

Self-Supervised Pretraining for Fine-Grained Plankton Recognition

Joona Kareinen¹, Tuomas Eerola¹, Kaisa Kraft², Lasse Lensu¹, Sanna Suikkanen², Heikki Kälviäinen^{1,3}

¹LUT University, Computer Vision and Pattern Recognition Laboratory, Lappeenranta, Finland

²Finnish Environment Institute, Helsinki, Finland

³Brno University of Technology, Faculty of Information Technology, Brno, Czech Republic

Abstract

Plankton recognition is an important computer vision problem due to plankton's essential role in ocean food webs and carbon capture, highlighting the need for species-level monitoring. However, this task is challenging due to its fine-grained nature and dataset shifts caused by different imaging instruments and varying species distributions. As new plankton image datasets are collected at an increasing pace, there is a need for general plankton recognition models that require minimal expert effort for data labeling. In this work, we study large-scale self-supervised pretraining for fine-grained plankton recognition. We first employ masked autoencoding and a large volume of diverse plankton image data to pretrain a general-purpose plankton image encoder. Then, we utilize fine-tuning to obtain accurate plankton recognition models for new datasets with a very limited number of labeled training images. Our experiments show that self-supervised pretraining with diverse plankton data clearly increases plankton recognition accuracy compared to standard ImageNet pretraining when the amount of training data is limited. Moreover, the accuracy can be further improved when unlabeled target data is available and utilized during the pretraining.

1. Introduction

Despite the advancements in vision foundation models, fine-grained recognition in the presence of dataset shift remains a challenging task [63]. The subtle differences between classes, combined with distribution shifts in class appearance and composition between training and target datasets, necessitate image representations that are both general enough to handle dataset shifts and specific enough to distinguish visually similar classes. These challenges are difficult to tackle with a general-purpose vision foundation model and call for tailored solutions. Plankton recognition offers an interesting and environmentally relevant application for studying and developing such methods. Recognizing taxonomically close plankton species is challenging

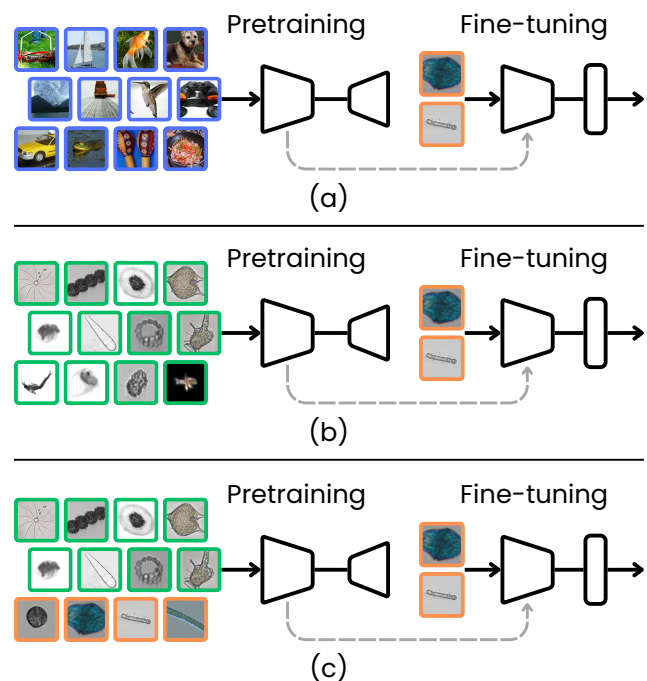


Figure 1. We evaluate three different self-supervised pretraining strategies: (a) pretraining on ImageNet-1k and fine-tuning on plankton data, (b) pretraining on a diverse plankton dataset and fine-tuning on unseen plankton data, and (c) pretraining on a diverse plankton dataset and fine-tuning on a subset of the same dataset.

and is further complicated by domain and class distribution shifts across the existing datasets.

Plankton are a collection of microscopic organisms that serve as the foundation of aquatic ecosystems. These organisms are broadly classified into two primary groups: phytoplankton and zooplankton. Phytoplankton are critical as primary producers at the base of the food chain, contributing approximately 50% of the world's oxygen production and around 40% of global carbon fixation [17, 18]. Similarly, zooplankton are critical in matter and nutrient cycling

from the primary producers to higher trophic levels, such as commercially significant fish species [5]. Furthermore, zooplankton transport carbon from surface to deep waters through feeding, daily migrations, and fecal pellets, playing a key role in global carbon cycles [2, 67].

In addition to being vital to the marine food web, plankton also serve as an indicator of ocean health. Plankton are sensitive to environmental changes: they respond rapidly to changes in temperature, nutrient availability, and water flows [22]. This makes them valuable for understanding the aquatic ecosystem dynamics, which in turn contributes to the prediction of environmental changes such as pollution and climate change [22].

Understanding plankton species distributions provides valuable information on the oceans' health, but monitoring plankton communities and composition requires sustained observations and is challenging due to their diversity and dynamic changes. To address this, various plankton imaging and analysis instruments have been developed for both *in-situ* and laboratory use [40]. Recent technological advancements have led to the development of both automated and semi-automated imaging systems that can be used to capture massive plankton datasets. However, the use of plankton imagery faces various challenges, such as the fine-grained nature of the identification task, the rarity of species, and the uncertainty of category labeling [15]. There is a growing need for automated plankton species identification systems, as they significantly reduce the need for manual identification and improve the efficiency of processing large amounts of data.

As plankton imaging instruments become more accessible, an increasing number of plankton image datasets are being collected. These datasets vary in both the imaging instruments used and the plankton species they contain. This causes dataset shifts between them, and therefore, a plankton recognition model trained on one dataset often fails when applied to another [4]. Manually labeling sufficient training data for each dataset to create dataset-specific models is impractical. Therefore, a more general model that minimizes the need for expert labeling would be highly beneficial. However, training a single general model using multiple datasets in a supervised manner is challenging due to the lack of a universally agreed-upon method for categorizing plankton, although connecting the classes to the World Register of Marine Species is routinely done by many users. The set of class labels in a plankton dataset depends on various factors, such as the environmental application for which the dataset was collected, the range of particle sizes the imaging instrument can capture, and the geographic region from which the water samples were collected. Moreover, the full taxonomical hierarchy of plankton is undergoing continuous revision as recent studies have deepened the understanding of their biodiversity [61]. Therefore, the

class labels between datasets are not always entirely comparable, making it difficult to use them simultaneously for supervised training. This issue is especially notable in classes that do not correspond to a specific species, as they are particularly affected by taxonomic revisions.

In this paper, we address the issue of incomparable class labels between datasets through self-supervised pretraining. First, we collect a large pool of public plankton image datasets and use unsupervised learning to obtain general-purpose plankton representations without the need to harmonize class labels across datasets. For this, we employ Masked Autoencoders (MAE) [24], which guide the encoder to learn plankton image representations by masking a large portion of input images and then learning to generate the missing parts. This masking technique allows the model to learn features that effectively describe the subtle visual differences between fine-grained plankton classes by forcing it to focus on small regions in the images. Next, we fine-tune the encoder with an incorporated classification head on plankton datasets with a limited amount of labeled training images.

In the experimental part of the work, we compare different pretraining methods on plankton data, as shown in Figure 1. We evaluate our models pretrained on plankton image data against standard ImageNet pretraining in varying labeled data scenarios. Our results show that domain-specific pretraining can obtain results comparable to ImageNet pretraining with significantly fewer images. Additionally, we show that domain-specific pretraining can surpass ImageNet pretraining in accuracy, particularly when labeled data is limited.

The main contributions of this study are as follows: (1) The first application of Masked Autoencoders for self-supervised pretraining on plankton image data, (2) Extensive evaluation of different pretraining strategies under varying labeled data conditions, and (3) Analysis of the benefits of domain-specific pretraining on fine-grained plankton recognition.

2. Related works

2.1. Plankton recognition

A large number of plankton recognition methods have been proposed, starting from traditional feature engineering methods to more modern deep learning methods [15]. Early plankton recognition methods relied on hand-picked features such as shape, texture, and size to obtain representative feature vectors, which were then classified using methods such as support vector machine or decision trees. In the past decade, convolutional neural networks (CNNs) have become the dominant method for plankton recognition. Plenty of different CNN architectures have been used for plankton recognition, including custom ones [6, 35, 43, 68].

More recently, vision transformers (ViTs) [34] have been shown to outperform CNNs on various plankton datasets [7, 44]. ViTs work by splitting the image into patches and converting these patches into vectors, which are then processed using self-attention mechanisms. The self-attention allows the model to capture relationships between different parts of the image, enabling it to identify the most informative regions. Kyathanahally *et al.* [38] tested ensembling multiple Data-efficient image Transformers (DeiT) [66] for various ecological datasets, including four plankton datasets. Maracani *et al.* [44] applied in-domain and out-of-domain transfer learning in plankton recognition. Notably, they demonstrate that a model pretrained on ImageNet-21K and then fine-tuned on plankton data achieves better results compared to a model trained from scratch using a single plankton dataset. Callejas *et al.* [7] tested different ViTs and CNNs in the classification of ZooScanNet [16] and WHOI-Plankton [49], the two largest publicly available plankton datasets, and reported that ViTs achieve the best accuracies. An extensive review of existing plankton recognition methods can be found in [15].

The vast majority of existing plankton recognition models have been trained and evaluated on one or a few, often in-house, datasets. Due to dataset shifts caused by different imaging instruments and varying class compositions, these models do not generalize well to previously unseen plankton datasets. Additionally, efforts have been made to develop more general plankton recognition methods using transfer learning [44, 48], domain adaptation [4], and open-set recognition [1, 28, 52].

2.2. Self-supervised learning

Self-supervised learning (SSL) refers to a family of techniques that enable models to learn meaningful representations from unlabeled data [27]. By reducing the dependency on manually annotated datasets, SSL improves data efficiency, especially in domains where labeled data is scarce or expensive to obtain.

Some of the early deep SSL methods in computer vision were Generative Adversarial Networks (GANs) [19] and autoencoders [57]. GANs consist of a generator that creates realistic-looking images and a discriminator that tries to detect generated images from real samples. Through this training process, GANs indirectly capture meaningful image representations but are primarily designed for image generation rather than representation learning. Autoencoders, on the other hand, encode input images into a compact latent representation and then reconstruct them from this latent space. While both GANs and autoencoders showcased the potential of learning from unlabeled data, they focused on generative tasks rather than learning transferable representations for downstream tasks.

A major advancement in SSL was contrastive learning,

in which a model learns representations by grouping similar (positive) samples closer while pushing dissimilar (negative) ones apart in the feature space. In the absence of labels, positive and negative pairs are typically defined using data augmentation: different augmented views of the same image are treated as positive, while all other images are considered negative. Unlike the generative approaches, contrastive learning directly optimizes for feature discrimination, making it well-suited for transfer learning.

Momentum Contrast (MoCo), proposed by He *et al.* [23], introduced a memory queue that stores a large number of negative samples across training batches, addressing the batch size limitations in contrastive learning. MoCo utilizes a momentum-updated encoder, which prevents representation collapse by updating a secondary encoder using an exponential moving average of the main encoder's parameters. A simple framework for contrastive learning (SimCLR) [9] simplified contrastive learning by removing the need for queue samples and instead relied on a large batch size to generate a diverse set of negative samples. SimCLR applies two random augmentations to each data sample, passing them through a shared encoder and projection head, and maximizes the similarity between the augmented views. After self-supervised training, the projection head is removed, and only the encoder is used for downstream tasks. SimCLR demonstrated that strong data augmentations in SSL significantly improve the quality of learned representations. Barlow Twins [71] is a novel approach to contrastive learning through redundancy reduction. Instead of using contrastive loss, Barlow Twins minimize the off-diagonal elements in a cross-correlation matrix between representations of two augmented views, effectively encouraging representations to be invariant to distortions.

Bootstrap Your Own Latent (BYOL) [21] further advanced SSL by showcasing that it can achieve state-of-the-art results without any negative pairs. Later, self-distillation with no labels (DINO) [8] introduced a self-distillation framework inspired by BYOL. DINO trains a student network using a cross-entropy loss between the student and a momentum-updated teacher network's predictions. DINO demonstrated that ViTs can learn strong representations through self-supervised training without requiring large labeled datasets. Further improvements to this method were introduced in DINO v2 [47].

Masked image modeling (MIM) has emerged as a recent approach to SSL, inspired by masked language modeling in natural language processing (*e.g.* BERT [12]). Instead of comparing augmented image pairs, MIM methods train a model to reconstruct missing parts of an image. BEIT [3] is a token-based masked model where the image is first tokenized into discrete visual embeddings, and the model learns to predict masked tokens. This approach was further

improved in BEIT v2 [51]. Masked Autoencoder [24] represents a more efficient patch-based masking strategy, reconstructing raw pixel patches using a lightweight decoder. SimMIM [70] simplified the approach even further by applying a direct pixel-wise loss.

2.3. Semi-supervised learning in fine-grained recognition

Semi-supervised learning approaches have evolved significantly in recent years. Traditional methods typically utilize both labeled and unlabeled data during the training process, whereas another approach is to first utilize self-supervised learning followed by supervised fine-tuning. The latter strategy has become more popular after Chen *et al.* [10] demonstrated that large self-supervised models excel when fine-tuned with small amounts of labeled data.

Su *et al.* [64] evaluated the existing semi-supervised learning methods for fine-grained image classification. The results revealed that semi-supervised learning methods can improve performance compared to training from scratch. Their research further showed that self-supervised pretraining can benefit significantly from out-of-distribution data. Kim *et al.* [32] introduced a novel open-set self-supervised learning in which the pretraining dataset contains instances of relevant or irrelevant domains to the target dataset. They proposed the SimCore algorithm that samples a subset of the open-set that is semantically similar to the target dataset and demonstrated enhanced representation learning performance in fine-grained recognition tasks. Duan *et al.* [13] proposed a pseudo-label selection method that encouraged pseudo labels to include likely ground truth labels while excluding noisy ones.

In the domain of plankton recognition, researchers have explored various self-supervised and semi-supervised learning methods that typically utilize clustering to allow the potential new-class discovery from unlabeled data. Early work by Salvesen *et al.* [53] introduced an autoencoder-based framework with an embedded clustering layer that learns a latent representation while simultaneously performing unsupervised clustering. Later, they extended the approach with rotation invariant features and spectral clustering to refine class separation [54]. Schmarje *et al.* [56] proposed a semi-supervised learning framework for handling fuzzy labels, where a small set of certain images is used to guide clustering of a large set of uncertain fuzzy images. Schröder *et al.* [58] compared multiple feature extraction methods under varying assumptions of label and data availability and tested two new semi-supervised learning methods tailored for MorphoCluster [59].

Recent advances have focused more on enhancing feature representation with Schanz *et al.* [55] applying SimCLR pretraining on CNN models while addressing the issue with imbalanced classes. Pastore *et al.* [50] proposed an

unsupervised learning method that combined features from multiple pretrained CNNs and ViTs and further compresses them to be used in clustering. In [45], the authors compared both transfer learning and DINO pretraining with CNNs and reported that DINO pretraining obtained better results.

3. Methods

3.1. Masked autoencoder

Masked autoencoder (MAE) is a self-supervised learning method proposed by He *et al.* [24]. MAE is designed to reconstruct missing portions of an input image by leveraging only small unmasked parts of the information. Like standard autoencoders, MAE consists of an encoder that maps the input image into a latent representation and a decoder that reconstructs the original input from this representation. However, unlike traditional autoencoder architectures, MAE uses asymmetric design where the encoder is significantly larger than the decoder. Recently, MAE has been shown to scale not only with model size but also with the size of the training dataset in larger models [62].

MAE follows a patch-based processing approach similar to ViTs, where the input image is divided into non-overlapping patches. During training, a large random subset of these patches is masked. The encoder is applied only to the unmasked patches, which significantly reduces the computational complexity and allows the training of larger encoders. The masking encourages the encoder to learn meaningful representations from the unmasked patches that are useful for reconstruction. After the visible patches have been encoded, the decoder reconstructs the masked patches using the learned latent representations.

The training objective in MAE is to minimize the reconstruction error between the original and the predicted masked patches. The loss function is formulated as

$$\mathcal{L} = \frac{1}{N_m} \sum_{i \in M} (x_i - \hat{x}_i)^2, \quad (1)$$

where N_m is the number of masked patches, M is the set of masked patch indices, x_i is an original patch, and \hat{x}_i is its reconstruction. This loss function, based on mean squared error (MSE), is computed only for the masked patches and not for the visible ones. In the original paper [24], the authors showed that a high masking ratio is required for the model to learn transferable feature representations. For downstream tasks, the decoder is removed from the network, and only the encoder is used as a feature extractor. This architecture allows fine-tuning the model for specific recognition tasks while leveraging the feature representations learned during self-supervised pretraining.

Table 1. Summary of the datasets used for pretraining.

Dataset	Imaging instrument(s)	Region	Plankton type	# of species	# of images
Kaggle-Plankton [11]	ISIS-2	Straits of Florida, U.S	zooplankton	121	130,000
Lake Zooplankton [37]	DSPC	Lake Greifensee, Switzerland	zooplankton	35	18,000
SYKE-Plankton-ZooScan_2024 [29]	ZooScan	Baltic Sea	zooplankton	20	24,000
PMID2019 [39]	Bright-field microscope	Jiaozhou Bay, China	phytoplankton	24	14,000
SYKE-Plankton-IFCB_2022 [36]	IFCB	Baltic Sea	phytoplankton	50	63,000
UDE Diatoms in the Wild 2024 [33]	Bright-field microscope	–	phytoplankton	611	84,000
DAPlankton [4]	IFCB, CS, FlowCam	Baltic Sea	phytoplankton	44	112,000
Total					443,000

4. Experiments

4.1. Data

To obtain a diverse dataset for self-supervised pretraining, we combined data from seven publicly available plankton datasets. These datasets include Kaggle-Plankton [11], Lake Zooplankton [37], SYKE-Plankton-ZooScan_2024 [29], PMID2019 [39], SYKE-Plankton-IFCB_2022 [36], UDE Diatoms in the Wild 2024 [33], and DAPlankton [4]. Together, these datasets contain both phyto- and zooplankton images and capture a wide range of species and imaging conditions. Class overlap between datasets is minimal, with the notable exception of DAPlankton and SYKE-Plankton-IFCB_2022, which share 33 taxa (Full class overlap in supplementary material). A summary of these datasets is presented in Table 1.

For Kaggle-Plankton, which contains a labeled training set with 30,000 images and an unlabeled test set with 130,000 images, we utilize only the larger test set for pretraining. DAPlankton contains two subsets: DAPlankton_{LAB} and DAPlankton_{SEA}. Both subsets were used in the pretraining dataset.

Our preprocessing strategy follows a similar approach to earlier works [42, 44]. To maintain the original aspect ratio while creating a square image, we pad the smaller dimension using a background color that matches the image. This background color matching is done by 1) computing the mode color from all edges of the image and 2) estimating the background noise by calculating the standard deviation from 20% of the edge pixels closest to the mode color. Dataset-specific preprocessing includes removing size indication legends from SYKE-Plankton-ZooScan_2024 images and utilizing the supplied ground truth bounding box information to crop individual samples from PMID2019. Example images from each dataset after preprocessing are shown in Figure 2.

For fine-tuning and testing, we utilized both subsets of DAPlankton: DAPlankton_{LAB} and DAPlankton_{SEA}. DAPlankton_{LAB} consists of 47,471 images from 15 phytoplankton species captured with three different imaging instruments: Imaging FlowCytobot (IFCB) [46], CytoSense

(CS) [14], and FlowCam (FC) [60]. DAPlankton_{SEA} consists of 64,453 images from 31 phytoplankton species captured using IFCB and CS. The DAPlankton_{SEA} is the more realistic of the two datasets as the images were collected *in-situ* from the Baltic Sea and the dataset is heavily imbalanced. The data compositions and the range of images per class are shown in Table 2.

Table 2. Summary of the DAPlankton subsets.

Dataset	Instrument	# of images	# of images per class
DAPlankton _{LAB}	IFCB	16,476	1,001 – 1,376
	CS	13,187	608 – 1,285
	FC	17,808	286 – 2,618
DAPlankton _{SEA}	IFCB	51,622	54 – 12,280
	CS	12,831	5 – 5,443

The DAPlankton dataset is unique among the publicly available plankton datasets as it contains the same shared classes across both partitions. Additionally, the laboratory partition has negligible label uncertainty as the cultures were grown in a laboratory and checked for cross-contamination. The shared taxonomy between different imaging instruments allows the evaluation of how different imaging instruments affect the model’s performance on identical taxonomic targets.

4.2. Design of experiments

4.2.1. Self-supervised pretraining

Our implementation of MAE was built using PyTorch, PyTorch Lightning, and LightlySSL. The ViT models were implemented using the PyTorch image models (timm) library [69].

Image Augmentations: The pretraining includes a series of image augmentations that make the pretraining dataset more versatile. The input images were resized to 256×256 pixels and converted to grayscale. The grayscale conversion was added to unify the plankton images as in any case, the color information in them is very limited. A random patch was selected with a scale ranging from 0.4 and 1.0 and resized to 224×224 pixels. Following this, we applied random horizontal and vertical flips and normalization

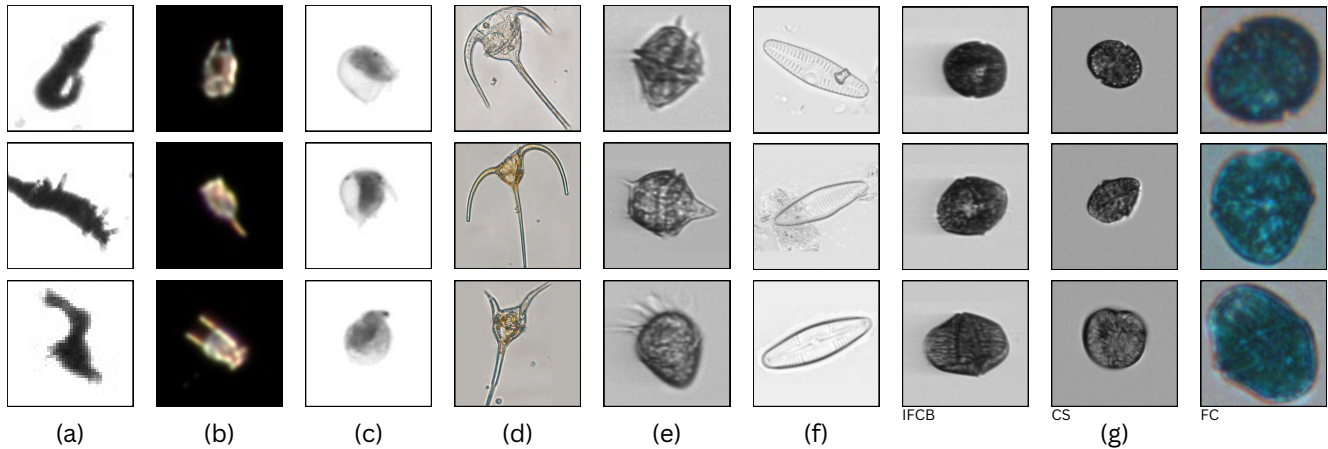


Figure 2. Example plankton images from different datasets: a) Kaggle-Plankton [11], b) Lake Zooplankton [37], c) SYKE-Plankton-ZooScan.2024 [29], d) PMID2019 [39], e) SYKE-Plankton-IFCB.2022 [36], f) UDE Diatoms in the Wild 2024 [33], g) DAPlankton [4]. The shown images are taken from three different, visually similar classes within each dataset, highlighting the fine-grained nature of the data. For DAPlankton, the same three classes are shown across all instruments.

using the dataset mean and standard deviation.

Architecture: We employed a ViT-L as our encoder backbone, following the original implementation from [24]. The decoder was implemented as a lighter transformer architecture with 8 layers and an embedding dimension of 512. During pretraining, we randomly masked 75% of the image patches, which was shown to be effective in the original MAE approach. The patch size was set to 16×16 , resulting in 196 patches per image.

Optimization: We utilized AdamW optimizer [41] with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. The learning rate followed the cosine decay schedule with linear warmup, where the base learning rate was set to $1.5e-4 \times (\text{eff_batchsize}/256)$ following the linear scaling law [20] with the weight decay set to 0.05. The reconstruction loss was computed between normalized pixel values between the reconstructed and the original masked patches.

We pretrained two different models from scratch: one using the complete pretraining dataset and another excluding the DAPlankton dataset and all classes present in both DAPlankton and SYKE-Plankton-IFCB.2022 datasets. This reduction resulted in a pretraining dataset of 280,000 images. The first model, referred to as ViT-L (with-daplankton), was trained for 800 epochs using 8 nodes with 4 NVIDIA V100 GPUs per node, while the second model, ViT-L (no-daplankton), was trained using only 4 nodes. Both models used an effective batch size of 4,096, with the smaller training setup utilizing gradient accumulation to match this batch size. The results showcased high-quality reconstruction, as shown in Figure 3.

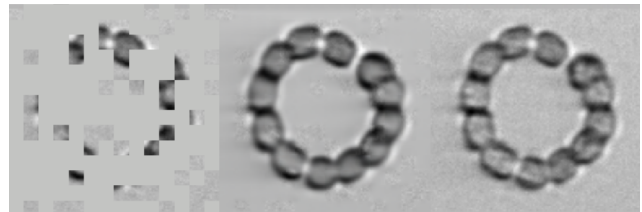


Figure 3. Masked autoencoder learns from the small unmasked patches (left) to reconstruct (middle) the original plankton image (right).

4.2.2. Fine-tuning

Dataset and evaluation: We fine-tuned our pretrained models using both DAPlankton_{LAB} and DAPlankton_{SEA} datasets. Additionally, we utilized ViT-L model trained with the ImageNet-1k dataset as a transfer learning baseline. Each experiment ran for 50 epochs, and we applied 5-fold cross-validation. For each fold, the dataset was first split into 20% testing and 80% training, of which 15% was selected as a validation set. For splitting the dataset, we utilized a stratified strategy, which ensured that each fold was similarly balanced. We applied the same set of augmentations for the training set as in pretraining, but for validation and testing, the vertical and horizontal flips were removed, and the crop was centered on the image. We evaluate our results by using accuracy as the main metric.

Model architecture: We adopted the bottleneck architecture from [44], where the encoder’s output is passed through a hidden layer of size 512, with LayerNorm and GELU activation [25], followed by the final classification layer. We utilized layer-wise learning rate decay [3] with a decay factor of 0.75 to gradually increase learning rates

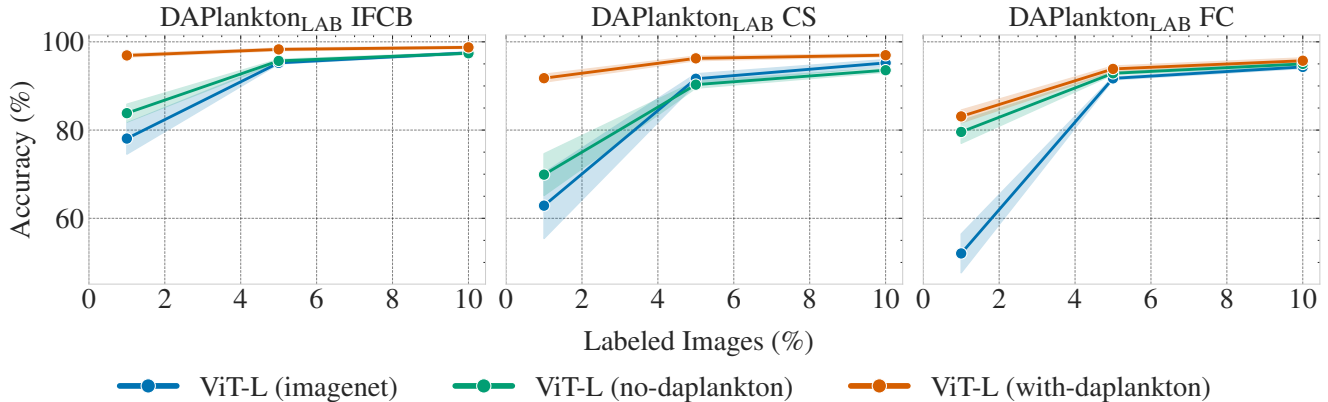


Figure 4. Mean accuracy and standard deviation for $\text{DAPlankton}_{\text{LAB}}$ across different labeled data subsets.

from deeper to shallower layers to preserve the pretrained features.

Optimization: During fine-tuning, we set the learning rate to $2e-3$ and used the AdamW optimizer with a weight decay of 0.01. The learning rate followed a cosine decay with a 5-epoch warmup period. As a loss function, we use cross-entropy loss with label smoothing [65] of 0.1. For further regularization, we applied stochastic depth [26] with a drop path rate of 0.2, following the fine-tuning strategy from [24]. We used a batch size of 128 and fine-tuned the models using a single NVIDIA A100 GPU.

4.3. Results

The results for fine-tuning using the full $\text{DAPlankton}_{\text{LAB}}$ dataset are presented in Table 3. The table shows the accuracies for all three pretrained models for all three partitions of the dataset. The results show that when using a large amount of training data for fine-tuning, there are no significant differences in accuracy between the pretraining approaches. This indicates that pretraining a ViT-L with a diverse plankton dataset does not provide a notable benefit compared to a model pretrained with ImageNet-1K, when enough labeled images in the target dataset are available. However, it should be noted that ImageNet-1K is 3-4 times larger than the used combinations of plankton datasets, implying that a smaller amount of data is sufficient for pretraining if it consists of plankton images. Moreover, in real plankton recognition applications, a large amount of labeled data for the target dataset is often not available, and the goal is to minimize the required labeling efforts. Therefore, it is more interesting to study the ability of the pretrained models to learn from a small number of labeled images.

In addition to evaluating the models with the full dataset, we experimented using only small subsets of labeled data. We run three different experiments for each model, utilizing only 1%, 5%, and 10% of training data for fine-tuning. To maintain class representation even in the smallest sub-

sets, we ensured a minimum of one sample per class in each experiment. The details of these dataset subsets for both $\text{DAPlankton}_{\text{LAB}}$ and $\text{DAPlankton}_{\text{SEA}}$ are presented in Table 4.

Table 3. Accuracy (in %) for full fine-tuning for $\text{DAPlankton}_{\text{LAB}}$.

Model	IFCB	CS	FC
ViT-L (imagenet)	99.29 ± 0.13	98.49 ± 0.23	98.18 ± 0.13
ViT-L (no-daplankton)	99.27 ± 0.20	98.42 ± 0.30	98.35 ± 0.11
ViT-L (with-daplankton)	99.32 ± 0.18	99.01 ± 0.09	98.21 ± 0.19

Table 4. Number of labeled samples across $\text{DAPlankton}_{\text{LAB}}$ and $\text{DAPlankton}_{\text{SEA}}$ subsets.

Dataset	1%		5%		10%	
	Per Class	Total	Per Class	Total	Per Class	Total
$\text{DAPlankton}_{\text{LAB}}$						
IFCB	6 – 9	104	34 – 46	555	68 – 93	1,115
CS	4 – 8	82	20 – 43	489	41 – 87	888
FC	1 – 17	113	9 – 89	598	19 – 178	1,204
$\text{DAPlankton}_{\text{SEA}}$						
IFCB	1 – 83	344	1 – 417	1,740	3 – 835	3,495
CS	1 – 37	95	1 – 185	430	1 – 370	862

Results for $\text{DAPlankton}_{\text{LAB}}$ are shown in Figure 4 for all three imaging instruments. The results show that the MAE models pre-trained with plankton data achieve strong performance even with limited labeled data, clearly outperforming ViT-L pretrained on ImageNet. For example, the model pretrained on all plankton data, ViT-L (with-daplankton), obtained 97% recognition accuracy on the IFCB subset with only 6-9 labeled training images per class. ViT-L (with-daplankton) outperforms other pretraining strategies as expected. However, it should be noted that this model has seen the images, albeit without the labels, in the target dataset during pretraining, giving it an advantage over the corresponding models. This corresponds to a scenario where the target dataset is available during the pre-

training stage, allowing the model to be exposed to the entire testing distribution during self-supervised pretraining.

The model pretrained on all plankton data except DAPlankton, ViT-L (no-daplankton), provides a more realistic scenario where pretraining is done only once, and the model is fine-tuned to target datasets that were not available during pretraining. As can be seen from the results, this model still clearly outperforms the ViT-L model pretrained on ImageNet-1K across all imaging instruments, especially when only 1% of the training data is used. The confusion matrix with class-specific accuracies is shown in Figure 5, where only 1% of the FC data is used. The results indicate that ViT-L (no-daplankton) achieves significantly higher class accuracies and exhibits less confusion overall. In contrast, ViT-L (ImageNet) performs well on only a subset of classes, and it struggles to distinguish between visually similar categories.

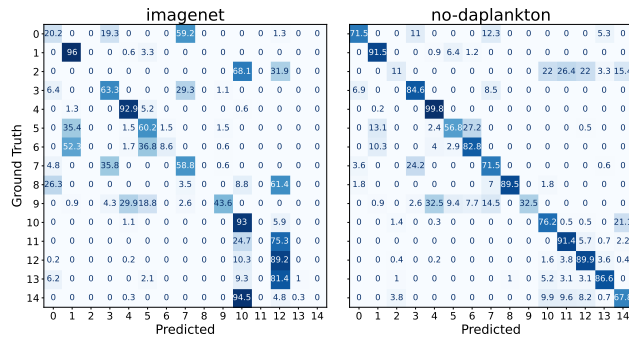


Figure 5. Confusion matrices for ViT-L (ImageNet) and ViT-L (no-daplankton), evaluated on 1% of labeled FC data from DAPlankton_{LAB}.

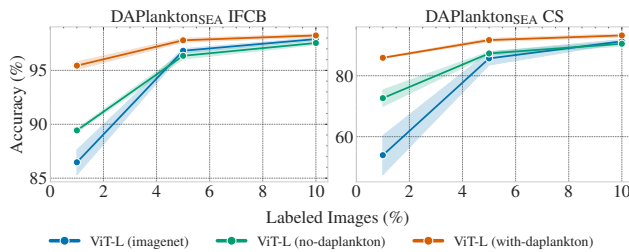


Figure 6. Mean accuracy and standard deviation for DAPlankton_{SEA} across different labeled data subsets.

Results for fine-tuning with the DAPlankton_{SEA} are shown in Figure 6. The results show that ViT-L (with-daplankton) obtains the best results, followed by ViT-L (no-daplankton) and ViT-L (imagenet). The performance difference is more notable in the CS subset, likely due to its higher class imbalance. Additionally, ViT-L (imagenet) tends to predict the most common classes and fails to generalize well, resulting in lower accuracy. Overall, the results

obtained with DAPlankton_{LAB} and DAPlankton_{SEA} demonstrate the benefit of creating custom pretrained models for plankton recognition, especially if only a small amount of labeled data is available. We have made our pretrained models available for other researchers [30, 31].

5. Conclusion

In this paper, we studied the benefits of self-supervised pretraining for fine-grained plankton recognition. We collected a large and diverse plankton image dataset comprising 443,000 images by combining multiple publicly available datasets captured with different imaging instruments. Due to the varying species compositions and labeling practices, the class labels do not correspond between the individual datasets, making the combined dataset unsuitable for supervised training. Therefore, we applied self-supervised learning using masked autoencoders to obtain general-purpose plankton image encoders.

We evaluated the obtained pretrained encoders by fine-tuning them for plankton recognition on individual datasets with varying amounts of labeled data. When a large amount of labeled training data was used for fine-tuning, the custom pretrained models did not provide a significant benefit over standard ImageNet pretraining. However, when only a small number of labeled training images was used for fine-tuning, the models pretrained on plankton data showed considerably higher recognition accuracy compared to the ImageNet encoder. This highlights the advantages and potential of custom-pretrained models for the fine-grained recognition of plankton images. The model that had access to the unlabeled target data showed the highest recognition accuracy, suggesting that it is beneficial to repeat the pretraining before analyzing new datasets when possible. Nevertheless, the custom model that did not include target data during pretraining also showed very promising results, indicating that self-supervised learning enables the models to learn efficient and general plankton image representations. The results demonstrate that self-supervised learning has the potential to significantly reduce the need for manual labeling of plankton images by experts.

6. Acknowledgments

The research was carried out in the FASTVISION and FASTVISION-plus projects funded by the Academy of Finland (Decision numbers 321980, 321991, 339612, and 339355). We wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

- [1] Ola Badredeen Bdawy Mohamed, Tuomas Eerola, Kaisa Kraft, Lasse Lensu, and Heikki Kälviäinen. Open-set plank-

- ton recognition using similarity learning. In *International Symposium on Visual Computing*, pages 174–183, 2022. 3
- [2] Karl Banse. Zooplankton: Pivotal role in the control of ocean production: I. biomass and production. *ICES Journal of Marine Science*, 52:265–277, 1995. 2
- [3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEIT: BERT pre-training of image transformers. <https://arxiv.org/abs/2106.08254>, 2021. 3, 6
- [4] Daniel Batrakhov, Tuomas Eerola, Kaisa Kraft, Lumi Haraguchi, Lasse Lensu, Sanna Suikkanen, María Teresa Camarena-Gómez, Jukka Seppälä, and Heikki Kälviäinen. DAPlankton: Benchmark dataset for multi-instrument plankton recognition via fine-grained domain adaptation. In *ICIP*, pages 158–164, 2024. 2, 3, 5, 6
- [5] Grégory Beaugrand, Keith M Brander, J Alistair Lindley, Sami Souissi, and Philip C Reid. Plankton effect on cod recruitment in the north sea. *Nature*, 426:661–664, 2003. 2
- [6] Jaroslav Bureš, Tuomas Eerola, Lasse Lensu, Heikki Kälviäinen, and Pavel Zemčík. Plankton recognition in images with varying size. In *ICPR Workshops and Challenges*, pages 110–120, 2021. 2
- [7] Sofía Callejas, Hernan Lira, Andrew Berry, Luis Martí, and Nayat Sanchez-Pi. No plankton left behind: Preliminary results on massive plankton image recognition. In *High Performance Computing*, pages 170–185, 2025. 3
- [8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 3
- [9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020. 3
- [10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *NeurIPS*, 33:22243–22255, 2020. 4
- [11] Robert K. Cowen, Su Sponaugle, Kelly L. Robinson, Jessica Luo, and Cedric Guigand. Planktonset 1.0: Plankton imagery data collected from f.g. walton smith in straits of florida from 2014-06-03 to 2014-06-06 and used in the 2015 national data science bowl (ncei accession 0127422). <https://doi.org/10.7289/v5d21vj2d>, 2015. 5, 6
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. 3
- [13] Yue Duan, Zhen Zhao, Lei Qi, Luping Zhou, Lei Wang, and Yinghuan Shi. Roll with the punches: expansion and shrinkage of soft label selection for semi-supervised fine-grained learning. In *AAAI Conference on Artificial Intelligence*, pages 11829–11837, 2024. 4
- [14] George BJ Dubelaar, Peter L Gerritzen, Arnout ER Beeker, Richard R Jonker, and Karl Tangen. Design and first results of cytobuoy: A wireless flow cytometer for in situ analysis of marine and fresh waters. *Cytometry*, 37:247–254, 1999. 5
- [15] Tuomas Eerola, Daniel Batrakhov, Nastaran Vatankhah Barazandeh, Kaisa Kraft, Lumi Haraguchi, Lasse Lensu, Sanna Suikkanen, Jukka Seppälä, Timo Tamminen, and Heikki Kälviäinen. Survey of automatic plankton image recognition: challenges, existing solutions and future perspectives. *Artificial Intelligence Review*, 57:114, 2024. 2, 3
- [16] Amanda Elineau, Corinne Desnos, Laetitia Jalabert, Marion Olivier, Jean-Baptiste Romagnan, Manoela Costa Brandao, Fabien Lombard, Natalia Llopis, Justine Courboulès, Louis Caray-Counil, Bruno Serranito, Jean-Olivier Irissou, Marc Picheral, Gaby Gorsky, and Lars Stemmann. Zooscan-net: plankton images captured with the zooscan. <https://doi.org/10.17882/55741>, 2024. 3
- [17] Paul G Falkowski. The role of phytoplankton photosynthesis in global biogeochemical cycles. *Photosynthesis research*, 39:235–258, 1994. 1
- [18] Christopher B Field, Michael J Behrenfeld, James T Randerson, and Paul Falkowski. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, 281:237–240, 1998. 1
- [19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014. 3
- [20] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. <https://arxiv.org/abs/1706.02677>, 2017. 6
- [21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. *NeurIPS*, 33:21271–21284, 2020. 3
- [22] Graeme Hays, Anthony Richardson, and Carol Robinson. Climate change and marine plankton. *Trends in Ecology & Evolution*, 20:337–344, 2005. 2
- [23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 3
- [24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2, 4, 6, 7
- [25] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). <https://arxiv.org/abs/1606.08415>, 2016. 6
- [26] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, pages 646–661, 2016. 7
- [27] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9:2, 2020. 3
- [28] Joona Kareinen, Annaliina Skyttä, Tuomas Eerola, Kaisa Kraft, Lasse Lensu, Sanna Suikkanen, Maiju Lehtiniemi, and

- Heikki Kälviäinen. Open-set plankton recognition. In *ECCV Workshops*, 2024. 3
- [29] Joonas Kareinen, Annaliina Skyttä, Tuomas Eerola, Kaisa Kraft, Lasse Lensu, Sanna Suikkanen, Maiju Lehtiniemi, and Heikki Kälviäinen. SYKE-plankton_ZooScan_2024. <https://doi.org/10.23729/fa115087-2698-4aa5-aedd-11e260b9694d>, 2024. 5, 6
- [30] Joonas Kareinen, Tuomas Eerola, Kaisa Kraft, Lasse Lensu, Sanna Suikkanen, and Heikki Kälviäinen. no-daplankton. https://huggingface.co/Jookare/no_daplankton_vit_large_patch16_224.mae, 2025. [Online; accessed April, 11, 2025]. 8
- [31] Joonas Kareinen, Tuomas Eerola, Kaisa Kraft, Lasse Lensu, Sanna Suikkanen, and Heikki Kälviäinen. with-daplankton. https://huggingface.co/Jookare/plankton_vit_large_patch16_224.mae, 2025. [Online; accessed April, 11, 2025]. 8
- [32] Sungnyun Kim, Sangmin Bae, and Se-Young Yun. Core-set sampling from open-set for fine-grained self-supervised learning. In *CVPR*, pages 7537–7547, 2023. 4
- [33] Michael Kloster, Andrea Burfeid-Castellanos, Mimoza Dani, Ntambwe Albert Serge Mayombo, Bank Beszteri, and Danijela Vidaković. UDE Diatoms in the Wild 2024, 2024. 5, 6
- [34] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3
- [35] Kaisa Kraft, Otso Velhonoja, Tuomas Eerola, Sanna Suikkanen, Timo Tamminen, Lumi Haraguchi, Pasi Ylöstalo, Sami Kielosto, Milla Johansson, Lasse Lensu, et al. Towards operational phytoplankton recognition with automated high-throughput imaging, near-real-time data processing, and convolutional neural networks. *Frontiers in Marine Science*, 9, 2022. 2
- [36] Kaisa Kraft, Otso Velhonoja, Jukka Seppälä, Heidi Hällfors, Sanna Suikkanen, Pasi Ylöstalo, Silvia Anglès, Sami Kielosto, Harri Kuosa, Sirpa Lehtinen, Johanna Oja, and Timo Tamminen. SYKE-plankton_IFCB_2022. <https://b2share.eudat.eu/records/abf913e5a6ad47e6baa273ae0ed6617a>, 2022. 5, 6
- [37] Sreenath P Kyathanahally, Thomas Hardeman, Ewa Merz, Thea Bulas, Marta Reyes, Peter Isles, Francesco Pomati, and Marco Baity-Jesi. Deep learning classification of lake zooplankton. *Frontiers in Microbiology*, 12:746297, 2021. 5, 6
- [38] Sreenath P Kyathanahally, Thomas Hardeman, Marta Reyes, Ewa Merz, Thea Bulas, Philipp Brun, Francesco Pomati, and Marco Baity-Jesi. Ensembles of data-efficient vision transformers as a new paradigm for automated classification in ecology. *Scientific Reports*, 12:18590, 2022. 3
- [39] Qiong Li, Xin Sun, Junyu Dong, Shuqun Song, Tongtong Zhang, Dan Liu, Han Zhang, and Shuai Han. Developing a microscopic image dataset in support of intelligent phytoplankton detection using deep learning. *ICES Journal of Marine Science*, 77:1427–1439, 2020. 5, 6
- [40] Fabien Lombard, Emmanuel Boss, Anya M. Waite, Meike Vogt, Julia Uitz, Lars Stemmann, Heidi M. Sosik, Jan Schulz, Jean-Baptiste Romagnan, Marc Picheral, Jay Pearlman, Mark D. Ohman, Barbara Niehoff, Klas O. Möller, Patricia Miloslavich, Ana Lara-Lpez, Raphael Kudela, Rubens M. Lopes, Rainer Kiko, Lee Karp-Boss, Jules S. Jaffe, Morten H. Iversen, Jean-Olivier Irisson, Katja Fennel, Helena Hauss, Lionel Guidi, Gaby Gorsky, Sarah L. C. Giering, Peter Gaube, Scott Gallager, George Dubelaar, Robert K. Cowen, François Carlotti, Christian Briseño-Avena, Léo Berline, Kelly Benoit-Bird, Nicholas Bax, Sonia Batten, Sakina Dorothée Ayata, Luis Felipe Artigas, and Ward Appeltans. Globally consistent quantitative observations of planktonic ecosystems. *Frontiers in Marine Science*, 6, 2019. 2
- [41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6
- [42] Alessandra Lumini and Loris Nanni. Deep learning and transfer learning features for plankton classification. *Ecological Informatics*, 51:33–43, 2019. 5
- [43] Alessandra Lumini and Loris Nanni. Deep learning and transfer learning features for plankton classification. *Ecological Informatics*, 51:33–43, 2019. 2
- [44] Andrea Maracani, Vito Paolo Pastore, Lorenzo Natale, Lorenzo Rosasco, and Francesca Odone. In-domain versus out-of-domain transfer learning in plankton image classification. *Scientific Reports*, 13:10443, 2023. 3, 5, 6
- [45] Ellen Oldenburg, Raphael M Kronberg, Barbara Niehoff, Oliver Ebenhöf, and Ovidiu Popa. DeepLOKI—a deep learning based approach to identify zooplankton taxa on high-resolution images from the optical plankton recorder LOKI. *Frontiers in Marine Science*, 10:1280510, 2023. 4
- [46] Robert J Olson and Heidi M Sosik. A submersible imaging-in-flow instrument to analyze nano- and microplankton: Imaging flowcytobot. *Limnology and Oceanography: Methods*, 5:195–203, 2007. 5
- [47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. <https://arxiv.org/abs/2304.07193>, 2024. 3
- [48] Eric C Orenstein and Oscar Beijbom. Transfer learning and deep feature extraction for planktonic image data sets. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1082–1088, 2017. 3
- [49] Eric C. Orenstein, Oscar Beijbom, Emily E. Peacock, and Heidi M. Sosik. Whoi-plankton- a large scale fine grained visual recognition benchmark dataset for plankton classification. <https://arxiv.org/abs/1510.00745>, 2015. 3

- [50] Vito Paolo Pastore, Massimiliano Ciranni, Simone Bianco, Jennifer Carol Fung, Vittorio Murino, and Francesca Odone. Efficient unsupervised learning of biological images with compressed deep features. *Image and Vision Computing*, 137:104764, 2023. 4
- [51] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. <https://arxiv.org/abs/2208.06366>, 2022. 4
- [52] Yuchun Pu, Zhenghui Feng, Zhonglei Wang, Zhenyu Yang, and Jianping Li. Anomaly detection for in situ marine plankton images. In *ICCV*, pages 3661–3671, 2021. 3
- [53] Eivind Salvesen, Aya Saad, and Annette Stahl. Robust methods of unsupervised clustering to discover new planktonic species in-situ. In *Global Oceans 2020: Singapore-US Gulf Coast*, pages 1–9, 2020. 4
- [54] Eivind Salvesen, Aya Saad, and Annette Stahl. Robust deep unsupervised learning framework to discover unseen plankton species. In *International Conference on Machine Vision*, pages 241–250, 2022. 4
- [55] Tobias Schanz, Klas Ove Möller, Saskia Rühl, and David S Greenberg. Robust detection of marine life with label-free image feature learning and probability calibration. *Machine Learning: Science and Technology*, 4:035007, 2023. 4
- [56] Lars Schmarje, Johannes Brünger, Monty Santarossa, Simon-Martin Schröder, Rainer Kiko, and Reinhard Koch. Fuzzy overclustering: Semi-supervised classification of fuzzy labels with overclustering and inverse cross-entropy. *Sensors*, 21:6661, 2021. 4
- [57] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015. 3
- [58] Simon-Martin Schröder and Rainer Kiko. Assessing representation learning and clustering algorithms for computer-assisted image annotation—simulating and benchmarking MorphoCluster. *Sensors*, 22:2775, 2022. 4
- [59] Simon-Martin Schröder, Rainer Kiko, and Reinhard Koch. MorphoCluster: efficient annotation of plankton images by clustering. *Sensors*, 20:3060, 2020. 4
- [60] Christian K Sieracki, Michael E Sieracki, and Charles S Yentsch. An imaging-in-flow system for automated analysis of marine microplankton. *Marine Ecology Progress Series*, 168:285–296, 1998. 5
- [61] Nathalie Simon, Anne-Lise Cras, Elodie Foulon, and Rodolphe Lemée. Diversity and evolution of marine phytoplankton. *Comptes Rendus Biologies*, 332:159–170, 2009. 2
- [62] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *ICCV*, pages 5484–5494, 2023. 4
- [63] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. BioCLIP: A Vision Foundation Model for the Tree of Life. In *CVPR*, pages 19412–19424, 2024. 1
- [64] Jong-Chyi Su, Zezhou Cheng, and Subhransu Maji. A realistic evaluation of semi-supervised learning for fine-grained classification. In *CVPR*, pages 12966–12975, 2021. 4
- [65] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 7
- [66] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021. 3
- [67] Jefferson T. Turner. Zooplankton fecal pellets, marine snow, phytodetritus and the ocean’s biological pump. *Progress in Oceanography*, 130:205–248, 2015. 2
- [68] Aäron van den Oord, Ira Korshunova, Jeroen Burms, Jonas Degraeve, Lionel Pigou, Pieter Buteneers, and Sander Dieleman. Classifying plankton with deep neural networks. <https://sander.ai/2015/03/17/plankton.html>, 2015. [Online; accessed February, 24, 2025]. 2
- [69] Ross Wightman. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, 2019. 5
- [70] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663, 2022. 4
- [71] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320, 2021. 3