Optimizing Plankton Image Classification with Metadata-Enhanced Representation Learning

Mojtaba Masoudi, Sarah L.C. Giering, Noushin Eftekhari, Miquel Massot-Campos, Jean-Olivier Irisson, and Blair Thornton *Member, IEEE*

***Abstract*—Automated camera-based sensors are widely used in vessel-based research to monitor plankton and marine particles. However, current methods suffer from the costly and time-consuming requirement of annotating data for fully supervised learning, especially in plankton grouping tasks characterized by long-tailed datasets. In response, we propose a novel self-supervised learning (SSL) framework that significantly reduces reliance on expensive human annotations by leveraging crucial metadata such as water depth and location. The method comprises three major steps: self-supervised training, innovative sampling, and final classification. It identifies key sample subsets from an unlabelled dataset using hierarchical clustering approach and incorporates an innovative balancing representative subsampling strategy that addresses the challenge of dataset imbalance and enhances generalisability across diverse plankton classes. Our approach prioritises discerning representation features observed in images that exhibit correlations with the patterns found in their associated metadata. Furthermore, our method introduce a novel grouping based on visual perspective selection method, enabling the identification of balanced subset views that depart from traditional class-based categorisation. Our experimental results showcase a significant enhancement in image classification accuracy, with a 23% improvement over methods that do not utilise metadata, and attains a macro F1-score of 54% for 10 populated species from a severely long-tailed dataset. This is achieved with a mere 0.3% of the entire dataset used for annotation.**

***Index Terms*—self-supervised learning, representation learning, convolutional neural network, plankton, marine imaging.**

I. Introduction

THE critical role of plankton as the foundation of aquatic food webs underscores the necessity of their monitoring [1], [2]. Traditionally (and still prevalent today), monitoring involves manual sampling with plankton nets and subsequent microscopic analysis, which is time-consuming and labour-intensive. These methods often result in years between sample collection and data availability, and sampling itself can be limited by weather conditions and access to sampling sites. To overcome these limitations and to make plankton monitoring more efficient and effective, significant efforts have been dedicated to developing automated quantitative imaging equipment (see references in [3] and [4]) like the Underwater Vision Profiler (UVP) [5], which provides detailed optical information on individual plankton, enhancing the efficiency and scope of monitoring efforts.

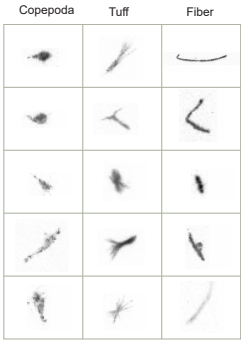


Fig. 1. Visual examples of plankton classes: Copepoda, Tuff-like, and Fiber-like. Given the low image quality, distinguishing across classes will be difficult.

Despite these technological advances, the processing and interpretation of data from imaging systems are fraught with challenges, as illustrated in Fig1, where the ambiguity in many images demonstrates the difficulty of plankton classification. Plankton images are frequently of poor quality due to the microscopic size of the organisms, optical water distortions, and movement from both the camera and the plankton, resulting in unclear images difficult for even skilled taxonomists to classify. Moreover, the presence of numerous non-plankton objects like 'marine snow' (detritus particles) creates a significant imbalance, with dominant classes overshadowing rare species that may be critical for ecosystem assessment. To efficiently handle the rapidly growing volumes of plankton imagery, there is an increasing reliance on computer-aided workflows that streamline the annotation process, enhancing the accuracy of plankton identification [6], [7].

This paper introduces a novel learning framework em-

Manuscript received February 12, 2024. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101000858 (TechOceanS). This output reflects only the author's view and the Research Executive Agency (REA) cannot be held responsible for any use that may be made of the information contained therein. (Corresponding author: Sarah L.C. Giering)

Mojtaba Masoudi and Sarah L.C. Giering are with the Department of Data, Science and Technology (DST), National Oceanography Centre, Southampton, SO14 3ZH, United Kingdom (e-mail: mojmas,s.giering@noc.ac.uk).

Noushin Eftekhari is with The Alan Turing Institute, London, NW1 2DB, United Kingdom (e-mail: neftekhari@turing.ac.uk).

Miquel Massot-Campos, and Blair Thornton are with the Faculty of Engineering and Physical Science, University of Southampton, Southampton, SO16 7QF, United Kingdom (e-mail: miquel.massot-campos,b.thornton@soton.ac.uk).

Jean-Olivier Irisson is with Laboratoire d'Océanographie de Villefranche, Sorbonne Université, 06230 Villefranche-sur-Mer, France (e-mail: jean-olivier.irisson@imev-mer.fr).

ploying self-supervised learning (SSL) techniques to enhance understanding of plankton data by leveraging crucial metadata such as water depth and location. The framework is designed to minimise human annotation reliance, integrating advanced machine learning strategies specifically tailored for plankton image recognition. By incorporating metadata directly into the learning process, the model significantly improves its ability to generate low-dimensional latent representations, aiding in the semantic interpretation of marine biodiversity. This approach not only ensures sensitivity to ecological variations across different marine environments but also facilitates a deeper and more robust classification of diverse and rare plankton species.

Our model addresses the challenge of imbalanced datasets by employing a novel balancing representative subsampling approach, which minimises the need for extensive manual annotations. It utilises a semi-supervised workflow that leverages both labelled and unlabeled data, with the latter being augmented by pseudo-labels generated through the model's self-learning capabilities. Through the strategic use of metadata and innovative hierarchical techniques, our framework establishes a new standard for automated plankton classification, offering substantial improvements in efficiency and accuracy while effectively tackling the challenges posed by the vast diversity and imbalance found in marine species datasets.

Here, the concept of 'views'—different visual perspectives or characteristics within the same category—is pivotal. By strategically selecting representative subsamples from each view, the model enhances its ability to generalise across the vast and heterogeneous plankton classes more effectively than traditional methods. This is particularly critical given the high degree of intra-class variability and the presence of rare species, which often complicate classification tasks. In summary, our proposed method makes the following key contributions.

- *Enhanced SSL with Metadata Regularisation:* We explore the effectiveness of SSL metadata regularisation in learning meaningful representations of plankton images. This process involves reconstructing images while embedding prior information to group visually similar patterns together, thus enhancing the robustness of the learned features.
- *Semantic Mapping and Retrieval Applications:* The learned features are utilised in semantic mapping applications, including hierarchical clustering and content-based image retrieval. This approach provides a richer contextual understanding of the images, enabling more nuanced interpretations and classifications.
- *Innovative Representative Sampling for Imbalance Management:* We introduce a novel representative sampling selection strategy designed to address significant imbalances in the data. This method prioritises discerning representation without extensive human-supervised labelling, substantially streamlining the annotation process and reducing the workload on human taxonomists.
- *Handling Rare and Diverse Species:* Our framework excels in identifying rare species and managing the substantial diversity and feature overlap inherent in plankton images.
- *Visual Perspective Selection for Improved Classification:* We leverage novel grouping based on visual perspective selection by introducing the concept of 'view' to achieve more accurate classification performance.

II. Background

A. Conventional methods

Traditional research in automatic plankton and particle classification has predominantly employed handcrafted methods that focus on low-level visual attributes like size, morphology, SIFT, and LBP, assuming these are key for distinguishing classes [8]–[15]. The rise of deep learning (DL) has, however, questioned the efficiency of these methods by introducing algorithms that transform raw data into feature vectors, capturing essential invariances that handcrafted features often miss [16]–[18]. DL's ability to learn abstract representations and adapt across diverse datasets without domain-specific expertise marks a significant shift towards more flexible and generalisable methods [19]. Supervised DL methods are particularly favored for their robust capacity to learn and distinguish complex visual features. These methods streamline the classification process by integrating data representation and classification into a single end-to-end workflow, thus bypassing the need for extensive feature engineering and parameter tuning [18], [20]–[25]. [20] applied ResNet-32 to improve throughput and efficiency in plankton image classification. Meanwhile, [21] conducted an extensive exploration of various CNN configurations, integrating an inception layer designed to handle multi-size input images. [22] proposed ZooplanktoNet, testing the effects of data augmentation and the number of convolutional layers. [23] introduced a model using multiple image views, combining the original image with versions processed through Gaussian filters to slightly improve accuracy.

However, supervised ML models demonstrate poor performance, generalisation failures, and biases to the extremely imbalanced dataset [26]. Many efforts, such as augmentation strategies, try to alleviate some of the problems but still, highly correlated data lead to unacceptable model performance [24]. Addressing class imbalance, [24] used a simplified CNN derived from AlexNet, pre-training with class-normalized samples to boost performance. Additionally, CGAN-like architectures [27] were employed to augment under-represented classes, and a two-stage training process was used where a CNN pre-trained on the least represented categories was later integrated into a full dataset training setup [25]. [28] used a method called background resampling, which involves hard negative mining to downsample background data, creating a more balanced dataset for training.

Despite advances in supervised learning, its application in plankton research faces significant challenges. A primary challenge is the lack of extensive annotated datasets. Researchers typically work in isolation, on relatively small datasets collected with different instruments or instrument settings, and following different taxonomic choices (e.g. differing taxonomic resolution or naming conventions). Key species can furthermore be rare [29], resulting in low numbers of training data for these critical groups. Moreover, in many real-world

scenarios, it is impractical to create extensive labelled training datasets because of the vast diversity and unpredictability of marine environments. To address the constraints on human resources, incorporating unlabeled data into the training process has become a common strategy.

Overall, these challenges— severe imbalanced data, constraints on human resources, and the complexities of dynamic marine environments— underscore the necessity for innovative approaches that can more effectively utilise existing data and enhance generalisation across diverse conditions and datasets.

B. Innovative methods for reducing annotation

Research into learning representations using unlabeled and 'few labelled' data has led to the development of several research fields, including unsupervised, semi-supervised, and self-supervised [30]. These fields aim to bypass the laborious task of data annotation by developing representations that can generalise across different learning tasks [31]. Particularly, self-supervised learning (SSL) leverages pre-existing or inherently available labels, eliminating the need for explicit human annotations. In SSL, features are acquired through a simple pretext or proxy task defined for the network to solve during training [32]. Typically, models like CNNs or increasingly Vision Transformers (ViTs) [33] are trained on a large corpus of unlabeled data by optimizing a self-generated objective. This process allows the model to capture high-level data representations without manual labels, which are then adaptable for supervised tasks in practical applications. [34] categorizes SSL approaches into three main types: generative, context-based, and contrastive methods.

Generative methods, including autoencoders [35] and generative adversarial networks (GANs) [27], focus on recreating or generating the input data, and have been applied to various data types such as multispectral and hyperspectral images [36], [37]. Context-based methods leverage the contextual features within images, with techniques designed to exploit aspects like context similarity [38] and spatial structure [39]. A pioneering method in this category involved predicting the relative positions of image patches to understand their spatial relationships [40]. Foundation models such as SpectralGPT [41], which are crafted for processing spectral data, can combine features of both generative and context-based methods within the SSL framework [42].

Contrastive methods improve a model's ability to identify similarities among semantically related inputs without relying on specific single pretext tasks. These methods train models by comparing semantically identical inputs, such as two augmented views of the same image, and encouraging similar representation in the embedding space. A notable advance in this field is the development of SimCLR [43], a simple framework for contrastive learning of visual representations, which has shown promising results in hyperspectral image classification with minimal labels [44].

Additionally, clustering-based SSL methods group similar features together in the embedding space using clustering algorithms like K-means, as demonstrated in DeepCluster [45]. This approach generates pseudolabels that help in training models to predict these labels effectively. Techniques such as training autoencoders with additional loss functions to enhance clustering in the latent space have also been developed, showcasing their utility in applications like seafloor imaging [46]–[48].

Metadata can offer contextual information that enhances data categorization by emphasizing relevant features and patterns [49]. For example, in conventional microscopy-based taxonomy, details such as the water depth and collection location of an organism assist in its classification. In the study by [46], the researchers introduced a Location Guided Autoencoder (LGA) that utilises horizontal location information to regulate learning.

Some researchers remain unclear whether the high-level representations shaped by category-level influences are comparable to those our visual system uses to recognise and distinguish the myriad of objects we encounter daily [50]. The specific knowledge governing these visual representations in the human brain is still not fully understood [51]–[53], leading some to question the suitability of category-level forces as proxies in understanding these representations and to consider alternative theories. Our analysis of plankton data supports this perspective, which reveals greater similarities between images of plankton and marine particles across different classes than within the same class. Consequently, we explore whether self-supervised grouping of a representative subset of images could improve classification accuracy. Throughout this paper, we refer to these groups as 'views' within each class, where 'view' in computer vision usually refers to the visual perspective or angle of image capture. This terminology is vital for our research as it aids in categorizing images by their visual presentation. Our model's hierarchical structure is devised to focus on distinguishing features within each plankton view, rather than forcing it to recognize features across all classes [54], [55].

III. METHODOLOGY

We developed a metadata-driven representation learning workflow using SSL to pinpoint key samples for understanding category distinctions of interest and apply this to plankton image classification problems, as illustrated in (Fig2). Initially, inputs are passed through an encoder-decoder network, leveraging a convolutional backbone to learn the representations, detailed in Section A. Next, a data sampling technique is employed on these latent representations to refine the number of dominant classes into a manageable number of representative samples, ensuring diverse coverage that includes a variety of types, particularly infrequent species, as described in Section B. These features are then organised into clusters to distinguish among various class views. An unsupervised clustering algorithm selects m key samples from these clusters, where $m \in (1...N_c)$ and N_c represents the total number of samples per cluster. We demonstrate that this approach can yield effective results even with smaller sample sizes. Subsequently, based on these key samples, pseudo-labels are assigned. These annotations are then used in the final phase of the downstream task, as outlined in Section C.

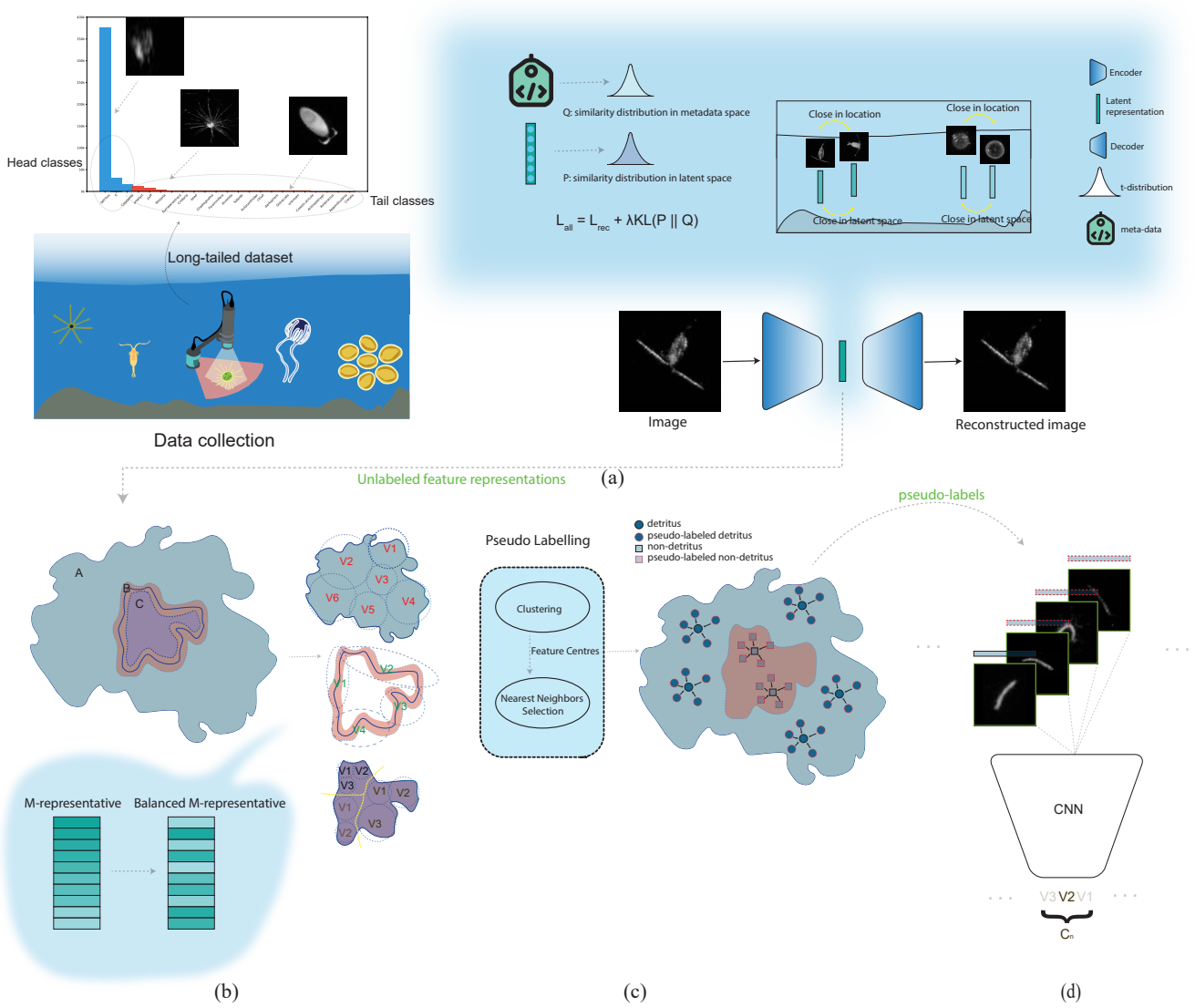


Fig. 2. An overview of the study design. (a) Following data collection, our augmented data was processed through an encoder-decoder network, enabling regularisation with the metadata. The KL divergence was assessed by evaluating the mutual information between the latent space's similarity distribution and the metadata space. (b) Here, the data distribution is depicted. This step categorises the representation into three groups, reducing the impact of the predominant detritus class, which constitutes 80% of the data. One group is entirely non-detritus, another is a borderline samples that is difficult to discern, and the third is composed solely of detritus samples. (c) Select representative samples were annotated by human experts, and pseudo-labels were assigned to all unlabeled samples. (d) These annotations and pseudo-labels were then incorporated into the CNN architecture for fine-tuning.

A. Self-supervision using metadata-guided autoencoder

The concept of autoencoders was initially presented in [35] as a neural network designed to learn the reconstruction of its input. The primary mechanism of an autoencoder involves two key processes: encoding and decoding. In the encoding phase, the network transforms the input data into a compressed latent space representation, and in the decoding phase, it attempts to reconstruct the input data from this compressed representation. The ultimate goal of an autoencoder is to minimise the reconstruction error, making the output as close as possible to the original input. Formally, as described in [56], the objective of an autoencoder can be encapsulated by the functions $\mathbf{A} : \mathbb{R}^n \rightarrow \mathbb{G}^p$ and $\mathbf{B} : \mathbb{G}^p \rightarrow \mathbb{R}^n$, where $\mathbf{A}$ is the encoder and $\mathbf{B}$ is the decoder. Here, $\mathbb{R}$ and $\mathbb{G}$ represent the sets from which the data vectors and their encoded representations are drawn, respectively, with n and p being positive integers such that $0 < p < n$. The effectiveness of these functions is measured through the minimisation of the expected value of a dissimilarity function Δ, which quantifies the difference between the original input and its reconstruction:

$$\arg\min_{A,B} \mathrm{E}[\Delta(\mathbf{B} \circ \mathbf{A}(x), x)] \tag{1}$$

In this equation, x represents the input data, $h = \mathbf{A}(x)$ is the encoded latent representation and $x_{rec} = \mathbf{B}(h)$ is the reconstructed data. The function E denotes the expectation over the distribution of x, and Δ is the dissimilarity function, typically

a loss function like mean squared error, that measures how well the autoencoder is performing in terms of reconstructing the original input from the encoded representation. While autoencoders primarily focus on accurate input reconstruction, it is equally crucial for the low-dimensional representation to encompass meaningful and generalisable features. Utilizing an auxiliary data-driven target function in the SSL context can accelerates the learning process within the embedding space [57]. For instance, in environmental monitoring, integrating geo-referenced information into the loss function acts as an effective regularisation strategy, enhancing model performance [46]. The overall loss function in such a scenario is designed to minimise both the reconstruction loss and a regularisation term. Specifically, it is defined as follows:

$$L_{all} = L_{rec} + \lambda KL(P \parallel Q) \tag{2}$$

Here, L_{rec} represents the reconstruction loss, and the term involving the Kullback-Leibler (KL) divergence functions as the regularisation component. The KL divergence measures the discrepancy between two probability distributions, P and Q, derived from the encoded latent space and the associated metadata, respectively. This regulariser aims to align the distribution P, characterizing the similarity between data points in the latent space, with the distribution Q, which is informed by external metadata.

Here, L_{rec} represents the reconstruction loss, and the term involving the KL divergence functions as the regularisation component. The KL divergence measures the discrepancy between two probability distributions, P and Q, derived from the encoded latent space and the associated metadata, respectively. This regulariser aims to align the distribution P, characterizing the similarity between data points in the latent space, with the distribution Q, which is informed by external metadata. An effective feature learner aims to minimise the distance between similar data points' embeddings, denoted as h_i and h_j, within the latent representation space. This means that if the original data, x_i and x_j, are similar, their corresponding embeddings should be positioned closer together. However, considering the similarity of just x_i and x_j might not fully represent their relationship; the associated metadata y_i and y_j also play a crucial role. Therefore, it is common to employ the Student's t distribution [58] as a kernel for quantifying the affinity or similarity between data points in a transformed, lower-dimensional space. This kernel helps ensure that points closer in the original space are also proximate in the embedded space, thus maintaining topological fidelity. The probability distribution P_{ij}, representing the relationship between embeddings of samples i and j, can be formally expressed as follows:

$$P_{ij} = \frac{(1 + \|\mathrm{h_i} - \mathrm{h_j}\|^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{i'}\sum_{j'}(1 + \|\mathrm{h_{i'}} - \mathrm{h_{j'}}\|^2/\alpha)^{-\frac{\alpha+1}{2}}} \tag{3}$$

For the metadata-driven distribution Q_{ij}, the following formulation is used to capture the similarity between metadata samples i and j:

$$Q_{ij} = \frac{(1 + \mathrm{sim}(\mathrm{y_i}, \mathrm{y_j})^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{i'}\sum_{j'}(1 + \mathrm{sim}(\mathrm{y_i}, \mathrm{y_{j'}})^2/\alpha)^{-\frac{\alpha+1}{2}}} \tag{4}$$

In this setup, α is fixed at 1, and $\mathrm{sim}(y_i, y_j)$ represents the similarity measure between metadata samples, computed here using Euclidean distance [59] due to its simplicity and interpretability, aligning well with the type of metadata available.

The autoencoder described is designed to compress the visual appearance of images while also regularising patterns based on associated metadata. This process is guided by minimising a specific loss function, referred to as Equation 2. Once trained, the encoder component of the autoencoder functions as a robust feature extractor. To combat the risk of overfitting, a strategic sampling method is employed. Specifically, a 1:1 sampling ratio is used to balance between images that meet certain user-defined metadata criteria and those that are randomly selected from the entire dataset. For example, within the scope of user-defined criteria, similarity among images is measured using Euclidean distance. Let's consider a scenario where the metadata criterion is depth, and the maximum allowable similarity distance is set at 10 meters. In practical terms, this means selecting the first image randomly, then choosing 50% of subsequent samples from within a 10-meter distance of first image, and the remaining 50% from across the entire dataset. This sampling strategy ensures a divers range of samples in the affinity matrices P, effectively preventing excessive regularisation due to skewed sample distributions. It also allows for equal consideration of both similar and dissimilar images during each batch iteration. By maintaining this balance, the model can learn from a varied set of examples, improving its ability to generalise and avoid biased representations.

B. Reduction of large class imbalance

In this part, we utilise latent representations derived from the previous encoder-decoder network, forming the foundation for our clustering and sampling strategies. These strategies are essential for creating a balanced dataset crucial for the subsequent analysis phases and the success of the supervised taxonomist-led training phase (Section C). A balanced dataset is crucial to mitigate overfitting and to ensure that the learning process is not skewed by the over-representation of any single class. In our study, the dataset is predominantly composed of the detritus class, which accounts for 80% of the images. This imbalance can adversely affect the performance of clustering algorithms, leading them to disproportionately select representative samples from the detritus class. Such a scenario often results in the misclassification of the majority class instances as minority classes, creating clusters of uniform size rather than clusters that reflect the true distribution of the data. This issue is known as the "uniform effect" in k-means clustering [60], presenting significant challenges in adequately representating rarer classes, such as specific types of plankton which are of particular interest in our study. To tackle this imbalance, we have developed a customized data sampling strategy. Considering the shape heterogeneity and the overlap between the detritus class and other plankton classes [61], [62], our approach employs an innovative undersampling method. This method strategically eliminates noisy and less informative examples from the detritus class. Our methodology comprises several steps (illustrated in Figure 2-b):

1) Unsupervised Clustering: We apply hierarchical k-means clustering (H-kmeans) to the latent data generated in section A to group similar images based on their visual and metadata characteristics. We set the number of clusters (k) to twice the number of expected classes ($k = 46$), as this has shown to enhance the clustering algorithm's sensitivity to nuanced patterns within each class based on preliminary tests (see section IV-B2).
2) Expert Annotation: The most representative image from each sub-cluster is selected for annotation by expert taxonomists. This step is crucial for accurately depicting the variability within each cluster, especially when the clustering resolution is challenged by ambiguous class boundaries. This process further divides each of the k clusters into smaller sub-clusters using a hierarchical approach, calculated by the ratio $[M_1/k]$ where M_1 equals 500, the total number of images designated for annotation. From these, the most central image in each sub-cluster is selected for detailed examination by expert taxonomists, ensuring representative samples are used, especially in cases with ambiguous class boundaries.
3) Reduction of Dominant Class Samples: To refine the representation of non-detritus classes, we reduce the number of detritus images by retaining only a small proportion (10% as determined in our tests) of images from near the center of each detritus-dominated sub-cluster.

Subsequently, each of the k clusters is categorized into one of three groups based on the composition of the images they contain:

- Group A (Detritus): Clusters composed exclusively of detritus class samples.
- Group B (Ambiguous): Clusters containing a mix of detritus and non-detritus samples, reflecting transitional characteristics that may blur the distinction between detritus and plankton, especially when image quality affects the clarity of class features.
- Group C (Non-Detritus): Clusters distinctly separate from the detritus class, exclusively containing non-detritus (plankton) samples.

This categorisation strategy effectively addresses the overwhelming presence of detritus in our dataset. Initially, most clusters fall into Group A, which is heavily dominated by detritus. This dominance restricts the adequate representation of Groups B and C. To manage this, we adopt a targeted undersampling strategy for Group A, retaining only a small portion of images from the central region of each heavily detritus-laden sub-cluster. By increasing the number of clusters and selectively reducing detritus samples, we improve the resolution of non-detritus classes, enabling the clustering algorithm to better distinguish previously overlooked non-detritus groups.

C. Human-led annotation for pseudo-labelling and final classification

In the latter stages of our methodology, after mitigating the dominance of the detritus class, we recalibrate and select a fresh set of M_2 representative images. The number of images is set to M_2 = 2000 based on experiments (see section IV-B3). Expert taxonomists annotate these images to provide key insights that are crucial for reorganisation of the dataset. For images that remain unannotated, we implement a nearest neighbor technique to assign pseudo-labels by matching them with the most similar annotated samples.

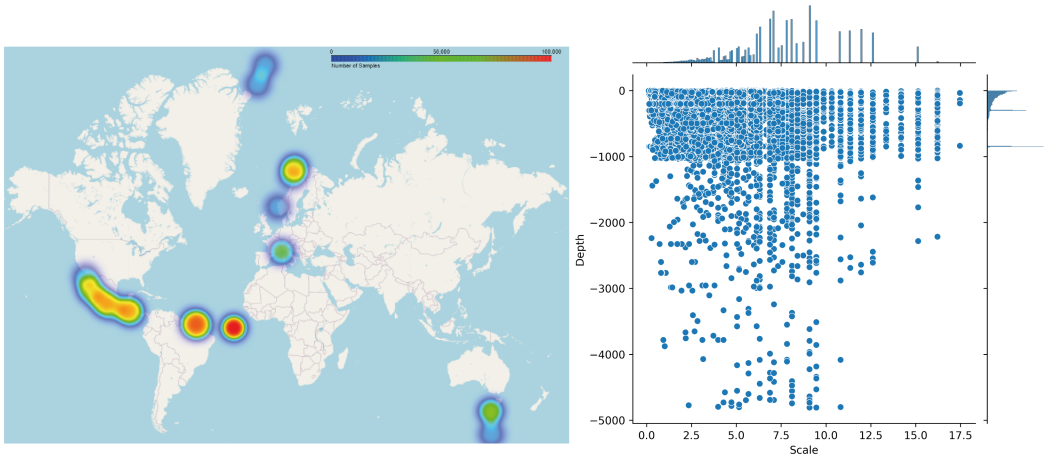
Next, we introduce a structured framework of 'views' within our methodology. Every cluster, redefined post-annotation, is regarded as a unique 'view'. This structure helps address challenges such as class imbalance and the presence of subclasses with subtle visual differences. The 'multi-views' strategy reorganizes the data, aligning it according to visual similarities and dissimilarities, both within and between classes. This is particularly important when some views in a class visually resemble other classes more than their own, necessitating a model that prioritises distinction over broad categorisation. This nuanced approach not only helps in identifying subtle features but also significantly boosts classification accuracy. While the 'multi-view' setup is advantageous, it necessitates careful management to avoid overfitting. The number of views is determined through empirical testing and ongoing evaluation of the model's performance during the fine-tuning phase. This carefully structured approach informs our final step of model training. We proceed to train a CNN classifier on this pseudo-labelled dataset using five classifiers for comparison: Random Forest, Decision Tree, Support Vector Machine, XGBoost and CNN (see section IV-B4.

IV. Experiments

In this section, we conduct a series of comprehensive comparative experiments using a benchmark dataset of plankton images collected by the UVP (Irisson et al., in prep.). After describing the dataset, we present a detailed analysis of the performance of our proposed workflow across various experimental parameter settings, aiming to establish the effectiveness of our approach. Each configuration was evaluated individually to assess its efficacy. We then briefly delve into a discussion of the obtained results, exploring the connections between our findings and the field of ocean ecology. This analysis will shed light on the insights gained from our research and their implications. Finally, we will identify potential avenues for further research and exploration in this domain.

A. Dataset

A large image data set collected using the UVP6 camera system [5] was used to develop and test the suggested method. The UVP6 system utilises a 5 Megapixel CMOS monochrome image camera sensor (Sony IMX264) where objects are illuminated by a collimated light beam positioned in front of the lens. To ensure accurate object detection, the processing unit software incorporates a zone-specific gain correction method that adjusts the gain settings of different image regions based on their specific lighting conditions. Subsequently, an automatic background subtraction and thresholding-cropping technique is applied. The camera system has been specially



(a) (b)

Fig. 3. Charting locations and analyzing depth-size distributions. (a) Global heatmap revealing dive locations across the globe. (b) Exploring the relationship between Depth and Scaling Factor with a bivariate and histogram analysis.

designed to cover a broad variety of particle sizes, ranging from about 100 µm to around 50 mm in diameter.

The geographical locations of our data set cover a broad range of oceanic regions, including the Mediterranean Sea, equatorial Atlantic Ocean, equatorial Pacific Ocean and polar regions (Fig. 3(a)). The data set comprises images from 980 dives between 0 and 5000 m depth (Fig. 3(b)), with the highest data density within the upper 1000 m of depth. A total of 451,379 images were recorded.

In order to examine the size distribution of the imaged particles and accurately categorize them, we employed a scaling factor calculation for each individual object as part of the metadata. This scaling factor is determined by comparing the original size of the specimen, measured in millimetres per pixel, to a standardized value such as a fixed patch size of the image (i.e. 227 pixels). This calculation plays a crucial role in differentiating between small and large plankton, which would otherwise be grouped together based solely on visual observations. We can calculate the size of each item by including the scaling factor; bigger plankton are indicated by greater scaling factor values. Scaling factors range from 0.1 to 17.5, indicating that we imaged plankton that spans two orders of magnitude in size (Fig. 3(b)).

Expert plankton taxonomists annotated all images in the dataset across 55 classes, with varying image counts per class ranging from 46 to 509k. To streamline the computer-led classification, we condensed the number of human-annotated classes to 23 by merging closely related and/or underrepresented classes (e.g. combining the visually almost identical order of *Collodaria* and family of *Aulosphaeridae* to clade level, Rhizaria). It is noteworthy that these annotations may contain a degree of human error in the classification (see discussion by [63], [64]).

B. Experimental Settings Description

For the self-supervised model, the autoencoder is structured using the Alexnet [65] architecture modified to incorporate batch normalization (BN) at each layer, serving as the encoder. The traditional AlexNet is adapted by altering the final fully connected layer to output a compressed 64-dimensional latent representation the input image, which measures 227 × 227 pixels. The decoder, designed as the inverse of the encoder, employs transposed convolutional layers to reconstruct the original image from its latent representation, effectively mirroring the encoding process. The training of this autoencoder model leverages the Adam optimizer [66], chosen for its efficiency with sparse gradients and adaptability to different data distributions. To ensure stability and precision in the learning process, a low learning rate of 1e-5 is used alongside a batch size of 256. The self-supervised model is trained for 200 epochs when weights are initialised with the value of AlexNet pre-trained on the ImageNet dataset. Each image in the dataset has been carefully cropped to feature only a single object against a dark background, enhancing the object's visibility. The images are stored with an 8-bit depth in two channels, providing a simplified yet effective data structure. To filter out irrelevant data, images captured before the device reaches the sea surface, which often contain artifacts and negative depth values, are automatically removed from the dataset. Additionally, conventional data transformation methods, such as rotating, flipping, and shifting are used as well as random image cropping and Gaussian noise. For example, the image is randomly flipped up and down, left and right, and then randomly, a zero-mean Gaussian noise is added to the actual ROI. During the training phase, further refinement of the input data includes the normalization of all metadata, with location data converted into the Universal Transverse Mercator (UTM) coordinate system to facilitate uniformity and precision in spatial computations. This normalization is crucial for the subsequent calculation of Euclidean distances between data points. A strategic approach to sampling is adopted to enhance the model's learning efficacy; half of the images in each training batch are selected based on their proximity in the metadata-defined space to a randomly chosen sample. This

selection is governed by a Gaussian distribution with varying sigma values, ensuring a diverse yet representative sample pool. The remaining images are selected randomly from across the entire dataset. Since classification performance is affected when dealing with a severely skewed dataset, we adopted an enhanced F1-measure evaluation for multi-class classification tasks, following the methodology of [67]. Specifically, we calculate F1-scores for each category using a one-vs-rest (OvR) approach [68], culminating in a macro-average that does not weigh classes by their sample size. This approach ensures that no class is deemed more significant than another regardless of its frequency in the dataset, promoting fairness in model evaluation. To rigorously test the generalisability of our model, we employed stratified 5-fold cross-validation (CV). This method maintains the relative distribution of each class across all folds, ensuring that our model is tested under varied conditions that closely mimic real-world distributions. Further, our experimental setup includes a comprehensive exploration of key parameters likely to affect feature learning and classification performance. We systematically varied:

1) The regulariser, defined by user-selected metadata criteria such as collection location, depth, or object size, each at four distinct levels.
2) The clustering algorithm, choosing between k-means, hierarchical k-means (H-kmeans), or a random assignment approach.
3) The number of annotations for generating pseudo-labels, tested at three different scales (500, 1000, and 2000) to evaluate robustness against varying amounts of training data.
4) The classifier type, where we tested several robust models including Support Vector Machine (SVM), Decision Tree, Random Forest, XGBoost, and CNN, to identify which performs best across diverse scenarios.

These steps ensure that our findings are not only statistically valid but also broadly applicable, enhancing the credibility of our results in real-world applications. Through this meticulous experimental design, we aim to demonstrate the robustness and scalability of our proposed solution in the face of varying data characteristics and analytical conditions.

1) Regulariser: In the feature extraction phase, each sample is passed through a deep CNN backbone and projected into a 64-dimensional latent point, incorporating regularisation with metadata information. The representation learning objective involves two key components. First, we enhance the similarity between these embedded points by adjusting them in the direction of the average representation among them. Simultaneously, we give more weight to data points with more confident assignments through an auxiliary target distribution (i.e. the metadata). To examine whether this auxiliary information reveals any critical structural similarity in latent space, we quantify how much the loss regulariser function contributes to classification success. The results are compared with those from an identical network architecture without the regulariser. We set different experiments with user-defined metadata criteria as follows: location distances as l = 10m, 100m, 1km, 10km, different depths as d = 1m, 5m, 10m, 100m, and different scale parameters as s = 0.2, 0.4, 0.6, 0.8. We only used one user-defined metadata criteria at a time as preliminary results on a smaller subset of the data suggested no improvement or even worsening of the results when multiple criteria are combined (data not shown), because of time limitations, and because discussion of the results is more straightforward. The score value results with these parameters for all experimental settings (including choice of clustering algorithm, number of annotations and classifier) are reported in Table I and Table II.

Based on our observation, incorporating the metadata auxiliary distribution as a regulariser in our deep autoencoder can generally enhance the learning process, resulting in accelerated classification performance compared to the absence of this regulariser (Table I). In terms of classification performance, the depth regulariser exhibited superior results compared to the location, scale, and no regulariser approaches. Specifically, in one equal configuration G3, C3, K3, and M3, the depth regulariser achieved an impressive accuracy of 46.9%. In contrast, the location, scale, and no regulariser approaches achieved accuracies of 32.1%, 31.2%, and 30.0% respectively. The impact of the regulariser across different metadata variations is depicted in Fig8-a through a box plot. The average F1-score for varying depth values consistently surpassed the performance of other regularisers. Notably, when the depth was specifically configured to 10m, it achieved the highest average F1-score of 50.5%. In terms of location, the performance demonstrated improvement as the distance increased. The highest efficiency was achieved with a distance of one kilometre, resulting in a 35.9% F1-score. However, the overall mean for scale parameters negatively affected performance across most configurations when compared to not using scale regularisation. The highest F1-scores achieved with scale regularisation were 36.2% and 32.0% respectively. Overall, these findings highlight the advantageous impact of the depth regulariser, while also showcasing the potential benefits of considering location in optimizing the model's performance. However, caution should be exercised when utilising scale parameters, as their inclusion often results in diminished performance.

Based on our observations, incorporating the metadata auxiliary distribution as a regulariser in our deep autoencoder resulted in a mixed response in classification performance depending on the criteria, level and classifier used. For the following reported values are based on the H-kmeans algorithm and an annotation number of M_2 = 2000. The depth regulariser enhanced classification performance at any level (1, 5, 10 and 100m) and all classifiers, with an average F1-score increase of 11.6 units (range: 2.8 - 20.3 %pt). On average, level 10 m resulted in the best classification results (score increase of 15.9$\pm$2.8 units). Specifically, for the non-deep-learning classifiers, it achieved the highest average F1-score (50.5%) when configured to 10 m. For the CNN classifier, it also achieved the highest average F1-score (54.4%) albeit for a configuration with d = 1 m. For the location regulariser, there was generally a positive trend between location distance level and F1-score improvement, though consistent F1-score improvements across all classifiers were only achieved for levels

TABLE I
PERFORMANCE OF VARIOUS

Non-deep Learning Methods on the UVP6 Benchmark Dataset: F1-Scores, Standard Deviations (%), and Best Performer for Each Metadata is Highlighted in Bold. All Results are Based on Multiview Grouping Technique. DT represents Decision Tree, RF for Random Forest, SVM for Support Vector Machines, and XG for XGBoost.

config	regularisation	classifier	**random**			**k-means**			**H-kmeans**		
			500	1000	2000	500	1000	2000	500	1000	2000
A1	Location:	DT	16.5±5.5	17.5±4.1	17.2±3.9	20.1±2.6	21.7±5.1	22.3±6.1	21.6±4.4	21.5±4.9	23.1±2.3
A2	10m	RF	19.4±4.1	20.6±3.8	20.1±4.1	22.4±7.7	24.2±4.9	25.2±2.4	20.4±5.3	22.3±3.2	22.1±4.1
A3		SVM	21.4±5.2	21.6±5.4	22.7±4.3	25.5±6.1	26.6±7.5	28.2±2.0	27.4±3.3	30.2±4.1	32.4±1.7
A4		XG	20.9±4.6	21.3±3.4	21.8±3.1	24.1±4.2	25.4±5.7	27.3±1.5	25.1±5.2	26.9±2.8	29.1±3.3
B1	Location:	DT	18.7±2.8	19.2±3.2	19.0±5.2	21.9±6.7	20.4±6.2	20.9±3.8	23.2±7.2	26.3±4.5	26.1±2.3
B2	100m	RF	18.1±3.0	18.5±3.8	19.1±4.4	21.3±5.3	21.4±4.1	21.1±4.5	25.7±3.9	26.8±6.2	28.8±4.5
B3		SVM	26.6±3.7	27.1±3.5	27.5±5.7	30.3±7.3	32.1±1.7	32.6±3.8	29.5±2.4	30.6±6.0	35.5±3.6
B4		XG	27.5±4.4	27.8±4.7	28.3±4.2	31.2±4.1	32.5±2.7	33.2±1.9	27.6±5.2	28.2±5.1	28.1±3.7
C1	Location:	DT	19.2±5.1	19.9±2.7	20.2±2.4	23.3±2.7	23.6±4.1	27.2±3.3	28.8±7.1	30.2±5.8	31.6±1.0
C2	1km	RF	18.5±4.5	19.8±3.4	19.4±2.2	22.4±5.2	26.5±6.2	26.6±4.6	24.7±6.1	25.9±6.5	27.2±3.4
C3		SVM	25.0±3.1	26.1±4.1	25.7±3.3	28.5±2.9	31.2±3.5	33.8±1.9	32.1±2.9	34.2±3.2	35.9±2.5
C4		XG	25.2±5.2	25.9±3.2	25.4±3.7	29.1±6.8	33.2±4.0	33.7±4.1	33.8±3.5	35.1±3.8	**36.2±1.6**
D1	Location:	DT	19.5±3.6	20.6±3.5	21.0±4.2	22.6±2.5	26.5±5.7	29.4±4.0	27.8±3.5	29.1±2.8	31.2±0.5
D2	10km	RF	17.3±3.7	18.9±3.4	18.6±2.7	21.2±1.4	23.6±3.2	23.9±1.7	25.3±2.6	25.9±4.7	28.8±2.2
D3		SVM	24.5±2.9	24.7±3.1	25.7±3.6	28.1±6.2	30.1±5.3	31.2±2.7	31.8±4.4	31.8±1.5	32.1±4.1
D4		XG	24.1±4.1	24.2±2.8	23.4±2.8	26.3±5.7	26.8±4.0	28.1±3.8	32.2±1.2	33.8±5.0	34.1±2.7
E1	Depth:	DT	23.1±4.8	24.2±2.8	23.4±2.8	28.2±1.3	29.5±2.4	32.1±1.6	30.3±4.1	36.5±1.2	39.3±0.8
E2	1m	RF	23.8±3.9	24.4±2.8	25.2±2.2	26.8±2.5	28.4±4.1	30.4±1.2	25.6±3.1	29.1±1.3	31.8±1.5
E3		SVM	33.1±4.4	34.5±1.9	34.6±2.1	36.5±5.9	40.7±1.9	41.4±1.3	44.1±2.6	46.3±2.8	49.5±2.6
E4		XG	32.4±5.1	33.5±2.6	33.2±3.1	35.4±4.7	38.1±2.8	39.7±1.9	39.6±1.7	40.1±5.1	42.8±0.2
F1	Depth:	DT	24.9±3.6	26.3±3.1	25.8±2.9	28.4±2.4	28.2±5.0	30.7±2.9	32.8±3.9	34.7±2.2	36.8±1.9
F2	5m	RF	26.3±4.2	27.4±2.7	27.0±3.2	29.5±3.3	30.2±2.1	33.9±0.2	28.3±3.2	28.5±3.8	29.6±2.7
F3		SVM	30.2±3.8	30.6±3.0	30.7±2.7	33.2±6.0	36.1±4.2	37.6±4.4	36.5±6.9	39.1±7.1	42.2±3.9
F4		XG	30.6±2.5	31.7±2.6	32.0±2.5	34.2±5.1	35.2±1.7	35.6±5.3	35.1±5.1	38.2±3.0	42.3±1.1
G1	Depth:	DT	29.3±2.8	30.1±2.4	30.4±2.6	33.4±5.4	38.2±2.7	39.1±0.4	39.8±5.4	41.9±3.3	43.2±1.7
G2	10m	RF	27.4±3.4	28.4±3.8	29.2±3.4	30.5±5.1	33.8±4.2	34.2±1.5	34.6±4.2	36.2±3.6	39.9±2.1
G3		SVM	38.4±2.8	38.8±4.1	39.5±2.6	41.5±2.4	44.4±3.4	45.2±2.0	46.9±3.2	47.2±2.0	**50.5±1.8**
G4		XG	36.9±1.7	38.6±2.8	38.0±1.1	40.2±5.1	41.1±1.1	42.0±0.4	45.6±4.5	46.2±1.5	47.1±2.5
H1	Depth:	DT	26.5±4.3	27.3±3.7	28.1±3.1	30.1±2.5	32.0±1.9	32.1±2.2	37.1±4.7	38.2±3.6	38.2±1.2
H2	100m	RF	29.1±2.8	29.8±2.6	30.7±2.9	32.5±3.1	33.5±2.0	34.1±2.6	35.4±3.6	36.2±2.9	37.1±2.2
H3		SVM	29.3±4.6	31.1±3.6	29.9±3.5	32.8±6.2	35.2±5.1	36.5±2.3	37.1±6.2	39.2±5.2	41.8±4.3
H4		XG	26.3±2.8	27.0±2.4	27.4±3.2	30.1±7.1	31.8±4.9	33.7±3.9	38.9±4.3	39.3±2.8	40.1±2.4
I1	Scale:	DT	15.8±2.9	16.9±2.8	16.3±4.2	18.3±2.4	17.7±5.1	18.5±2.8	20.1±3.1	23.6±2.4	24.6±1.5
I2	0.2	RF	12.8±3.5	13.6±3.8	13.4±3.4	16.2±6.8	18.9±3.0	19.4±1.7	18.3±2.0	20.9±2.5	22.4±0.7
I3		SVM	19.8±4.2	21.2±2.8	20.3±2.8	23.2±4.6	24.1±1.6	25.2±1.8	24.4±5.1	25.8±2.9	25.5±3.1
I4		XG	17.2±3.5	18.2±5.2	18.0±4.3	21.8±1.5	21.5±2.0	22.0±1.7	25.5±6.2	25.1±4.1	26.6±3.3
J1	Scale:	DT	15.5±2.7	16.8±2.1	17.1±3.6	18.3±3.2	20.7±1.2	20.5±1.3	19.4±2.9	19.6±1.1	20.0±1.2
J2	0.4	RF	12.7±4.1	13.8±4.1	13.9±2.7	16.6±2.3	18.6±2.4	20.8±1.6	16.2±1.4	17.9±1.5	19.6±0.9
J3		SVM	18.6±2.8	19.2±2.8	20.9±3.9	21.4±5.6	22.3±2.6	22.2±2.4	23.1±2.5	25.1±3.1	25.7±2.7
J4		XG	20.6±2.9	21.5±3.1	21.3±3.3	22.4±1.1	23.1±3.0	24.8±2.5	20.2±5.1	21.3±3.5	22.5±0.9
K1	Scale:	DT	19.8±2.2	21.7±2.6	20.9±2.9	22.5±1.7	26.5±1.1	26.2±0.5	24.5±2.3	26.8±2.2	28.0±1.6
K2	0.6	RF	19.8±3.7	21.1±1.9	21.0±3.3	23.2±2.8	25.2±2.3	28.9±2.2	26.1±2.8	29.2±4.2	30.4±1.4
K3		SVM	27.6±5.2	29.1±3.5	29.7±2.5	30.2±5.2	31.3±4.5	32.8±0.8	31.2±7.8	33.2±6.1	**36.2±2.1**
K4		XG	21.9±2.7	24.4±4.4	25.5±4.0	26.2±2.3	26.6±3.4	29.2±2.1	28.6±1.8	30.5±2.1	31.5±3.2
L1	Scale:	DT	20.5±4.2	21.0±3.9	21.6±2.6	23.2±2.9	25.3±1.1	26.5±1.3	22.6±3.9	26.3±1.8	29.7±2.4
L2	0.8	RF	19.5±2.6	21.0±2.7	19.9±3.7	21.6±3.4	22.9±3.1	24.6±2.2	22.3±1.5	24.7±1.2	26.1±1.3
L3		SVM	23.4±1.3	24.4±2.6	25.5±3.5	25.7±6.0	27.1±3.4	29.1±1.6	24.4±3.5	26.7±5.2	26.1±2.6
L4		XG	23.4±2.8	26.1±3.8	25.2±2.4	25.5±5.2	26.8±3.1	28.2±0.2	25.9±6.3	28.5±3.5	30.2±1.9
M1	No	DT	22.2±4.3	23.1±3.8	22.7±3.6	26.1±3.0	26.9±2.2	27.2±2.1	27.2±2.5	26.7±2.7	30.3±0.8
M2	regularisation	RF	20.0±3.5	20.3±4.2	20.5±3.1	24.1±2.8	24.8±2.3	25.8±2.5	25.2±2.9	25.5±2.1	26.8±1.6
M3		SVM	22.7±4.2	23.7±3.6	23.5±2.6	26.9±3.5	28.4±4.1	29.5±2.6	30.0±6.2	31.7±3.2	**32.0±1.6**
M4		XG	23.9±3.8	24.8±2.7	24.9±3.2	26.2±3.9	26.5±4.3	27.0±0.2	29.0±3.3	29.4±4.7	31.0±3.0

1 km and 10 km (average 3.4±2.5 and 2.7±2.9 units improvement, respectively). The scale regulariser generally worsened classification performance with a mean F1-score difference of -3.15 units (range -10.3 to 6.0 %pt). The impact of the regulariser across different metadata variations is depicted in Fig8-a through a box plot. Overall, these findings highlight the advantageous impact of the use of metadata-based regularisers in optimising the model's classification performance. However, caution should be exercised when choosing the metadata parameters as they can also - as shown by the scale criteria - result in diminished classification performance. Potential ecological reasons for the observed change in classification performance when using the different criteria are discussed below (Section V).

2) Clustering algorithm: We undertook a comprehensive comparison of different methods for selecting representative samples from our dataset. Specifically, we explored the effectiveness of k-means clustering, H-kmeans clustering, and random selection. Our primary objective was to identify a specific data group suitable for fine-tuning while ensuring that

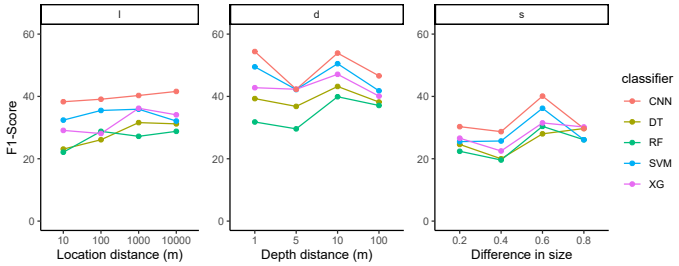


Fig. 4. Evaluation of classification performance based on the regularisation. Difference in F1-score of the different regulariser configurations (using H-kmeans and M_2 = 2000) compared to no regularisation. Panels show regularisation criteria: location (l), depth (d), and scale (s). Colours indicate classifier as shown in the legend.

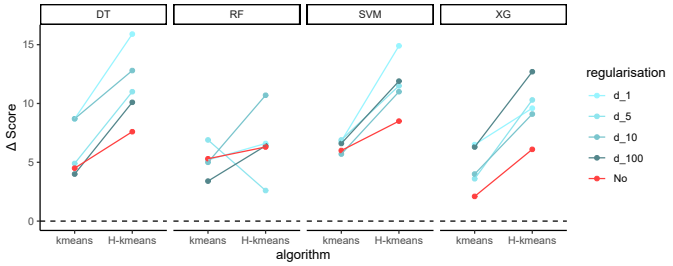


Fig. 5. Evaluation of classification performance based on the clustering algorithm. Difference in F1-score of the non-deep-learning classifiers with either depth-regularisation or no regularisation (M_2 = 2000) compared to the random control (same regularisation and number of annotated images). Panels show classifiers: Decisiton Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and XGBoost (XG). Colours indicate regularisation level as shown in the legend.

all classes present in the entire dataset were included. To do that, we took into account the taxonomy tree of plankton and focused on 23 coarse classes that represented the entire dataset. Instead of performing a computationally expensive grid search considering all possible values, we opted for a fixed number of 46 clusters (twice the number of classes) to provide the algorithm with enough flexibility to consider smaller subsets.

For the reduction technique aimed at decreasing the number of majority-class samples (i.e. detritus images), we determine the number of images to retain from each majority-class cluster (group A) by testing a range (10 values between 1 and 50%) of different ratios between the number of samples belonging to group A ($Size_A$) and the combined size of groups B and group C ($Size_{BC}$):

$$R^i = Size^i_A / Size^i_{BC} \tag{5}$$

where i denotes the iteration. A larger value of R^i indicates a cluster with more samples from group A and fewer from group BC, aligning it with the majority class. Conversely, a smaller R^i suggests a cluster with more Group-BC class samples, deviating from majority class characteristics. Our goal is to identify a trade-off value for R^i that balances these considerations [69]. To quantify the trade-off, we continue reducing until all our 10 target classes are found in representative samples. The selected reduction value of 90% was determined through experimentation and yielded the best trade-off based on the number of non-detritus samples found in representative samples selected by H-kmeans. This 90% reduction value is consistently applied across all configurations in cross-validation tests to ensure a fair comparison.

To ensure fairness and draw conclusive comparisons about which clustering algorithm performs best, we maintained consistent configurations for our data selection algorithms. For instance, we utilised the same set of 64-dimensional features extracted from a pre-trained deep autoencoder across all approaches. In the case of k-means, both the hierarchical and non-hierarchical variants, we applied identical parameter settings. To account for the potential impact of initialization, we executed each algorithm ten times, employing different centroid seeds in each run. Additionally, we took into consideration the class frequencies of the inputs to balance the weighting for clustering. By adhering to consistent configurations, running multiple trials, and accounting for class frequencies, we aimed to minimise any potential sources of bias and obtain reliable and conclusive comparisons between the different data selection algorithms.

Both clustering algorithms, k-means and H-kmeans, improved classification accuracy compared to the random selection with an average improvement of 5.1±1.9 %pt and 8.0±3.4 %pt, respectively (all configurations, n = 52). H-kmeans yielded better classification performance in 45 out of the 52 tested configurations, with an average performance difference of 2.8±2.9 %pt. The highest classification performance gains were observed for the regularisation with depth (4.6±2.9 %pt compared to k-means and 10.4±3.3 %pt compared to random, all levels), specifically for the level d = 1m for the classifiers SVM and DT (respectively +8.1 %pt and +7.2 %pt compared to k-means) (Fig 5). The gains for the other regularisers, compared respectively to random and k-means, were 7.8±3.2 %pt and 2.3±3.2 %pt for location, 5.9±2.6 %pt and 1.6±2.5 %pt for scale, and 7.1±1.1 %pt and 2.6±1.3 %pt for no regularisation.

3) Number of Annotated Samples: Before running predictions on the unlabeled data, manual human-led annotations were performed on a subset of representative sample views, referred to as M_2-selected samples. F1-scores are reported for different numbers of annotated samples, specifically M_2 = 500, 1000, 2000 (Tables I and II). Increasing the number of annotated samples generally results in higher scores: a change from 500 to 2000 samples (H-kmeans and all model configurations) resulted in an average increase of 3.7 %pt (range 0.3 - 16.1 %pt; Tables I and II). However, for some model configurations the additional 1000 images from M_2 = 1000 to M_2 = 2000, which effectively doubles the workload for the human expert taxonomist, starts to have a diminishing return (e.g., for "No regularisation", "d = 5m", "l = 10m", and "s = 0.8" using the CNN classifier; Fig6-a). For the best model configuration with the CNN classifier (d = 1 m and 10 m), more annotation would likely be a good investment of time as there is no tailing off yet, indicating that further annotation could improve classification accuracy. Nevertheless, our ultimate objective is to achieve satisfactory performance while minimising the number of annotations

TABLE II
RESULTS OF THE CNN METHOD USING SINGLE OR MULTIPLE VIEWS PER CLASS WITH HIERARCHICAL K-MEANS DATA SELECTION ON THE UVP6 BENCHMARK DATASET: F1-SCORES AND STANDARD DEVIATIONS (%). BEST PERFORMING METRICS ARE HIGHLIGHTED IN BOLD.

config	regularisation	classifier	Single view			Multi-view		
			500	1000	2000	500	1000	2000
N1	Location: 10m	CNN	29.7±1.8	33.6±3.8	36.0±0.6	31.7±2.2	36.5±1.3	38.3±2.4
N2	Location: 100m	CNN	30.8±2.1	33.4±2.6	36.4±1.2	32.6±1.2	35.7±2.2	39.1±2.3
N3	Location: 1km	CNN	33.5±1.1	35.6±1.5	38.4±0.3	35.3±2.7	37.6±2.6	40.3±2.4
N4	Location: 10km	CNN	34.9±2.6	35.1±1.1	39.0±0.1	36.4±2.4	37.3±2.5	**41.6±1.2**
O1	Depth: 1m	CNN	45.6±1.2	46.3±2.2	52.9±1.4	48.3±2.3	49.2±1.1	**54.4±0.5**
O2	Depth: 5m	CNN	34.9±1.4	39.5±1.3	39.3±0.2	37.3±0.5	41.6±2.4	42.3±2.1
O3	Depth: 10m	CNN	46.8±1.4	47.0±1.6	50.7±0.8	49.3±1.5	49.5±1.7	53.9±1.6
O4	Depth: 100m	CNN	36.3±3.2	40.0±2.1	44.4±0.6	39.2±2.6	42.3±1.8	46.6±1.5
P1	Scale: 0.2	CNN	26.4±2.2	27.5±3.4	29.1±1.5	28.7±3.5	29.3±2.2	30.3±0.2
P2	Scale: 0.4	CNN	25.2±2.3	25.0±0.2	25.9±0.8	25.2±2.1	26.7±1.3	28.7±1.2
P3	Scale: 0.6	CNN	30.3±3.4	34.8±2.3	38.8±1.2	34.5±1.5	37.2±2.5	**40.1±2.3**
P4	Scale: 0.8	CNN	24.3±1.2	27.6±1.2	27.7±1.4	26.4±2.1	30.1±2.6	29.7±1.5
Q1	No regularisation	CNN	30.9±2.4	32.4±1.7	33.3±1.5	31.9±2.5	33.4±1.4	34.1±1.6

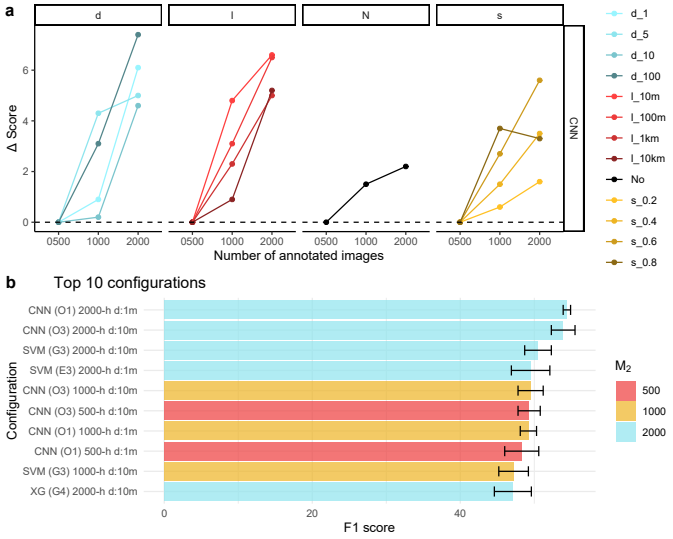


Fig. 6. Evaluation of classification performance based on the number of manually annotated images (M_2). **(a)** Difference in F1-score of all CNN + H-kmeans configurations compared to the CNN + H-kmeans control (no regularisation, 500 annotated images). Panels show regularisation criteria depth (d), location (l), no regularisation (N) and scale (s). Colours indicate regularisation level as shown in the legend. **(b)** Top performing model configurations in terms of overall F1-score. Colours indicate number of annotated samples as red = 500, yellow = 1000 and blue = 2000. Configuration details are stated on the y-axis. Error bars show the standard deviation of model re-runs.

required. While employing 2000 annotations yields a high score, certain configurations outperform others even with a smaller number of annotations. For example, in the top 10 configurations in terms of classification performance (out of 585 tested configurations), two configurations used only 500 manual annotations (CNN + hierarchical + d = 10m and SVM hierarchical + d = 10m; Fig6-b), hence outperforming the majority of configurations with M_2 = 2000. These findings underscore that, with an optimized configuration, even with very limited input from human expert taxonomists, our semi-supervised metadata-guided model achieves high classification performance.

4) Classifier: We conducted a comparative study involving various well-known classic machine learning techniques, which are commonly employed to advance machine learning models. These methods can be broadly categorized into two distinct groups:

Group 1: The first group encompasses commonly used methods that operate directly on latent space vectors. We utilised Support Vector Machine (SVM), a popular algorithm that learns a non-linear model using the kernel trick [70]. To ensure effective classification boundaries, we specifically chose the Radial Basis Function (RBF) kernel with a large degree. We incorporated Decision Trees (DT), which make predictions by traversing from the root node to a leaf node based on feature conditions [71]. DTs are known for their interpretability and can capture complex decision boundaries. Another approach employed in our study is Random Forest (RF), which involves constructing an ensemble of decision trees that are randomly generated [72]. This randomness and aggregation of multiple trees often lead to improved predictive accuracy compared to a single decision Lastly, we employed a gradient-boosting algorithm called XGBoost (XG). This powerful classification model combines the predictions of multiple base estimators to enhance overall robustness and performance [73]. XG leverages the greedy boosting strategy to iteratively refine the model's predictions.

Overall, we observed a trend where the two SVM and XG classifiers tended to outperform the two DT and RF classifiers(Fig3 in Appendix). For the specific configuration with a depth-regularisation of d = 1 m (E configuration), hierarchical clustering, and M_2 = 500 samples, we observed notable variations in performance among the classifiers. In this setup, SVM outperformed the other three approaches, achieving an F1-score of 44.1%. In comparison, the DT obtained an F1-score of 30.3%, RF scored 25.6%, and XG achieved a score of 39.6% (Table I). In the comparison between SVM and XG, SVM consistently demonstrated slightly superior performance compared to XG. However, there were specific setups where XG exhibited an average advantage of approximately 1%pt in mean scores. For instance, when considering the specified configuration for clustering and the number of samples, SVM outperformed 8 out of 13 samples across all metadata variations. An example of the superiority of XG can be observed in the C configuration (location regularisation l = 1 km), where

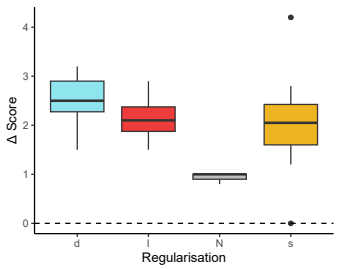


Fig. 7. Effect of the multi-view setting on classification performance. Difference in F1-score of all CNN + H-kmeans configurations with multi-view compared to single-view for regularisation criteria depth (*d*; blue), location (*l*; red), no regularisation (N; grey) and scale (*s*; yellow).

XGBoost achieved an F1-score of 33.8%, while SVM obtained a slightly lower F1-score of 32.1% (see Table I and Fig8(b) for detailed results).

Group 2: The second group comprises methods that are based on deep convolutional networks. In our study, we employed the ResNet18 architecture for fine-tuning our model. This architecture introduces skip connections to preserve the gradient flow, leading to a significant performance boost [74]. By comparing the single and multi-view suggested approaches, we showcased the classification performance of various configurations utilising the CNN model within the N - Q setup. Overall, the most remarkable performances are observed in configurations O1 and O3, corresponding to a CNN configuration with depths of 1m (with an F1-score of 54.4± 0.5%) and 10m (with an F1-score of 53.9± 1.6%), respectively. Notably, the performance difference between these configurations is minimal and within the standard errors.

Finally, we investigated the effect of multiple views compared to a single-view setting for the CNN classifier (using H-kmeans). Across all settings with various regularisers and numbers of annotated images, multi-view outperformed single-view by an average of 2.1± 0.8 %pt (range: 0 - 4.2 %pt; Fig 7). Additionally, we demonstrated the benefits of utilising multiple views for specific classes by examining the performance per class (Fig1 in Appendix). Notably, the utilisation of multi-view data has resulted in improved average scores for the detritus and *Rhizaria* classes, which exhibit a wide range of variability. As a result, the overall average score has seen an increase of approximately 2-3 %pt. This finding highlights the advantage of incorporating multiple views in the classification process, particularly for classes that exhibit greater variation within their samples.

V. Discussion

1) Use of metadata to help classification: For traditional microscope-based zooplankton identification, metadata provides the taxonomist with vital information that aids the correct species identification, such as the size of the organism and the location and water depth where it was found. In contrast, however, the usefulness of metadata for our unsupervised classification framework was mixed (Fig 4). We suspect that size was not useful information as our classification categories are too broad, leading to a considerable overlap in the sizes of different categories. For example, copepods range in size from <0.1 to 18 mm, hence spanning 4 orders of magnitude, while the size range of amorphous detritus is even larger, practically spanning from microscopic to several cm. Hence, at broad classification levels like ours, size appears to hinder rather than help the classification. In settings with higher taxonomic resolution that include classes with distinct size ranges (e.g. species level), however, we expect size to increase classification accuracy. Location appeared to be more useful at larger search distances (>1 km radius), possibly reflecting the patchiness of plankton in the ocean in terms of geographical distribution. Different to traditional taxonomy, we did not use absolute location (e.g. equatorial Pacific vs North Atlantic) to aid classification but rather the relative location of two objects to each other. Such an approach should be sufficiently sensitive to capture geographical changes in plankton distribution caused by, e.g., climate change [75], [76], without enforcing strict location constraints that could conceal distribution changes. Finally, depth information greatly improved classification accuracy likely because plankton of the same species often swarm, co-locating in relatively narrow depth bands of just a few meters (e.g. [77]–[79]). A class-based analysis shows that depth information is particularly useful for aiding the classification of copepods and *foraminifera* (Fig9), both of which are often found in distinct depth layers [79], [80]. For detrital particles, their shape and type typically also change with depth (e.g. [81], [82]) as particles are reworked and become more refractory the further they are away from the surface ocean, from where they originate. An interesting future addition in metadata would be the combination of the sampling depth and the time of day when sampling occurred (relative to sunset/sunrise) as many zooplankton species undergo diel vertical migration, feeding near the surface during night and resting at depth (often >500 m depth) during the day. Yet, as shown by size, any metadata criteria used for regularisation has to be chosen with care and checked for ecological meaning.

2) Reduced need for human expert annotation: Our model framework dramatically reduced the work for human expert annotation to $\tilde{1}$000 images ($M_1 + M_2$ = 500 + [500, 1000, 2000]) for a dataset of $\tilde{4}$50,000 images. The question is whether the resulting classification is useful for ecological purposes. In all configurations, our self-supervised model algorithms detected 10 common classes (detritus, fibre/filaments, *copepoda*, artefact, puff, *rhizaria*, *eumalacostrata*, *chaetognatha*, *foraminifera* and *salipda*). These common classes are among the 13 most abundant classes in our dataset. Certain rare classes, such as *ostracoda*, *actinopterygii* (ray-finned fish), or *appendicularia* (which were only represented by 28 samples), were also detected in specific configurations. Our best-performing configuration (O1) found 16 of the 24 classes (it also found *larvacean* houses, *aulacanthidae*, *ostracoda*, *Creseis acicula*, *actinopterygii*, and *Aulatractus*). Hence, the SSL appears adept in recognising the main structure of the plankton community across our dataset. The F1-scores for individual classes varied widely between model configurations and classes (Fig9-k). Notably, the class 'artefact', attained the highest average F1-score of 75%, underscoring its utility in data quality control and in scenarios where the data is

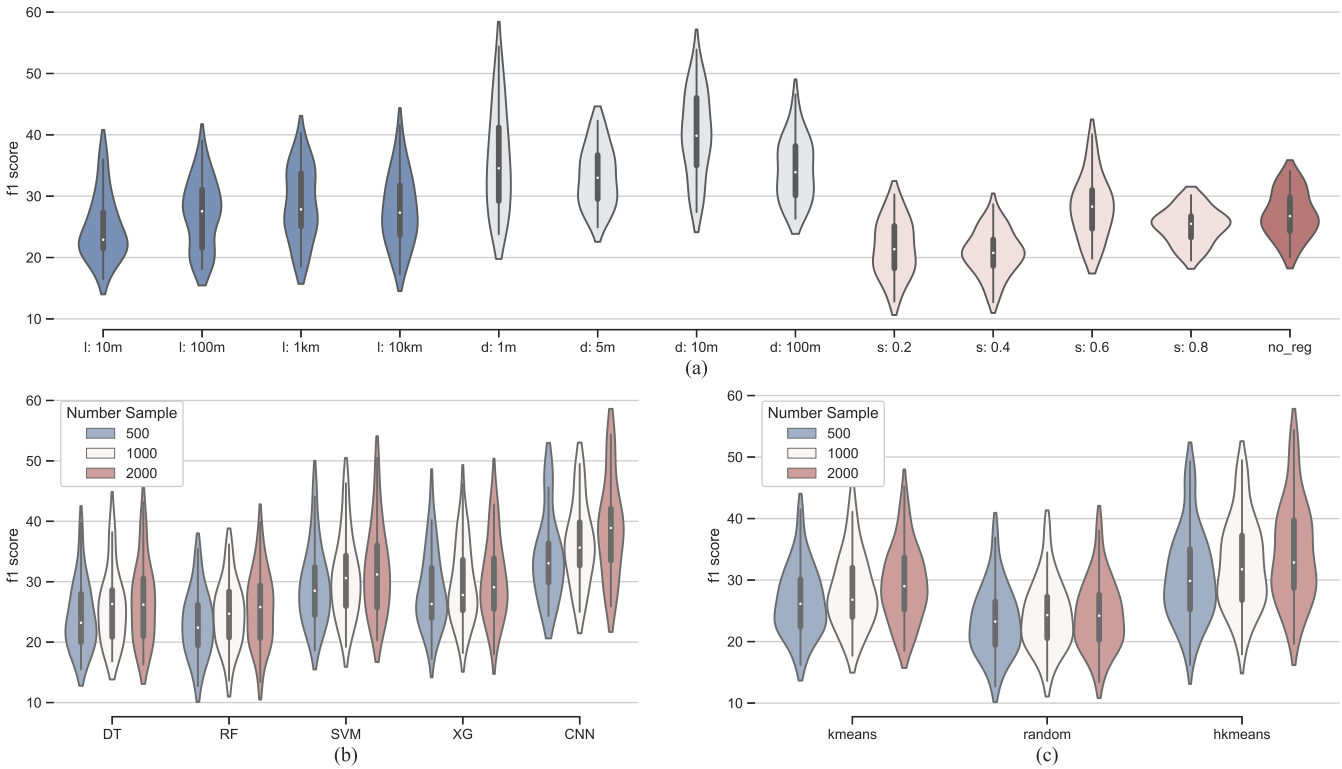


Fig. 8. Effect of the input parameters in box plots where the first quartile is at the bottom, and the third quartile is at the top: Box plots of the f1-score for various input regularisers, classifiers, and data selection methods are shown in Figures (a), (b), and (c), respectively.

employed for class-independent analyses, such as community size spectra [83].

To demonstrate how well our model configurations perform, we calculated the F1-score for two classification options: (1) the classifier assigns a label at random with an equal chance for all classes, or (2) the classifier assigns detritus to all images. For the random labelling, the F1-score ranges from 0.01 - 8.3% per class with a mean of 1.2%. For the all-detritus, though the F1-score for the detritus class is 90.8%, the mean score is 3.9%. Hence, both models perform much worse than our best-performing configuration, O1 (mean 54.4%; albeit having a mean of $\tilde{3}4.7\%$ if we consider the F1-score for undetected classes as 0%).

The ability to detect classes and the overall classification performance are dependent on the number of images for each class in the representative samples for human expert annotation (M_1 and M_2). M_1 and M_2 are selected by H-kmeans, whereby the algorithm tries to find key features for detecting and distinguishing classes, and their composition is hence less linked to the relative abundance of the individual classes in the entire dataset. As a result, many of the configurations showed the beginning of 'levelling off' as M_2 increases, e.g., the control configuration (no regularisation; Fig 6). For our best-performing model configuration (O1 with M_2 = 2000), we found a weak relationship between the overall abundance and F1-score for each class in the entire dataset (linear regression: R^2 = 0.35, p = 0.055, n = 9, when not including detritus), indicating the importance of selecting representative samples that capture the diversity of features across classes. When M_2 was increased further, from 2000 to 3000, the F1-score only slightly improved (data not reported), hence showing the 'levelling off' associated with the model finding more of the relevant representative samples. Overall, we demonstrate that SSL is effective at learning the key features relevant for plankton classification and can hence dramatically reduce the required input from human expert taxonomists.

In the rapidly evolving domain of automated annotation, the effectiveness of different methodologies can be quantitatively compared by their ability to save human-time and enhance throughput. Traditional manual annotation techniques typically process between 300 and 1000 objects per hour (pers. comm. and [7]). Although supervised platforms like EcoTaxa [6] improve on this by allowing for a sorting speed ranging from 300 to 15,000 objects per hour, depending on the level of automation and manual validation involved [7], these numbers are still modest when compared to more sophisticated methods. For instance, the interactive semi-supervised approach used by MorphoCluster [7] markedly increases efficiency, reaching

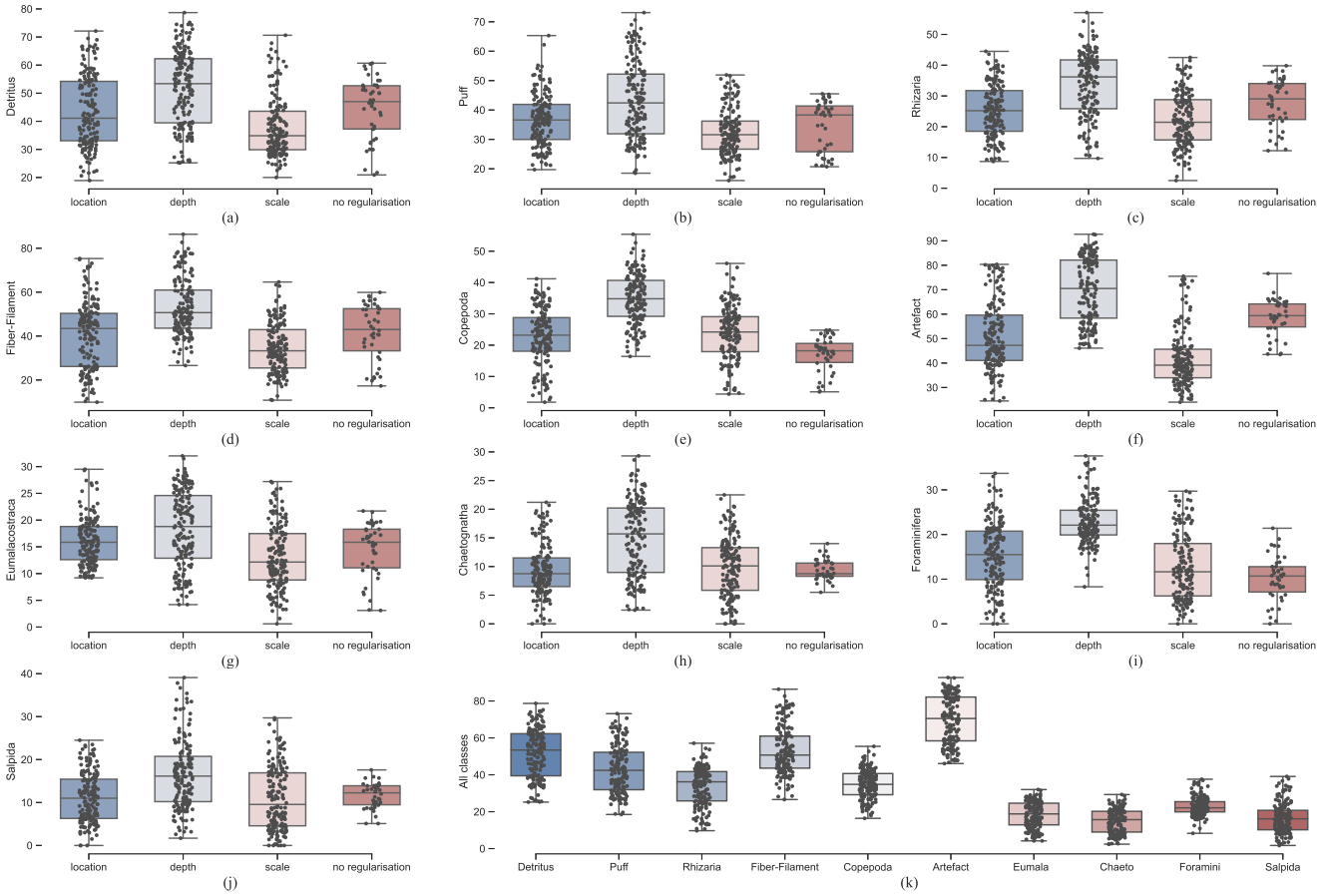


Fig. 9. Boxplots showing the F1-score from all model configurations for the 10 classes. Boxplots indicate median (line), upper and lower quantile (box), minimum and maximum (error bars). and individual data points (solid dots). (a-j) Comparison of F1-score with regularisation for location, depth, scale and no regularisation. Note the difference in the y-axis scale. The F1-scores for the best model, under the O1 configuration, are as follows: Detritus = 75.2, Puff = 73.1, Rhizaria = 48.5, FI (Fiber-Filament) = 77.6, Copepoda = 46.6, Artefact = 92.5, Eumalacostraca = 27.2, Chaetognatha = 29.3, Foraminifera = 36.1, and Salpida = 37.8. (k) Comparison of the F1-score of all classes using all model configurations.

speeds of approximately 17,000 objects per hour of human annotation. Our method, however, represents a significant leap forward, demonstrating the ability to process around 200,000 to 450,000 objects per hour of human annotation, as illustrated in Fig10. In terms of class management, large number of classes (e.g. see MorphoCluster [7]) often include the same class in different orientations, or mixed classes (in between two pure classes, like copepods and detritus), which are often subsequently merged by the plankton researcher for their analysis. In our approach, such 'pseudo-classes' would be merged, which is not only computationally efficient but also aligns closely with the needs of final ecological interpretation, where such distinctions are often unnecessary and combined for broader ecological insights.

The methodology aims to diminish the reliance on extensive human annotation, which is particularly beneficial for large-scale monitoring projects where manual annotation of vast amounts of data is impractical. By automating this process, resources can be reallocated to other critical tasks within marine research. The model is well-suited for deployment on edge devices commonly used in marine environments, such as AUVs, floats, or gliders. These devices benefit from the model's real-time processing capabilities, which reduce the need for extensive data transmission back to shore-based systems, allowing for immediate decision-making directly in the field. However, it is critical to recognize that the model's operational efficiency comes with a trade-off in the granularity of data classification. Operating with a minimal amount of data embedding—while advantageous for conserving bandwidth and reducing data transmission over satellite—inevitably leads to a reduction in the resolution of data classification. This resolution reduction predominantly affects the classification of

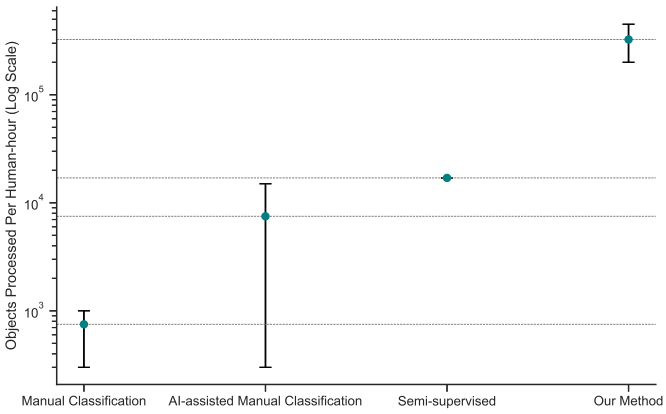


Fig. 10. Comparison of object processing speeds across different annotation methods. This plot illustrates the processing speeds of human manual classification, AI-assisted manual classification (Ecotaxa [6]), semi-supervised (MorphoCluster [7]) and our proposed method, showcasing the average number of objects processed per human-hour and the range for each method. Error bars indicate the minimum and maximum throughput.

less common species, which could be a significant limitation for ecologists requiring detailed taxonomic analysis across a diverse range of plankton types. However, the use of more powerful platforms that can afford higher computational resources and energy, could be considered. Such devices would enable the use of more complex models without the constraints of lesser-equipped platforms, thereby enhancing both the depth and breadth of marine ecological studies.

In our study, variability in plankton image classification primarily stems from fluctuations in environmental conditions such as lighting, water clarity, or sensor discrepancies, similar to spectral variability observed in many remote sensing scenarios [84]. Such variability significantly impacts the performance of traditional image classification models by altering the visual appearance of plankton, leading to inconsistencies and inaccuracies in species identification and quantification. To address this, our approach integrates metadata to enhance representation learning, enabling the model to contextualize each image and adapt its classification decisions to specific environmental conditions. This methodology not only increases the model's resilience against variable conditions but also ensures effectiveness under diverse and challenging scenarios. Applied to datasets characterized by high levels of variability, we anticipate that the model would demonstrate superior resilience compared to traditional methods, as the use of environmental metadata facilitates effective normalization of imaging differences. Looking forward, investigating the impacts of additional environmental variables such as water temperature, salinity, and light intensity on classification accuracy presents a promising avenue for future research. Further, developing adaptive algorithms that dynamically adjust to real-time environmental changes could significantly advance the field of aquatic bio-imaging, providing robust solutions for accurate plankton classification under varying conditions.

VI. Conclusion

In this paper, we present a novel representation learning model designed for plankton image classification, taking inspiration from metadata information that serves as valuable guidance for class grouping. Our approach involves multiple stages to achieve efficient representation learning and improve classification performance. Firstly, we employ an encoder-decoder network that optimizes the reconstruction loss and employs a regularised loss function integrated with location, depth, and scale information. This fusion of metadata information helps in effectively grouping species, enhancing the model's understanding of the underlying patterns. Secondly, we focus on data reduction for the majority class and select representative samples from the latent space data. These selected representative samples are then labelled and grouped based on different class views, as suggested by our encoder-decoder architecture. This approach enables us to handle data with class imbalances and efficiently leverage the labelled samples for training. Lastly, we utilise a CNN network to fine-tune the model for the downstream classification task, leveraging the representative annotated samples. This fine-tuning step further refines the model's capabilities and enhances its classification performance.

While our suggested framework demonstrates promising results, there is substantial scope for advancing its performance through more sophisticated representation learning techniques. A notable direction involves integrating stronger constraints, such as cosine similarity in contrastive learning, could enhance the network's capability to fully utilise metadata information without relying heavily on human annotations. The scalability and efficiency of this approach are particularly valuable for deployment in real-time systems on autonomous vehicles and remote sensing platforms, where rapid and reliable image analysis is critical. These methods could potentially reduce the risk of overfitting, particularly when the data includes noise or extraneous details. As a result, the model becomes more adaptable to various scenarios. Furthermore, they are efficient in environments needing quick similarity computations, like real-time recommendation engines or interactive systems. Future studies could explore these aspects, potentially transforming the way we process and utilise marine imagery in dynamic, resource-constrained environments. By advancing these techniques, we aim to minimise reliance on extensive human annotations, further automating and enhancing the accuracy of ecological monitoring and conservation efforts.

Additionally, it's important to acknowledge that our current dataset is limited to a specific instrument and exhibits narrow variability (i.e. primarily concentrated near the coast). This constraint hinders our model's ability to uncover broader patterns and insights about the data. To overcome this limitation, future investigations could focus on datasets with greater variability, providing the model with more diverse and comprehensive data for learning. Comprehensive testing on newer version camera systems could also be considered to assess performance across diverse datasets. Through these endeavors, future studies aim to verify whether the model remains effective and agnostic to different systems, sustaining

its effectiveness in plankton image classification tasks across various setups.

ACKNOWLEDGMENTS

This work was supported through the ANTICS project, receiving funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No 950212), and through the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 101000858 (TechOceanS).

REFERENCES

[1] G. Beaugrand, A. McQuatters-Gollop, M. Edwards, and E. Goberville, "Long-term responses of north atlantic calcifying plankton to climate change," *Nature Climate Change*, vol. 3, no. 3, pp. 263–267, 2013.

[2] K. Kalloniati, E. Christou, A. Kournopoulou, J. Gittings, I. Theodorou, S. Zervoudaki, and D. Raitsos, "Long-term warming and human-induced plankton shifts at a coastal eastern mediterranean site," *Scientific Reports*, vol. 13, no. 1, p. 21068, 2023.

[3] F. Lombard, E. Boss, A. M. Waite, M. Vogt, J. Uitz, L. Stemmann, H. M. Sosik, J. Schulz, J.-B. Romagnan, M. Picheral *et al.*, "Globally consistent quantitative observations of planktonic ecosystems," *Frontiers in Marine Science*, vol. 6, p. 196, 2019.

[4] S. L. C. Giering, E. L. Cavan, S. L. Basedow, N. Briggs, A. B. Burd, L. J. Darroch, L. Guidi, J.-O. Irisson, M. H. Iversen, R. Kiko *et al.*, "Sinking organic particles in the ocean—flux estimates from in situ optical devices," *Frontiers in Marine Science*, vol. 6, p. 834, 2020.

[5] M. Picheral, C. Catalano, D. Brousseau, H. Claustre, L. Coppola, E. Leymarie, J. Coindat, F. Dias, S. Fevre, L. Guidi *et al.*, "The underwater vision profiler 6: an imaging sensor of particle size spectra and plankton, for autonomous and cabled platforms," *Limnology and Oceanography: Methods*, vol. 20, no. 2, pp. 115–129, 2022.

[6] M. Picheral, S. Colin, and J.-O. Irisson, "Ecotaxa, a tool for the taxonomic classification of images," 2017.

[7] S.-M. Schröder, R. Kiko, and R. Koch, "Morphocluster: Efficient annotation of plankton images by clustering," *Sensors*, vol. 20, no. 11, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/11/3060

[8] F. Zhao, F. Lin, and H. S. Seah, "Binary sipper plankton image classification using random subspace," *Neurocomputing*, vol. 73, no. 10-12, pp. 1853–1860, 2010.

[9] P.-H. Liu, S.-F. Su, M.-C. Chen, and C.-C. Hsiao, "Deep learning and its application to general image classification," in *2015 International Conference on Informative and Cybernetics for Computational Social Systems (ICCSS)*. IEEE, 2015, pp. 7–10.

[10] G. Gorsky, M. D. Ohman, M. Picheral, S. Gasparini, L. Stemmann, J.-B. Romagnan, A. Cawood, S. Pesant, C. García-Comas, and F. Prejger, "Digital zooplankton image analysis using the zooscan integrated system," *Journal of plankton research*, vol. 32, no. 3, pp. 285–303, 2010.

[11] H. M. Sosik and R. J. Olson, "Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry," *Limnology and Oceanography: Methods*, vol. 5, no. 6, pp. 204–216, 2007.

[12] J. Ellen, H. Li, and M. D. Ohman, *Quantifying California current plankton samples with efficient machine learning techniques*. IEEE, 2015.

[13] P. F. Culverhouse, R. Simpson, R. Ellis, J. Lindley, R. Williams, T. Parisini, B. Reguera, I. Bravo, R. Zoppoli, G. Earnshaw *et al.*, "Automatic classification of field-collected dinoflagellates by artificial neural network," *Marine Ecology Progress Series*, vol. 139, pp. 281–287, 1996.

[14] M. B. Blaschko, G. Holness, M. A. Mattar, D. Lisin, P. E. Utgoff, A. R. Hanson, H. Schultz, E. M. Riseman, M. E. Sieracki, W. M. Balch *et al.*, "Automatic in situ identification of plankton," in *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1*, vol. 1. IEEE, 2005, pp. 79–86.

[15] Z. Liu, J. Watson, and A. Allen, "Efficient affine-invariant fourier descriptors for identification of marine plankton," in *OCEANS 2017-Aberdeen*. IEEE, 2017, pp. 1–9.

[16] J. Y. Luo, J.-O. Irisson, B. Graham, C. Guigand, A. Sarafraz, C. Mader, and R. K. Cowen, "Automated plankton image analysis using convolutional neural networks," *Limnology and Oceanography: methods*, vol. 16, no. 12, pp. 814–827, 2018.

[17] P. González, A. Castaño, E. E. Peacock, J. Díez, J. J. Del Coz, and H. M. Sosik, "Automatic plankton quantification using deep features," *Journal of Plankton Research*, vol. 41, no. 4, pp. 449–463, 2019.

[18] E. C. Orenstein and O. Beijbom, "Transfer learning and deep feature extraction for planktonic image data sets," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 1082–1088.

[19] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[20] J. Cui, B. Wei, C. Wang, Z. Yu, H. Zheng, B. Zheng, and H. Yang, "Texture and shape information fusion of convolutional neural network for plankton image classification," in *2018 Oceans-MTS/IEEE Kobe Techno-Oceans (OTO)*. IEEE, 2018, pp. 1–5.

[21] O. Py, H. Hong, and S. Zhongzhi, "Plankton classification with deep convolutional neural networks," in *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference*. IEEE, 2016, pp. 132–136.

[22] J. Dai, R. Wang, H. Zheng, G. Ji, and X. Qiao, "Zooplanktonet: Deep convolutional network for zooplankton classification," in *OCEANS 2016-Shanghai*. IEEE, 2016, pp. 1–6.

[23] B. Guo, L. Nyman, A. R. Nayak, D. Milmore, M. McFarland, M. S. Twardowski, J. M. Sullivan, J. Yu, and J. Hong, "Automated plankton classification from holographic imagery with deep convolutional neural networks," *Limnology and Oceanography: Methods*, vol. 19, no. 1, pp. 21–36, 2021.

[24] H. Lee, M. Park, and J. Kim, "Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3713–3717.

[25] C. Wang, Z. Yu, H. Zheng, N. Wang, and B. Zheng, "Cgan-plankton: Towards large-scale imbalanced class generation and fine-grained classification," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 855–859.

[26] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data mining and knowledge discovery*, vol. 28, pp. 92–122, 2014.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[28] J. L. Walker and E. C. Orenstein, "Improving rare-class recognition of marine plankton with hard negative mining," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3672–3682.

[29] S. L. Giering, P. F. Culverhouse, D. G. Johns, A. McQuatters-Gollop, and S. G. Pitois, "Are plankton nets a thing of the past? an assessment of in situ imaging of zooplankton for large-scale ecosystem assessment and policy decision-making," 2022.

[30] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1476–1485.

[31] G.-J. Qi and J. Luo, "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2168–2187, 2020.

[32] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[34] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 4, pp. 213–247, 2022.

[35] D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations*. MIT press, 1986.

[36] L. Madhuanand, F. Nex, and M. Y. Yang, "Self-supervised monocular depth estimation from oblique uav videos," *ISPRS journal of photogrammetry and remote sensing*, vol. 176, pp. 1–14, 2021.

[37] R. Kemker and C. Kanan, "Self-taught feature learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2693–2705, 2017.

[38] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-modalnet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 167, pp. 12–23, 2020.

[39] D. Hong, J. Chanussot, N. Yokoya, U. Heiden, W. Heldens, and X. X. Zhu, "Wu-net: A weakly-supervised unmixing network for remotely sensed hyperspectral imagery," in *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*. IEEE, 2019, pp. 373–376.

[40] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.

[41] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia *et al.*, "Spectralgpt: Spectral remote sensing foundation model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[42] M. R. S. B. DATA, "Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing big data in earth observation," *Innovation*, vol. 2, no. 1, p. 100055, 2024.

[43] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[44] L. Zhao, W. Luo, Q. Liao, S. Chen, and J. Wu, "Hyperspectral image classification with contrastive self-supervised learning under limited labeled samples," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[45] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 132–149.

[46] T. Yamada, A. Prügel-Bennett, and B. Thornton, "Learning features from georeferenced seafloor imagery with location guided autoencoders," *Journal of Field Robotics*, vol. 38, no. 1, pp. 52–67, 2021.

[47] T. Yamada, M. Massot-Campos, A. Prügel-Bennett, S. B. Williams, O. Pizarro, and B. Thornton, "Leveraging metadata in representation learning with georeferenced seafloor imagery," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7815–7822, 2021.

[48] T. Yamada, M. Massot-Campos, A. Prügel-Bennett, O. Pizarro, S. B. Williams, and B. Thornton, "Guiding labelling effort for efficient learning with georeferenced images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 593–607, 2022.

[49] J. S. Ellen, C. A. Graff, and M. D. Ohman, "Improving plankton image classification using context metadata," *Limnology and Oceanography: Methods*, vol. 17, no. 8, pp. 439–461, 2019.

[50] T. Konkle and G. A. Alvarez, "A self-supervised domain-general learning framework for human ventral stream representation," *Nature communications*, vol. 13, no. 1, p. 491, 2022.

[51] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science*, vol. 293, no. 5539, pp. 2425–2430, 2001.

[52] N. Kanwisher, "Functional specificity in the human brain: a window into the functional architecture of the mind," *Proceedings of the National Academy of Sciences*, vol. 107, no. 25, pp. 11 163–11 170, 2010.

[53] H. P. O. de Beeck, I. Pillet, and J. B. Ritchie, "Factors determining where category-selective areas emerge in visual cortex," *Trends in cognitive sciences*, vol. 23, no. 9, pp. 784–797, 2019.

[54] B. Long, C.-P. Yu, and T. Konkle, "Mid-level visual features underlie the high-level categorical organization of the ventral stream," *Proceedings of the National Academy of Sciences*, vol. 115, no. 38, pp. E9015–E9024, 2018.

[55] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[56] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proceedings of ICML workshop on unsupervised and transfer learning*. JMLR Workshop and Conference Proceedings, 2012, pp. 37–49.

[57] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan, "Learning features by watching objects move," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2701–2710.

[58] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[59] A. Darmochwał, "The euclidean space," *Formalized Mathematics*, vol. 2, no. 4, pp. 599–603, 1991.

[60] H. Xiong, J. Wu, and J. Chen, "K-means clustering versus validation measures: a data distribution perspective," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 779–784.

[61] M. Zimmer, "Detritus," in *Encyclopedia of Ecology*, S. E. Jørgensen and B. D. Fath, Eds. Oxford: Academic Press, 2008, pp. 903–911. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780080454054004754

[62] M. Stukel, K. Mislan, M. Décima, and L. Hmelo, "Detritus in the pelagic ocean," in *Eco-DAS IX Symposium Proceedings*. Association for the Sciences of Limnology and Oceanography Waco, TX, 2014, pp. 49–76.

[63] P. F. Culverhouse, R. Williams, B. Reguera, V. Herry, and S. González-Gil, "Do experts make mistakes? a comparison of human and machine indentification of dinoflagellates," *Marine ecology progress series*, vol. 247, pp. 17–25, 2003.

[64] P. F. Culverhouse, N. Macleod, R. Williams, M. C. Benfield, R. M. Lopes, and M. Picheral, "An empirical assessment of the consistency of taxonomic identifications," *Marine Biology Research*, vol. 10, no. 1, pp. 73–84, 2014.

[65] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

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

[67] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3573–3587, 2018.

[68] R. P. Espíndola and N. F. Ebecken, "On extending f-measure and g-mean metrics to multi-class problems," *WIT Transactions on Information and Communication Technologies*, vol. 35, 2005.

[69] S.-J. Yen and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5718–5727, 2009.

[70] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[71] B. Kamiński, M. Jakubczyk, and P. Szufel, "A framework for sensitivity analysis of decision trees," *Central European journal of operations research*, vol. 26, no. 1, pp. 135–159, 2018.

[72] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[73] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

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

[75] B. Gregory, L. Christophe, and E. Martin, "Rapid biogeographical plankton shifts in the north atlantic ocean," *Global change biology*, vol. 15, no. 7, pp. 1790–1803, 2009.

[76] E. Villarino, G. Chust, P. Licandro, M. Butenschön, L. Ibaibarriaga, A. Larrañaga, and X. Irigoien, "Modelling the future biogeography of north atlantic zooplankton communities in response to climate change," *Marine Ecology Progress Series*, vol. 531, pp. 121–142, 2015.

[77] K. Bandara, S. L. Basedow, G. Pedersen, and V. Tverberg, "Mid-summer vertical behavior of a high-latitude oceanic zooplankton community," *Journal of Marine Systems*, vol. 230, p. 103733, 2022.

[78] H. Ueda, A. Kuwahara, M. Tanaka, and M. Azeta, "Underwater observations on copepod swarms in temperate and subtropical waters," 1983.

[79] J. W. Ambler, "Zooplankton swarms: characteristics, proximal cues and proposed advantages," *Hydrobiologia*, vol. 480, pp. 155–164, 2002.

[80] T. Pados and R. F. Spielhagen, "Species distribution and depth habitat of recent planktic foraminifera in fram strait, arctic ocean," *Polar Research*, vol. 33, no. 1, p. 22483, 2014.

[81] S. E. Wilson, D. K. Steinberg, and K. O. Buesseler, "Changes in fecal pellet characteristics with depth as indicators of zooplankton repackaging of particles in the mesopelagic zone of the subtropical and subarctic north pacific ocean," *Deep Sea Research Part II: Topical Studies in Oceanography*, vol. 55, no. 14-15, pp. 1636–1647, 2008.

[82] S. L. Giering, B. Hosking, N. Briggs, and M. H. Iversen, "The interpretation of particle size, shape, and carbon flux of marine particle images is strongly affected by the choice of particle detection algorithm," *Frontiers in Marine Science*, vol. 7, p. 564, 2020.

[83] T. Platt and K. Denman, "Organisation in the pelagic ecosystem," *Helgoländer Wissenschaftliche Meeresuntersuchungen*, vol. 30, pp. 575–581, 1977.

[84] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1923–1938, 2018.

Mojtaba Masoudi received a BEng in software engineering from Shahrood University of Technology, Iran, followed by an MSc in Artificial Intelligence and Robotics from the School of Computer Science and Engineering, Ferdowsi University of Mashhad, Iran, in 2014 and 2017, respectively. Presently, he serves as a research scientist at the National Oceanography Centre, Southampton, UK, with a keen interest in the field of Ocean AI.

Sarah L.C. Giering is a Marine Biogeochemist interested in the role of plankton and particles in the ocean carbon cycle, recognised for having made key discoveries which are critical to understanding earth system functioning. She obtained a BSc (Hons.) in Marine Biology at the University of Aberdeen, UK, in 2009 and a PhD in Marine Biogeochemistry at the University of Southampton, UK, in 2013. She leads an interdisciplinary research group that aims to improve state-of-the-art technologies and methodologies to fill the biggest knowledge gaps in our understanding of the ocean carbon cycle.

Noushin Eftekhari holds a Ph.D. in artificial intelligence and is a post-doctoral research associate at the Alan Turing Institute. Her research encompasses a broad spectrum, primarily focusing on computer vision, cancer research, multi-modal data integration, and biodiversity monitoring. In cancer research, she specializes in developing deep learning models to explore personalized medicine. Additionally, she contributes to advancing biodiversity monitoring by creating innovative tools to enhance the monitoring process.

Miquel Massot-Campos received a MEng from BarcelonaTech, Spain in 2011, MSc and PhD from the University of the Balearic Islands, Spain in 2013 and 2019, respectively. He is currently with the University of Southampton, UK. He is a member of Centre for In Situ and Remote Intelligent Sensing and is interested in the scalability of autonomous underwater vehicle's missions.

Jean-Olivier Irisson is a computational ecologist. After graduating from the École Normale Supérieure in Paris, he studied the behaviour of tropical fish larvae and its implication for population connectivity for his Masters and PhD at the École Pratique des Hautes Études in Perpignan (France) as well as for his post-doc at the University of Miami. Since 2009, he is an associate professor at Sorbonne Université, in the Laboratoire d'Océanographie de Villefranche (Southern France). He leads the plankton ecology team and is involved in long term monitoring programs as well as research on plankton ecology, at the interface between environmental and computer sciences.

Blair Thornton (M'07) obtained a BEng in Naval Architecture and PhD in Underwater Robotics from Southampton University in 2002 and 2006, respectively. In 2003 he joined the Underwater Robotics and Application Lab., Institute of Industrial Science, UTokyo, before rejoining Southampton in 2016 where he is Professor of Marine Autonomy. He has spent 450+ days at sea deploying robotic systems and is dedicated to generating data and insight in marine science through improved sensing and autonomy.