Automatic estimation of lipid content from in situ images of Arctic copepods using machine learning

Frédéric Maps, Piotr Pasza Storożenko, Jędrzej Świeżewski, Sakina-Dorothée Ayata

OXFORD UNIVERSITY PRESS | Journal of Plankton Research

Automatic estimation of lipid content from in situ images of Arctic copepods using deep learning

Journal:	*Journal of Plankton Research*
Manuscript ID	JPR-2023-061
Manuscript Type:	Original Article
Date Submitted by the Author:	23-Jun-2023
Complete List of Authors:	Maps, Frederic; Université Laval, Biologie Storożenko, Piotr; Appsilon, Data for good Świeżewski, Jędrzej; Appsilon, Data for good Ayata, Sakina-Dorothée; Sorbonne Université, Laboratoire d'Océanographie et du Climat (LOCEAN/IPSL - UMR 7159)
Keywords:	Arctic, Copepods, Lipid, Imagery, Machine learning

Automatic estimation of lipid content from in situ images of Arctic copepods using deep learning

Frédéric Maps[1,2], Piotr Pasza Storożenko[3], Jędrzej Świeżewski[3], Sakina-Dorothée Ayata[4]

[1] Département de biologie, Université Laval, Québec (QC) Canada

[2] International Research Laboratory Takuvik (LRI 3376), Université Laval (Canada) – CNRS (France)

[3] Appsilon Data for Good, Warsaw, Poland

[4] Sorbonne Université, CNRS, IRD, MNHN, Laboratoire d'Océanographie et du Climat : Expérimentations et Approches Numériques (LOCEAN-IPSL), Paris, France

Abstract

In the Arctic, large planktonic copepods form a crucial hub of matter and energy owing to the role their energy-rich lipid stores play for the biological carbon pump and for marine trophic networks. Until now, such lipid stores could be estimated manually from pictures of individuals sampled via plankton nets. Unfortunately, with this traditional approach, the link with environmental information would at best be crude, at worst lost. Since the past ~15 years, in situ imaging devices provide images whose resolution allows to estimate an individual copepod's lipid sac volume and reveal a wealth of ecological information inaccessible otherwise. However, when done manually, weeks of work are needed by trained personnel to obtain such information for a handful of samples. We removed this hurdle by training a machine learning algorithm to estimate the lipid content of individual Arctic copepods from in situ images. This algorithm obtains such information at a speed (a few minutes) and a resolution (individuals, over half a meter on the vertical) allowing us to revisit historical datasets of in situ images to better understand the dynamics of lipid production and distribution and to develop efficient monitoring protocols at a moment when marine ecosystems are facing rapid upheavals and increasing threats.

Introduction

Since the turn of the 21st century, the use of in situ optical methods to sample plankton individuals and communities has grown exponentially. The number of samples, as well as the pace of sampling has exponentially increased year after year. Today, all major oceanic basins and many regional seas have been explored by these modern tools (Irisson et al., 2022). Moreover, tens of different devices have been designed and deployed to collect images at all scales relevant for plankton ecology, from microns to metres, from high vertical resolution from in situ tows to large horizontal scales thanks to autonomous gliders or ship-based continuous flow instruments (Lombard et al., 2019). This has led to a dramatic increase in the quantity of images to analyse that definitely requires the help of automated methods to treat all this precious information.

The quantity and quality of information that plankton imagery can reveal is arguably as momentous as what satellite ocean colour imagery has become about 30 years ago for our ability to understand and simulate oceanic biogeochemical systems (Groom et al., 2019). While the recent and systematic use of images revealed previously inaccessible ecological patterns (e.g., Drago et al., 2022; Sonnet et al., 2022; Trudnowska et al., 2021; Vilgrain et al., 2021) and strengthened a trait-based approach of marine ecology (Martini et al., 2021), they were hindered from the start by the bottle-neck of human's implication in image processing. As even the basic taxonomic identification of individuals sampled in net tows by experts has always been a lengthy process, it was obvious that orders of magnitude increase in the inflow of data could not be handled by "traditional" approaches.

In the meantime, machine learning approaches soared in many spheres of our societies, and their potential for research stirred a lot of interest in the communities of researchers who were already instrumenting many biological systems with, among other data loggers, cameras. The first application in imagery in which machine learning proved to be both useful and efficient was taxonomic identification and sorting (Irisson et al., 2022). This major improvement already allowed several studies to reveal how more precise and detailed our understanding of marine ecosystems would be by using in situ imaging data in conjunction with widespread measurements of environmental variables (e.g., Schmid et al., 2018, 2016; Schmid and Fortier, 2019). However, machine learning can obviously do much more than simply help categorize images into broad taxonomic groups: it is particularly well suited to provide measurements and estimates of individual properties visible from the images (Orenstein et al., 2022).

Among the many individual properties measurable from images, the visual estimation of lipid stores is an obvious research target, owing to the important role energy-rich lipids play for the biological carbon pump (Pinti et al., 2023; Record et al., 2018) and for marine trophic networks up to humans (Belton and Thilsted, 2014). Productive, lipid-rich food webs in the ocean depend on a handful of large pelagic copepod species that form a hub of matter and energy between the intense but often short-lived microbial primary production and upper trophic levels (Kattner and Hagen, 2009). In this study, we decided to revisit the first study that provided a very detailed account of the spatio-temporal vertical organisation of a community of lipid-rich copepods from a productive environment. Schmid et al. (2018) studied the dynamics of planktonic copepod assemblages in the North Water Polynya between Greenland and Canada (named Pikialasorsuaq by Inuit communities) by using the Lightframe On-sight Keyspecies Investigation (LOKI) imaging system. A machine learning method (a random forest algorithm) was then used to *identify* and *classify* species and development stages of the dominant copepods in the system. However, estimates of individual total lipid content (mg) or lipid fullness (% of biovolume) had to be done "manually", by a human clicking with a mouse on individual images imported in the ImageJ software.

In Schmid et al. 2018, the first author measured a total of 822 images chosen randomly to be representative of the copepod community sampled by their device. The area of the lipid sac (mm^2) was measured for each individual following the manipulation procedure described above, and then used in a published empirical relationships that allowed for estimating total lipids (Vogedes et al., 2010), while lipid fullness was simply the proportion of the surface of the prosome of the individual occupied by the lipid sac area. Then, this information obtained at the individual level was related to environmental variables such as temperature and chlorophyll *a* fluorescence to produce high spatial resolution profiles of 1-m bins. The authors detected diel vertical migration patterns specific to development stages and species, as well as ontogenetic migrations related to the accumulation of lipid stores induced by the diapause life-cycle strategies of *Calanus* congeneric species. Obviously, while providing a very interesting proof of concept, such a time consuming procedure is not suitable for the exploration of the extensive images databases already constituted, and even less for an eventual operationalization of such an approach in the current context of expanding monitoring of marine ecosystems (Lombard et al., 2019). As a result, we have developed and validated a machine learning approach based on convolutional neural networks (CNN) for estimating the total lipid content (mg) of oil sacs in individual images of these crucial components of the Arctic pelagic ecosystems.

Methods

Images and annotation

Sampling took place between the 15th of August and the 2nd of September 2013 in the Northwater Polynya in the Canadian Arctic (see Schmid et al. 2016 for more details). From a total of 16 vertical tows of the LOKI, 14,558 zooplankton images have been validated (i.e. artifacts and duplicated images removed) and sorted by a trained random forest algorithm into many classes that contained information of species but also on development stages and even orientation of the images (e.g. dorsal, lateral, etc.). Schmid et al. (2018) then used a subset of 822 images taken during a Lagrangian drift experiment to estimate visually, following Vogedes et al. (2010), the lipid content from advanced development stages of the three dominant calanoid species known to accumulate significant lipid stores within an oil sac inside their transparent prosome: *C. hyperboreus* stages C3 to adult female (F), *C. glacialis* stages C4 to F and *Metridia longa* stages C5 to F (Fig. 1).

Within the original images, we found 2,920 that fit these criteria. We reviewed each one of them, and we found 2,309 that presented an oil sac visible enough to be annotated by a human. Moreover, since large images contained many LOKI-introduced artifacts (glares, borders, etc.), we filtered them out and left only those with both height and width smaller than 600 pixels to reach a count of 2,216 images left. As machine learning models typically require input of a constant size, and LOKI images have a dark background, we aligned each image in the center and pad its borders with black so the resulting image is of a constant size of 600 by 600 pixels. It is important to note that each pixel in a LOKI image corresponds to a size of 23 μm. Annotation, which is an important step of the process, has been done as follows: (i) open original LOKI images in the ImgLab Open Source application (https://imglab.in; https://github.com/NaturalIntelligence/imglab), (ii) trace the contour of the oil sac with the *polygon* tool, (iii) and export the annotations in the COCO JSON format, which stored the contour as a matrix of vertex coordinates.

Machine learning algorithm

We used the fastai library (Howard et al., 2018) implementation of the Convolutional Neural Network (CNN) model U-Net (Ronneberger et al., 2015), with a ResNet34 backbone (i.e. the initial layers; He et al., 2015), pretrained on the ImageNet dataset (Deng et al., 2009). The library, the model and the dataset are all publicly available. Using pretrained models is a common practice in research situations such as ours; it is called transfer learning (Orenstein and Beijbom, 2017). Indeed, when training a

neural network classifier, the initial layers are trained to detect simple features such as curves and slant lines, whatever the objects to identify are. It is only the final layers of a network classifier that eventually learns to identify classes relevant for the project. Moreover, retraining a very complex network model such as ResNet34 from scratch would require a number of annotated images that would far exceed what we could provide (ImageNet contains more than 14 millions of labeled images), an issue that is quite common in environmental and engineering research as well (e.g., Robbes and Janes, 2019).

However, while the ResNet34 network has been trained to classify whole images (e.g., whether the animal in the image is a cat or a dog), we used it for segmentation, i.e., to predict for each pixel of the image whether it belongs to an oil sac or not. The U-Net neural network architecture provides a framework for image segmentation that can leverage many different pretrained classification CNN, including ResNet34 (Ronneberger et al., 2015).

Data split

The dataset was divided into a train set and test set following a 90%-10% split proportion, while also taking into account the particularities of the data obtained at each towing location. Indeed, the LOKI device can, on rare occasions, create almost-duplicated images that should be split when constructing the training and validation datasets. It is important to avoid any data-leakage (information flow) between both datasets. After the split, the training and test set contained 1991 and 225 images, respectively.

Training

We used the default fine-tuning procedure of the fastai library for training our model. We optimized the Cross Entropy loss function. For determining the necessary length of training, we used the early stopping technique. Eventually, the model trained for 26 epochs that took just under 4 hours on a single Nvidia T4 GPU. To find the optimal learning rate, we used the one cycle policy introduced by Smith and Topin (2017), implemented in the fastai library. A typical neural network training procedure includes data augmentation, especially for models that learn on images. During training, we used the standard fastai augmentation methods, i.e. random flipping, rotating, zooming and re-lighting.

Validation

It is a common practice in machine learning to optimize models on one metric that has convenient mathematical properties, and assess model quality on another that is easily interpretable. In this case we used the intersection over union (IoU) metric, also known as the Jaccard index, which is the most used metric for evaluating the performance of tasks such as segmentation or object detection (Jaccard, 1901; Taha and Hanbury, 2015). Given the prediction for a single LOKI image, we can count the number of pixels in the oil sac. Then, we can combine the fact that all LOKI images have pixels of the same size to estimate the volume of the oil sac and the mass of lipids of each individual, following Vogedes et al. (2010), as did Schmid et al. (2018).

Results

The vast majority of selected images were well suited for an automated image analysis, owing to their transparency, resolution and orientation (Fig. 1). Individuals had a visible lipid sac, but with a wide range in size (lipid fullness). For example, we can see a *C. glacialis* C5 individual with a very full lipid sac (Fig. 1B), as well as a very well-defined *C. hyperboreus* female with a much depleted lipid sac (Fig. 1H).

The performance of the model is generally satisfactory, with an overall median IoU of 0.82 and a highly negatively skewed distribution (Fig. 2). However, IoU values distribution changed significantly according to the individual's species (Table 1): while median IoU values are above 0.8 in both the training and validation sets for the *Calanus* congeners, but they are significantly lower for *M. longa* individual images (0.64 and 0.41, respectively).

The model produced individual lipid content estimates that varied widely, from a minimum of $4.9\ 10^{-3}$ mg of total lipid, to a maximum of 3.68 mg. *M. longa* individuals dominate values lower than 0.1 mg, followed by *C. glacialis* and *C. hyperboreus*. The latter are the only individuals with lipid content larger than 1 mg. The relative errors in the prediction of individual lipid content from the automatic image analysis are small, but they still reach > 5% for the small *M. longa* individuals (Table 2). Moreover, our estimates coming from both annotations and model predictions compare well with the values previously obtained in Schmid et al. (2018), except for *M. longa* for which we detected more lipids (Table 2).

Discussion

Limitations and potential improvements to the approach

Our model performed remarkably well, as illustrated by a vast majority of IoU values being higher than 0.75 and only a handful of them spreading towards zero. It is interesting, nonetheless, to explore the images for which the model performed the worst. For 49 images (2.3%), the IoU value was lower than 0.3. Among those, two broad categories of images could be described. The first (c.a. 22% of the bad fit) is formed by good quality images of large individuals whose lipid sac has been identified and annotated correctly by the human operator, while the model did not predict the correct area (Fig. 3A & 3B). It could have been lured by contrasting features in the image, such as a strong glare at the bottom of the lipid sac. Such outcome could probably be addressed via a more careful pre-treatment of the image, such as modifying the contrast, brightness or hue (Shorten and Khoshgoftaar, 2019). The second category (c.a. 78% of the bad fit) is composed by images of such quality that a human cannot reliably identify a lipid sac. A lipid sac location had nonetheless been identified (most likely wrongly so), and hence the model evaluated poorly while being potentially right in identifying the lipid sac (Fig. 3C & 3D). This could be fixed by putting more effort into a pre-selection phase before submitting the images to such a model, even though this may not be a trivial task. Another option would be reinforcement learning, i.e. to add a second step of annotation, to correct for the initial errors revealed by the model for the poorly-labelled images, and then run the model again. This is an avenue we will explore in a future project for which we plan to expand the approach to past datasets (see below).

Ecological relevance of fast and accurate lipid content estimations for individual copepods

Some long-chained fatty acids that are essential for all animals (e.g. docosahexaenoic acid - DHA - and eicosapentaenoic acid - EPA) must be obtained from planktonic primary producers via the hub of matter and energy formed by copepod communities toward higher trophic levels, including human population (Parrish, 2009; Record et al., 2018). More generally, lipid is the currency of productive marine trophic networks that rely on short-lived but intense periods of primary production (blooms), since it is the most efficient form of energy and carbon storage, easily transferable among the different actors of the network. Lipid accumulation by zooplankton even plays a significant role in the marine biological carbon pump and the global carbon cycle (Jónasdóttir et al., 2015). Pinti et al. (2023) showed that five species of lipid-rich copepods spread around the globe in productive, high-latitude ecosystems, contribute to almost 1% of total carbon export, and up to 3% of carbon sequestration

mediated by the global biological pump. They conclude by stating that including more information on other species, as well as more precise information in terms of lipid content, vertical and seasonal distribution for these species could help reaching better estimates of the global carbon cycle.

Our approach was particularly well suited to estimate accurately and efficiently the amount of lipids accumulated within the highly productive (sub)polar and upwelling ecosystems. The global error in the lipid biomass predicted by the model relative to annotation remained small, at 1%. Moreover, while the frequency distributions of individual lipid content revealed the well-known disproportionate role played by large individuals of the *Calanus* congeners, they also provided information impossible to obtain by usual sampling approaches (e.g., net tows). For instance, these distributions could be presented at a high vertical resolution and correlated to environmental measurements to provide new insights into fine-scale ecosystem regulations (similar to Schmid et al. 2018). Moreover, changes over time in the ranges and median values of species-specific distributions of lipid content could reveal changes in copepod community phenology, in communities compositions, in planktonic ecosystem functions or any combination of these (see, for example, the changes hypothesized in the modelling study of Renaud et al., 2018).

Future developments

Our approach could provide new information if applied to historical datasets collected with LOKI instruments since the beginning of the 2010s. The LOKI has been deployed on many campaigns since its first deployment (e.g., Hildebrandt et al., 2017; Massicotte et al., 2020; Niehoff et al., 2017; Schmid and Fortier, 2019) and the widespread use of centralized databases such as EcoTaxa (https://ecotaxa.obs-vlfr.fr); e.g. (Drago et al., 2022) will greatly facilitate the reanalysis of existing data. Even though such datasets are relatively recent, the accelerating pace of climate change impacts on (sub)arctic marine ecosystems in particular makes such an approach particularly relevant. It could significantly enhance our ability to finely monitor marine ecosystems (e.g. Cornils et al., 2022), while keeping the operational burden and monetary investments at a sustainable level. Many in situ imagers can be deployed on rosettes or to replace traditional nets, but the treatment of the huge amount of information collected has long been the limiting factor when considering a more widespread deployment of these approaches (Lombard et al., 2019). For example, Irisson et al. (2022) estimated based on the past few years of routine operations that a single ZooScan instrument produces about 1 billion pixels containing ~2 million objects per year, while approximately 100 of which are now distributed worldwide. Cornils et al. (2022) showed recently from a case study in the Fram Strait how

these instruments can be used for monitoring purposes by providing abundance data and taxonomic resolutions that are comparable to microscopic analyses with a fraction of its human cost and effort. However, efforts have still to be invested to improve the speed and accuracy of traits identification and measurements derived from individual images analysis in order to gain a finer understanding of marine ecosystems functioning while both their forcing and responses are rapidly changing (e.g. Panaïotis et al., 2022; Orenstein et al. 2021).

Acknowledgement

FM acknowledges that part of this work (first images annotations and preliminary development of machine learning algorithm) was conducted as part of Alexandra Mercier Master's thesis at the Département de génie électrique et génie informatique, Université Laval, Québec (QC) Canada. AM has decided not to complete her degree. It has not been possible for the first author to reach her, despite several attempts made to offer her co-authorship. FM acknowledges the contribution of Moritz Schmid who provided the images from a previous work, coauthored with FM.

Funding

This work was supported by an NSERC Discovery Grant (RGPIN-2021-03876) to F.M. and by the Institut des Sciences du Calcul et des Données (ISCD) of Sorbonne Université (IDEX SUPER 11-IDEX-0004) through the sponsored project-team FORMAL (From ObseRving to Modelling oceAn Life) F.M. and S.D.A belong to. This work is a contribution to the research program of the strategic cluster of oceanography research Québec Océan and of the IRL Takuvik (3376). Appsilon funded the development of the machine learning model and the costs of training it.

Bibliography

Belton, B., Thilsted, S.H. (2014) Fisheries in transition: Food and nutrition security implications for the global South. Glob. Food Secur. 3, 59–66.

Cornils, A., Thomisch, K., Hase, J., Hildebrandt, N., Auel, H., Niehoff, B. (2022) Testing the usefulness of optical data for zooplankton long-term monitoring: Taxonomic composition, abundance, biomass, and size spectra from ZooScan image analysis. Limnol. Oceanogr. Methods 20, 428–450.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database, in: IEEE Computer Vision and Pattern Recognition (CVPR).

Drago, L., Panaïotis, T., Irisson, J.-O., Babin, M., Biard, T., Carlotti, F., Coppola, L., Guidi, L. et al., 2022. Global Distribution of Zooplankton Biomass Estimated by In Situ Imaging and Machine Learning. Front. Mar. Sci. 9.

Groom, S., Sathyendranath, S., Ban, Y., Bernard, S., Brewin, R., Brotas, V., Brockmann, C., Chauhan, P., et al. (2019) Satellite Ocean Colour: Current Status and Future Perspective. Front. Mar. Sci. 6.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem, 770–778.

Hildebrandt, N., Henke, T., Niehoff, B., 2017. Optical methods in zooplankton studies – how efficient is LOKI (Lightframe On-sight Key species Investigation) in analyzing Arctic zooplankton communities?

Howard, J., others, 2018. fastai. GitHub. https://github.com/fastai/fastai

Irisson, J., Ayata, S., Lindsay, D.J., Karp-Boss, L., Stemmann, L., 2022. Machine Learning for the Study of Plankton and Marine Snow from Images. Annu. Rev. Mar. Sci. 14, 277–301.

Jaccard, P., 1901. Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines. Bull. Soc. Vaudoise Sci. Nat. 37, 241–72.

Jónasdóttir, S.H., Visser, A.W., Richardson, K., Heath, M.R., 2015. Seasonal copepod lipid pump promotes carbon sequestration in the deep North Atlantic. Proc. Natl. Acad. Sci. 112, 12122–12126.

Kattner, G., Hagen, W., 2009. Lipids in marine copepods: latitudinal characteristics and perspective to global warming, in: Kainz, M., Brett, M.T., Arts, M.T. (Eds.), Lipids in Aquatic Ecosystems. Springer, New York, NY, pp. 257–280.

Lombard, F., Boss, E., Waite, A.M., Vogt, M., Uitz, J., Stemmann, L., Sosik, H.M., Schulz, J., et al. (2019) Globally Consistent Quantitative Observations of Planktonic Ecosystems. Front. Mar. Sci. 6.

Martini, S., Larras, F., Boyé, A., Faure, E., Aberle, N., Archambault, P., Bacouillard, L., Beisner, B.E., et al. (2021) Functional trait-based approaches as a common framework for aquatic ecologists. Limnol. Oceanogr. 66, 965–994.

Massicotte, P., Amiraux, R., Amyot, M.-P., Archambault, P., Ardyna, M., Arnaud, L., Artigue, L., Aubry, C., et al. (2020) Green Edge ice camp campaigns: understanding the processes controlling the under-ice Arctic phytoplankton spring bloom. Earth Syst. Sci. Data 12, 151–176.

Niehoff, B., Köhler, V., Hildebrandt, N., 2017. Using the optical plankton recorder LOKI (Lightframe On-sight Key species Investigations) to elucidate high-resolution vertical distribution patterns of Arctic zooplankton species in Fram Strait.

Orenstein, E.C., Ayata, S., Maps, F., Becker, É.C., Benedetti, F., Biard, T., de Garidel-Thoron, T., Ellen, J.S., et al. (2022) Machine learning techniques to characterize functional traits of plankton from image data. Limnol. Oceanogr. 67, 1647–1669.

Orenstein, E.C., Beijbom, O., 2017. Transfer Learning and Deep Feature Extraction for Planktonic Image Data Sets, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). Presented at the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1082–1088.

Panaïotis, T., Caray–Counil, L., Woodward, B., Schmid, M.S., Daprano, D., Tsai, S.T., Sullivan, C.M., Cowen, R.K., Irisson, J.-O., 2022. Content-Aware Segmentation of Objects Spanning a Large Size Range: Application to Plankton Images. Front. Mar. Sci. 9.

Parrish, C.C., 2009. Essential fatty acids in aquatic food webs, in: Kainz, M., Brett, M.T., Arts, M.T. (Eds.), Lipids in Aquatic Ecosystems. Springer, New York, NY, pp. 309–326.

Pinti, J., Jónasdóttir, S.H., Record, N.R., Visser, A.W., 2023. The global contribution of seasonally migrating copepods to the biological carbon pump. Limnol. Oceanogr. 1–14.

Record, N.R., Ji, R., Maps, F., Varpe, Ø., Runge, J.A., Petrik, C.M., Johns, D., 2018. Copepod diapause and the biogeography of the marine lipidscape. J. Biogeogr. 45, 2238–2251.

Renaud, P.E., Daase, M., Banas, N.S., Gabrielsen, T.M., Søreide, J.E., Varpe, Ø., Cottier, F., Falk-Petersen, S., Halsband, C., Vogedes, D., Heggland, K., Berge, J., 2018. Pelagic food-webs in a changing Arctic: a trait-based perspective suggests a mode of resilience. ICES J. Mar. Sci. 75, 1871–1881.

Robbes, R., Janes, A., 2019. Leveraging Small Software Engineering Data Sets with Pre-Trained Neural Networks, in: 2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER). IEEE, pp. 29–32.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation 9, 16591–16603. https://doi.org/10.48550/arXiv.1505.04597

Schmid, M.S., Aubry, C., Grigor, J., Fortier, L., 2016. The LOKI underwater imaging system and an automatic identification model for the detection of zooplankton taxa in the Arctic Ocean. Methods Oceanogr. 15–16, 129–160.

Schmid, M.S., Fortier, L., 2019. The intriguing co-distribution of the copepods *Calanus hyperboreus* and *Calanus glacialis* in the subsurface chlorophyll maximum of Arctic seas. Elem Sci Anth 7, 50. https://doi.org/10.1525/elementa.388

Schmid, M.S., Maps, F., Fortier, L., 2018. Lipid load triggers migration to diapause in Arctic Calanus copepods—insights from underwater imaging. J. Plankton Res. 40, 311–325.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on Image Data Augmentation for Deep Learning. J. Big Data 6, 60.

Smith, L.N., Topin, N., 2017. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates 36.

Sonnet, V., Guidi, L., Mouw, C.B., Puggioni, G., Ayata, S., 2022. Length, width, shape regularity, and chain structure: time series analysis of phytoplankton morphology from imagery. Limnol. Oceanogr. 1–15.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. Imaging 15, 29.

Trudnowska, E., Lacour, L., Ardyna, M., Rogge, A., Irisson, J.O., Waite, A.M., Babin, M., Stemmann, L., 2021. Marine snow morphology illuminates the evolution of phytoplankton blooms and determines their subsequent vertical export. Nat. Commun. 12, 2816.

Vilgrain, L., Maps, F., Picheral, M., Babin, M., Aubry, C., Irisson, J., Ayata, S., 2021. Trait-based approach using in situ copepod images reveals contrasting ecological patterns across an Arctic ice melt zone. Limnol. Oceanogr. 66, 1155–1167.

Vogedes, D., Varpe, Ø., Søreide, J.E., Graeve, M., Berge, J., Falk-Petersen, S., 2010. Lipid sac area as a proxy for individual lipid content of arctic calanoid copepods. J. Plankton Res. 32, 1471–1477.

Tables

Table 1. IoU scores in training and validation datasets, split by copepod species.

Species	Image set	Median IoU	N
Calanus hyperboreus	Training	0.86	1131
	Validation	0.81	141

Calanus glacialis	Training	0.82	415
	Validation	0.80	40
Metridia longa	Training	0.64	349
	Validation	0.41	35

Table 2. Median total lipid content per individual (TL, in mg ind^{-1}) estimated from the area of the lipid sac, following Vogedes et al. (2010). Data are presented for annotated areas, areas predicted by the model, the relative error between both, as well as values from the original study of Schmid et al. (2018).

Species	TL annotated	TL predicted	Relative error	Schmid et al. (2018)
C. hyperboreus	672.1	679.9	1.2 %	699
C. glacialis	411.7	412.4	0.17 %	357
M. longa	94.5	99.9	5.7 %	55.6

Figures

Figure 1. Examples of model performance on selected LOKI images. A) *Metridia longa* adult female, B) *Calanus glacialis* C4, C) *C. glacialis* C5, D) *C. glacialis* adult female, E) *Calanus hyperboreus* C3, F) *C. hyperboreus* C4, G) *C. hyperboreus* C5, H) *C. hyperboreus* adult female. Green: annotation of lipid sac location. Red: lipid sac location estimated by the model. Yellow: overlap between both (IoU, see text).

Figure 2. Frequency distribution of individual A) IoU values and B) lipid estimates produced by the model, according to each species of copepod analyzed: *Metridia longa* (smallest individuals, c.a. 2 to 3 mm), *Calanus glacialis* (large individuals, c.a. 2.5 to 4 mm) and *C. hyperboreus* (largest individuals, c.a. 3 to 7 mm). Vertical line in A) shows median IoU = 0.82.

Figure 3. Examples of LOKI images for which the model performed badly. A) *C. hyperboreus* adult female (IoU=0.059) for which no obvious problem in either the image or the annotation is identified, B) *C. hyperboreus* C4 (IoU=0.102) for which a glare in the image could be a problem, C) (IoU=0) & D) (IoU=0.229) *Metridia longa* adult female for which the lack of contrast (dark image in C; bright

image in D) could be problematic. Moreover, the annotation in C is most likely wrong. Green: annotation of lipid sac location. Red: lipid sac location estimated by the model. Yellow: overlap.

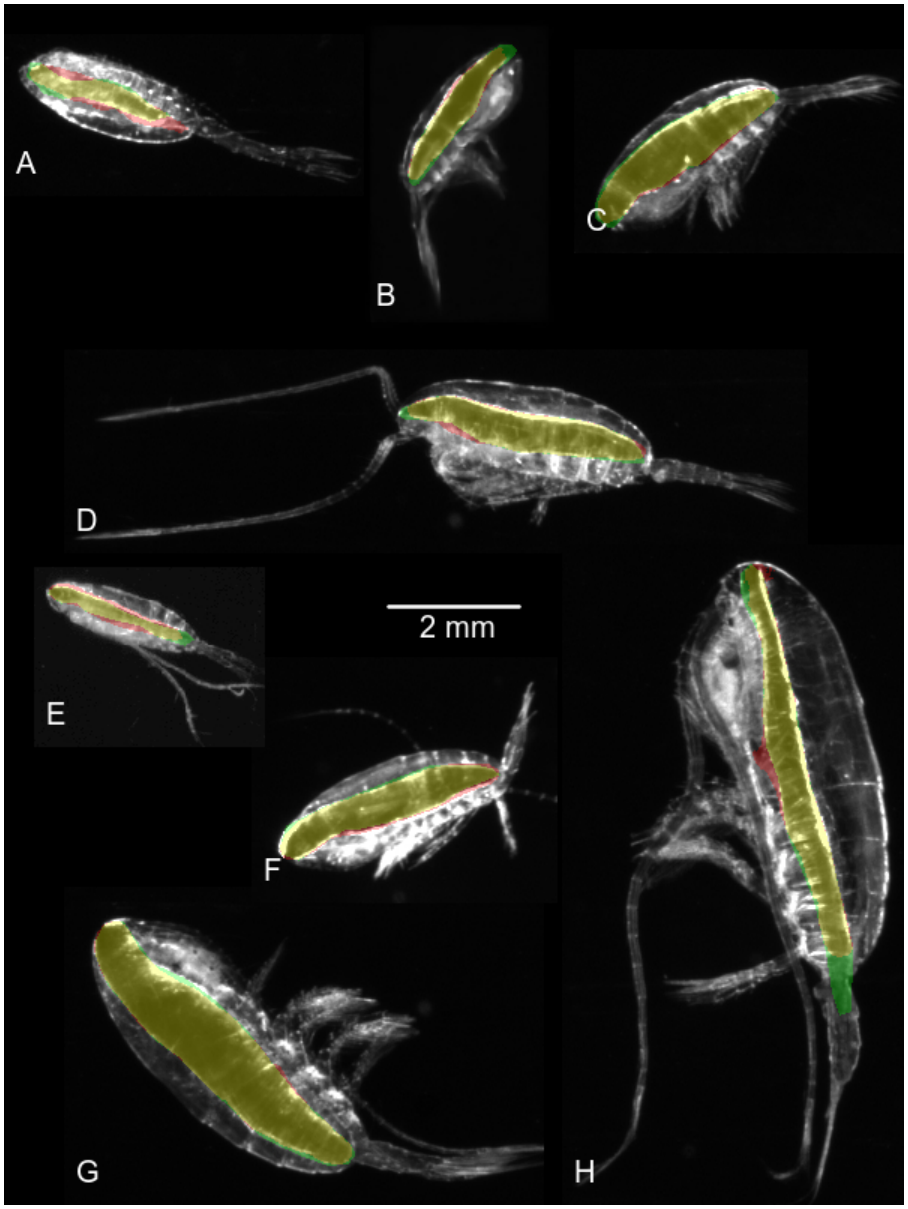

Figure 1. Examples of model performance on selected LOKI images. A) Metridia longa adult female, B) Calanus glacialis C4, C) C. glacialis C5, D) C. glacialis adult female, E) Calanus hyperboreus C3, F) C. hyperboreus C4, G) C. hyperboreus C5, H) C. hyperboreus adult female. Green: annotation of lipid sac location. Red: lipid sac location estimated by the model. Yellow: overlap between both (IoU, see text).

50x67mm (300 x 300 DPI)

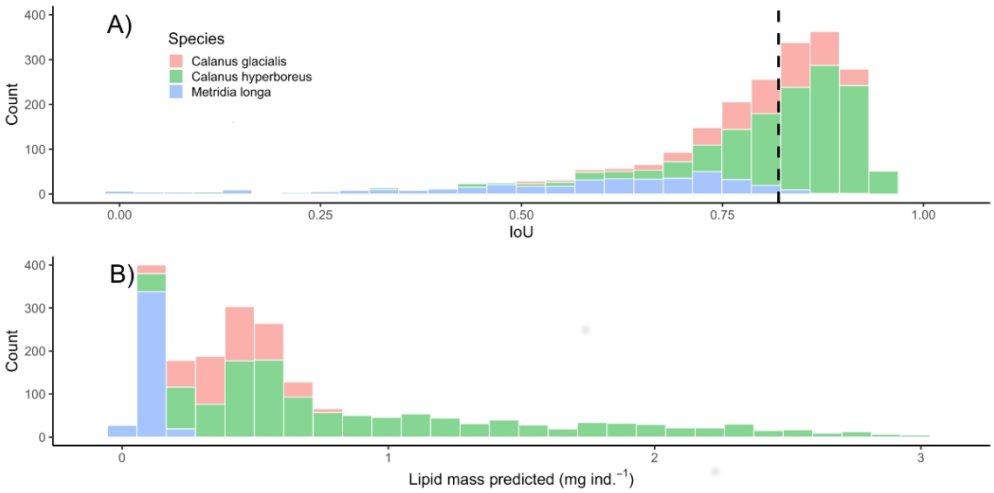

Figure 2. Frequency distribution of individual A) IoU values and B) lipid estimates produced by the model, according to each species of copepod analyzed: Metridia longa (smallest individuals, c.a. 2 to 3 mm), Calanus glacialis (large individuals, c.a. 2.5 to 4 mm) and C. hyperboreus (largest individuals, c.a. 3 to 7 mm). Vertical line in A) shows median IoU = 0.82.

101x50mm (300 x 300 DPI)

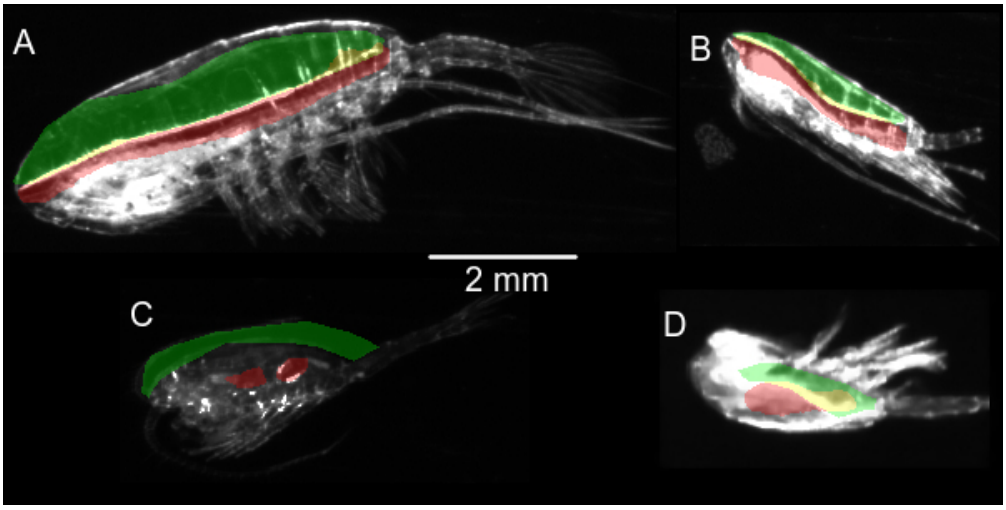

Figure 3. Examples of LOKI images for which the model performed badly. A) C. hyperboreus adult female (IoU=0.059) for which no obvious problem in either the image or the annotation is identified, B) C. hyperboreus C4 (IoU=0.102) for which a glare in the image could be a problem, C) (IoU=0) & D) (IoU=0.229) Metridia longa adult female for which the lack of contrast (dark image in C; bright image in D) could be problematic. Moreover, the annotation in C is most likely wrong. Green: annotation of lipid sac location. Red: lipid sac location estimated by the model. Yellow: overlap.

50x25mm (300 x 300 DPI)

Statement of significance

Large planktonic copepods form a crucial hub of matter and energy in productive pelagic ecosystems by converting primary production blooms into voluminous stores of lipid that are available all year long for higher trophic levels and contribute for a significant portion of the biological carbon pump. Hence it is crucial to make accurate and timely measurements of these lipid stores. Until now, such information could be retrieved from binocular pictures of individuals sampled via plankton nets. Unfortunately, with this traditional approach, the link with environmental information, both abiotic (e.g. depth, temperature, light level, oxygen concentration, etc.) and biotic (e.g. chlorophyll concentration, plankton community structure, etc.) would at best be crude, at worst lost. Since the past ~15 years, in situ imaging devices provide images whose resolution allows to estimate an individual copepod's lipid sac volume and reveal a wealth of ecological information inaccessible otherwise. This has been pioneered by Schmid, Fortier & Maps 2018 (doi: 10.1093/plankt/fby012) despite a serious methodological bottleneck: estimating the surface area occupied by a lipid sac within a copepod's prosome had to be done manually, so that weeks of work were needed by trained personnel to obtain such information for a handful of samples. With this article, we removed this hurdle by training a machine learning algorithm to estimate the lipid content of individual Arctic copepods from in situ images. This work conducted on the same dataset than in Schmid et al. (2018) now allows us to obtain such information at a speed (a few minutes) and a resolution (individuals, over half a meter on the vertical) unheard of. We think such solutions are necessary to revisit historical datasets of in situ images to better understand the dynamics of lipid production and distribution in pelagic ecosystems and to put in place efficient monitoring protocols at a moment when marine ecosystems are facing rapid upheavals and increasing threats.

Authors Contribution

FM conceived and designed the study and redacted the article.

PPS developed the machine learning algorithm and contributed to the redaction.

JŚ conceived the numerical experiments, contributed to the development of the machine learning algorithm and contributed to the redaction.

SDA contributed to the design of the study, provided the collaboration with Appsilon and contributed to the redaction.