarejo-Mayor, and J. R. Núñez-Álvarez, "Monte carlo dropout for uncertainty estimation and motor imagery classification," *Sensors*, vol. 21, no. 21, p. 7241, 2021.

[80] H. Asgharnezhad, A. Shamsi, R. Alizadehsani, A. Khosravi, S. Nahavandi, Z. A. Sani, D. Srinivasan, and S. M. S. Islam, "Objective evaluation of deep uncertainty predictions for covid-19 detection," *Scientific Reports*, vol. 12, no. 1, pp. 1–11, 2022.

[81] A. Brando, J. A. Rodríguez-Serrano, M. Ciprian, R. Maestre, and J. Vitrià, "Uncertainty modelling in deep networks: Forecasting short and noisy series," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2018, pp. 325–340.

[82] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[83] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[84] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[85] ——, "Efficientnetv2: Smaller models and faster training," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 096–10 106.

[86] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[87] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[88] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[89] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[90] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, 2020.

the