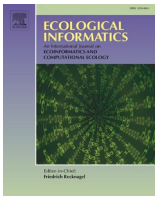
Ecological Informatics 80 (2024) 102535

Contents lists available at ScienceDirect

Ecological Informatics

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



Convolutional neural networks for hydrothermal vents substratum classification: An introspective study

Pedro Juan Soto Vega [a,*], Panagiotis Papadakis [b], Marjolaine Matabos [c], Loïc Van Audenhaege [c,d], Annah Ramiere [c], Jozée Sarrazin [c], Gilson Alexandre Ostwald Pedro da Costa [e]

[a] University Brest, LaTIM, INSERM, UMR 1101, 29200 Brest, France
[b] IMT Atlantique, Lab-STICC, UMR 6285, Team RAMBO, F-29238 Brest, France
[c] University Brest, CNRS, Ifremer, UMR6197 Biologie et Ecologie des Ecosystèmes marins Profonds, 29280 Plouzané, France.
[d] National Oceanography Center, Southampton, United Kingdom.
[e] Institute of Mathematics and Statistics, State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil

ARTICLE INFO

Keywords:
Image classification
Deep learning
Hydrothermal vents
Uncertainty analysis

ABSTRACT

The increasing availability of seabed images has created new opportunities and challenges for monitoring and better understanding the spatial distribution of fauna and substrata. To date, however, deep-sea substratum classification relies mostly on visual interpretation, which is costly, time-consuming, and prone to human bias or error. Motivated by the success of convolutional neural networks in learning semantically rich representations directly from images, this work investigates the application of state-of-the-art network architectures, originally employed in the classification of non-seabed images, for the task of hydrothermal vent substrata image classification. In assessing their potential, we conduct a study on the generalization, complementarity and human interpretability aspects of those architectures. Specifically, we independently trained deep learning models with the selected architectures using images obtained from three distinct sites within the Lucky-Strike vent field and assessed the models' performances on-site as well as off-site. To investigate complementarity, we evaluated a classification decision committee (CDC) built as an ensemble of networks in which individual predictions were fused through a majority voting scheme. The experimental results demonstrated the suitability of the deep learning models for deep-sea substratum classification, attaining accuracies reaching up to 80% in terms of F1-score. Finally, by further investigating the classification uncertainty computed from the set of individual predictions of the CDC, we describe a semiautomatic framework for human annotation, which prescribes visual inspection of only the images with high uncertainty. Overall, the results demonstrated that high accuracy values of over 90% F1-score can be obtained with the framework, with a small amount of human intervention.

1. Introduction

The increasing anthropogenic impact in the deep sea imposes an urgent need to evaluate the health and status of marine ecosystems via a better understanding of their dominant processes (Halpern et al., 2015; Levin and Le Bris, 2015).

For benthic ecosystems, specifically, images of the seabed retrieved by underwater sensors have been commonly used to assess the spatial distribution of fauna and substrata, e.g., (Marcon et al., 2014; Schoening et al., 2018; van den Beld et al., 2017). The latter provides a visual proxy for small-scale habitat characteristics, e.g., topographic complexity and substratum hardness, that can strongly shape the composition of faunal communities (Boulard et al., 2022; Simon-Lledó et al., 2019b). Additionally, substratum classification based on visual interpretation of seabed properties remains essential for calibrating and interpreting the acoustic response acquired from multibeam echo sounders (Lucieer et al., 2013b).

Recent developments of underwater platforms and cameras have provided images with improved pixel resolution covering increasing extents of seafloor surface, surveyed over space and time (Meyer et al.,

* Corresponding author.
E-mail addresses: sotovega@univ-brest.fr (P.J.S. Vega), panagiotis.papadakis@imt-atlantique.fr (P. Papadakis), Marjolaine.Matabos@ifremer.fr (M. Matabos), loicva@noc.ac.uk (L. Van Audenhaege), Jozee.Sarrazin@ifremer.fr (J. Sarrazin), gilson.costa@ime.uerj.br (G.A.O.P. da Costa).

https://doi.org/10.1016/j.ecoinf.2024.102535
Received 10 October 2023; Received in revised form 16 February 2024; Accepted 16 February 2024
Available online 22 February 2024

2019; Simon-Lledó et al., 2019a; Taylor et al., 2017). Nevertheless, manual annotation of seabed images is costly and biased due to operator subjectivity and self-learning when annotating large datasets (Schoening et al., 2016; Schoening et al., 2017a). Those limitations led to the emergence of a scientific community dedicated to devising computer tools capable of automatically analyzing the related, ever-increasing volume of data.

To date, several methods have been proposed for automatically processing large volumes of data. These methods combine computer vision (CV) and machine learning (ML) techniques, some of which are used in underwater studies, such as (Achanta et al., 2012; Faillettaz et al., 2016; Lucieer et al., 2013a; Osterloff et al., 2016; Schmid et al., 2016; Schoening et al., 2012). Regardless of their specific objectives, such efforts were mainly based on traditional image analysis approaches that rely on hand-crafted features and shallow-learning techniques. Although many different feature extraction methods have been evaluated in the past, e.g., (Achanta et al., 2012; Hossain and Chen, 2019; Sivic and Zisserman, 2003), the reported results are poor, primarily due to the deficiency of such methods in representing fundamental image properties for proper pattern recognition.

More recently, deep learning (DL) techniques for image analysis, particularly those based on convolutional neural networks (CNNs), have evolved in such a way that they now represent the state-of-the-art in many application fields, mainly due to their ability to learn discriminative features directly from data (Bengio et al., 2013).

Following that trend, DL has also been successfully employed in deep-sea applications, as seen in several research works, e.g., Villon et al. (2018); Song et al. (2019); Durden et al. (2021); Xue et al. (2021); Piechaud and Howell (2022); Katija et al. (2022). These studies mainly aimed to segment underwater images or detect specific objects within them. However, specifically for substrata labeling of marine images, DL-based techniques have not yet been widely explored, and most annotation procedures are still supported by visual interpretation (Gerdes et al., 2019; Neufeld et al., 2022).

Commonly, when annotating substrata in seafloor images, a categorical label is assigned to the whole image, considering specific characteristics of features observed within the image space (Althaus et al., 2013). In such a context, a trained human eye can identify a substratum through characteristics related to color, texture, relief, particle granulometry, or specific shapes and geomorphologies. Those features are overly complex for conventional automated recognition procedures, which renders the discrimination among different characteristics nearly unfeasible (Filippo et al., 2021).

Due to its success in modeling complex problems, DL, characterized by neural networks encompassing more hidden layers and learning parameters than their predecessors, represents a promising alternative. Additionally, the increasing availability of optical imagery datasets organized into large-scale temporal habitat maps is very convenient for evaluating the performance of DL-based techniques in deep-water studies. However, the limited availability of labeled data for properly training DL models in that context still remains a major obstacle.

Motivated by the scenario described above, this work investigates deep learning-based techniques for substratum classification in hydrothermal vent environments. Hydrothermal vents are high-temperature fluid emissions that arise on the seafloor and that sustain unique faunal communities. Our main goals are to reduce the human effort involved in substratum characterization; alleviate the subjectivity inherent to human photo interpretation; and contribute to reduce the costs and time involved in monitoring the deep sea.

With those purposes in mind, we devised a DL-based solution for automatically labeling deep-sea images, which employs different CNN architectures and allows human photo interpreters to audit the DL-based predictions, considering only a small fraction of the complete image dataset.

The proposed solution is designed to be employed in a human-in-the-loop approach, in which a learning algorithm can iteratively query a user to label some of the test data. The underlying system could then proactively select a subset of samples to be manually labeled from the set of new unlabeled data by analyzing the uncertainty scores of the predictions produced by the ensemble of neural networks.

In summary, the contribution of this work is four-fold:

- We evaluate six state-of-the-art convolutional neural network architectures, namely VGG, ResNet18 V1 & V2, ResNet50 V1 & V2, and Xception, in an image classification task for seafloor substratum discrimination in hydrothermal vent environments. The study reveals the generalization potential of general-purpose architectures under visual domain and site change.
- We assess the accuracy of a classification ensemble comprising all the implemented convolutional neural networks combined within a decision committee.
- We provide an uncertainty analysis conducted over the ensemble of the before-mentioned network architectures to assess the feasibility of using uncertainty in a semi-automatic image annotation scheme.
- We provide a visual interpretability analysis that provides insights into the decision making of models within the application of interest.

The DL models were evaluated using three different deep-sea image datasets, containing images acquired in different locations at the 1700 m deep Lucky Strike vent field region (Langmuir et al., 1993). Additionally, to analyze the generalization capacity of the studied models, a cross-evaluation was conducted using one of the three datasets for training and the others for testing.

It is worth mentioning that, considering the aforementioned experimental setup, the use of more recent approaches such as Vision Transformers (ViT) (Dosovitskiy et al., 2021) would not be feasible due to the scarcity of labeled training data.

As a subsidiary contribution, we also provide the code[1] used in the experiments, thus enabling further research and comparative evaluation. The datasets can also be provided upon request to the authors.

The remainder of this paper is organized as follows. Section 2 summarizes the previously proposed approaches for substrata classification using marine images. Section 3 presents the CNNs architectures investigated in this work. Section 4 describes the study site locations, the datasets used in our experimental analysis, the experimental setup, the networks' implementations, and the adopted performance metrics. Section 5 presents the obtained results, and finally, Section 6 presents conclusions and directions for future research.

2. Related works

Following the increasing availability of image data captured in deep-sea environments, many interesting machine learning-based computer vision methods have been proposed in the last few decades for marine and deep-sea applications. In the sequel, we mention some exemplary works in the field of interest.

Vandromme et al. (2012) made use of the Random Forest (RF) (Breiman, 1996) algorithm for the classification of zooplankton at pixel level. Schoening et al. (2012) proposed a semi-automatic image analysis system for assessing megafaunal densities at the Artic Deep Sea Observatory. The system comprises an ensemble of Support Vector Machine (SVM) classifiers, each associated with a particular species. A Maximum Likelihood Classifier (MLC) combined with two decision tree methods – Quick Unbiased Efficient Statistical Tree (QUEST) (Loh and Shih, 1997) and Classification Rule with Unbiased Interaction Selection and Estimation (CRUISE) (Kim and Loh, 2001) – were employed in Ierodiaconou et al. (2011) for detecting benthic biological communities, using video imagery among other capturing systems. Also, using different imaging systems, Schmid et al. (2016); Faillettaz et al. (2016), employed RF for

[1] https://github.com/pjsoto/IFREMER-ABYSSES.git

zooplankton analysis. The spatio-temporal distribution of shrimps was the objective of Osterloff et al. (2016). In that work, images were automatically pre-processed using a super-pixel segmentation algorithm named Simple Linear Iterative Clustering (SLIC) (Achanta et al., 2012), and RF was used to classify the super-pixels. Sharma et al. (2010) used shallow Artificial Neural Networks (ANN) to estimate deep-sea minerals using seafloor images.

Specifically for substratum studies, Lucieer et al. (2013a) proposed an Object-Based Image Analysis (OBIA) (Hossain and Chen, 2019) approach which relies on the *k*-nearest neighbor (*k*NN) algorithm to classify the substratum associated with image segments. Schoening et al. (2017b) applied a set of nodule compactness heuristics to delineate the polymetallic nodules in deep-sea images. A technique based on Bag of Visual Words (BoW) (Sivic and Zisserman, 2003), applied to Local Binary Pattern (LBP) features (Ojala et al., 2002) extracted from seafloor images, was proposed for substratum characterization in Kalmbach et al. (2016). Pillay et al. (2020) recently proposed a substrata seafloor characterization tool using advanced processing multibeam bathymetry, backscatter, and side scan sonar, jointly with Random Forest, Decision Trees, and K-means clustering.

The methods mentioned so far are based on shallow features, most of which are hand-crafted and potentially deficient in representing fundamental substratum properties, which might limit their performances. On the other hand, DL methods can learn complex and semantically rich representations of the input data and classes of interest, favoring classification performance. Villon et al. (2018) used (CNNs) to identify coral reef fish species, while in Durden et al. (2021) CNNs were trained to classify fauna in seabed images. Xue et al. (2021) studied the performance of several state-of-the-art DL architectures for identifying deep-sea debris. A method for recognition and tracking deep-sea organisms was proposed in Lu et al. (2020) using the YOLO (You Only Look Once) model (Redmon et al., 2016) as an object detector. For another deep-sea application, i.e., visual monitoring, Gradient Generation Adversarial Networks (GGAN) were proposed in Ma et al. (2021) to restore noisy images from the bottom of the sea. Juliani and Juliani (2021) employed a model based on the U-Net architecture (Ronneberger et al., 2015) for segmenting seafloor mounts directly over the raw bathymetry data. More recently, Katija and co-authors introduced the FathomNet (Katija et al., 2022), which provides annotated and localized imagery for developing ML algorithms. They also provide a set of ML models trained to detect the fauna present in the image data.

Considering the substratum characterization problem, McEver et al. (2023) introduced the DUSIA (Dataset for Underwater Substrata and Invertebrate Analysis) large-scale dataset, which was meant to train, validate, and test methods for localizing four underwater substrata and 59 underwater invertebrate species temporally and spatially. The inclusion of different substrata in DUSIA aimed at improving the localization of invertebrates along videos captured by Remotely Operated Vehicles (ROVs). The substrata analysis was accomplished using DL techniques based on a ResNet-inspired architecture (He et al., 2016). To the best of our knowledge, that was the only previous DL-based method proposed for substrata classification from deep-sea images.

However, regarding the substrata analysis dimension, McEver et al. (2023) represents a shallow study since substrata can be characterized

Table 1
Summary of related works previously published.

Papers	Comparison fields						
	Application	Data		Features		Classifier	Task
		Sound	Imagery	Hand crafted	Deep learning		
Sharma et al. (2010)	Deep-sea minerals estimation	×	✓	×	×	Shallow ANN	Image Classification
Ierodiaconou et al. (2011)	Benthic communities detection	✓	×	✓	×	MLC	Semantic Segmentation
Vandromme et al. (2012)	Zooplankton semantic segmentation	×	✓	✓	×	RF	Semantic Segmentation
Schoening et al. (2012)	Megafaunal densities recognition	×	✓	✓	×	SVM	Image Classification
Lucieer et al. (2013b)	Substratum classification	✓	×	✓	×	SVM	Image Classification
Lucieer et al. (2013a)	Substratum classification	×	✓	✓	×	*k*NN	Image Classification
Schmid et al. (2016)	Zooplankton analysis	×	✓	✓	×	RF	Image Classification
Faillettaz et al. (2016)	Zooplankton analysis	×	✓	✓	×	RF	Image Classification
Osterloff et al. (2016)	Shrimps distributions analysis	×	✓	✓	×	RF	Semantic Segmentation
Kalmbach et al. (2016)	Substratum classification	×	✓	✓	×	BoW	Image Classification
Schoening et al. (2017b)	Deep-sea polymetallic nodules	×	✓	✓	×	–	Semantic Segmentation
Villon et al. (2018)	Coral fish species identification	×	✓	×	✓	CNN	Semantic Segmentation
Pillay et al. (2020)	Substratum classification	✓	×	✓	×	RF, K-means, Decision Trees	Semantic Segmentation
Lu et al. (2020)	Deep-sea organism tracking	×	✓	×	✓	YOLO	Object Detection
Durden et al. (2021)	Fauna classification	×	✓	×	✓	CNN	Image Classification
Xue et al. (2021)	Deep-sea debris identification	×	✓	×	✓	CNN	Image Classification
Ma et al. (2021)	Noisy image restoration	×	✓	×	✓	–	Image Restoration
Juliani and Juliani (2021)	Mineral mounts segmentation	×	✓		✓	U-Net	Semantic Segmentation
Katija et al. (2022)	Fauna detection	×	✓	×	✓	CNN	Video Motion Analysis
McEver et al. (2023)	Invertebrates detection	×	✓	×	✓	CNN	Object Detection

considering more specific geological variables, such as lithology or morphology. Furthermore, no substratum characterization studies have been carried out on the evaluation of different DL architectures, assessing their generalization capacity considering domain shift problems. Additionally, no DL-based studies in this application field have focused on understanding the learned features, considering interpretability and classification uncertainty.

Table 1 provides an informative summary of the works mentioned above. It offers useful insights regarding comparisons in terms of the tasks, data types (sound or images), feature types (hand-crafted or automatically learned), and classification techniques. An analysis of the table allows one to easily identify the main characteristics that differentiate the works and draw useful conclusions for future research. It is worth mentioning that although DL has been used in some deep-sea applications, the number of these works is low, indicating that DL has not been widely explored in this field.

In this work, we evaluate six different DL architectures in the substrata classification task, considering different characterization criteria, namely, morphological, lithological and the number of shells and white fragments. Additionally, we evaluate a way to merge predictions from multiple DL models (combined into a decision committee) and calculate uncertainty scores for each image sample. These scores can then be used in an audit scheme, which can substantially increase overall classification performance and reduce human effort in the substrata characterization process. We also provide an evaluation of the behavior of each architecture and of the decision committee considering the domain shift problem. Finally, we conduct an interpretability analysis with the aim of providing insights into networks' decision-making.

3. Methods

This section outlines the main characteristics of the CNN architectures selected for this study. We note that diverse other image classification architectures have been proposed to date, some of which with outstanding performance on different applications. To the best of the authors' knowledge, however, no specific architecture has been designed so far for the particular application of substratum image classification. For that reason, we opted to use seminal, general-purpose models, extensively used and well-documented, which we believe would ease the reproduction of our experiments and results. Notwithstanding, the underlying processing chain could very well comprise alternative architectures.

The selected CNNs typically contain an encoder stage, often called a feature extractor, which reduces the spatial resolution of the input through convolution and pooling operations in consecutive layers. The encoder is then followed by fully connected layers of neurons that predict the input image class based on the previously extracted features to perform image classification.

After briefly introducing the chosen CNN architectures, we will discuss the decision fusion process. That process is based on a decision committee that collects the predictions of the respective CNN-trained models. We will also explain uncertainty assessment based on model predictions.

3.1. VGG

To date, the VGG (Visual Geometry Group) network is one of the most popular image classification architectures, and pre-trained VGG models are commonly used in transfer learning (e.g., fine-tuning) schemes. It was proposed in Simonyan and Zisserman (2015), which aimed at investigating the effects of increasing convolutional network depth. Evaluated on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015) in 2015, the VGG16 model outperformed all previous participants, which comprised several state-of-the-art architectures.

Variants of the VGG network have demonstrated that increasing network depth can improve classification accuracy by enabling the learning of semantically enriched features. In the present work, we adopted a particular architecture inspired by VGG that has 14 layers (13 convolutional and one dense layer). A detailed description of the architecture implemented in this work can be found in Section 7.

3.2. ResNet

Residual Networks (ResNet), introduced by He et al. (2016), aimed at improving convergence issues during the training of very deep network architectures. ResNet is a deep learning model that aims to alleviate two main problems in training neural networks. The first problem is the vanishing gradient, which makes it difficult to optimize the model during training. The second is the degradation problem, which occurs when adding more layers to a deep neural network leads to greater training errors. ResNet solved those problems by using residual learning blocks between layers of the network, allowing for better optimization and higher training accuracy.

The degradation problem suggested that when the network's full capacity was underused for solving a particular task, the optimization process would have difficulties in approximating nonlinear layers into identity mappings, which could automatically adjust network depth. Then, instead of hoping that every few stacked layers directly fit a desired underlying mapping, e.g., identity, ResNet explicitly lets those layers fit a residual mapping, which is easier to optimize. In this work, we implemented four variants of the ResNet, with 18 and 50 layers, using two versions of residual blocks, namely, ResNet18 V1, ResNet18 V2, ResNet50 V1, and ResNet50 V2. The main difference between the residual blocks in the V1 and V2 versions is that in V2 the blocks are not followed by a ReLU activation function.

3.3. Xception

Proposed by François Collet (Chollet, 2017), Xception is an improved version of the InceptionV3 architecture (Szegedy et al., 2015; Szegedy et al., 2016). Its main innovation is the use of depth-wise separable convolutions instead of regular convolutions. A depth-wise separable convolution consists of a depth-wise convolution followed by a point-wise convolution (Vanhoucke, 2014). Additionally, Xception's design benefited from several prior efforts, such as the previous innovations brought by the Inception family (Szegedy et al., 2015, 2016) and residual connections (He et al., 2016). Details about the Xception architecture implemented in this work are shown in the Appendix (see Section 7).

3.4. Classification decision committee

Besides assessing the performance of DL networks whose architectures were mentioned in the previous sections, we further investigate the result of fusing the different models' decisions, i.e., considering the set of classifiers as an ensemble, hereinafter denoted *Classification Decision Committee* (CDC).

The basic idea is to employ M different classifiers $h_1, \cdots, h_M$ for the same task and combine their individual outcomes. In this work, we implemented the majority voting scheme for combining the decisions of the individual classifiers that comprise the ensemble. Thereby, the final prediction is given by the class that achieves the majority of votes of the individual classifiers. Fig. 1 shows a schematic representation of that strategy. Each image x_n is evaluated by each trained classifier h_i. Then, the majority voting scheme uses each classifier prediction $\widehat{y}_i$ to release the final prediction $\widehat{y}$ of the input image x_n. Such a scheme consists of computing the mode of the set of predictions as represented in the following equation:

$$\widehat{y} = mod\{\widehat{y}_1, \ldots, \widehat{y}_i, \ldots, \widehat{y}_M\} \quad (1)$$

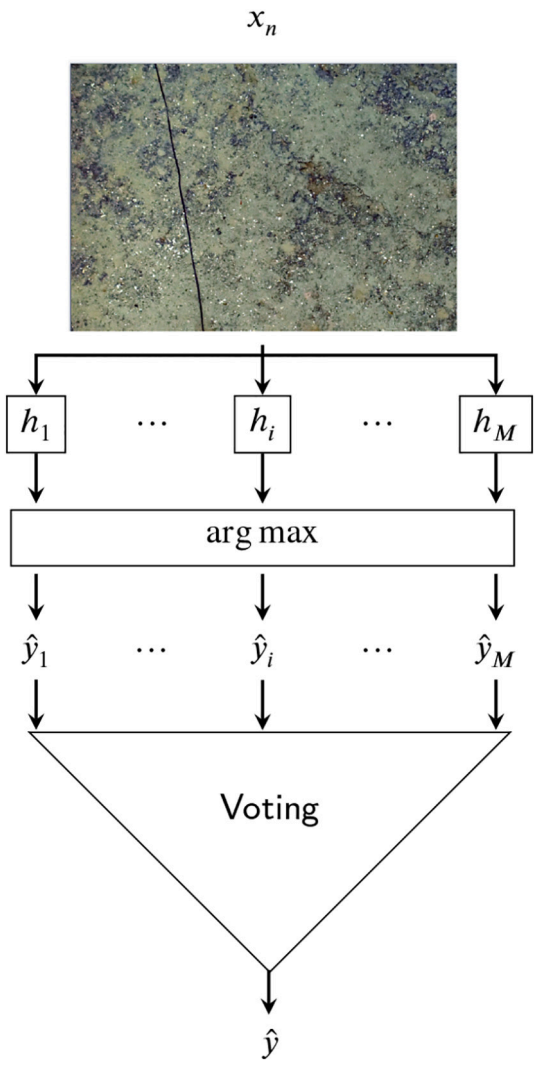


Fig. 1. Representation of the classification decision committee (CDC) evaluated in this work. The voting scheme is represented by Eq. (1).

where M represents the number of networks that compose the ensemble, and *mod* stands for the mode of the classifier predictions.

As anticipated in previous sections, the number of networks considered in this work is $M = 6$, so the computed mode may have two or more values. When that is the case, we take as a final prediction the category where the average probability among the networks is maximum.

3.5. Uncertainty estimation

Uncertainty studies in DL have become an important tool to address the lack of expressiveness and transparency of deep neural network predictions, the inability to distinguish between in- and out-of-domain samples and sensitivity to domain shifts (Abdar et al., 2021). During the past years, several metrics have been proposed to quantify uncertainty, among which *Predictive Entropy*, is widely used.

Formally, let $\mathbf{s} = \{s^{(i)}\}_{i=1}^{M}$ be the set of M probability predictions for the image x_n for all C classes. Let also, $s_c^{(i)}$ be the c-th element of $s^{(i)}$ corresponding to the prediction for class c of the input image x_n. The final prediction $\bar{s}_c$ for an image x_n and class c is the average overall M predictions $s_c^{(i)}$:

$$\bar{s}_c = \frac{1}{M}\sum_{i=1}^{M} s_c^{(i)} \tag{2}$$

Then, the *Predicitve Entropy* $H(\mathbf{s}|x_n)$ is obtained by computing the entropy of the average prediction over x_n:

$$H(\mathbf{s}|x_n) \approx -\frac{1}{C}\sum_{c=1}^{C} \bar{s}_c log(\bar{s}_c) \tag{3}$$

It is worth noting that the majority of uncertainty metrics already proposed in the literature consider a set of M predictions about the same image. Such a number (M) can be obtained in different ways, depending on the method employed for the uncertainty estimation.

According to Gawlikowski et al. (2023), uncertainty estimation methods can be categorized into four groups, considering the number (single or multiple) and the nature (deterministic or stochastic) of the deep neural networks (DNNs) used. *Bayesian methods* cover all kinds of stochastic DNNs, where two forward passes for the same sample generally lead to different results. *Ensemble methods*, on the other hand, combine the predictions of several different deterministic networks at inference time. *Test-Time Augmentation methods* adopt one single deterministic network to produce a prediction but augment the input data at test-time, aiming at generating several predictions to compute the uncertainty score. Finally, *Single Deterministic methods* produce the predictions based on a single forward pass within a deterministic network.

In this work, we implemented the ensemble-based alternative for uncertainty estimation and selected the *Predictive Entropy* as the uncertainty metric. Considering the classification setting addressed in this work, where the output of a model is a conditional probability, we computed the *Predictive Entropy* measure from the set of individual predictions of the different DL models.

3.6. Gradient-weighted class activation mapping (grad-CAM)

In view of the variety of potential applications and the complexity of deep learning models, interpretability techniques have become essential for building systems whose decisions can be explainable to human experts.

Among several alternatives, Grad-CAM, introduced by Selvaraju and co-authors (Selvaraju et al., 2017), exploits the gradients associated with the computation of convolutional features to understand and visualize which parts of the input image were the most important for its classification. The Grad-CAM algorithm computes the importance map by taking the derivative of the reduction layer output for a given convolutional feature map.

To obtain a class-discriminative activation map, $L^c_{Grad-CAM} \in \mathbb{R}^{u \times v}$ where u and v represent the width and height of activation maps in a specific network layer for any class c, Grad-CAM first computes the gradient of the score for class c, $\bar{h}^c$, before applying the *softmax* function, with respect to the $k-th$ feature maps A^k of a convolutional layer, i.e., $\frac{\partial \bar{h}^c}{\partial A^k}$. Then, the back-propagated gradients are globally average-pooled to obtain the neurons' importance weights w_k^c, defined as:

$$w_k^c = \frac{1}{Z}\sum_{i}^{u}\sum_{j}^{v} \frac{\partial \bar{h}^c}{\partial A_{ij}^k} \tag{4}$$

where Z represents the total number of neurons in the $k-th$ feature map.

According to Selvaraju et al. (2017), the weights w_k^c constitute a partial linearization of the deep network downstream from A^k and capture the importance of feature map k for a target class c.

Finally, Grad-CAM performs a weighted combination of forwarded activation maps, followed by a *ReLU* activation function as:

$$L^{c}_{Grad-CAM} = ReLU\left(\sum_{k} w^{c}_{k} A^{k}\right) \quad (5)$$

4. Experimental analysis

The experiments conducted in this work aimed at verifying the effectiveness of DL-based models in the context of a particular image classification problem, namely, ridge and hydrothermal vent substrata classification. The datasets used in this work comprise images taken from three different locations on a particular vent field (Fig. 2).

We basically performed two sets of experiments: (i) we evaluated the DL models individually and as a member of a classification ensemble, the so-called classification decision committee (CDC), which combines all models in a decision committee; and (ii) we performed cross-evaluation experiments in which the models were trained using one

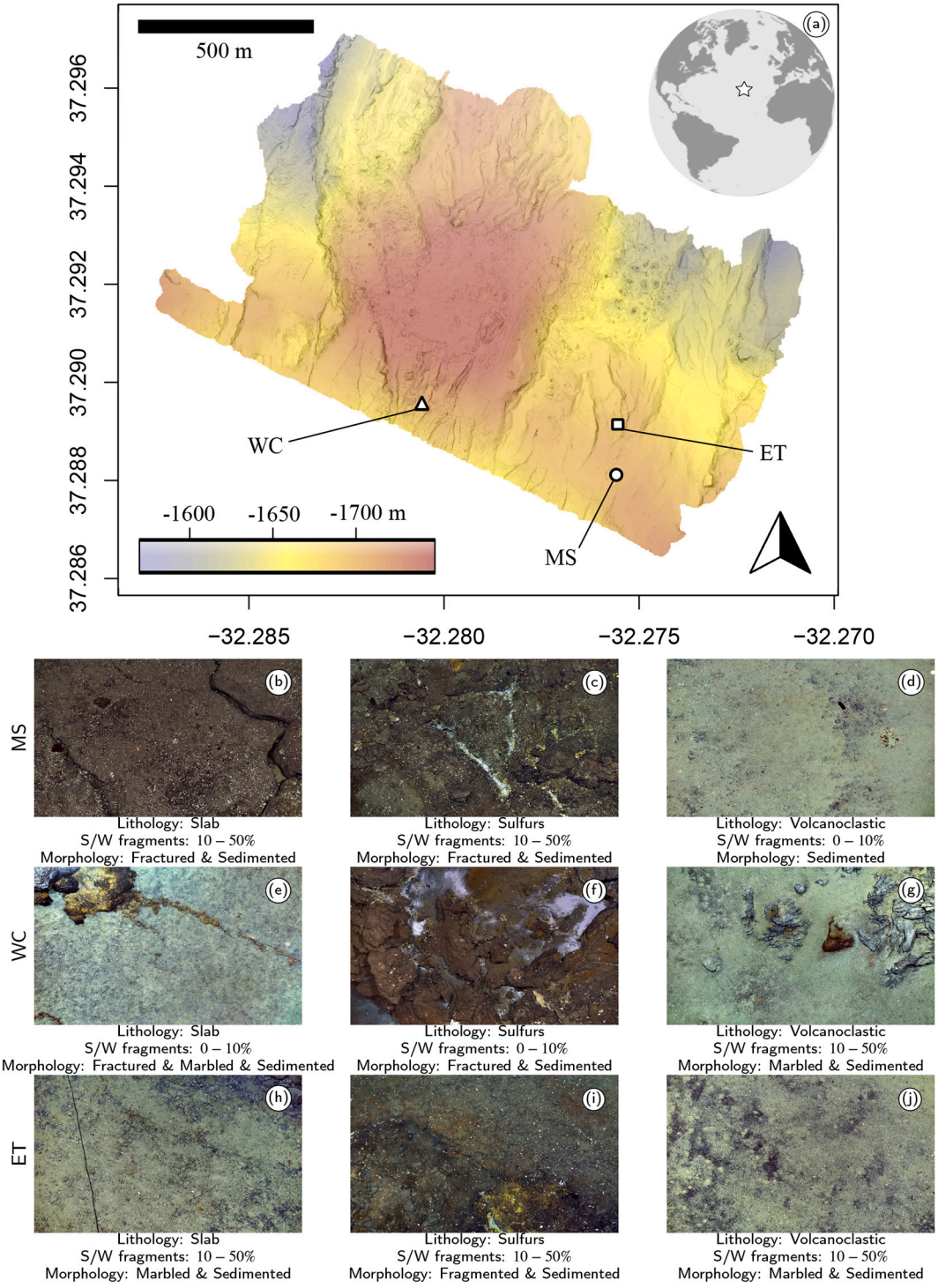


Fig. 2. Image samples and map of the Lucky Strike vent field (northern Mid-Atlantic Ridge). In the map (a), the color gradient corresponds to the bathymetry. Active vent fields are indicated on the map with a triangle (MS), a circle (MS), and a square (ET). Images from Montsegur edifice (MS) are represented in (b) (c) (d); White Castle (WC): (e)(f)(g); and Eiffel Tower (ET): (h)(i)(j). Below each image is listed its respective classes regarding lithology, shell and white (S/W) fragments, and morphology.

dataset and tested on each of the other datasets in turn.

4.1. Study areas

The study areas are located at the Lucky Strike (LS) vent field along the Mid-Atlantic Ridge (MAR, 37°17*N*, 32°16*W*). LS is a basalt-hosted hydrothermal vent field located near the Azores Triple Junction on the slow-spreading MAR at a depth of approximately 1700 m (Langmuir et al., 1993). This large hydrothermal field extends over more than 1 km^2 and lies at a seamount's summit, harboring a central fossilized lava lake surrounded by three volcanic cones and faults (Ondréas et al., 2009). Emissions of hydrothermal vent fluids are distributed all around the lava lake (Barreyre et al., 2012). They can occur over vertical vent edifices that build up through the accumulation of minerals, including sulfides. In the flatter periphery of edifices, the hydrothermal activity can expand over a seafloor composed of a slab of basaltic fragments indurated by silica and barite (Cooper et al., 2000).

The ecology of benthic communities at LS has been thoroughly investigated since its discovery in 1992, and, more recently, since the deployment of the EMSO-Azores observatory (Matabos et al., 2022). However, while it is known that *Bathymodiolus* vent mussels dominate the benthic communities of the sulfide edifice of Eiffel Tower (ET) (Husson et al., 2017), little is known about the distribution of vent specialists and non-vent fauna outside the edifices. This gap in ecological knowledge motivated the retrieval of a large set of seabed images focusing on edifices and their peripheries at LS. Fig. 2 shows the LS localization as well as the positions of the Eiffel Tower (ET), Montsegur (MS), and White Castle (WC) vent edifices.

4.2. Dataset

The dataset comprises RGB images collected during the MoMARSAT 2018 cruise (Cannat and Sarradin, 2018) using the Remotely Operated Vehicle (ROV) *Victor6000* over and around the following edifices (areas hereafter called sites): Montsegur (MS, see Marticorena et al. (2021)), White Castle (WC) and Eiffel Tower (ET, see Girard et al. (2020)). Each image has a dimension of 4000 × 6000 pixels with a spatial resolution of 0.001 m/pixel. Images of the seabed have been acquired at one image every three seconds with a downward-looking HD camera OTUS2 with navigation tracks. Constant ROV altitude (5 ± 2 m) planned in parallel transects spaced 1.8 m apart, to ensure overlap between each captured image at a constant speed of 0.5 $m.s^{-1}$. Next, the image sets were pre-processed in the following order. First, blurred and obscured samples were removed. Second, a non-overlapped set of pictures was selected using the MATISSE 3D software (Arnaubec et al., 2015) (Ifremer). The MATISSE 3D computes image overlaps through geo-referencing, using the ROV's navigation parameters and camera positions. Third, the set of non-overlapped images of each site was corrected by attenuating the blue color and homogenizing the light conditions, contrast and saturation in MATISSE 3D.

Fig. 3 shows the categories the experts consider in labeling each image based on the substratum type. The manual labeling procedure was done at the image scale, i.e., one label for the entire image according to three criteria: lithology, morphology, and the amount of mussel shells and unidentified white fragments (S/W fragments) contained in the image.

Lithology relates to information on the nature and origin of the substrata, associated with mineral composition and hardness, derived from image colors and geomorphologic features. The different lithologies of the rock found at LS are basalts, sulfures, hydrothermal indurated slab, and volcanoclastic sediments (see Fig. 2). Morphology relates to the shape of the substrata encompassing categories such as fractured, marbled, scree/rubbles, and brecciated/pillow (B/P) lava. Finally, in areas where S/W fragments cover the seafloor, the images were annotated based on the percentage of the covered image space: 0–10%, 10–50%, and 50–100%. Table 2 shows the number of images in each category and site. Regarding the number of labels, for lithology and S/W fragments, each image is associated with a unique label. However, as distinct morphological features can occur within a single image, it can be labeled with more than one morphology class. That explains why the total number of images provided in the last row of Table 2 is the sum of per-class images for lithology and S/W fragments, but not for morphology.

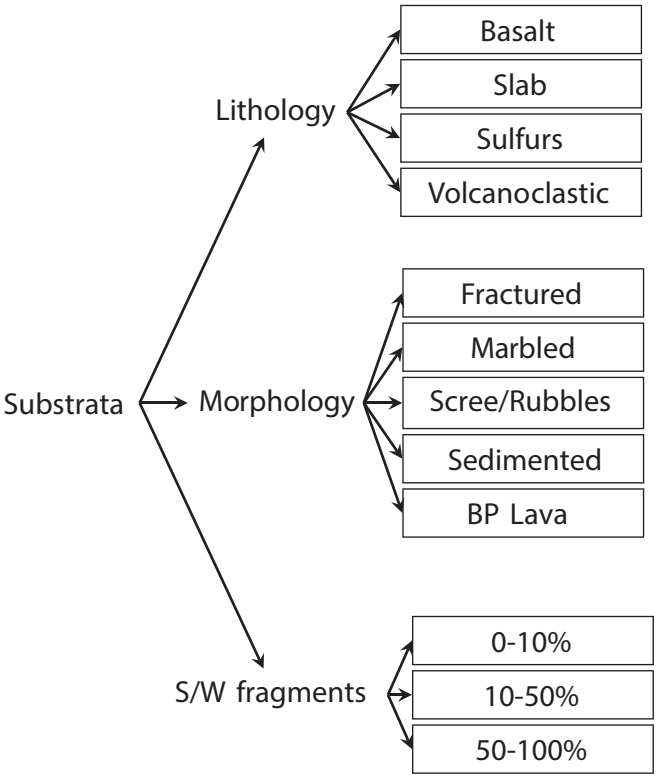


Fig. 3. Substratum categories studied in this work for model training. In the figure, S/W fragments stand for Shells and White fragments while BP Lava for Brecciated-Pillow Lava.

It is worth noting that a homogenization procedure was performed during the expert annotation to ensure spatial coherence and reduce uncertainties. First, uncertain labels were replaced by the nearest neighbor's corresponding ones, considering the previously calculated geographical localization. Second, the subset of images that did not form a group of at least three contiguous images underwent a similar procedure, assigning the dominant surrounding label to the images in the subset. Fig. 2 shows some image samples of the three sites.

Table 2
Number of images in each domain in and class.

Criteria	Classes	Sites		
		MS	WC	ET
S/W fragments	0–10%	166	170	443
	10–50%	105	47	270
	50–100%	19	3	17
Lithology	Basalt	–	43	104
	Slab	142	91	384
	Sulfurs	123	42	78
	Volcanoclastic	25	44	164
Morphology	Fractured	210	94	236
	Marbled	158	179	534
	Scree/rubbles	24	102	311
	Sedimented	274	220	718
	BP Lava	–	43	104
# of images		290	220	730

As shown in Table 2, the site-specific datasets are imbalanced regarding the number of samples per class. For instance, MS does not contain basalt samples and consequently no B/P lava in lithology and morphology, respectively. On the other hand, ET contains more images than MS and WC, while 50–100% S/W fragments is the less represented class among all categories in all sites.

At this point, it is important to observe that, if not considered explicitly in the training procedure, such a problem (i.e., class imbalance) can introduce undesirable bias that may affect the performance of a classifier, as it will be prone to predict the over-represented classes more often due to the smaller impact of the errors associated with their samples in computing overall accuracy. In the next section we describe how we dealt with class imbalance in this work.

4.3. Experimental setup

For the accuracy assessment of substratum characterization, we used a *k-fold* scheme with $k = 3$. More specifically, the set of images from each site was divided into three disjoint subsets containing randomly chosen image samples from all classes. We used $k - 1$ folds (two subsets) for training and the remaining one for testing. The accuracy metrics values reported in Section 5 are then averages across the testing folds.

Considering the manual labeling process, we adopted similar protocols for training the networks. First, we created classifiers that characterize the images according to lithology, morphology, and the amount of shell and white (S/W) fragments separately. In simple terms, for each of the feature extractor plus classifier architecture (mentioned in the Appendix section and described in Fig. 9), one network was trained for identifying solely the lithology classes; another network was trained to identify morphology classes; and another for S/W fragments classes. Considering the different ResNet variants (see Section 3.2), networks based on six different architectures (i.e., VGG, ResNet18 V1, ResNet50 V1, ResNet18 V2, ResNet50 V2, and Xception) were trained for each substratum type and site.

Second, we adopted single-label image classification for lithology and S/W fragments since, for those substratum categorizations, the images in the datasets have been annotated as belonging to a unique class. In those cases, the CNNs' outputs are computed with a *Softmax* function, as represented in Fig. 9, and the class corresponding to the neuron the higher activation is selected. Conversely, a multi-label image classification approach has been adopted for morphology, as the dataset images may have been associated with more than one morphology class. For that reason, the output layers of each network use a *Sigmoid* activation function instead of the *Softmax*. That allows more than one output neuron to be activated for a given input image.

To compensate for class imbalance (see Table 2), we adopted a *weighted cross-entropy* cost function to train the networks. The intention was to force the CNNs not to be biased towards the over-represented classes by assigning larger weights to the underrepresented ones. Eq. (6) represents the loss function ($\mathscr{L}_{ic}$) employed in the training of the networks for lithology and S/W fragments. Eq. (7) represents the loss function ($\mathscr{L}_{ml}$) for training the morphology specialized networks.

$$\mathscr{L}_{ic} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c(y_n log(h(x_n)\,)\,) \tag{6}$$

$$\mathscr{L}_{ml} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c(y_n log(h(x_n)\,)\,) + (1-y_n) log(1-h(x_n)\,) \tag{7}$$

In both equations, N stands for the number of training images, x_n is the $n-th$ training image, while y_n represents the true label (or labels) codified in a one-hot vector of x_n. Furthermore, $h(x_n)$ corresponds to a vector comprising the predicted likelihood values for each class of x_n, computed with the learned function $h(\cdot)$. Additionally, $w_c = \frac{N}{N_c}$ is the weight of each class $c \in C$, which comprises N_c images.

During training, the inputs to the networks were patches with dimensions $224 \times 224 \times 3$, extracted from the original full-resolution dataset images. The patches were extracted using a sliding window procedure, with an overlap of 25% in each direction. As in applications such as those addressed in (Soto et al., 2022; Soto Vega et al., 2021), splitting the images into patches functions as a data augmentation strategy, and also eases GPU memory handling. Another benefit of training the CNNs using image patches is that it contributes to learning the frequency of occurrence of particular characteristics of the various classes, thus improving the overall understanding of original images. During test time, the networks make a prediction by convolving over the full-size image.

During training, the cost function was minimized using the Adam optimizer (Kingma, 2017), with an initial learning rate μ_0 and momentum β_1 equal to 0.0001 and 0.9, respectively. Aiming at better convergence during training, we adopted a learning rate decay procedure proposed in Ganin et al. (2017) by implementing the following equation:

$$\mu_e = \frac{\mu_0}{(1+\alpha p)^{\beta}} \tag{8}$$

where $p = \frac{e}{\#Epochs}$, and e is the current training epoch. Following Ganin et al. (2017), α and β were set to 10 and 0.75, respectively.

The batch size was 32, and the early stopping procedure was used to avoid over-fitting. The patience parameter, which controls the number of epochs without improvements in the validation loss, was set to 10. Each network, with a particular architecture and considering a specific substratum characterization, was trained and executed three times, each time with a different (random) initialization of the trainable parameters and with a different data fold. As already mentioned, the results shown in the next section are averages of those three executions. Data augmentation was applied to all extracted patches: a 90° rotation and vertical and horizontal flips.

4.4. Performance metrics

As already mentioned, the accuracy metrics values reported in Section 5 are averages across the testing folds. Therefore, the performance of the classifiers in all experiments is expressed in terms of the average F1-scores computed for each individual class. Specifically, for each class the F1-score is expressed by the harmonic mean of Precision (P_c) and Recall (R_c) as follows:

$$F1 - score_c = \frac{2 \times P_c \times R_c}{P_c + R_c}, \tag{9}$$

where

$$P_c = \frac{t_p}{t_p + f_p}. \tag{10}$$

$$R_c = \frac{t_p}{t_p + f_n}. \tag{11}$$

In Eqs. (10) and (11), t_p is the number of images correctly assigned to the class c (true positives), f_p represents the number of images erroneously classified as the current class c (false positives). Similarly, f_n corresponds to the number of images incorrectly classified as non-class c (false negatives).

5. Results and discussion

In this section, we present the results of the experiments carried out in this study. We start by comparing the performances of the DL models on each site. We considered in such comparison the outcome of the models trained and tested with data from the same site; and the outcome of the models trained on a particular site, but tested with data from other sites (cross-site evaluation). Next, we present results obtained with the

decision committee, which combines the outcomes of the different network architectures into a single decision. We then visually assess some of the networks' activation maps, using an explainable artificial intelligence (XAI) technique to better understand the inference process. Finally, we analyze the classification uncertainty derived from the set of individual predictions of the committee members and show how the uncertainty measures can be exploited in a semiautomatic decision process.

5.1. Accuracy of the deep learning networks

Table 3 shows the F1-scores obtained in the image classification experiments. Since the MS dataset does not contain basalt samples in lithology nor B/P lava in morphology, for the comparative analysis to be fair, such classes were not taken into account during training and testing.

In Table 3, the results obtained with the different DL networks are represented as matrices, in which the diagonal contains the results obtained by training and testing with data from the same site. The off-diagonal values are associated with the cross-site evaluation scheme. The values in bold represent the best results obtained with each network for the different configurations of training/testing sites.

It can be observed by inspecting the values in Table 3 that, in most cases, the different networks achieved similar scores when trained and tested on the same pair of sites. Indeed, the predictions of the different networks are not so dissimilar in terms of F1-scores, regardless of the substratum category. There are, however, some site configurations in which the accuracy variation is more significant, as in the case of the ET/WC (training/testing) configuration according to S/W fragments categorization, but those can be considered exceptions.

It is also interesting to observe that the accuracies obtained when the networks were trained and tested with data from the same site were, in the majority of cases, higher than those associated with the cross-site evaluations. The larger differences in that regard were observed for the lithology categorization. The smallest differences, which represent a better generalization capacity, occur for the morphology categorization.

In terms of the cross-site evaluations, the accuracies achieved for the shells and white (S/W) fragments categorization were in between those obtained for lithology and morphology. Moreover, considering the cross-site evaluation results obtained for S/W fragments and morphology, it can be verified in Table 3 that the highest accuracies are associated with the networks trained on the ET dataset.

We hypothesize that a classifier trained on ET is more efficient in discerning morphology and S/W fragments classes because ET is a larger dataset (see Table 2). However, the same behavior was not observed for lithology, where the classifiers trained on ET performed similarly to those trained on the remaining sites, WC and MS, and vice-versa. Those results deserve further investigation.

5.2. Accuracy of the classification decision committee

In this section, we present the results of the classifier decision committee (CDC), composed of the different DL networks, employing the majority voting decision fusion strategy explained in Section 3.4 and represented in Fig. 1.

Table 4 shows the CDC results in terms of F1-scores for all substratum characterization criteria and sites. The table shows the average F1-score of three rounds of experiments as in the previous experiments. Similarly, the bold values represent the best results obtained with the committee.

In general, the ensemble of networks reached similar scores to those obtained by each network individually. Specifically, considering evaluations in sites where the classifiers were trained, i.e., results described in the diagonal of Table 4, CDC reached results that are superior to most of those obtained with the individual models.

Notwithstanding, the CDC substantially improved the classification accuracy of cross-site evaluations, notably in S/W fragments, e.g., results obtained by the CDC when models were trained in ET and evaluated in MS and WC. Regarding the lithology categorization, the committee followed the same trend observed in the results of the individual networks. The latter reflects that all networks are affected by the site/domain shift phenomenon, at least from the lithological point of

Table 4
Average F1-score (%) for three executions (one per fold) of the DL decision committee.

Criteria	Training on:	Testing on:		
	Sites	MS	WC	ET
S/W fragments	MS	**69.8**	52.8	57.2
	WC	55.5	57.4	54.0
	ET	65.8	**75.5**	**69.2**
Lithology	MS	**76.8**	57.3	54.1
	WC	54.2	**74.5**	46.5
	ET	54.8	54.2	**75.4**
Morphology	MS	**69.3**	64.5	58.2
	WC	59.4	**78.9**	66.5
	ET	63.0	76.9	**76.4**

Table 3
Average F1 (%) for three executions (one per fold) of the DL networks for the substrata characterization.

Criteria:			S/W fragments			Lithology			Morphology		
Architectures	Sites		Testing on:								
			MS	WC	ET	MS	WC	ET	MS	WC	ET
VGG	Training on:	MS	**61.6**	52.5	57.1	**68.0**	50.9	56.2	**67.4**	75.5	70.2
		WC	49.8	59.6	57.9	57.0	**72.4**	48.0	65.2	78.3	70.4
		ET	56.6	**67.2**	**64.9**	53.8	51.1	**72.0**	67.1	**78.4**	**74.5**
ResNet18 V1		MS	**70.2**	65.7	62.2	**73.2**	50.6	51.4	**68.3**	72.2	64.1
		WC	49.0	64.2	55.2	49.3	**64.4**	41.8	67.0	78.2	72.7
		ET	63.9	**72.1**	**67.5**	50.7	51.5	**71.6**	66.1	**78.5**	**74.7**
ResNet50 V1		MS	**64.2**	**66.4**	59.3	**70.7**	51.6	51.1	**68.6**	75.0	67.5
		WC	41.0	45.1	45.3	43.4	56.9	30.9	67.2	**80.0**	71.9
		ET	53.3	65.6	**69.3**	50.1	**57.2**	**70.5**	64.9	78.5	**75.2**
ResNet18 V2		MS	**67.5**	40.0	55.1	**66.8**	50.7	45.4	**68.5**	71.5	64.2
		WC	39.5	49.6	52.0	54.0	**75.2**	42.1	63.9	**80.2**	68.7
		ET	54.3	**54.6**	**61.9**	49.4	49.5	**72.4**	65.8	79.0	**75.0**
ResNet50 V2		MS	**67.8**	40.1	51.8	**75.3**	55.3	55.7	**69.4**	74.2	66.9
		WC	49.0	46.0	52.1	48.1	**71.7**	46.4	65.8	**79.4**	71.5
		ET	64.9	**57.2**	**67.7**	54.1	53.4	**70.5**	65.6	78.3	**75.6**
Xception		MS	**66.9**	62.7	53.5	**66.1**	62.7	48.4	**68.8**	73.0	66.5
		WC	55.0	53.3	54.3	52.6	**64.4**	42.9	66.0	**79.9**	70.8
		ET	63.8	**74.0**	**70.1**	56.9	59.8	**76.5**	66.3	78.0	**73.6**

view. That behavior is probably related to the subtlest differences in substratum texture among the three sites. Fig. 2 shows examples of images from the three sites and representations of each lithology class. The figure shows considerable differences among images of the same classes belonging to different sites.

To enrich the current analysis, we present in Fig. 4 the confusion matrices associated with the CDC results for all site combinations according to the lithology classification, and in Fig. 5 the confusion matrices according to S/W fragments classification. Given that morphology categorization entails multiple labels, the analysis of confusion matrices becomes more complex, as each class requires computation separately. That's why we limited this discussion to the lithology and S/W fragments categories. In the figures, the matrices in the diagonal show the results obtained for the networks trained and tested on the same site. The remaining matrices represent cross-site results.

Specifically, for lithology (see Fig. 4), the CDC had problems discerning between slab and sulfurs in all evaluated scenarios. Moreover, the committee predicted a reasonable number of slab images as volcanoclastic, mainly when it was trained or evaluated on ET.

Regarding S/W fragments, the observed misclassifications occurred mainly between classes 0–10% and 10–50%, which is unsurprising since estimating the amount of shell and white fragments is problematic, driven by subjective visual quantification. Additionally, the worst results were for the 50–100% class. We believe the reason for that behavior lies in the under-representation of that category in the training dataset.

5.3. Classification visual analysis

This section presents a visual assessment of internal network representations to better understand the criteria or the specific features that seem to be taken into account by the deep learning networks in substratum classification. For that purpose, we used the Gradient-weighted

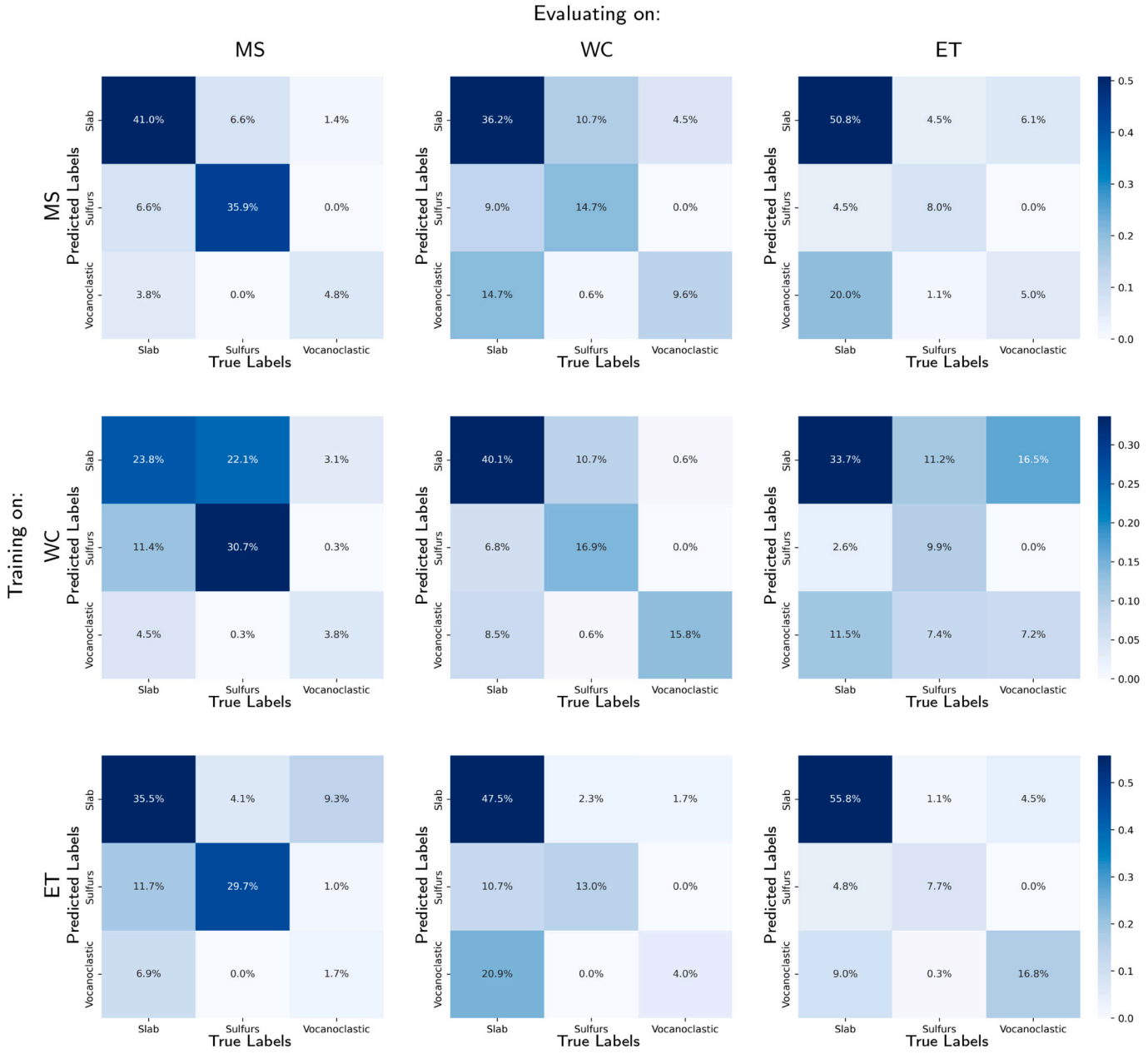


Fig. 4. Confusion matrices of the classification decision committee's predictions according to lithology categorization, in percentage and white to dark blue gradient. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

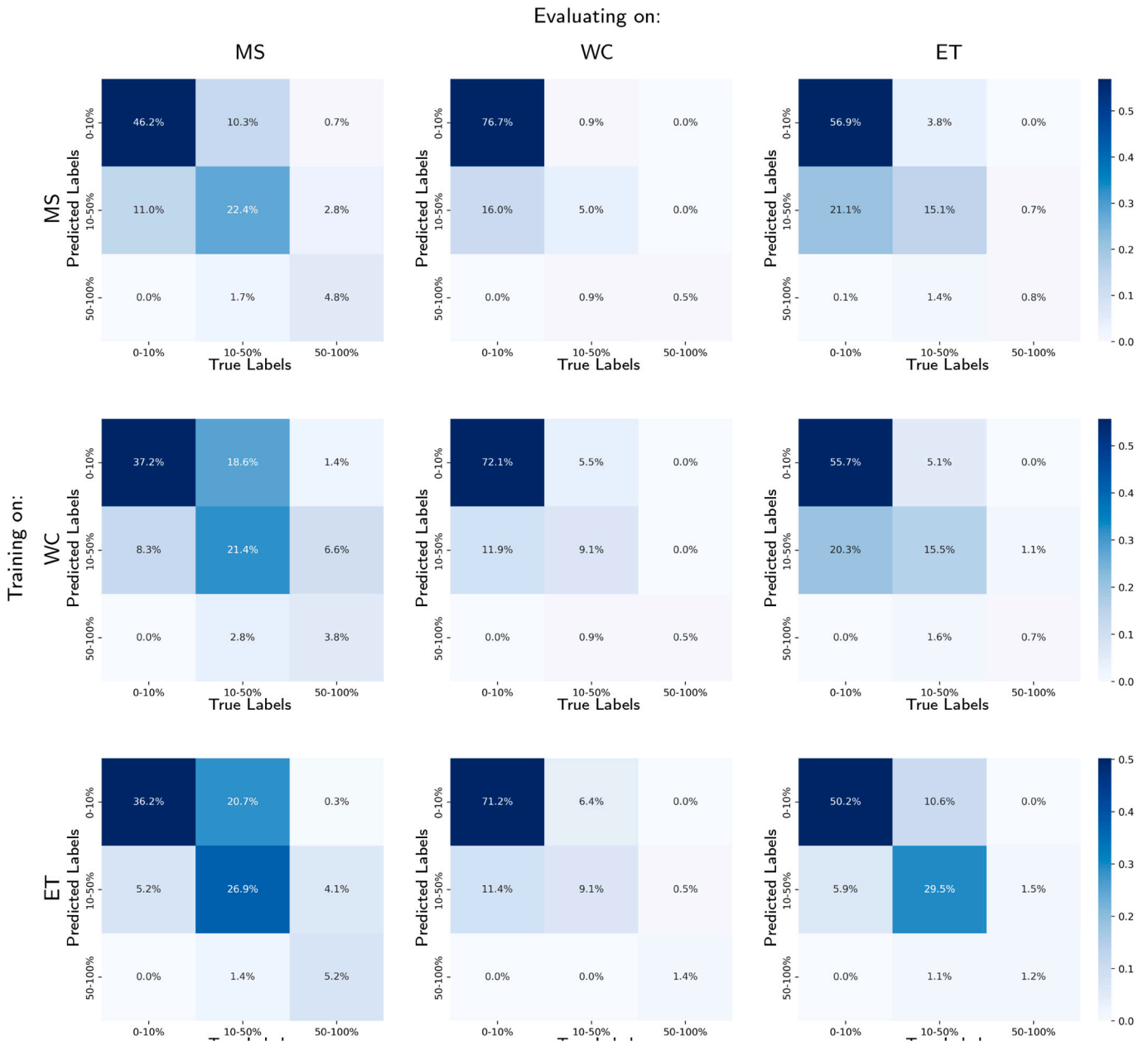


Fig. 5. Confusion matrices of the classification decision committee's predictions according to S/W fragments categorization, in percentage and white to dark blue gradient. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Class Activation Mapping (Grad-CAM)(Selvaraju et al., 2017) technique.

Fig. 6 shows examples of the activation maps computed by Grad-CAM over the VGG network trained and evaluated in the Montsegur (MS) site. In the figure, each column represents a substratum categorization, i.e., Fig. 6(a)(d)(g) corresponds to S/W fragments; Fig. 6(b)(e)(h) to lithology; and Fig. 6(c)(f)(i) to morphology. Below each image, we indicate the true label (TL) and the predicted label (PL). Additionally, the importance map is overlaid on the original images. The places where the gradients in such a map are strong, i.e., red-colored, represent precisely the information the network considers more important to make its prediction.

For S/W fragments, Fig. 6(a)(d)(g), it is easy to observe that the network focused only on the smaller white spots in the images (highlighted in green circles). Note that the network has not considered all the white groups of pixels (see Fig. 6(a)), which means that it has acquired a proper understanding of that class, e.g., quantifying small shells and white fragments, although not doing it with the desired precision in all cases, see Fig. 6(g).

Regarding lithology, the examples shown in Fig. 6(b)(e)(h) indicate that the network focuses on certain image regions to predict slab, sulfur, and volcanoclastic sediments. Regarding the slab, Fig. 6(b) clearly shows how the network looks for small boulders and rubbles notably harboring iron oxide deposits commonly found on the slab. Seabed slabs may be larger than the image extent, so the model may have looked for more detailed features within the images. Regarding sulfides, Fig. 6(e) indicates a focus on the spatial distribution of the fauna that usually relies on chemosynthetic primary production, permitted by the presence of hydrothermal vent emissions in the proximity of that substratum. For volcanoclastic sediments, as represented in Fig. 6(h), the network concentrates more on the presence of sediment patches harboring different colors. The visual representation under consideration reveals distinct features that are more conspicuous in certain regions of the image. Red circles have been employed to designate those regions. It is postulated that these regions are more likely to exhibit the said features compared

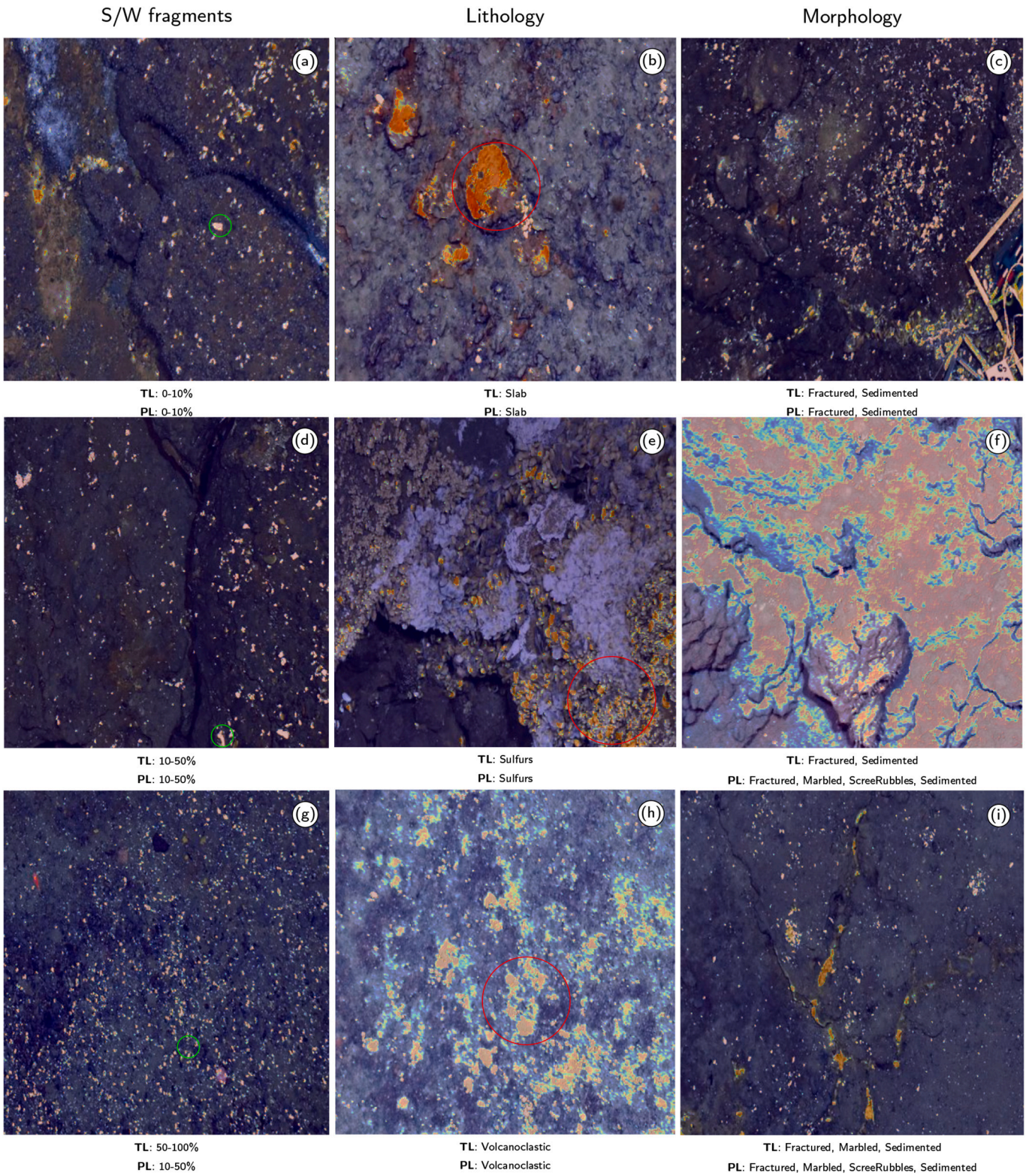


Fig. 6. Grad-CAM representations extracted from VGG architecture, on samples from all the studied regions and covering all categories, overlaid on the original images. The red-colored areas represent the information the network considers more important to make a prediction. In the figure, **TL** and **PL** refer to True Label and Predicted Label, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to other parts of the images.

Seabed substrata can be intrinsically made up of different morphological features. While those may be correlated with lithological characteristics, annotating several overlapping classes may prevent the model from clearly detecting the key features. While we considered geomorphological classes independently, this may not represent the whole image variability. The fact that they overlap provides much more combination and possible arrangement of critical features within the image, which is challenging to analyze in the current experiment.

5.4. Uncertainty analysis

The current section presents an uncertainty analysis of the substratum classification performed with the CDC. The analysis is based on the Predictive Entropy uncertainty measure, $H(\mathbf{s}|x_n)$, which indicates how confident the ensemble is with respect to its predictions. In simple terms, if most models in the ensemble agree with a particular prediction, the uncertainty associated with the respective sample is low; if most models disagree, the uncertainty is high. Furthermore, we hypothesize that uncertainty correlates negatively with classification accuracy, i.e., the higher the uncertainty, the higher the chance of the CDC decision being incorrect.

In this analysis, we hypothetically investigate a way to use the uncertainty computed for each sample (image) as part of an audit scheme

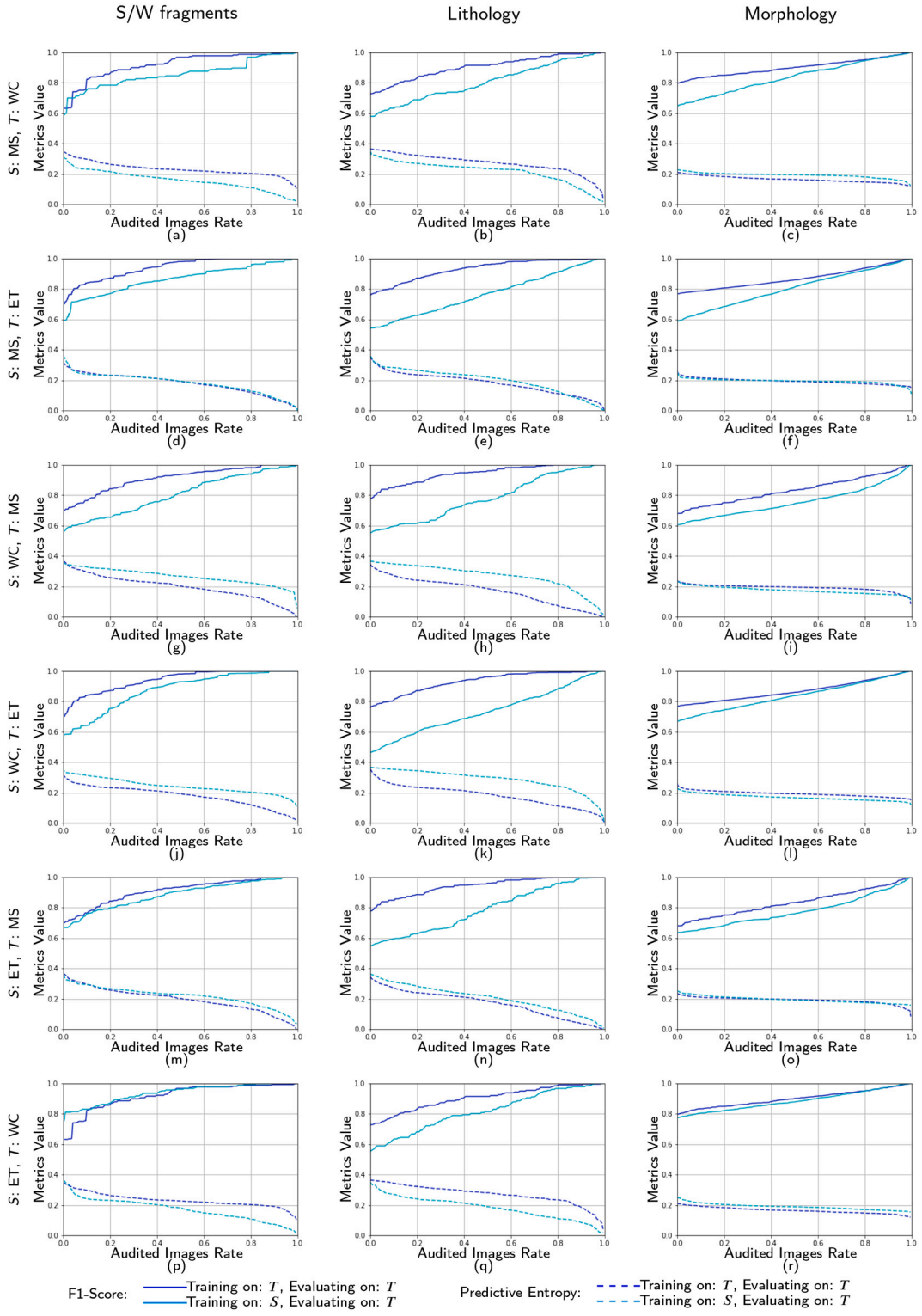


Fig. 7. Uncertainty analysis results. The curves represent the F1-score (solid) and Predictive Entropy (dashed) values versus the Audited Image Rate for the different substratum categorizations and site (training/testing) combinations.

that could be used within a semiautomatic classification procedure that may deliver arbitrarily high accuracy rates (assuming that human classification is always right). The basic idea is to submit only the samples with uncertainty above a specific threshold for the inspection of a human expert. We observe that such a procedure can be used to strongly reduce the human effort associated with creating an annotated image dataset, limiting the manual labeling task to the images with high uncertainty, according to an arbitrary threshold value.

In the analysis, we define a so-called Audited Image Rate (AIR), which represents the number of images with uncertainties above the threshold divided by the total number of images. First, we compute the uncertainty for each sample (image), and then we vary the threshold value from zero to the highest observed uncertainty value. For each uncertainty threshold value, we compute the correspondent AIR and F1-score. But note that when computing the F1-score, we consider that the samples whose predictions have uncertainties that lie above the uncertainty threshold value, i.e., samples to be audited, are correctly classified (once again, assuming that the manual labeling is always correct). Therefore, when the threshold value equals zero, all samples are audited, i.e., AIR equals one, and the corresponding F1-score is 100%. We present in Fig. 7 F1-score curves (continuous lines) jointly with the Predictive Entropy measures (dashed lines) versus the AIR.

In the figure, each column contains results for each substratum characterization criteria, i.e., the amount of S/W fragments, Fig. 7(a)(d)(g)(j)(m)(p); lithology, Fig. 7(b)(e)(h)(k)(n)(q); and morphology, Fig. 7(c)(f)(i)(l)(o)(r). Each row represents a different site combination for training and testing the CDC, which permits the assessment of the generalization capacity of the ensemble in cross-site evaluations. The dark blue curves represent the accuracy and uncertainty of the ensemble when it was trained and tested using images from the same site. The light blue curves represent cross-site combinations in which the CDC was trained with data from one site and evaluated with data from a different site.

In general, the CDC achieved F1-score values of about 80% when less than 20% of the images needed to be audited, with the ensemble trained and tested with data from the same site (dark blue curves in Fig. 7). Additionally, in most of those cases, F1-scores of over 90% were obtained with an AIR of 40%. Taking into account the result when no images are audited, i.e., the first point in the curve, those numbers indicate that the devised auditing scheme can significantly improve the overall classification accuracy by visually inspecting relatively small groups of the set of images.

Moreover, considering the cross-evaluation results (light blue curves), the classification accuracy also increases considerably as the number of audited images increases, reaching over 80% F1-score for an audited rate of 50% in almost all cases. The curves also show that the poorer generalization across sites occurs for the lithology categorization.

Specifically about the uncertainty values (dashed line curves in Fig. 7), in most cases, lower values were obtained when the classifiers were trained and tested in the same site, as expected. However, a different behavior was observed when the ensemble was evaluated on WC but trained on MS or ET from the lithology and S/W fragments points of view (see Fig. 7(a)(b)(p)(q)). Higher uncertainty was obtained when training and testing on WC than when training on MS or ET and testing on WC. That means the ensemble trained on WC was less confident than when trained on MS or ET. We hypothesize that such behavior has to do with the number of samples available in WC, which is lower than in the other sites. The former is consistent with the results presented in Table 4. Furthermore, being WC the most remote site, compared with the distance between ET and MS, it may harbor benthic substrata that may more strongly differ from those of ET and MS Barreyre et al. (2012).

Although expected, the ensemble predictions were highly confident and strongly correlated to the uncertainty measures, which confirms the usefulness of the proposed auditing scheme for efficiently annotating large image sets while limiting human intervention.

5.5. Uncertainty visual analysis

The visual analysis of image samples, considering their true and predicted labels, along with the associated predictive entropy, can provide some information on the features that might lead to wrong classification, representing sources of uncertainty. In Fig. 8, the top line (Fig. 8(a)(b)(c)) depicts pictures with correct predicted labels and low uncertainty, which clearly contain the features used by the CDC are highlighted in Fig. 6, such as the presence of small mussels for “sulfurs” (Fig. 8(b)). All three images come out as typical of the class they belong to. The second line (Fig. 8(d)(e)(f)) depicts images not properly labeled with an expected higher uncertainty. Finally, the last line (Fig. 8(g)(h)(i)) shows images properly labeled that appear characteristic of their class but are, nevertheless, associated with high predictive entropy, providing invaluable information on the interpretation of the uncertainty value.

When considering the S/W fragments category (Fig. 8(a)(d)(g)), uncertainty appears related to the size of the fragments and the contrast with the background. This contrast can be related to the color of the substrate, i.e., white volcanoclastic sediment vs. dark slab (Fig. 8(d)(g)), but also illumination. In Fig. 8(d), the sediment color and white fragments are hard to distinguish because of an evident lack of contrast. In Fig. 8(g), the strong gradient in colors with the presence of white microbial patches could explain the high predictive entropy despite a reasonable classification and the obvious presence of white fragments properly detected.

Regarding Lithology (Fig. 8(b)(e)(h)), the presence of outcropping slab among patches of volcanoclastic sediment supported the volcanoclastic predicted label with high uncertainty (Fig. 8(e)). Any human observer could easily classify this picture out of environmental context as volcanoclastic. Only the presence of emerging rocks in some parts of the image indicates the possible occurrence of slab. Again, the uncertainty analysis might easily assign this class to a higher predictive entropy because slabs are larger than the image extent. The last image (Fig. 8(h)), associated with high Predictive Entropy (PE), is harder to interpret, but considering the network concentrates on sediment patches of different colors for volcanoclastic sediments (refer to Fig. 6), the uncertainty in this case could be related to the size of patches with the presence of strong contrast over reduced areas, creating uncommon small features for this class.

Finally, morphology displays higher entropy compared to the two other categories ((Fig. 8(c)(f)(i))). This might result from the fact that several classes can be attributed to the same image, increasing the chances of inaccurate predictions. Fig. 8(f) depicts an old sulfide structure covered by sediment. The sediment cover could result in less pronounced texture corresponding to rubbles, or alternatively, the absence of flat/sedimented patches in the image prevents the CDC from detecting rubbles due to a lack of clear textural boundaries. Indeed, variation in color contrast and size of clear patches appear to be an important factor of certainty and most likely also account for the high PE value in Fig. 8(i). In this image, small patches associated with the dark areas, or alternatively the presence of brownish spots, might constitute important features for the network to consider, leading to high uncertainty despite the obvious presence of marbled sediment.

This result strongly highlights the importance of the human capacity to take into account a more general context of the spatial extent and the ability to distinguish between an unusual pattern and a characteristic feature. We observe that the ROV altitude can strongly affect the lighting and contrasts, leading to underexposed or overexposed images. While this is easily captured by the human eye, contrast variation in such a complex environment with fine/subtle changes between classes can strongly affect the networks’ decisions and lead to high uncertainty. In addition, the spatial scale of the categories is usually of one order of magnitude greater than the size of a single image. Hence, the human observer tends to classify an image by taking into account its environmental context. This is most likely the case in Fig. 8(e), which

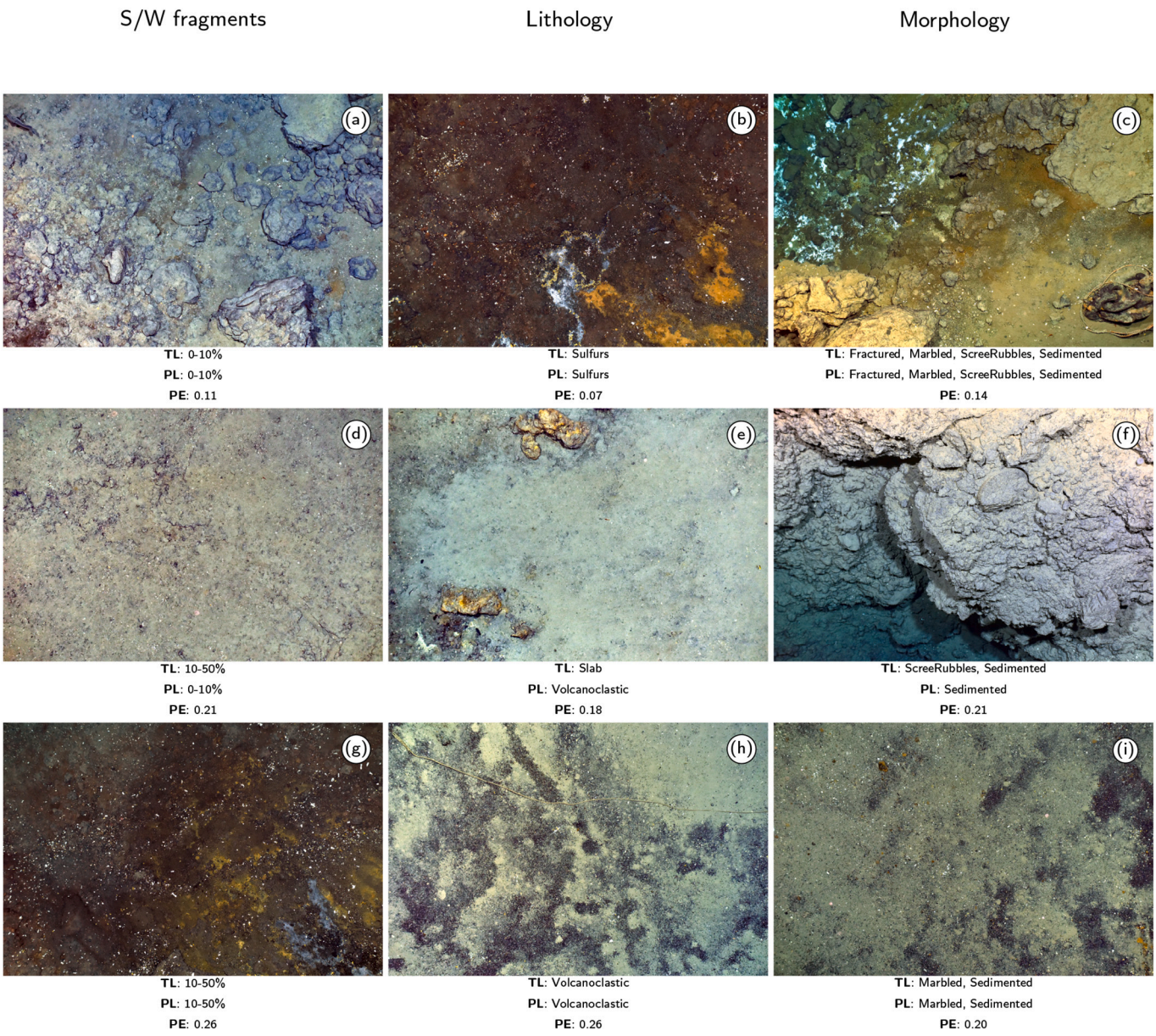


Fig. 8. Visual analysis of uncertainties. In the figure, **TL**, **PL**, and **PE** refer to True Label, Predicted Label, and Predictive Entropy, respectively.

corresponds to a transition zone between bare slab and slab covered by volcanoclastic sediments. This environmental surrounding context is currently not included in the decision scheme and could be added in the future to improve predictions.

6. Conclusions

In this work, six state-of-the-art deep learning architectures were evaluated for the problem of deep-sea substratum characterization. Additionally, a classification decision committee (CDC) constituted by an ensemble of networks with those architectures, in which the individual predictions are fused through a majority voting mechanism, was proposed and evaluated.

All deep learning models were evaluated on single and multi-label classification problems, and three different sites were considered in the experiments. Besides the conventional training scheme in which data from a single site is used for training and testing the models, we also evaluated the models in cross-site scenarios, aiming to assess the generalization capacity of the different architectures and that of the ensemble.

The experimental analysis, considering all class categories, demonstrated the suitability of the deep learning models for deep-sea substratum classification. Considering the lithology and morphology characterization, higher scores were achieved when the models were trained and tested on the same site. Regarding the cross-site evaluations, all results were impacted negatively by the inversion of the role between the training and the testing sets. However, the same trend was not observed for the shells and white (S/W) fragments substratum category. The issues underlying such behavior demand further investigation, as they could arise from a particular site and category characteristics. We believe such a generalization problem constitutes the main weakness of the proposed model in substrata image classification.

In this work, we also investigated the uncertainty in the classification conducted by the ensemble of networks regarding Predictive Entropy. Computed from the set of individual predictions of the CDC components, the uncertainty values helped to identify images with a higher chance of being misclassified, i.e., for which a high uncertainty value was obtained. Relying on the uncertainty information, we investigated a

possible semiautomatic procedure in which high-uncertainty images could be submitted to manual labeling, and after retraining the classifiers, accuracy would be improved. Indeed, the results demonstrated that high accuracy values could be obtained with such a procedure, with a relatively small amount of human intervention. In practice, the latter means that the procedure can assist a deep-sea expert annotator in a way that would not be necessary to review a complete amount of unlabeled images but only a small set of images with higher uncertainty.

The experimental results suggest that there might be room to improve the generalization capacity of the deep learning classifiers. That could be achieved by better exploiting the training data, e.g., with additional data augmentation techniques, using recent advances in unsupervised learning techniques, such as self-supervised methods, or by just simplifying the classifiers in terms of the number of parameters, thus reducing the risk of overfitting.

Another direction for continuing this research is to exploit the classification uncertainty further in an active learning context. The uncertainty measure could be used in interactively training the deep learning models, selecting high-uncertainty samples in the datasets, and increasing their importance in the computation of the loss function.

CRediT authorship contribution statement

Pedro Juan Soto Vega: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Panagiotis Papadakis:** Investigation, Conceptualization, Writing – review & editing, Writing – original draft, Supervision. **Marjolaine Matabos:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Investigation, Funding acquisition, Data curation. **Loïc Van Audenhaege:** Writing – original draft, Validation, Data curation. **Annah Ramiere:** Data curation. **Jozée Sarrazin:** Writing – original draft, Data curation. **Gilson Alexandre Ostwald Pedro da Costa:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We thank the captain and crew of the RV L'Atalante and the pilots of the ROV *Victor6000*. We also thank chief scientists Mathilde Cannat and Julien Legrand for co-leading the Momarsat 2018 cruise.

This work was supported by the European Union's Horizon 2020 research and innovation project iAtlantic under Grant Agreement No. 818123. This output reflects only the authors' view, and the European Union cannot be held responsible for any use that may be made of the information contained therein. The work is also supported by the region of Bretagne as part of the ABYSSES project. We also acknowledge financial support from the EU Project EMSO (http://www.emso-eu.org/), EMSO-France, and the French Observatory EMSO-Azores, funded by IFREMER and CNRS.

Finally, we acknowledge the support of the Brazilian Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

Appendix A. Network architectures

The network architectures of the models evaluated in the experiments are described in detail in Fig. 9. The light green areas in that figure represent feature extractor modules based on the different architectures, i.e., VGG, ResNet, and Xception. The orange area represents the architecture of the classifier module, to which the outputs of each feature extractor are submitted.

Each rectangle indicates the operations performed at a block of layers. *Conv* and *SConv* stand for regular convolution and depth-wise separable convolution, respectively. The values that follow such operations indicate the number of filters, filter size, stride, and dilation rate. Regarding the *maxpooling* operation, the values correspond to the kernel dimension and stride. *Dropout* refers to the number of neurons randomly turned off during each training inference. The number of neurons in the dense layer of the classifier module corresponds to the number of classes in the dataset.

All architectures were modified mainly in the block of fully connected layers, where we opted to use just the output layer after the feature extractor instead of two fully connected layers. Additionally, residual blocks were used at the architectures' stems rather than in the full feature extractor. All those modifications were experimentally determined.

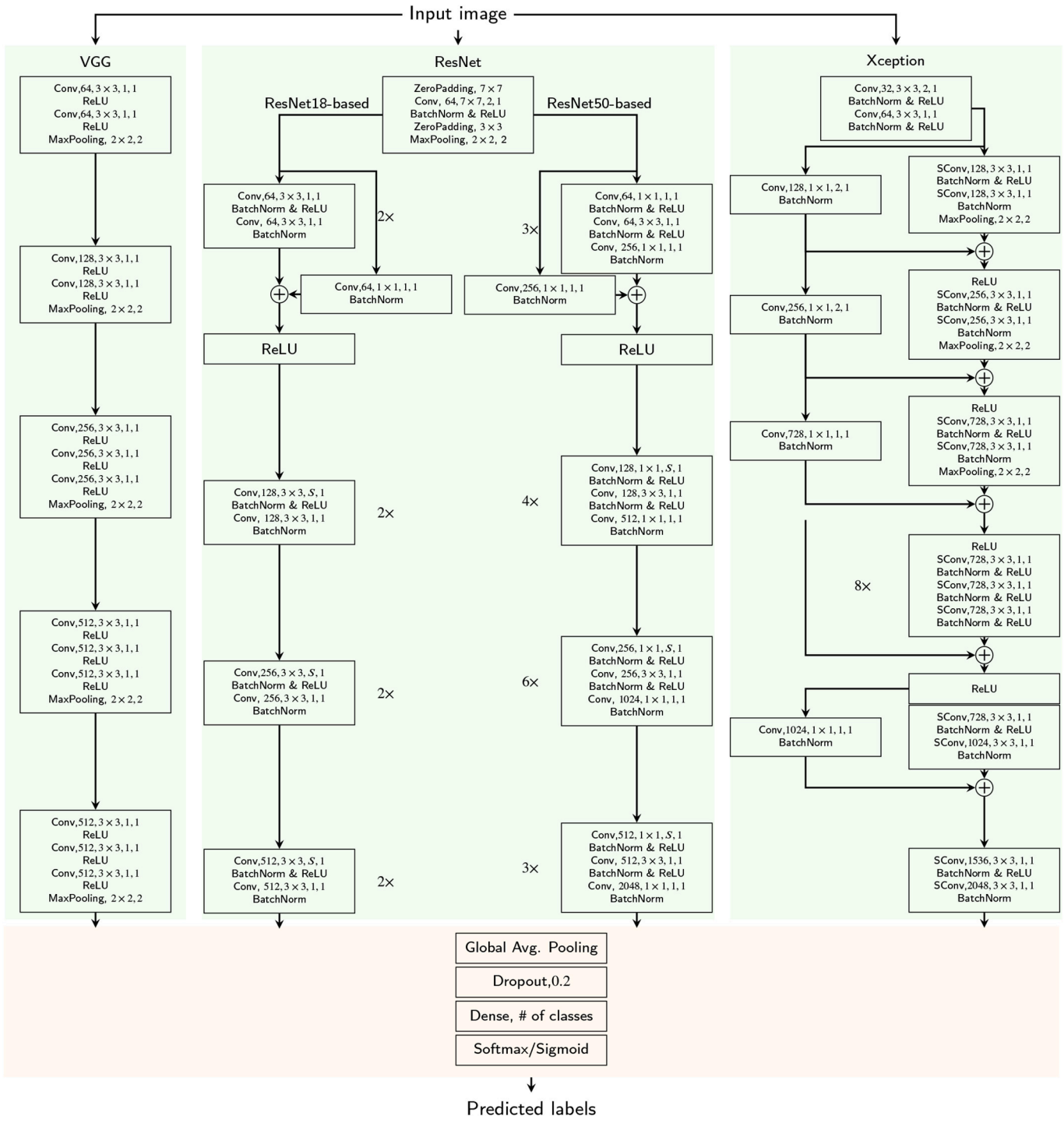


Fig. 9. Architecture details of the deep learning-based classifiers evaluated in this work. In the figure, each network architecture is represented in the green blocks separately. The orange block shows the fully connected layers at the end of all networks. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al., 2021. A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inform. Fusion 76, 243–297.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34, 2274–2281. https://doi.org/10.1109/TPAMI.2012.120.

Althaus, F., Hill, N., Edwards, L., Ferrari, R., Case, M., Colquhoun, J., Edgar, G., Fromont, J., Gershwin, L., Gowlett-Holmes, K., et al., 2013. Catami classification scheme for scoring marine biota and substrata in underwater imagery—a pictorial guide to the collaborative and annotation tools for analysis of marine imagery and video (catami) classification scheme. Version 1.3.

Arnaubec, A., Opderbecke, J., Allais, A.G., Brignone, L., 2015. Optical mapping with the ariane hrov at ifremer: the matisse processing tool. In: OCEANS 2015 - Genova, pp. 1–6. https://doi.org/10.1109/OCEANS-Genova.2015.7271713.

Barreyre, T., Escartín, J., Garcia, R., Cannat, M., Mittelstaedt, E., Prados, R., 2012. Structure, temporal evolution, and heat flux estimates from the lucky strike deep-sea

hydrothermal field derived from seafloor image mosaics. Geochem. Geophys. Geosyst. 13.
Bengio, Y., Yao, L., Alain, G., Vincent, P., 2013. Generalized denoising auto-encoders as generative models. Adv. Neural Inf. Proces. Syst. 1–9.
Boulard, M., Lawton, P., Baker, K., Edinger, E., 2022. The effect of small-scale habitat features on groundfish density in deep-sea soft-bottom ecosystems. In: Deep Sea Research Part I: Oceanographic Research Papers, p. 103891.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Cannat, M., Sarradin, P., 2018. Momarsat 2018 cruise, rv l'atalante. French Oceanogr. Cruis. 10, 18000514.
Chollet, F., 2017. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258.
Cooper, M., Elderfield, H., Schultz, A., 2000. Diffuse hydrothermal fluids from lucky strike hydrothermal vent field: evidence for a shallow conductively heated system. J. Geophys. Res. Solid Earth 105, 19369–19375.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: transformers for image recognition at scale arXiv: 2010.11929.
Durden, J.M., Hosking, B., Bett, B.J., Cline, D., Ruhl, H.A., 2021. Automated classification of fauna in seabed photographs: the impact of training and validation dataset size, with considerations for the class imbalance. Prog. Oceanogr. 196 https://doi.org/10.1016/j.pocean.2021.102612.
Faillettaz, R., Picheral, M., Luo, J.Y., Guigand, C., Cowen, R.K., Irisson, J.O., 2016. Imperfect automatic image classification successfully describes plankton distribution patterns. Methods Oceanogr. 15-16, 60–77. https://doi.org/10.1016/j.mio.2016.04.003.
Filippo, M.P., Gomes, O.D.F.M., da Costa, G.A.O.P., Mota, G.L.A., 2021. Deep learning semantic segmentation of opaque and non-opaque minerals from epoxy resin in reflected light microscopy images. Miner. Eng. 170, 107007.
Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2017. Domain-adversarial training of neural networks. Adv. Comp. Vision Patt. Recogn. 17, 189–209. https://doi.org/10.1007/978-3-319-58347-1_10.
Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al., 2023. A survey of uncertainty in deep neural networks. Artificial Intelligence 56, 1513–1589.
Gerdes, K., Arbizu, P.M., Schwentner, M., Freitag, R., Schwarz-Schampera, U., Brandt, A., Kihara, T., 2019. Megabenthic assemblages at the southern central indian ridge–spatial segregation of inactive hydrothermal vents from active-, periphery-and non-vent sites. Mar. Environ. Res. 151, 104776.
Girard, F., Sarrazin, J., Arnaubec, A., Cannat, M., Sarradin, P.M., Wheeler, B., Matabos, M., 2020. Currents and topography drive assemblage distribution on an active hydrothermal edifice. Prog. Oceanogr. 187, 102397.
Halpern, B.S., Frazier, M., Potapenko, J., Casey, K.S., Koenig, K., Longo, C., Lowndes, J. S., Rockwood, R.C., Selig, E.R., Selkoe, K.A., et al., 2015. Spatial and temporal changes in cumulative human impacts on the world's ocean. Nat. Commun. 6, 1–7.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-Decem, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
Hossain, M.D., Chen, D., 2019. Segmentation for Object-Based Image Analysis (OBIA): A Review of Algorithms and Challenges from Remote Sensing Perspective. https://doi.org/10.1016/j.isprsjprs.2019.02.009.
Husson, B., Sarradin, P.M., Zeppilli, D., Sarrazin, J., 2017. Picturing thermal niches and biomass of hydrothermal vent species. Deep-Sea Res. II Top. Stud. Oceanogr. 137, 6–25.
Ierodiaconou, D., Monk, J., Rattray, A., Laurenson, L., Versace, V.L., 2011. Comparison of automated classification techniques for predicting benthic biological communities using hydroacoustics and video observations. Cont. Shelf Res. 31 https://doi.org/10.1016/j.csr.2010.01.012.
Juliani, C., Juliani, E., 2021. Deep learning of terrain morphology and pattern discovery via network-based representational similarity analysis for deep-sea mineral exploration. Ore Geol. Rev. 129, 103936.
Kalmbach, A., Hoeberechts, M., Albu, A.B., Glotin, H., Paris, S., Girdhar, Y., 2016. Learning deep-sea substrate types with visual topic models. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1–9. https://doi.org/10.1109/WACV.2016.7477600.
Katija, K., Orenstein, E., Schlining, B., Lundsten, L., Barnard, K., Sainz, G., Boulais, O., Cromwell, M., Butler, E., Woodward, B., et al., 2022. Fathomnet: a global image database for enabling artificial intelligence in the ocean. Sci. Rep. 12, 1–14.
Kim, H., Loh, W.Y., 2001. Classification trees with unbiased multiway splits. J. Am. Stat. Assoc. 96, 589–604.
Kingma, D.P., 2017. Variational Inference & Deep Learning: A New Synthesis. PhD Thesis, pp. 1–162.
Langmuir, C., Jean-Luc, Charlou, Colodner, D., Daniel, Desbruyeres, Desonie, D., Emerson, T., Fornari, D., Yves, Fouquet, Humphris, S., Fiala-Medioni, A., Saldanha, L., Sours-Page, R., Thatcher, M., Tivey, M.K., Van Dover, C., von Damm, K., Wiese, K., Wilson, C., 1993. Lucky strike - a newly discovered hydrothermal site on the Azores platform. RIDGE Events 4, 3–5. URL: https://archimer.ifremer.fr/doc/00070/18096/.
Levin, L.A., Le Bris, N., 2015. The deep ocean under climate change. Science 350, 766–768.
Loh, W.Y., Shih, Y.S., 1997. Split selection methods for classification trees. Stat. Sin. 7, 815–840.
Lu, H., Uemura, T., Wang, D., Zhu, J., Huang, Z., Kim, H., 2020. Deep-sea organisms tracking using dehazing and deep learning. Mobile Netw. Appl. 25, 1008–1015.
Lucieer, V., Hill, N.A., Barrett, N.S., Nichol, S., 2013a. Do marine substrates 'look' and 'sound' the same? Supervised classification of multibeam acoustic data using autonomous underwater vehicle images. Estuar. Coast. Shelf Sci. 117, 94–106. https://doi.org/10.1016/j.ecss.2012.11.001.
Lucieer, V., Hill, N.A., Barrett, N.S., Nichol, S., 2013b. Do marine substrates 'look'and 'sound'the same? Supervised classification of multibeam acoustic data using autonomous underwater vehicle images. Estuar. Coast. Shelf Sci. 117, 94–106.
Ma, C., Li, X., Li, Y., Tian, X., Wang, Y., Kim, H., Serikawa, S., 2021. Visual information processing for deep-sea visual monitoring system. Cognit. Robot. 1, 3–11.
Marcon, Y., Sahling, H., Allais, A.G., Bohrmann, G., Olu, K., 2014. Distribution and temporal variation of mega-fauna at the r egab pockmark (n orthern c ongo f an), based on a comparison of videomosaics and geographic information systems analyses. Mar. Ecol. 35, 77–95.
Marticorena, J., Matabos, M., Ramirez-Llodra, E., Cathalot, C., Laes-Huon, A., Leroux, R., Hourdez, S., Donval, J.P., Sarrazin, J., 2021. Recovery of hydrothermal vent communities in response to an induced disturbance at the lucky strike vent field (mid-Atlantic ridge). Mar. Environ. Res. 168, 105316.
Matabos, M., Barreyre, T., Juniper, S.K., Cannat, M., Kelley, D., Alfaro-Lucas, J.M., Chavagnac, V., Colaço, A., Escartin, J., Escobar, E., et al., 2022. Integrating multidisciplinary observations in vent environments (imove): decadal progress in deep-sea observatories at hydrothermal vents. Front. Mar. Sci. 660.
McEver, R.A., Zhang, B., Levenson, C., Iftekhar, A., Manjunath, B., 2023. Context-driven detection of invertebrate species in deep-sea video. Int. J. Comput. Vis. 131, 1367–1388.
Meyer, H.K., Roberts, E.M., Rapp, H.T., Davies, A.J., 2019. Spatial patterns of arctic sponge ground fauna and demersal fish are detectable in autonomous underwater vehicle (auv) imagery. Deep-Sea Res. I Oceanogr. Res. Pap. 153, 103137.
Neufeld, M.B., Metaxas, A., Jamieson, J.W., 2022. Non-vent megafaunal communities on the endeavour and middle valley segments of the juan de fuca ridge, northeast pacific ocean. Front. Mar. Sci. 804.
Ojala, T., Pietikäinen, M., Mäenpää, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24, 971–987. https://doi.org/10.1109/TPAMI.2002.1017623.
Ondréas, H., Cannat, M., Fouquet, Y., Normand, A., Sarradin, P.M., Sarrazin, J., 2009. Recent volcanic events and the distribution of hydrothermal venting at the lucky strike hydrothermal field, mid-atlantic ridge. Geochem. Geophys. Geosyst. 10.
Osterloff, J., Nilssen, I., Nattkemper, T.W., 2016. A computer vision approach for monitoring the spatial and temporal shrimp distribution at the LoVe observatory. Methods Oceanogr. 15-16, 114–128. https://doi.org/10.1016/j.mio.2016.03.002.
Piechaud, N., Howell, K.L., 2022. Fast and accurate mapping of fine scale abundance of a vme in the deep sea with computer vision. Eco. Inform. 71, 101786. URL. https://www.sciencedirect.com/science/article/pii/S1574954122002369. https://doi.org/10.1016/j.ecoinf.2022.101786.
Pillay, T., Cawthra, H., Lombard, A., 2020. Characterisation of seafloor substrate using advanced processing of multibeam bathymetry, backscatter, and sidescan sonar in table bay, South Africa. Mar. Geol. 429, 106332.
Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: unified, real-time object detection. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 779–788.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y.
Schmid, M.S., Aubry, C., Grigor, J., Fortier, L., 2016. The LOKI underwater imaging system and an automatic identification model for the detection of zooplankton taxa in the Arctic Ocean. Methods Oceanogr. 15-16, 129–160. https://doi.org/10.1016/j.mio.2016.03.003.
Schoening, T., Bergmann, M., Ontrup, J., Taylor, J., Dannheim, J., Gutt, J., Purser, A., Nattkemper, T.W., 2012. Semi-automated image analysis for the assessment of megafaunal densities at the artic deep-sea observatory HAUSGARTEN. PLoS One 7. https://doi.org/10.1371/journal.pone.0038179.
Schoening, T., Osterloff, J., Nattkemper, T.W., 2016. Recomia—recommendations for marine image annotation: lessons learned and future directions. Front. Mar. Sci. 3, 59.
Schoening, T., Durden, J., Preuss, I., Branzan Albu, A., Purser, A., De Smet, B., Dominguez-Carrió, C., Yesson, C., de Jonge, D., Lindsay, D., et al., 2017a. Report on the Marine Imaging Workshop 2017.
Schoening, T., Jones, D.O., Greinert, J., 2017b. Compact-morphology-based poly-metallic nodule delineation. Sci. Rep. 7, 1–12.
Schoening, T., Köser, K., Greinert, J., 2018. An acquisition, curation and management workflow for sustainable, terabyte-scale marine image analysis. Sci. Data 5, 1–12.
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: visual explanations from deep networks via gradient-based localization. Proc. IEEE Int. Conf. Comp. Vision 618–626.
Sharma, R., Sankar, S.J., Samanta, S., Sardar, A.A., Gracious, D., 2010. Image analysis of seafloor photographs for estimation of deep-sea minerals. Geo-Mar. Lett. 30, 617–626. https://doi.org/10.1007/s00367-010-0205-z.
Simon-Lledó, E., Bett, B.J., Huvenne, V.A., Köser, K., Schoening, T., Greinert, J., Jones, D. O., 2019a. Biological effects 26 years after simulated deep-sea mining. Sci. Rep. 9, 1–13.

Simon-Lledó, E., Bett, B.J., Huvenne, V.A., Schoening, T., Benoist, N.M., Jones, D.O., 2019b. Ecology of a polymetallic nodule occurrence gradient: implications for deep-sea mining. Limnol. Oceanogr. 64, 1883–1894.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, pp. 1–14.

Sivic, J., Zisserman, A., 2003. Video google: a text retrieval approach to object matching in videos. In: Proceedings of the IEEE International Conference on Computer Vision. Institute of Electrical and Electronics Engineers Inc, pp. 1470–1477. https://doi.org/10.1109/iccv.2003.1238663.

Song, W., Zheng, N., Liu, X., Qiu, L., Zheng, R., 2019. An improved u-net convolutional networks for seabed mineral image segmentation. IEEE Access 7, 82744–82752.

Soto Vega, P.J., Costa, G.A., Feitosa, R.Q., Ortega Adarme, M.X., Almeida, C.A.D., Heipke, C., Rottensteiner, F., 2021. An unsupervised domain adaptation approach for change detection and its application to deforestation mapping in tropical biomes. ISPRS J. Photogramm. Remote Sens. 181, 113–128. URL. https://linkinghub.elsevier.com/retrieve/pii/S092427162100232X. https://doi.org/10.1016/J.ISPRSJPRS.2021.08.026.

Soto, P.J., Costa, G.A., Feitosa, R.Q., Ortega, M.X., Bermudez, J.D., Turnes, J.N., 2022. Domain-adversarial neural networks for deforestation detection in tropical forests. IEEE Geosci. Remote Sens. Lett. 19 https://doi.org/10.1109/LGRS.2022.3163575.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.

Taylor, J., Krumpen, T., Soltwedel, T., Gutt, J., Bergmann, M., 2017. Dynamic benthic megafaunal communities: assessing temporal variations in structure, composition and diversity at the arctic deep-sea observatory hausgarten between 2004 and 2015. Deep-Sea Res. I Oceanogr. Res. Pap. 122, 81–94.

van den Beld, I.M., Bourillet, J.F., Arnaud-Haond, S., de Chambure, L., Davies, J.S., Guillaumont, B., Olu, K., Menot, L., 2017. Cold-water coral habitats in submarine canyons of the bay of Biscay. Front. Mar. Sci. 4, 118.

Vandromme, P., Lars, S., Garcìa-Comas, C., Berline, L., Sun, X., Gorsky, G., 2012. Assessing biases in computing size spectra of automatically classified zooplankton from imaging systems: a case study with the ZooScan integrated system. Methods Oceanogr. 1-2, 3–21. https://doi.org/10.1016/j.mio.2012.06.001.

Vanhoucke, V., 2014. Learning visual representations at scale. In: ICLR Invited Talk 1. URL. http://vincent.vanhoucke.com.

Villon, S., Mouillot, D., Chaumont, M., Darling, E.S., Subsol, G., Claverie, T., Villéger, S., 2018. A deep learning method for accurate and fast identification of coral reef fishes in underwater images. Eco. Inform. 48, 238–244. https://doi.org/10.1016/j.ecoinf.2018.09.007.

Xue, B., Huang, B., Chen, G., Li, H., Wei, W., 2021. Deep-sea debris identification using deep convolutional neural networks. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 14, 8909–8921. https://doi.org/10.1109/JSTARS.2021.3107853.