


Improving in situ real-time classification of long-tail marine plankton images for ecosystem studies

Noushin Eftekhari¹, Sophie Pitois², Mojtaba Masoudi³, Robert E. Blackwell¹ , James Scott², Sarah L.C. Giering³, and Matthew Fry¹

¹ The Alan Turing Institute, British Library, 96 Euston Rd., London NW1 2DB, UK

² Centre for Environment Fisheries and Aquaculture Science, UK

³ National Oceanography Centre, European Way, Southampton SO14 3ZH, UK

Abstract. The escalating complexity of image classification tasks in ecological monitoring has highlighted the limitations of conventional machine learning models, particularly when faced with long-tailed data distributions typical of natural environments. This paper introduces a comprehensive framework that leverages a novel dataset obtained from the Plankton Imaging (Pi-10) instrument designed to enhance plankton image classification accuracy in a real-time application. We employ cutting-edge image classification architectures, including pre-trained Vision Transformers (ViT) and BEiT. We integrate the Label-Aware Smoothing Model into our training process to address the challenges of long-tailed data distributions often encountered in ecological datasets. Further, we innovate with dynamic label-aware smoothing, which adjusts smoothing factors based on attention scores from ViTs to tailor model confidence to the significance of different image regions. The results demonstrate marked improvements in classification performance on the Pi-10 dataset, effectively handling long-tail distribution challenges and setting new benchmarks for real-time image classification in ecological research. This approach advances the state of ecological imaging and provides a scalable solution adaptable to other domains encountering similar distributional challenges.

Keywords: Transformer · Long-Tail Recognition · Plankton Image Analysis

1 Introduction

Plankton, microscopic organisms in marine and freshwater environments, are key indicators of aquatic ecosystem health. Their sensitivity to environmental changes and short life cycles make them crucial for monitoring. By observing plankton, we gain insights into the biogeochemistry and productivity of the oceans, as their presence and behavior reflect the impacts of changing conditions [33, 38]. Image processing technologies offer a promising approach to enhancing the accuracy and efficiency of plankton monitoring. These technologies allow for

detailed insights into plankton distribution across various scales, reducing the reliance on traditional, labour-intensive methods that involve manual sampling and identification [2, 11, 18, 27].

However, applying machine learning and image processing techniques to plankton data introduces significant challenges, such as data set shifts (DSS), when the characteristics of the plankton images in the training set may not fully represent the diversity and variability present in the open ocean particles that the real-time classifier will encounter during deployment [31]. The DSS complicates the application of conventional machine learning models, as the feature distribution in training sets often fails to mirror real-world conditions [35].

Traditional deep-learning approaches to plankton image classification are particularly challenging due to the high variability of features within classes and the similarity between different classes. These factors complicate practical model training and accurate classification. However, given their significant success in computer vision, Vision Transformers (ViTs) are becoming increasingly popular and widely utilised in visual recognition tasks [14]. In response, a comprehensive study shows that ViTs outperform convolutional neural networks (CNN)-based methods [26]. However, real-world data often exhibits significant class imbalance that can severely skew outcomes in data-driven deep neural networks, making the task of Long-Tailed Recognition (LTR) particularly challenging, where common species (head) are over-represented relative to rarer species (tail). Traditional techniques - such as random downsampling of majority classes or oversampling of minority classes - can be used to address class imbalance, though may lead to the loss of important information or overfitting [3, 46]. To overcome these issues, LTR and transfer learning are increasingly utilised to enhance model adaptability and accuracy [15, 19, 26, 37].

This paper introduces a new dataset from the Plankton Imaging (Pi-10) instrument, a state-of-the-art high-speed colour line scan imaging device, alongside a real-time framework designed to elevate image classification efficacy, particularly addressing the complexities introduced by long-tailed data distributions. We have adapted leading-edge, pre-trained image classification architectures, notably the ViT [14] and BEiT [1], which are renowned for their exceptional performance across various tasks. Additionally, We have integrated the advanced Label-Aware Smoothing model (LAS) [51] into our training process to enhance efficacy, boost real-time classification capabilities, and effectively address LTR challenges [20, 40]. To further improve classification on highly imbalanced datasets, we propose a novel dynamic label-aware smoothing method that adjusts smoothing factors based on the attention scores computed for each patch by ViTs. This innovation ensures that the label smoothing factors are dynamically tuned, reflecting the varying significance of different image parts as perceived by the transformer architecture. In summary, our contributions are (i) the adoption of ViT and BEiT models to achieve superior image classification performance, (ii) the integration of dynamic label-aware smoothing to address long-tail data challenges effectively, and (iii) the effective use of transfer learning to enhance model performance on the newly introduced Pi-10 dataset.

2 Related work

Transformers The Vision Transformer (ViT) model represents a significant departure from traditional image classification methods, which predominantly utilised Convolutional Neural Networks (CNNs). The ViT model adopts the Transformer architecture, originally designed for natural language processing tasks [45], and adapts it for visual data. This approach allows the model to effectively handle long-range dependencies between image patches [14], introducing a novel methodology for image analysis. Complementing the ViT, the Bidirectional Encoder Representation from Image Transformers (BEiT) [1] employs a self-supervised learning approach, inspired by BERT [13], for pre-training vision transformers. This is achieved through a masked image modeling task, enabling the model to learn valuable visual representations without the need for labeled data, thus significantly improving its efficacy in downstream tasks.

The BEiT model was pre-trained on extensive datasets [42], and subsequently fine-tuned in a supervised fashion on ImageNet [12]. BEiT demonstrates superior capabilities in downstream image classification tasks, potentially surpassing traditional CNNs [5, 47]. ViT and BEiT have been used for plankton image analysis and showed promising results in enhancing ecology studies [8, 26]. Here, we incorporate ViT and BEiT with LTR strategies, representing significant advancements over traditional CNNs. Our results demonstrate that the ViT model is competitive with the BEiT model when applied to long-tail data, particularly after incorporating dynamic label-aware smoothing.

Transfer learning in plankton image classification Previous work on plankton image classification has demonstrated that out-of-domain transfer learning, using models pre-trained on large-scale natural image datasets such as ImageNet1K or ImageNet22K, yields good results [26]. Hence, we use a pre-trained ResNet-18 model with ImageNet as a baseline model. We then employ the ViT model, which uses the Transformer architecture (see 4.2 for more details), which was initially trained on ImageNet-21k and subsequently fine-tuned on ImageNet 2012. Furthermore, we use the BEiT model, pre-trained in a self-supervised manner on the same dataset as ViT (also known as ImageNet-21k [12]). This dataset consists of 14 million images and 21,841 classes, at a resolution of 224x224, and the model is subsequently fine-tuned on the same dataset at the same resolution.

Long-tail learning The challenges posed by imbalanced datasets in machine learning are critical, especially for tasks that rely on accurate representation of minority classes [20, 28, 40]. Various strategies have been developed to tackle this issue, including Post-hoc correction methods, which adjust a model’s outputs after training to better represent minority classes [3, 9, 17, 23, 36]. Data modification strategies alter the training dataset by methods like oversampling the less represented classes or undersampling the overrepresented ones, thereby aiming to balance the training environment [6, 7, 48]. Additionally, Loss weighting involves adjusting the loss function to impose heavier penalties for misclassifications of minority classes, thus directing the model’s focus towards these classes [10, 10, 16, 22, 29]. Lastly, Margin modification techniques, including logit

adjustment, tailor the decision boundary by modifying class logits to enhance the classifier’s fairness towards minority classes [21, 24, 30, 43, 49, 50].

The challenge posed by the skewed distribution of class frequencies in our plankton dataset was met with a novel integration of LTR strategies during the training phase. We implemented a Label-Aware Smoothing model (LAS) for ViT and compared it with a range of approaches including Class-Balanced (CB), Label-Distribution-Aware Margin (LDAM), Balanced Cross-Entropy (Bal_CE), and an adjusted form of Binary Cross-Entropy (BCE) loss [4, 10, 30]. The LAS loss was the most accurate, significantly enhancing model performance across minority classes. Additionally, transfer learning played a pivotal role in refining model accuracy and training efficiency. By incorporating the newly introduced Pi-10 dataset and fine-tuning pre-existing models over a concise series of training epochs, we achieved rapid model convergence and marked improvement in accuracy.

3 Plankton imager system

The Plankton Imager Pi-10⁴ is a high-speed color line scan imaging instrument [11, 34, 39]. The instrument is connected to the clean sea water underway system (inlet at 4 m depth), which supplies water continuously (flow rate of 34 L min⁻¹). It captures images of passing particles within a size range of 180 μm to 3.5 cm at a resolution of 10 μm . Images are taken in RGB color using an EPIX E8 frame store, and processed to extract a region of interest, saved in TIFF format with a timestamp and unique identifier. Raw images are stored in 12-bit resolution and then converted to 8-bit for viewing and analysis. The Pi-10 operates continuously during surveys, capturing images of all particles passing through the flow cell, with only mesozooplankton (200 μm – 2 cm) processed and saved due to file-size constraints and operational needs. Owing to plankton’s skewed distributions towards smaller species, most of the captured images show small plankton and particles [44]. The annotation process involves a trained taxonomist sorting through the collection of images, categorizing them into ecologically relevant groups (supplementary materials). The plankton dataset was gathered during a research cruise on the Research Vessel Cefas Endeavour at a specific site (Fig.1) in the North Sea, utilising the Plankton Imager (Pi-10 version).

Realtime access to plankton imager data We developed a real-time classification pipeline where images are taken and immediately classified. This pipeline bypasses the often long periods between image collection, big data transferring, and image analysis. The real-time data collection pipeline is equipped with a Pi-10 instrument. For in-situ image processing and classification tasks, we utilise the "Edge-AI" The NVIDIA Jetson AGX Orin⁵ is a compact and powerful edge device equipped with a GPU explicitly designed for AI and machine learning applications in resource-constrained environments. We chose this device

⁴ More details can be found at www.planktonanalytics.com

⁵ www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-agx-orin/

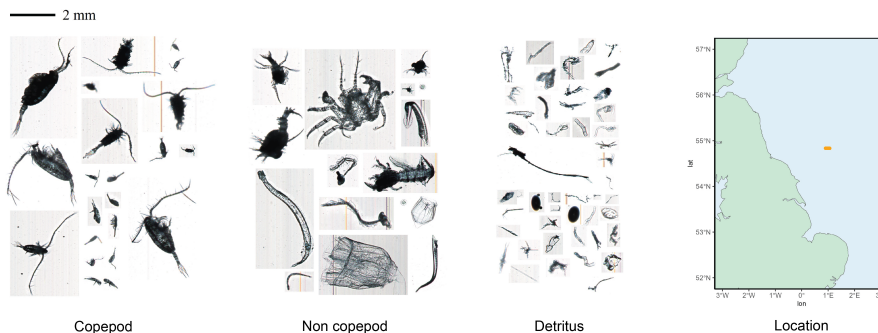


Fig. 1: Sample diversity captured by the Plankton Imager (Location of the Research Vesse): a comparative display of copepods, non-copepods, and detritus

because of its optimal balance of performance and size, which makes it ideal for deploying advanced AI workloads directly at the edge without relying on cloud infrastructure. This system processes the images to categorise them into counts of copepods, non-copepods, and detritus via real-time classification using a transformer model. Summarized data are then transmitted via the ship’s broadband satellite communication systems to a cloud-hosted digital dashboard⁶, enabling remote access and analysis. The Pi-10 imaging system captures images that are processed in real-time and simultaneously stored on a hard drive for further analysis. In this study, we utilised a comprehensive dataset of 55,570 samples from the Pi-10 dataset gathered during a fisheries survey conducted in the North Sea in August 2023. The data, meticulously labeled by taxonomists, was used to train our proposed model, ensuring accuracy and relevance in the classification process. The evaluation of the final model occurs in two stages: initially, it was assessed using a designated test set; subsequently, it will be integrated into the real-time pipeline for monitoring marine plankton. This dual-phase testing ensures both the efficacy and practical applicability of the model, culminating in its deployment to monitor plankton via a live dashboard actively.

Train/validation/test split To ensure robust model performance assessment, we employed a data splitting strategy: 70% of the dataset was allocated for training and 10% for validation purposes, while the remaining 20% was reserved for testing. The total number of tests is 11,114, comprising 2,451 for copepod, 8,203 for detritus, and 460 for non-copepod.

4 Methodology

4.1 Task definition

In studying visual recognition with skewed class distributions, we analyse a dataset $D = \{X, Y\}$ consisting of N entries divided among K different classes.

⁶ More details can be found at <https://planktonapi-dev.cefastest.co.uk>.

Each data point x_i in X corresponds to a label y_i from Y , with y_i ranging from 1 to K . The representation of each class K_i differs markedly, with n_i instances per class, characterizing the dataset’s long-tail distribution. We denote the disparity in class representation by the imbalance ratio $\gamma = \frac{n_{\max}}{n_{\min}}$. To address this imbalance, we propose to develop a model $M = \{F_{\theta_f}, W_{\theta_w}\}$, incorporating a feature encoder F_{θ_f} and a classifier W_{θ_w} , tailored to effectively handle these disparities.

4.2 Classification model

The classification process begins by segmenting each input image into smaller patches. Each patch is linearly transformed and embedded with positional encodings to preserve spatial relationships within the image. These encoded patches are then passed through multiple layers of the Vision Transformer (ViT) encoder, where they undergo multi-headed self-attention and feed-forward processing. The output from the final encoder layer is fed into a classification head that predicts the image class. The operations within the transformer utilised by the ViT are mathematically described by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q , K , and V represent the query, key, and value matrices derived from the input embeddings, and d_k is the dimensionality of the keys.

4.3 Label-aware smoothing

The label-aware smoothing (LAS) [51] model is crucial for addressing issues related to the over-confidence of predictions typically seen with cross-entropy loss and the varying distributions of predicted probabilities across classes.

The LAS loss function is defined as:

$$l(q, p) = -\sum_{i=1}^K q_i \log p_i, \quad (2)$$

where q_i is the target distribution after label-aware smoothing:

$$q_i = \begin{cases} 1 - \epsilon_y = \frac{\epsilon_y}{K-1} f(N_j), & \text{if } i = y, \\ \frac{\epsilon_y}{K-1}, & \text{otherwise.} \end{cases} \quad (3)$$

Here, ϵ is the base smoothing parameter, y is the true class, and N_y represents the number of instances for class y , influencing the function $f(N_y)$. The label-aware smoothing function $f(N_y)$ is designed to vary the smoothing effect dynamically according to class frequency, making the model more sensitive to less frequent classes. This function is defined in several forms, such as linear,

concave, and convex. In this study, we employ the concave form, which has demonstrated promising outcomes in initial experiments:

$$f(N_y) = \epsilon_K + (\epsilon_1 - \epsilon_K) \sin\left(\frac{\pi(N_y - N_K)}{2(N_1 - N_K)}\right) \quad (4)$$

where:

- ϵ_1 and ϵ_K are the maximum and minimum smoothing parameters, respectively.
- N_1 and N_K are the maximum and minimum class frequencies in the dataset, respectively.

The concave form of label smoothing increases regularisation more substantially for classes with fewer instances, effectively strengthening regularisation where it is most needed. This method aims to mitigate bias toward frequently occurring classes by moderating the confidence levels of the predictions, thereby enhancing performance in less common classes. The formula for cross-entropy loss after applying the softmax function is defined as

$$l(y; p) = -\log(p_y) = -w_y^T x + \log\left(\sum_i \exp(w_i^T x)\right), \quad (5)$$

where $y \in \{1, 2, \dots, K\}$ denotes the label. Here, the feature vector x resides in \mathbb{R}^M and is input into the classifier, and w_i represents the i -th column vector of the weight matrix W . The ideal scenario occurs when $w_y^* x \rightarrow \infty$ and $w_i^T x$ for $i \neq y$ remains comparatively low. This optimization is supported by the adjusted softmax weights:

$$w_i^* x = \begin{cases} \log\left(\frac{(K-1)(1-\epsilon_y)}{\epsilon_y}\right) + c, & \text{if } i = y, \\ c, & \text{otherwise,} \end{cases} \quad (6)$$

where c is a calibration constant.

4.4 Dynamic label-aware smoothing for ViTs

Label-aware smoothing in traditional CNNs adjusts the confidence levels assigned to class labels to mitigate the model’s overconfidence. For ViTs, which interpret the global context of an image differently through a self-attention mechanism, we propose a dynamic label-aware smoothing method that adjusts smoothing factors based on the attention scores computed for each patch. The modified label smoothing formula is given by:

$$q_i = \begin{cases} 1 - \epsilon_y \cdot \alpha_i & \text{if } i = \text{true class,} \\ \frac{\epsilon_y \cdot \alpha_i}{K-1} & \text{otherwise,} \end{cases} \quad (7)$$

where α_i represents the normalised attention weight assigned to the patch corresponding to the true class, and ϵ_y is the base smoothing factor, chosen based

on preliminary experiments to optimise model performance. This adjustment ensures that the smoothing factor is contextually relevant, reflecting the varying significance of different image parts as perceived by the transformer.

4.5 Attention-weighted smoothing factor

In ViTs, the self-attention mechanism assigns different weights to input image patches, reflecting their importance. These weights, $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_N\}$ for N patches, are derived using the softmax function to ensure they are normalised and sum to one, essential for their subsequent use as dynamic factors in label smoothing. The normalised attention weight for each patch i is calculated as:

$$\alpha_i = \frac{\exp(a_i)}{\sum_{j=1}^N \exp(a_j)} \quad (8)$$

where a_i is the raw attention score for patch i . These weights are then used to scale the label smoothing factors dynamically. This dynamic adjustment ensures that patches deemed more significant by the attention mechanism have a greater influence on the smoothed labels, thus aligning the model’s learning process more closely with the intrinsic data distribution and enhancing its performance on imbalanced datasets.

5 Experiment details

5.1 Model training

For our experiments, we utilised two pre-trained models: Google’s ViT [47] and Microsoft’s BEiT [12], both adapted to classify the three distinct categories depicted in Fig. 1. The input images underwent preprocessing with transformations, including random resized cropping to (224×224) , random horizontal flipping, and rotations of up to 10 degrees. Additionally, brightness, contrast, saturation, and hue were moderately adjusted to augment the dataset. We applied normalisation using the default ImageNet means and standard deviations. The dataset was split into training (70%) and validation (10%) sets. Class weights were computed and normalised to ensure balanced learning across the categories.

To handle the long-tail distribution, we evaluated several techniques, including LAS loss and Bal_CE loss, with parameters fine-tuned according to the class weights and specific model requirements. Optimisation was conducted using the Adam optimiser with an initial learning rate of 0.00005, with training extending up to 30 epochs or until convergence was observed on the validation set.

The optimal model configurations were determined by achieving the highest validation accuracy. We fine-tuned hyperparameters such as learning rate and batch size in subsequent training runs, with detailed configuration tracking using Weights & Biases. To ensure reproducibility, three random seeds were used across multiple runs. The results are reported as the best validation accuracy obtained during training, demonstrating that the models—especially

ViT—performed well, outperforming traditional CNN-based models and offering competitive results compared to BEiT.

5.2 Hyperparameter grid search

To optimise the performance of our models, we conducted a hyperparameter grid search, systematically exploring multiple combinations to identify the best configuration. For the ResNet model, the grid included the number of epochs $\{20, 30, 100\}$ and learning rates $\{0.01, 0.001, 0.0001\}$. The optimal configuration, determined based on validation accuracy, was a learning rate of 0.0001 and 30 epochs. For the fine-tuned ViT and BEiT models, we explored learning rates $\{1 \times 10^{-4}, 5 \times 10^{-5}, 1 \times 10^{-5}\}$, epochs $\{10, 20, 30\}$, a training batch size of 32, an evaluation batch size of 32, and gradient accumulation steps set to 4. The optimiser used was Adam with β values $\{0.9, 0.999\}$ and an epsilon of 1×10^{-8} . We employed a linear learning rate scheduler with a warmup ratio of 0.1 to improve training stability. These hyperparameters were meticulously selected to ensure robust training and high accuracy.

5.3 Metrics

We evaluated each model by calculating the macro-averaged F1-score, which assigns equal importance to all classes, regardless of their frequency. This approach ensures that the F1-score, recall, and precision metrics are calculated in a way that considers each class equally. By using macro-averaging, the method is sensitive to underperformance in less frequent classes, highlighting areas where the model may struggle. Additionally, we normalised the confusion matrix to show percentages, offering a clear visual representation of the model’s strengths and weaknesses across different categories (supplementary materials).

6 Results

6.1 Classifier

Table 1 compares the different models and loss functions based on their macro-averaged and non-copepod-specific F1-scores. Our proposed LAS loss function consistently outperforms alternatives, particularly when integrated with the ViT architecture. The ViT-LAS model achieves the highest macro-average F1-score (0.98 ± 0.2) and non-copepod F1-score (0.95 ± 0.5), setting a new benchmark for balanced multi-class classification in this domain. While models utilising ST_CE and Bal_CE demonstrate competitive performance, they encounter challenges with the non-copepod class, where class imbalance is more severe. BEiT and ViT models trained with BCE and LDAM losses also achieve strong results but exhibit greater variability across multiple runs. In summary, our findings highlight the LAS loss function’s robustness in addressing class imbalance and improving classification performance, particularly for the underrepresented non-copepod class. The ViT-LAS model stands out as the top performer, making it the most dependable choice for this classification task.

Table 1: Comparative Performance Metrics of Deep Learning Models: This Table Displays F1-Scores and Standard Deviations for Various Models Under Different Loss Functions and Training Epochs, Covering Both Overall Macro Averages and Minority Class-Specific Scores for Non-Copepods.

Model Name	Loss Function	Number of Epochs	F1-Score (macro avg)	F1-Score (non-copepod)
ResNet	ST_CE	30	0.93 ± 0.07	0.83 ± 0.2
Beit	Bal_CE	20	0.95 ± 0.05	0.88 ± 0.5
Beit	LDAM	20	0.96 ± 0.2	0.90 ± 0.4
Beit	LAS (ours)	20	0.97 ± 0.3	0.93 ± 0.7
Beit	BCE	10	0.96 ± 0.1	0.91 ± 0.8
Beit	ST_CE	20	0.94 ± 0.5	0.90 ± 0.3
Beit	CB_CE	20	0.94 ± 0.7	0.86 ± 0.2
ViT	ST_CE	30	0.93 ± 0.4	0.90 ± 0.1
ViT	BCE	20	0.97 ± 0.5	0.93 ± 0.6
ViT	Bal_CE	20	0.97 ± 0.2	0.92 ± 0.6
ViT	CB_CE	10	0.97 ± 0.7	0.92 ± 0.5
ViT	LDAM	10	0.97 ± 0.4	0.94 ± 0.5
ViT	LAS (ours)	30	0.98 ± 0.2	0.95 ± 0.5

7 Discussion

The experimental results emphasise the crucial role of Transformers and advanced loss functions in addressing the challenges of plankton image classification. Our approach demonstrated a 0.12% improvement in the F1-score for the minority class—a notable achievement given the severe class imbalance typically found in such datasets. This gain underscores the effectiveness of integrating ViT models with the Label-Aware Smoothing (LAS) loss function, highlighting their capability to manage complex, imbalanced data distributions.

Marine particle classification, especially differentiating between the key classes copepod, detritus, and non-copepod presents a significant challenge due to the skewed distribution and high variability within these categories. Our methodology capitalised on cutting-edge techniques such as transfer learning and long-tail learning strategies to tackle this highly imbalanced dataset. By integrating pre-trained ViT and BEiT architectures with advanced long-tail recognition (LTR) loss functions, our model achieved substantial performance gains. Specifically, the proposed ViT-LAS approach achieved an overall F1-score of 98%, with an impressive 95% F1-score for the tail class (non-copepods), a considerable improvement over the traditional ResNet model. In comparison, while ResNet attained an overall F1-score of 93%, it struggled with the non-copepod category, achieving only 83%.

These results align with existing literature comparing transformer-based models and convolutional neural networks (CNNs) in imbalanced classification tasks [8, 26, 32]. Research consistently shows that ViT models, leveraging their global attention mechanisms, excel at extracting features from underrepresented classes,

leading to superior performance in settings characterised by severe class imbalance. Additionally, the LAS technique complements this by reducing bias towards majority classes and mitigating overfitting, which aligns with our observed improvements.

The integration of ViT with LAS and LTR strategies within an end-to-end trainable framework not only enhances overall classification accuracy but also significantly improves the performance of minority classes. This is particularly valuable in scenarios involving complex visual data with high intra-class variability.

8 Conclusion

Marine scientists studying pelagic ecosystems face significant challenges in managing and analysing vast imagery datasets from diverse habitats and multiple imaging systems. Automated camera-based sensors are extensively utilised in vessel-based research to monitor plankton and marine particles. Despite their utility, annotating data for fully supervised learning in plankton grouping tasks is costly and time-intensive. In response, real-time and open-source software can be adapted to classify images of other marine objects and species, helping transform how scientists study the oceans. This opens the door to a new era of monitoring beyond plankton, where measured variables can be seen in real-time, thus allowing for sampling to be adapted according to visible changes. Studying changes in community structures and biodiversity as they happen will help further our understanding of what drives changes in biodiversity and community structures within the environment, thus increasing our ability to use our seas sustainably.

Looking ahead, we plan to enhance our methods for analysing marine plankton in real-time at a higher taxonomic resolution and revising biodiversity from a computer vision perspective to tackle the DSS problem. This project focuses on developing a sensor-agnostic method to generalise edge-AI performance across different ecological habitats and adapt to changing and upgraded camera systems. By utilising foundation models and self-supervised learning methods, we aim to address the limitations of traditional machine learning approaches that require retraining to accommodate new camera systems. These advanced techniques will enable consistent categorisation of particles, even when camera systems change, without the need for extensive retraining of machine learning models. In future work, we will go up in the taxonomist tree and use the self-supervised learning method, which was more robust than the supervised method [25, 41].

9 Code availability

The code we used to train and evaluate our models is available at <https://github.com/noushineftekhari/ViT-LASNet>

Acknowledgements

This work was funded by The Alan Turing Institute.

References

1. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
2. Benfield, M.C., Grosjean, P., Culverhouse, P.F., Irigoien, X., Sieracki, M.E., Lopez-Urrutia, A., Dam, H.G., Hu, Q., Davis, C.S., Hansen, A., et al.: Rapid: research on automated plankton identification. *Oceanography* **20**(2), 172–187 (2007)
3. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks* **106**, 249–259 (2018)
4. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems* **32** (2019)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: *European conference on computer vision*. pp. 213–229. Springer (2020)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16**, 321–357 (2002)
7. Chen, J., Su, B.: Instance-specific semantic augmentation for long-tailed image classification. *IEEE Transactions on Image Processing* (2024)
8. Ciranni, M., Murino, V., Odone, F., Pastore, V.P.: Computer vision and deep learning meet plankton: Milestones and future directions. *Image and Vision Computing* p. 104934 (2024)
9. Collell, G., Prelec, D., Patil, K.: Reviving threshold-moving: a simple plug-in bagging ensemble for binary and multiclass imbalanced data. arXiv preprint arXiv:1606.08698 (2016)
10. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 9268–9277 (2019)
11. Culverhouse, P., Gallienne, C., Williams, R., Tilbury, J.: An instrument for rapid mesozooplankton monitoring at ocean basin scale (2015)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*. pp. 248–255. Ieee (2009)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Ellen, J.S., Graff, C.A., Ohman, M.D.: Improving plankton image classification using context metadata. *Limnology and Oceanography: Methods* **17**(8), 439–461 (2019)
16. Fan, Y., Lyu, S., Ying, Y., Hu, B.: Learning with average top-k loss. *Advances in neural information processing systems* **30** (2017)
17. Fawcett, T., Provost, F.J.: Combining data mining and machine learning for effective user profiling. In: *KDD*. vol. 96, pp. 8–13 (1996)
18. Giering, S.L., Sanders, R., Lampitt, R.S., Anderson, T.R., Tamburini, C., Boutrif, M., Zubkov, M.V., Marsay, C.M., Henson, S.A., Saw, K., et al.: Reconciliation of the carbon budget in the ocean’s twilight zone. *Nature* **507**(7493), 480–483 (2014)

19. Guo, G., Lin, Q., Chen, T., Feng, Z., Wang, Z., Li, J.: Colorization for in situ marine plankton images. In: European Conference on Computer Vision. pp. 216–232. Springer (2022)
20. He, H., Garcia, E.A.: Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* **21**(9), 1263–1284 (2009)
21. Iranmehr, A., Masnadi-Shirazi, H., Vasconcelos, N.: Cost-sensitive support vector machines. *Neurocomputing* **343**, 50–64 (2019)
22. Jamal, M.A., Brown, M., Yang, M.H., Wang, L., Gong, B.: Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7610–7619 (2020)
23. King, G., Zeng, L.: Logistic regression in rare events data. *Political analysis* **9**(2), 137–163 (2001)
24. Li, M., Cheung, Y.m., Lu, Y., Hu, Z., Lan, W., Huang, H.: Adjusting logit in gaussian form for long-tailed visual recognition. *IEEE Transactions on Artificial Intelligence* (2024)
25. Liu, H., HaoChen, J.Z., Gaidon, A., Ma, T.: Self-supervised learning is more robust to dataset imbalance. arXiv preprint arXiv:2110.05025 (2021)
26. Maracani, A., Pastore, V.P., Natale, L., Rosasco, L., Odone, F.: In-domain versus out-of-domain transfer learning in plankton image classification. *Scientific Reports* **13**(1), 10443 (2023)
27. Martin, A., Boyd, P., Buesseler, K., Cetinic, I., Claustre, H., Giering, S., Henson, S., Irigoien, X., Kriest, I., Memery, L., et al.: The oceans’ twilight zone must be studied now, before it is too late. *Nature* **580**(7801), 26–28 (2020)
28. Masoudi, M., Giering, S.L., Eftekhari, N., Massot-Campos, M., Irisson, J.O., Thornton, B.: Optimizing plankton image classification with metadata-enhanced representation learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* pp. 1–18 (2024)
29. Menon, A., Narasimhan, H., Agarwal, S., Chawla, S.: On the statistical consistency of algorithms for binary classification under class imbalance. In: International Conference on Machine Learning. pp. 603–611. PMLR (2013)
30. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. arXiv preprint arXiv:2007.07314 (2020)
31. Moreno-Torres, J.G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N.V., Herrera, F.: A unifying view on dataset shift in classification. *Pattern recognition* **45**(1), 521–530 (2012)
32. Orenstein, E.C., Ayata, S.D., Maps, F., Becker, É.C., Benedetti, F., Biard, T., de Garidel-Thoron, T., Ellen, J.S., Ferrario, F., Giering, S.L., et al.: Machine learning techniques to characterize functional traits of plankton from image data. *Limnology and oceanography* **67**(8), 1647–1669 (2022)
33. Pitois, S., Yebra, L.: Contribution of marine zooplankton time series to the united nations decade of ocean science for sustainable development. *ICES Journal of Marine Science* **79**(3), 722–726 (2022)
34. Pitois, S.G., Tilbury, J., Bouch, P., Close, H., Barnett, S., Culverhouse, P.F.: Comparison of a cost-effective integrated plankton sampling and imaging instrument with traditional systems for mesozooplankton sampling in the celtic sea. *Frontiers in Marine Science* **5**, 5 (2018)
35. Plonus, R.M., Conradt, J., Harmer, A., Janßen, S., Floeter, J.: Automatic plankton image classification—can capsules and filters help cope with data set shift? *Limnology and Oceanography: Methods* **19**(3), 176–195 (2021)

36. Provost, F.: Machine learning from imbalanced data sets 101. In: Proceedings of the AAAI'2000 workshop on imbalanced data sets. vol. 68, pp. 1–3. AAAI Press (2000)
37. Pu, Y., Feng, Z., Wang, Z., Yang, Z., Li, J.: Anomaly detection for in situ marine plankton images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3661–3671 (2021)
38. Ratnarajah, L., Abu-Alhaija, R., Atkinson, A., Batten, S., Bax, N.J., Bernard, K.S., Canonico, G., Cornils, A., Everett, J.D., Grigoratou, M., et al.: Monitoring and modelling marine zooplankton in a changing climate. *Nature Communications* **14**(1), 564 (2023)
39. Scott, J., Pitois, S., Close, H., Almeida, N., Culverhouse, P., Tilbury, J., Malin, G.: In situ automated imaging, using the plankton imager, captures temporal variations in mesozooplankton using the celtic sea as a case study. *Journal of Plankton Research* **43**(2), 300–313 (2021)
40. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. *Journal of big data* **6**(1), 1–48 (2019)
41. Stevens, S., Wu, J., Thompson, M.J., Campolongo, E.G., Song, C.H., Carlyn, D.E., Dong, L., Dahdul, W.M., Stewart, C., Berger-Wolf, T., et al.: Bioclip: A vision foundation model for the tree of life. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19412–19424 (2024)
42. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision. pp. 843–852 (2017)
43. Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., Yan, J.: Equalization loss for long-tailed object recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11662–11671 (2020)
44. Team, D.S.G.: Data study group final report: Centre for environment, fisheries and aquaculture science (2022). <https://doi.org/10.5281/zenodo.6799166>, <https://doi.org/10.5281/zenodo.6799166>
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
46. Walker, J.L., Orenstein, E.C.: Improving rare-class recognition of marine plankton with hard negative mining. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3672–3682 (2021)
47. Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., Vajda, P.: Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677 (2020)
48. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Feature transfer learning for deep face recognition with long-tail data. arXiv preprint arXiv:1803.09014 **1**(2) (2018)
49. Zhang, X., Fang, Z., Wen, Y., Li, Z., Qiao, Y.: Range loss for deep face recognition with long-tailed training data. In: Proceedings of the IEEE international conference on computer vision. pp. 5409–5418 (2017)
50. Zhao, Y., Chen, W., Tan, X., Huang, K., Zhu, J.: Adaptive logit adjustment loss for long-tailed visual recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 3472–3480 (2022)
51. Zhong, Z., Cui, J., Liu, S., Jia, J.: Improving calibration for long-tailed recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16489–16498 (2021)