Heliyon 10 (2024) e25367

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Phytoplankton detection and recognition in freshwater digital microscopy images using deep learning object detectors

Jorge Figueroa [a,b,*], David Rivas-Villar [a,b], José Rouco [a,b], Jorge Novo [a,b]

[a] Centro de investigacion CITIC, Universidade da Coruña, 15071 A Coruña, Spain
[b] Grupo VARPA, Instituto de Investigacion Biomédica de A Coruña (INIBIC), Universidade da Coruna, 15006 A Coruña, Spain

ARTICLE INFO

Dataset link: http://varpa.org/research/phytoplankton#FMPD

Keywords:
Microscopy images
Phytoplankton detection
Deep learning
Faster R-CNN
RetinaNet

ABSTRACT

Water quality can be negatively affected by the presence of some toxic phytoplankton species, whose toxins are difficult to remove by conventional purification systems. This creates the need for periodic analyses, which are nowadays manually performed by experts. These labor-intensive processes are affected by subjectivity and expertise, causing unreliability. Some automatic systems have been proposed to address these limitations. However, most of them are based on classical image processing pipelines with not easily scalable designs. In this context, deep learning techniques are more adequate for the detection and recognition of phytoplankton specimens in multi-specimen microscopy images, as they integrate both tasks in a single end-to-end trainable module that is able to automatize the adaption to such a complex domain. In this work, we explore the use of two different object detectors: Faster R-CNN and RetinaNet, from the one-stage and two-stage paradigms respectively. We use a dataset composed of multi-specimen microscopy images captured using a systematic protocol. This allows the use of widely available optical microscopes, also avoiding manual adjustments on a per-specimen basis, which would require expert knowledge. We have made our dataset publicly available to improve the reproducibility and to foment the development of new alternatives in the field. The selected Faster R-CNN methodology reaches maximum recall levels of 95.35%, 84.69%, and 79.81%, and precisions of 94.68%, 89.30% and 82.61%, for W. naegeliana, A. spiroides, and D. sociale, respectively. The system is able to adapt to the dataset problems and improves the results overall with respect to the reference state-of-the-art work. In addition, the proposed system improves the automation and abstraction from the domain and simplifies the workflow and adjustment.

1. Introduction

Phytoplankton are one of the most important organisms in the aquatic ecosystems of the planet. They are photoautotrophs and produce the majority of the oxygen in sub-aquatic ecosystems. They also play a key role in the sub-aquatic food chain. However, there are some specific species of phytoplankton that produce toxins. These species threaten water health if their development rises to uncharacteristic concentrations, situations which are known as blooms. There are many factors [1] that can cause these blooms, for instance, the excess of nutrients (eutrophication) linked to climate change effects on calm and warm waters. With the contemporary rise of the effects of climate change, phytoplankton blooms will become more common and of higher impact in the near future.

* Corresponding author at: Centro de investigacion CITIC, Universidade da Coruña, 15071 A Coruña, Spain.
E-mail addresses: jorge.figueroa@udc.es (J. Figueroa), david.rivas.villar@udc.es (D. Rivas-Villar), jrouco@udc.es (J. Rouco), jnovo@udc.es (J. Novo).

https://doi.org/10.1016/j.heliyon.2024.e25367
Received 14 June 2023; Received in revised form 13 December 2023; Accepted 25 January 2024

Available online 30 January 2024

Cyanobacteria [2] are one of the most common toxic subgroups of phytoplankton. The toxins they produce are named cyanotoxins. Cyanotoxin blooms can be dangerous for humans and any other animal. They are not only harmful if they are ingested, but also if there is prolonged exposure to them. They can also be harmful [3] through inhalation of particles that are suspended in the air, even if there is no direct contact with the water. Their effects go from soft gastroenteritis to cancer or damage to the liver or the nervous system. One of the most hazardous groups is microcystins, which are hepatoxic, gravely affecting humans and all mammals in general.

Water analysis is an important task that must be performed periodically to prevent the previously mentioned health risks. Reservoirs, ponds, and lakes are the most common critical sites in terms of cyanobacterial blooms. A study [4] shows that, in Europe, it is estimated that 35% to 48% of the reservoirs pass through a cyanobacterial dominance at some time over the year, and around in 75% of those cases it implies toxicity. Therefore, local authorities are usually forced to perform periodic water analyses and create reports to detect and follow these possible toxic blooms, guaranteeing the potability and safety of the water.

Nowadays, toxic phytoplankton analyses of water are performed manually by experts. The process to do this analysis is as follows. First, water samples are extracted. The most common approach for water sampling involves the experts extracting the samples in situ, at different depths, ensuring their representativity. The samples are then transported to the laboratory, where they are prepared and manually inspected using microscopy imaging techniques. This microscopy analysis usually consists on species taxonomy and counting specimens to estimate the biomass. The whole process is time-consuming and prone to factors compromising the reliability of the analysis. In this sense, a study [5] shows that the performance of manual analysis by the experts can be as low as 75% recall in water samples with a high debris load. Additionally, inter and intra-expert agreements are severely affected by the varying complexity of the target species [6], as well as by the expertise of the trained taxonomists [7]. This variability can be mitigated with the establishment of standardized protocols and criteria. Other human factors, such as tiredness or boredom caused by repeating the same task numerous times, can also affect the quality of the analyses. Thus, automating the analysis task is desirable, as it would allow us to avoid all these quality disadvantages related to the human factor, along with the time-optimization of the whole process.

In the related bibliography, there have been several works aiming to automatize different aspects of the water phytoplankton analyses. Namely, the automatization of the water sample processing and imaging, the identification of individual phytoplankton specimens, and the automatic taxonomy, identifying the species of each specimen.

Regarding sample processing and image acquisition, some works have proposed the use of specialized hardware. Some of these devices, such as VPR [8], SIPPER [9], KRIP [10], and FlowCytobot [11], consist of submersible devices, pulled by a boat or other mechanism, that allow to gather and image the water samples in situ. The imaging process consists of flow control mechanisms, forcing that only one specimen passes at a time through the imaging sensor. This allows to obtain single-specimen images, i.e. effectively solving the identification of individual phytoplankton specimens, that are readily available for the subsequent taxonomy analysis process. However, in order to do that, it is necessary the use of different flow cells [12], depending on the size of the target species to be analyzed. This flow imaging mechanism is also used by other devices such as FlowCam [13–16]. In this case, the water sample gathering and processing are performed manually, and the specialized hardware is used in the laboratory to automate the imaging and specimen separation.

However, the flow imaging devices are usually expensive and application dependent. Instead, the most commonly available imaging devices are regular optical microscopes equipped with digital cameras. These microscopes are usually manually operated by adjusting the magnification and focus required for each specimen, which are of varying size and at a different depth within the water in the sample slide. Given the convenience of using regular microscopes, some works [17–20] aimed at providing a systematic microscopy imaging protocol that may release the experts from this manual per-specimen adjustment. This usually results in multi-specimen images that require further processing, i.e. detecting and separating the regions occupied by each specimen as a previous stage to species recognition. In these works, this is performed using classical image processing methods. In the case of PLASA [17] and PlanktoVision [18], the methods take advantage of fluorescence imaging, along with the use of multiple magnifications and focal points for each slide. This eases the identification of specimens but complicates the image capturing. Instead, Rivas et al. [19,20] proposed to use single magnification and single focus optical images. This simplifies the imaging process, which is convenient, while being unattended and systematic. However, this poses more challenging conditions for image processing and analysis, such as the more prevalent appearance of partially out-of-focus specimens, and lower overall magnification to be able to capture the larger species. These challenges need to be addressed by the posterior computational processes. In this work, we also follow this imaging approach.

In the context of phytoplankton species recognition, most of the works depart from single-specimen images as input. These single-specimen images were either captured using specific automatic hardware or previously extracted from multi-specimen images by performing a detection task using classical image processing techniques. The majority of these species recognition works [21–26] use classical image processing techniques to extract the features for the classification. Alternatively, some recent works [27–29] have explored the use of deep learning techniques to automate the feature extraction, usually obtaining better results than classical approaches.

Deep learning has been successfully applied to a wide variety of domains, such as medical [30], prediction of climate conditions [31], or tracking of objects in videos [32], among many other fields. Despite the relevant advances, and specifically regarding the use of deep learning methods for recognition, the use of deep learning has been barely explored for the phytoplankton specimen detection tasks. This is relevant, as the detection task in multi-specimen phytoplankton images is a complex process, potentially benefiting from advanced adjustments provided by deep learning to each specific water domain, imaging conditions, and morphology of the considered species. Moreover, state-of-the-art deep learning object detection techniques allow the unification of both tasks in

a single end-to-end trainable module, simplifying the adjustment process. This can result in an improved level of abstraction from the domain, as well as a potential performance enhancement with respect to classical image processing approaches.

In this work, we perform a comprehensive study covering both one-stage and two-stage families of detectors. We propose to test RetinaNet [33] and Faster R-CNN [34] from one-stage and two-stage respectively. We perform an in-depth analysis of the detection and recognition performance, both globally and for each target species. This allows to compare the suitability of both methods for the toxic phytoplankton detection and recognition problem. In addition, we provide a comparative study with the work of Rivas-Villar et al. [19]. This method is the reference state-of-the-art for phytoplankton detection and recognition in multi-specimen images. It consists on a fully automatic approach, based on classical image processing techniques, which performs the detection and recognition tasks in separate stages. Their work is developed over the same dataset as our method, therefore allowing for a direct and fair comparison environment. We summarize our contributions as follows:

- We present a comprehensive study of RetinaNet and Faster R-CNN for toxic phytoplankton detection and recognition, covering both one-stage and two-stage detectors and providing a more complete study than previous related works.
- We center our study on high recall detection regimes for relevant toxic species, providing an exhaustive species-wise performance analysis, aligned with the application goals.
- We apply a simplified systematic protocol for image acquisition, that actually allows to release the experts from manually adjusting each image. This, however, introduces additional challenges to the posed recognition problem, that have not been considered by previous recognition studies.
- We provide a performance comparison between our work and the work by Rivas-Villar et al. [19], the only work to date over the same dataset, showing the advantages of our proposed methods with respect to classical approaches.

This manuscript is structured in sections as follows: Section 2 Related Work, where we explore the different SOTA works; Section 3 Methodology, in which we explain our dataset characteristics, both used techniques (Faster R-CNN and RetinaNet), and the training details; Section 4 Results and discussion, where we illustrate the achieved performance and discuss it, as well as compare it with other SOTA work; Section 5 Conclusion, in which we summarize the research and findings of this work.

2. Related work

The phytoplankton species recognition problem is explored by a wide variety of works [21–29,35]. These methods only focus on species recognition, as they work over single-specimen images which are usually extracted with automatic image capturing devices [8–11,13–16]. While remarkable results have been achieved for each of the considered single-specimen imaging procedures, these works do not propose any solution for the specimen detection problem. In contrast, we address both detection and recognition in a single end-to-end trainable module.

A limited number of works center their study solely on specimen location. The algorithm proposed in KRIP [10] extracts regions of interest using classical image processing techniques. Other works [35,36] also follow a classical approach to segment and locate the specimens. However, in addition to not considering the species recognition problem, the methodology of these works consists of ad hoc procedures with assumptions based on the morphology of specific species, which introduces a considerable amount of domain dependencies.

With the aim of addressing the whole phytoplankton analysis process, some works [17–20] propose to cover both detection and recognition problems as separate modules of the proposed pipeline. These works operate over multi-specimen images but still base the detection on classical image processing techniques. PLASA [17] and PlanktoVision [18] are fully classical approaches that utilize fluorescence images, as well as multiple magnifications during the image acquisition, which significantly complicates the capturing process. On the other hand, the work by Rivas-Villar et al. [19] is the only work performed over the same dataset as our proposed work, consisting of images acquired using a simplified systematic protocol and optical microscopes. They perform the detection using classical image processing techniques, involving additional steps such as a machine learning-based candidate selection and a colony merging step, which results in a complex workflow. After the detection step, they follow a classical image machine learning-based pipeline for species recognition over the resulting single-specimen images. In contrast with these works, we propose to group both detection and classification in a single end-to-end trainable module simplifying the adjustment process and automating the feature extraction, which can improve the level of abstraction from the domain. Moreover, we address both detection and recognition with deep learning techniques, potentially improving the overall performance results with respect to classical approaches.

In this work, we provide a performance comparison with the work by Rivas-Villar et al. [19]. This is because our work and the one from Rivas-Villar et al. use the same dataset, which enables direct comparison. In the case of other methods, the use of different datasets impedes comparison. Even if the methods' pipeline is similar, the differences among the datasets difficult directly comparing approaches. Firstly, our dataset is from freshwater phytoplankton, which has higher biodiversity than marine water (the most common in phytoplankton analyses) [19]. Furthermore, the abundance of species and their morphology can vary from freshwater to marine environments, due to geographical locations, climate, seasons, etc. Moreover, the imaging device and its settings greatly impact the aspect of the images and thus the methods and their results. For instance, using an optical microscope, like in our dataset, several settings distinctly impact the images. Our approach captures multi-specimen images with fixed magnification, focal point and illumination differentiating it from other approaches which modify these settings on a per image or specimen basis, reducing the difficulty of the problem to automate at the cost of increasing the work of the experts. Thus, we can only compare our method to other ones using this dataset, which in this case, is just the work by Rivas-Villar et al. [19].

Several deep learning object detectors have been proposed for broad-domain applications over the last years. In this work, we selected to test Faster R-CNN [34] and RetinaNet [33], from the two-stage and one-stage paradigms respectively, due to their architectural similarities and the fact that they can employ the same backbone. The functioning of the R-CNN two-stage family of detectors [34,37,38] is based on two well-differentiated steps dedicated to region proposal and object classification, respectively. Due to this, they are considered two-stage detectors. Within this family, Fast R-CNN [38] introduces a multi-task loss that allows to train the detection and recognition networks at the same time, and Faster R-CNN [34] introduces the Region Proposal Network (RPN), dedicated to processing the regions generated by strided windows different shapes and resolutions, called anchors, and decide if there is an object in the region. An alternative is the one followed by one-stage detectors, which propose a single-step approach to detection and recognition, avoiding the region proposal step. Within this latter family, You Only Look Once (YOLO) [39] divides the image space in a grid and predicts the presence of objects of each candidate size in each cell. In contrast, Single-Shot Detector (SSD) [40] reduces the dimensionality of the feature maps to a considerably low resolution and performs the detection at each position. Finally, RetinaNet [33] generates regions of interest by applying the same concept of anchors as in Faster R-CNN [34]. However, instead of dedicating part of the network to discard the background regions, RetinaNet uses a focal loss that minimizes the impact of these regions on the adjustments. Both RetinaNet [33] and Faster R-CNN [34] utilize similar backbone architectures. Therefore, due to these design concepts and structure similarities, we propose to test RetinaNet [33] and Faster R-CNN [34] using the same backbone architecture in both cases.

Although some recent works [41–43] explore the use of deep learning for phytoplankton tasks, they are based on marine phytoplankton and do not approach both detection and recognition tasks. Two related works have explored the use of deep learning object detectors for phytoplankton detection and recognition in freshwater samples. Baek et al. [44] use Fast R-CNN [38] to perform cyanobacteria cell counting, and Qian et al. [45] use Faster R-CNN [34] to identify the taxonomic information of different species of algae. These works are adjusted on different datasets, which are captured with varying magnification and focus in order to adjust it to each specimen. Moreover, they consider different target species, creating a completely different experimental environment that impedes a fair comparison between the performance of theirs and our work. In addition, the experimental and analytical purpose of these studies differ from our proposed work. These works only explore the use of two-stage detectors, whereas we also contemplate one-stage detectors with RetinaNet [33], providing a more comprehensive study of the field. The dataset of Baek et al. [44] was captured with variations in magnification and focal point, which were selected ad hoc for each specimen, requiring an expert taxonomist manually operating the microscope. On the other hand, Qian et al. [45] do not provide specifications about the image capturing process. Conversely, our images are captured with constant magnification and focal point, improving the repeatability and reliability of the process, as well as the practical implications on the use of the developed system. Moreover, this increases the presence of partially out-of-focus specimens, as well as specimens of reduced size, which increases the complexity of the detection and recognition problem and aligns it with the requirements of a truly unattended pipeline for the imaging to the final result. Finally, the analyses provided in these works do not evaluate the reliability of the detection of target toxic species. In the case of Qian et al. [45], the evaluation is only approached with mean Average Precision (mAP) for each species, without considering the evaluation of the possible high recall regime performance of the toxic species. Instead, Baek et al. [44] focus their evaluation on the toxic species, but their analysis is centered on cell counting. In this work, in addition to covering a wider spectrum of object detectors, we perform a more exhaustive detection evaluation for the toxic species, which is essential in a problem that involves health risks.

3. Methodology

In this section we explain in-depth the proposed methodologies, as well as the dataset and experimental conditions. Overall, the proposed method consists of a deep learning object detector that is trainable end-to-end. The object detector receives an image as input and directly outputs the bounding boxes and species of the predicted specimens. These object detectors are able to identify each individual specimen without the necessity of additional processing techniques.

3.1. Dataset

The used dataset [19] consists of 293 color microscopy images with multiple phytoplankton specimens of varying species in each one. They have an image size of 2080×1540 pixels and were captured with a magnification of 10×, providing a resolution of 0.33 μm per pixel. The samples were all taken from lake Doniños (Ferrol, Galicia, Spain) (UTM 555593 X, 4815672 Y; Datum ETRS89), at different depths, and at different times of the year (2017-2019). In this regard, the dataset represents the lake biomass under different conditions. Each image of the dataset was manually labeled by an expert taxonomist, determining the species and bounding box for each specimen. In contrast with datasets in previous multi-specimen approaches [44,45], the images were taken with invariant focal point and magnification without making any ad hoc adjustment with respect to the specimens, removing the necessity of an expert taxonomist operating the microscope and resulting in a significant workload reduction for the experts. Moreover, the simplification of the capturing process facilitates the acquisition of new samples for inference tasks. This unattended imaging protocol, however, implies the presence of several out-of-focus specimens that may be confused with background or any other element. This introduces additional challenges in the detection and recognition problem. Although this increases the complexity of the problem, the flexibility improvements of the unattended imaging using regular microscopes are essential for the practical automation of the whole analysis process. In order to facilitate comparisons with our approaches and to ease access to our dataset, it has been made public.[1] As

[1] http://varpa.org/research/phytoplankton#FMPD.

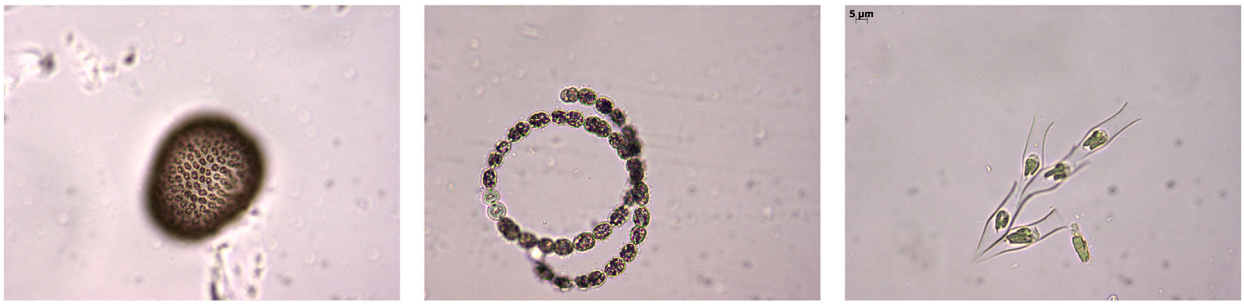


Fig. 1. Specimen example for each of the target species. These images were captured with higher magnification to present the target species. (a) W. naegeliana, (b) A. spiroides, (c) D. sociale.

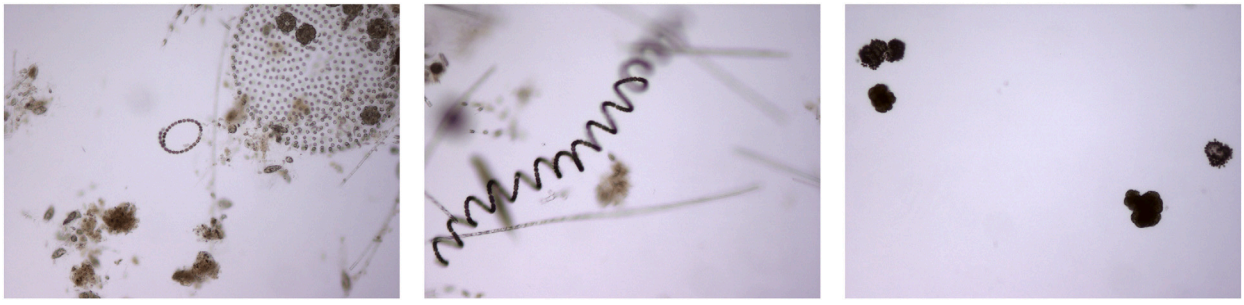


Fig. 2. Input image examples. In (a) and (b) we show several out-of-focus and overlapping specimens, as well as two A. spiroides specimens of significantly different size and appearance. In (c) we can see multiple specimens that are similar in shape and color but belong to different species.

mentioned, this dataset is unique due to its image capture approach, following a systematic and repeatable approach which reduces the burden on the experts.

From the dataset images, we selected three phytoplankton species as target classes. The reason for selecting two of them is due to their high degree of toxicity, making them the most important species for the application. Additionally, one more species was chosen to be the target of our method, since it was abundant in the dataset, and it presents a unique set of additional challenges for species identification due to its morphology. The unidentified instances of non-target species are gathered in an additional, generic, fourth class, “Other phytoplankton”. Most of these species in the “Other phytoplankton” class are either non-toxic or too scarce to be considered their own class. In summary, the following three phytoplankton species are treated as target classes for the proposed method:

- **Woronichinia naegeliana:** formed by cells usually grouped in an oval shape (Fig. 1a). Their toxins can produce damage to the nervous system and liver. There are 233 specimens in the dataset.
- **Anabaena spiroides:** formed by cells grouped in an elongated shape (Fig. 1b). They have similar toxic characteristics as W. naegeliana. This is the least represented species in the dataset with 58 specimens, so the behavior of this species is one of the main focuses of the performance analysis.
- **Dinobryon sociale:** it does not produce toxins, however, it can deteriorate the water quality. These specimens have a green nucleus and a translucent capsule, making them a challenging species, as it could generate confusion with background and other particles (Fig. 1c). In addition, some zooplankton species also have a translucent capsule and can be easily confused with D. sociale. This is the most abundant species in the dataset, with 354 specimens.

Fig. 1 shows an example of each target species. Phytoplankton, and especially the species represented in our dataset, are characterized by having significantly large inter and intra-class variability. In addition, there are other non-phytoplankton organisms and particles that can complicate the detection, as well as additional challenges of the dataset such as out-of-focus and reduced-size specimens. Fig. 2 depicts three representative images from the dataset that illustrate these challenges. For instance, we can see the disparities in specimen size and appearance between the A. spiroides specimens of Figs. 2a and 2b, as well as the presence of out-of-focus specimens. Fig. 2c shows the presence of multiple specimens with similar appearances, where two of them are W. naegeliana specimens and the rest are Botryococcus braunii specimens. This illustrates the challenge that the inter-class similarity implies.

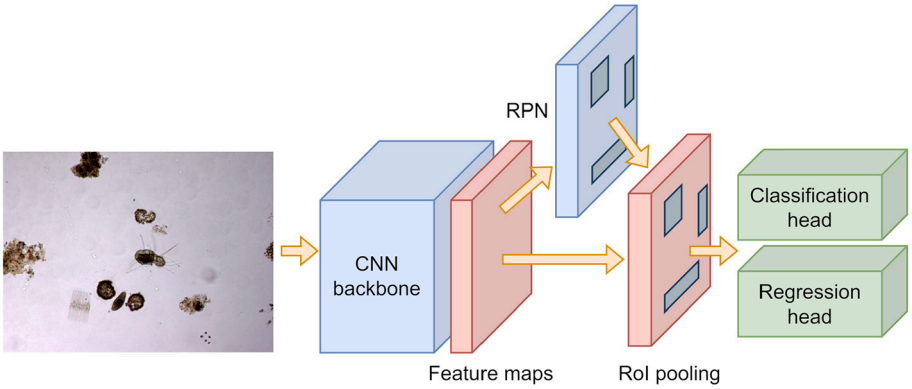


Fig. 3. Overview of the Faster R-CNN [34] architecture.

3.2. Two-stage detection: faster R-CNN

Faster R-CNN [34] is the most relevant model following the two-stage detection paradigm due to its wide use. Two-stage detectors are characterized by having two well-differentiated steps. The first step consists on region proposal whereas the second step consists on processing those Regions Of Interest (ROI) to predict the exact bounding box and class of the object. Faster R-CNN is an evolution from previous two-stage object detectors such as R-CNN [37] and Fast R-CNN [38].

Fig. 3 shows an overview of the Faster R-CNN architecture. It features the use of a Convolutional Neural Network (CNN) feature extraction backbone and an ROI regression and classification head, equivalent to that of Fast R-CNN [38]. The initial regions are generated by strided windows of predefined shape and size, called anchors, at each location of the backbone feature maps. The Region Proposal Network (RPN), as a first stage, is in charge of selecting between these candidate object regions, for which the backbone features are pooled. Then, in the second stage, these features, for each object candidate, are fed into the classification and regression head, which returns the object class and an accurate estimation of the object's bounding box. The classifier considers all the possible object classes plus an additional background class to reject false positive RPN detections.

The network is trained end-to-end using a multi-task loss [34] over the training object samples. This multi-task loss is defined as

$$L(p,u,t^u,v) = L_{cls}(p,u) + \lambda[u \geq 1]L_{loc}(t^u,v), \tag{1}$$

where L_{cls} and L_{loc} are the classification and regression losses, p is the class probability output vector, u is the true class, t^u denotes the predicted bounding box coordinates for the true class, v denotes the ground truth coordinates, $[u \geq 1]$ denotes an indicator function that nullifies the regression for background regions, and λ is a weighting factor to balance the tasks. The classification loss consists on the cross-entropy loss

$$L_{cls} = -\log(p_u), \tag{2}$$

where p_u denotes the predicted probability for the true class. The regression loss is defined in Equation (3) as

$$L_{loc}(t^u,v) = \sum_{i\in\{x,y,w,h\}} smooth_{L_1}(t_i^u - v_i), \tag{3}$$

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases} \tag{4}$$

where $smooth_{L_1}$ (Equation (4)) is a robust L_1 loss that is designed to be less sensitive to outliers, easing the avoidance of exploding gradients [46,47]. The loss balancing parameter λ is usually set to 1. The indicator function $[u \geq 1]$ is used to be able to provide feedback to the regression network only when there is an object in the processed region, allowing to train both classification and regression sub-nets at the same time without conflicts.

3.3. One-stage detection: RetinaNet

RetinaNet [33] is one of the most relevant object detectors in the one-stage group [33,39,40], characterized by performing the region proposal and object identification in a single step. We selected to test RetinaNet in this work due to its similarities in structure and features with Faster R-CNN despite following a different detection paradigm. An overall scheme of the RetinaNet architecture is shown in Fig. 4. RetinaNet architecture is based on a Feature Pyramid Network [48] that provides feature maps at different resolutions. The FPN consists of a multiscale top-down decoder that upsamples and integrates the multiple levels of a CNN backbone that is used for bottom-up feature extraction from the input image. Then, several regions of interest of predefined shape and size are extracted from each pyramid level by sliding anchors over the feature maps, following the same concept as in Faster R-CNN [34].

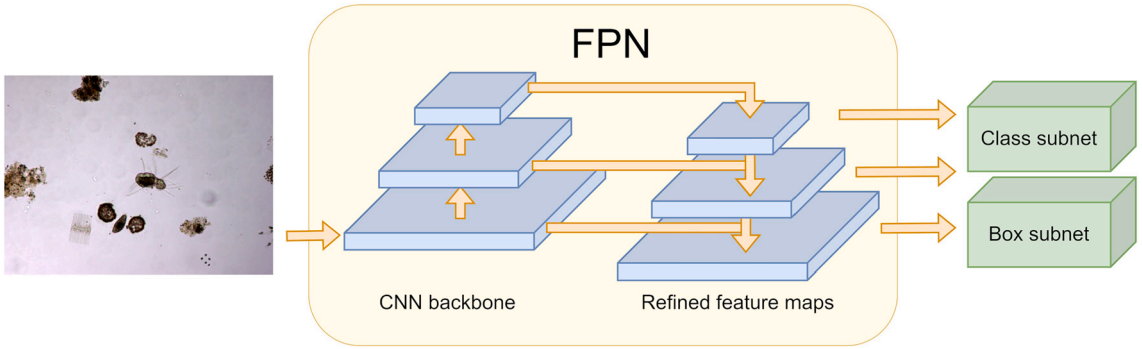


Fig. 4. Overview of the RetinaNet [33] architecture.

These regions are directly processed by two CNN subnet heads, dedicated to classifying the object and refining its bounding box coordinates. These subnets have the same weights regardless of the pyramid level.

The network is trained with the same multi-task loss as Faster R-CNN [34], described in Equation (1). However, in the case of RetinaNet, none of the background cases are discarded by an RPN as it would happen in Faster R-CNN [34]. This generates a significant class imbalance since most of the regions belong to the background. This is solved by using the focal loss [33] for the classification sub-net output, instead of the regular cross-entropy loss in Equation (2). The focal loss is defined as

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma} log(p_t), \tag{5}$$

where p_t is the probability inferred by the net for the class t, α_t is a weighting factor for addressing the imbalance of the target classes, and γ controls the relevance of the easy samples in the loss. Increasing the value of γ results in less impact on the loss from the samples that were classified with a high probability, i.e. the easy samples for which the model shows high confidence at each training iteration. Note that choosing $\gamma = 0$ and removing the weighting factor α_t corresponds to the regular cross-entropy loss, as in Equation (2) in Faster R-CNN [34].

3.4. Training details

Both Faster R-CNN and RetinaNet models were pre-trained in COCO [49]. For both methods, we selected an FPN [48] based on ResNet-50 [50] as CNN backbone, pre-trained in ImageNet [51]. A k-fold cross-validation split of $k = 5$ is used for the refinement on the phytoplankton dataset. We consider the global distribution of the dataset to be representative of the overall populations of phytoplankton species in this lake, as it was obtained from water sampled at different points of the year, accounting for the seasonal distribution of the species. Therefore, in order to keep this distribution in the 5-fold splits, the multi-specimen nature of the dataset must be taken into account as each image has an arbitrary number of specimens belonging to an arbitrary number of species. This means that the global distribution is unlikely to be replicated completely accurately due to this per-image constraint. Thus, we generated several random 5-fold splits and selected the set whose distributions were closer to the global dataset distribution. For each experiment, 4 folds are used for training and the remaining one for testing. We do not use a validation subset as we directly use a fixed number of epochs. This allows to maximize the training data used on each repetition, while also minimizing the chance of inconsistent training lengths among the folds due to the early stopping patience value. Thus, we created a controllable experimental environment.

The multi-task loss balancing factor is set to $\lambda = 1.0$, which is the default. For RetinaNet, the focal loss parameters (see Equation (5)) are set to $\gamma = 2.0$ and $\alpha = 0.25$ following the recommendations on the original work [33]. The same training details are used for both architectures in order to set a fair comparison environment. The starting learning rate was set to 0.001 and it was decreased by a factor of 0.1 at epochs 50 and 100. These values are sufficient for the obtainment of satisfactorily trained models, guaranteeing their convergence. Both models were trained with the SGD optimizer with momentum [52,53], setting its weighting parameter to the default value $\beta = 0.9$. We additionally apply weight decay regularization [54,55] to improve generalization and avoid vanishing and exploding gradients [46,47]. The weight decay parameter is set to 0.0001, relatively smaller than the learning rate but sufficient to potentially provide minor improvements. Online data augmentation is performed at each iteration. Combinations of multiple random transforms are applied with a probability of 0.5, consisting of vertical and horizontal flipping, and random brightness and contrast modifications with factors between 0.8 and 1.2. The selected augmentation parameters were selected in order to obtain new images that visually resemble the original samples from the dataset.

For inference in test, non-max suppression is performed to filter nearby detections. For each prediction, we check if there is any other prediction that overlaps with an Intersection over Union (IoU) of 0.5 or higher, and we select the one with higher class probability. The IoU is computed as the ratio between the area in which two bounding boxes intersect and the total area that both joint boxes cover.

Table 1
Mean AP and F1 score with their respective deviation over the folds. Best values are highlighted in bold.

	AP		F1 score	
	Faster R-CNN	RetinaNet	Faster R-CNN	RetinaNet
W. naegeliana	**97.30**±2.76	87.52 ± 8.54	**96.64**±1.90	94.60 ± 3.10
A. spiroides	**90.09**±7.78	65.36 ± 14.91	**89.19**±5.74	59.95 ± 17.98
D. sociale	**84.49**±3.57	74.08 ± 7.47	**82.36**±2.62	81.09 ± 5.58
Other	**67.87**±3.02	67.41 ± 5.41	**67.17**±4.77	66.65 ± 3.40
Global	**76.22**±2.38	70.35 ± 4.10	**75.89**±3.62	74.58 ± 2.97

3.5. Evaluation metrics

Although both object detectors integrate the detection and recognition tasks in a single module, different metrics are extracted to evaluate the response in each of these tasks. We consider as true positives the predictions matching the ground truth class and with an IoU of 0.5 or higher with respect to the ground truth bounding box. In contrast, the state-of-the-art work by Rivas-Villar et al. [19] uses the overlapping to decide if a prediction is a true positive, considering as correct the predictions whose overlapping area with the specimens of the ground truth species surpasses a threshold. This evaluation approach can consider colonies, i.e. single detections that cover multiple specimens, as a true positive. In this way, our IoU approach is more demanding, as it measures the performance on the individual specimens, requiring a more exact specimen-wise detection process to obtain satisfactory results. In the comparative study with Rivas-Villar et al. [19] that we provide in this work, we evaluate both methods following our IoU approach.

For the evaluation of specimen detection performance, we propose to use precision and recall. Precision measures the ratio between the correctly detected objects and the total detected objects. On the other hand, recall measures the capability of the model to detect all the objects, calculated as the ratio between the correctly detected objects and the total existent objects. Precision and recall are computed both globally and for each class. This allows us to perform an exhaustive analysis of the response to each class. The precision-recall curves are constructed by averaging the individual curves over the folds. Since each fold model obtains different operating points, the curves are interpolated in a common recall space. We construct precision-recall curves with the aim of obtaining a better estimation of the global model behavior regardless of the specific operating points. Average precision (AP) is also computed globally and for each class. The AP is calculated as the mean over the precision at each operating point. This metric gives a quick overview of the precision-recall curve. We compute the average and standard deviation of the AP along all the folds.

In order to analyze the classification performance (i.e. identification of species), we employ confusion matrices. These matrices are computed globally and for each class. Confusion matrices provide information about the classification error between object classes, allowing to assess the misclassification trends. The confusion matrices are computed using the predictions for all the test folds instead of performing an average.

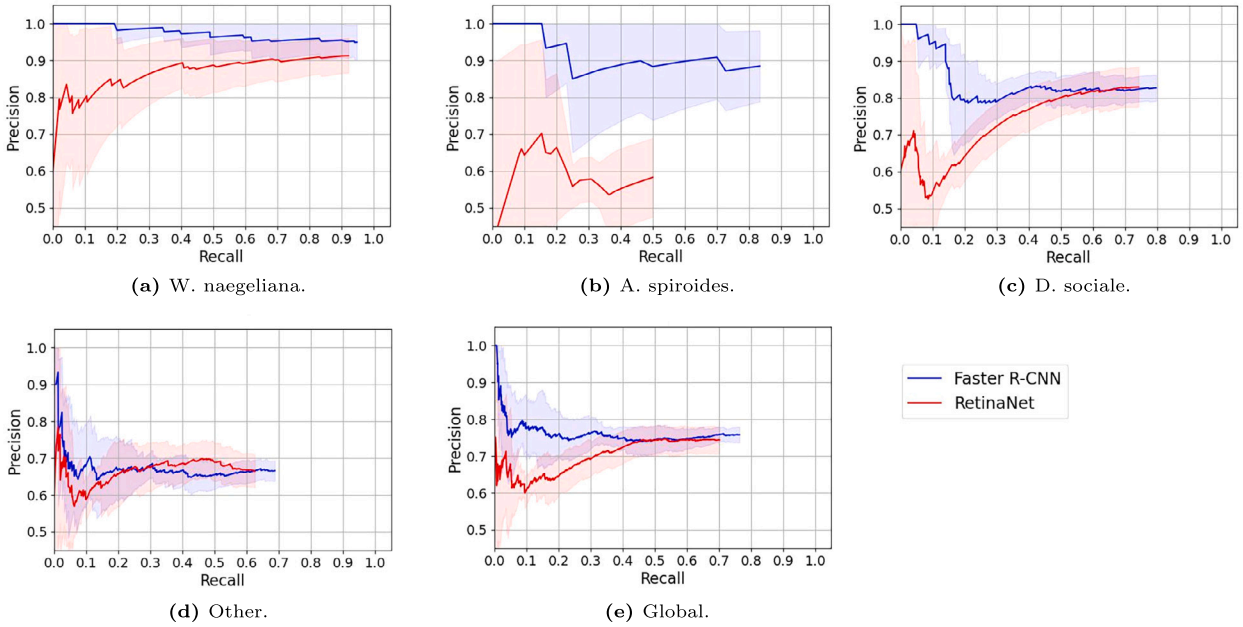
4. Results and discussion

4.1. Overall performance analysis

Table 1 shows the AP for all the classes and both models. Faster R-CNN outperforms RetinaNet AP in A. spiroides by a notable margin of 25% on average. In the rest of the target species, Faster R-CNN also surpasses RetinaNet by about 10% margin. This reveals that Faster R-CNN is more precise than RetinaNet overall, especially in A. spiroides, the class with fewer samples. In this context, we observe that Faster R-CNN presents some architectural advantages, as the RPN is able to filters most of the background samples, reducing the class imbalance that they generate for the classification and regression heads. In comparison to Faster R-CNN, the lack of a dedicated part of the RetinaNet architecture to filter the background samples implies more difficulties to address this class imbalance. In RetinaNet, the classification head processes all the possible ROIs and have an impact on the focal loss. These ROIs include several non-phytoplankton specimens (i.e. background areas) with similar features to the target classes in the dataset, as well as subareas from specimens with homogeneous textures and appearance that can be misconceived as specimens themselves. Thus, a considerable amount of difficult background samples make an impact on the focal loss and consequently in the adjustment of the model. This reduces, in general, the relevance of the specimens from the target species, especially the less represented species. Thus, considering that the AP reduction in RetinaNet is larger in A. spiroides, the less represented class, than in other classes, the focal loss from RetinaNet is more sensitive to these issues taking into account the complexity of the problems related to the background. In contrast, Faster R-CNN provides structural advantages with the RPN, being more adequate to correctly discard background samples without significant secondary effects in the subsequent classification.

4.2. Precision-recall analysis

Figs. 5 shows the precision-recall curves for the target classes (Figs. 5a, 5b, 5c), other species (Fig. 5d), and the global curves (Fig. 5e) for both models. We observe that Faster R-CNN can achieve higher recall in all the cases. The most remarkable improvement is achieved for the A. spiroides class. In this case, Faster R-CNN surpasses 80% recall while RetinaNet obtains around 50%, which



(a) W. naegeliana. (b) A. spiroides. (c) D. sociale.

(d) Other. (e) Global.

Fig. 5. Precision/Recall curves for each class, for Faster R-CNN (blue) and RetinaNet (red).

Table 2
Precisions reached by the models at recall levels of 70%, 80% and 90%. The cells with "-" indicate that the model does not reach that recall level. Best values are highlighted in bold.

	Faster R-CNN Precision (%)			RetinaNet Precision (%)		
	70% Recall	80% Recall	90% Recall	70% Recall	80% Recall	90% Recall
W. naegeliana	**95.30**±4.88	**95.84**±4.34	**95.50**±4.64	90.16 ± 5.84	90.64 ± 4.71	91.23 ± 4.71
A. spiroides	**90.87**±11.40	**88.07**±9.99	-	-	-	-
D. sociale	82.02 ± 3.43	-	-	**82.88**±5.32	-	-
Global	**75.85**±2.06	-	-	74.40 ± 3.73	-	-

is not acceptable in a problem where missing toxic specimens can be harmful. This recall difference in A. spiroides between both models reveals again that the RPN of Faster R-CNN generates more adequate conditions for the classification head. In contrast, the number of background samples that affect the focal loss leads to a performance reduction in RetinaNet, being remarkable in A. spiroides but also noticeable in W. naegeliana.

Table 2 shows the precisions achieved at 70%, 80%, 90% recall by the compared models. We observe that the precisions in W. naegeliana for Faster R-CNN are superior at all the selected high recall levels. Despite this, the results in W. naegeliana are satisfactory for both models and in the expected range, since two-stage detectors are more precise by nature. Both models achieve similar precision at 70% recall for D. sociale, the most represented class in the dataset. This shows that the background issues of RetinaNet have a less noticeable impact on the species with more samples. In contrast, the achieved precision at high recall levels in A. spiroides is around 90% for Faster R-CNN while RetinaNet does not reach any high recall level, remarking the imbalance issues of RetinaNet in this particular problem.

Table 3 shows the precision at the maximum achieved recall. Faster R-CNN surpasses RetinaNet recall in W. naegeliana by around 2.7%, while also improving the precision at the maximum recall by 3.3%. Regarding A. spiroides, we observe that Faster R-CNN surpasses the maximum recall of RetinaNet by around 32%, and even improves the precision at this maximum recall by around 27%, remarking the superiority of Faster R-CNN in this class and the issues of RetinaNet regarding scarce species. In D. sociale, Faster R-CNN provides a 5% recall improvement with respect to RetinaNet despite the similarities in precision. The main problem regarding D. sociale is related to the presence of out-of-focus specimens and other non-phytoplankton specimens with similar morphological features. This recall improvement shows that Faster R-CNN has a superior discriminative capability in these cases, being more robust to these dataset problems. This is again linked to the structural advantages of Faster R-CNN, as it simplifies the classification environment by previously discarding most of the background samples with the RPN. We prioritize reaching high recall over high precision, therefore this 5% improvement is more relevant than a minor precision increment of less than 1% in RetinaNet, therefore Faster R-CNN is superior also for D. sociale.

In general, the architectural and functional advantages of Faster R-CNN with respect to RetinaNet are the cause of the superior precision and recall results. The dedication of part of the architecture, i.e. the RPN, to discard background samples results in more adequate inputs for the classification head, which has to deal with a lower amount of background samples. This leads to precision

Table 3
Precisions at maximum recall levels, calculated as the average among the folds. The standard deviation is also provided. Best values are highlighted in bold.

	Faster R-CNN		RetinaNet	
	Precision (%)	Max. Recall (%)	Precision (%)	Max. Recall (%)
W. naegeliana	**94.68**±5.00	**95.35**±2.31	91.35 ± 4.78	92.76 ± 3.18
A. spiroides	**89.30**±8.96	**84.69**±8.91	61.94 ± 12.96	52.60 ± 21.31
D. sociale	82.61 ± 2.93	**79.81**±7.19	**83.26**±5.85	74.84 ± 5.98
Other	66.51 ± 2.49	**68.94**±1.71	**66.54**±47.43	62.93 ± 3.77
Global	**75.56**±1.77	**76.39**±1.83	74.59 ± 4.01	70.19 ± 3.37

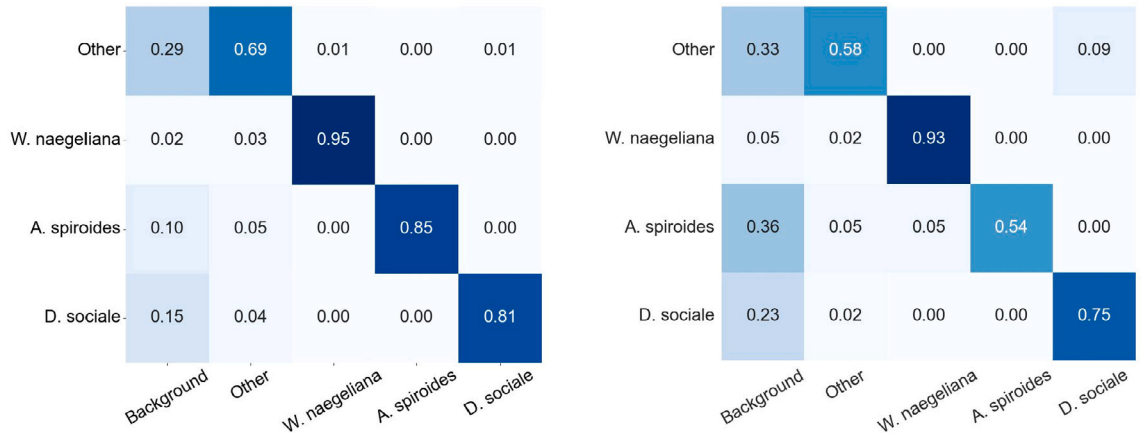


Fig. 6. Confusion matrices for (a) Faster R-CNN and (b) RetinaNet, calculated over all the detections. The rows represent the ground truths, and the columns the predictions.

and recall improvements in W. naegeliana and A. spiroides, being remarkable in the latter, while also providing a recall improvement in D. sociale. In contrast, the difficult background samples that are processed by the classification head of RetinaNet, i.e. subareas of specimens with homogeneous textures and appearance that have similar features to the complete specimen, are more frequent. This affects the model adjustment, decreasing the relevance of the target species, especially the less represented ones such as A. spiroides, resulting in an inferior overall performance. Therefore, Faster R-CNN is a more adequate option for the phytoplankton detection and recognition problem, in which there is a considerable amount of difficult background elements that produce undesired effects on the adjustment of RetinaNet.

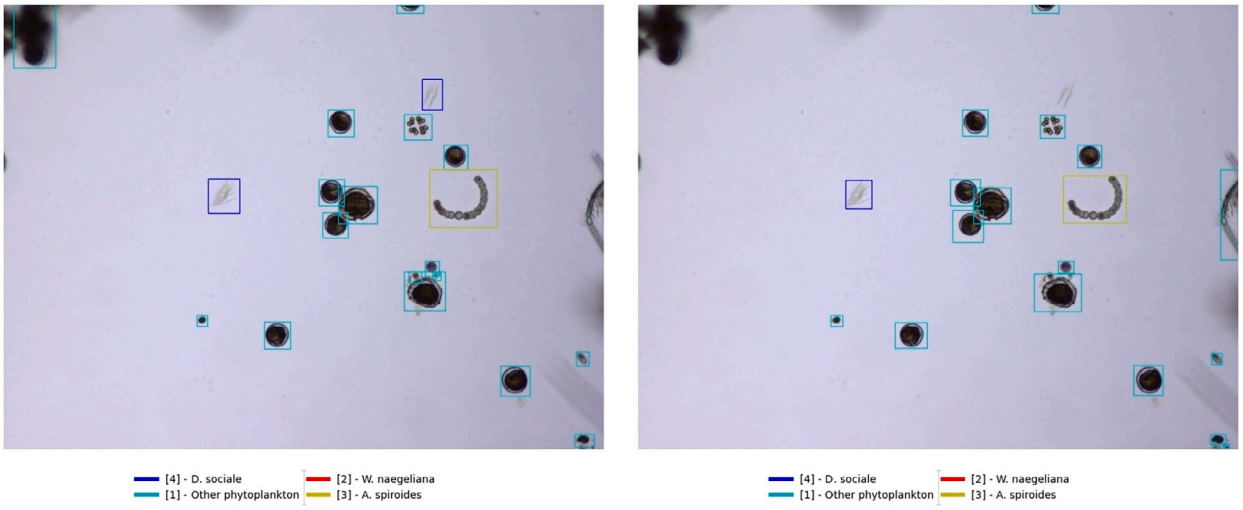
4.3. Classification performance analysis

Fig. 6 shows the confusion matrices for Faster R-CNN (Fig. 6a) and RetinaNet (Fig. 6b). We observe that the majority of the confusions of the target species in both models are either with background or with other non-target species. The confusions with other species have a similar impact on all the target species and usually involve specimens with similar morphological features. However, these confusions have less relevance than the background confusions, i.e. specimens classified as background, which are more frequent in both models. We observe that RetinaNet classifies a 36% as background for A. spiroides and a 23% for D. sociale, while Faster R-CNN only a 10% and a 15%, respectively. In W. naegeliana, as an exception, the background confusions are not impactful in any of the models.

To complement the confusion matrices, Table 4 shows the number of background samples classified as a target class for both models, corresponding with the unnormalized first row of the confusion matrices. We decided to separately represent it with the raw number of predictions rather than a percentage, since there is a variable number of predictions and no background ground truth, so the percentages are not representative. Faster R-CNN classifies more background samples as D. sociale than RetinaNet, which is the reason behind Faster R-CNN obtaining a mildly lower precision in this class. Fig. 7 depicts a Faster R-CNN prediction (Fig. 7a) in which an element was identified as D. sociale but is not a phytoplankton specimen (see Fig. 7b). Note the morphological similarities between both specimens, such as having the same translucent capsule. These similarities are strengthened by the fact that they are partially out of focus. The same causes were found for the D. sociale confusions in RetinaNet since these issues are related to the dataset features. Faster R-CNN classifies a lower number of background samples as W. naegeliana. However, this does not imply a considerable impact on the performance given the abundance of W. naegeliana specimens in the dataset. Faster R-CNN achieves a similar superiority regarding background samples classified as A. spiroides but, in this case, it has an impact on the overall model

Table 4
Number of false positives, i.e. background samples classified as a target class, for Faster R-CNN and RetinaNet.

	Other	W. naegeliana	A. spiroides	D. sociale
Faster R-CNN	254	6	8	53
RetinaNet	234	15	17	47



(a) Prediction. (b) Ground Truth.

Fig. 7. Example of Faster R-CNN confusing an out-of-focus non-phtyoplankton specimen with D. sociale.

performance since the number of A. spiroides specimens is considerably low. In addition, Figure 7 shows the inherent capability of the model to individually identify overlapping specimens without the necessity of further processing techniques.

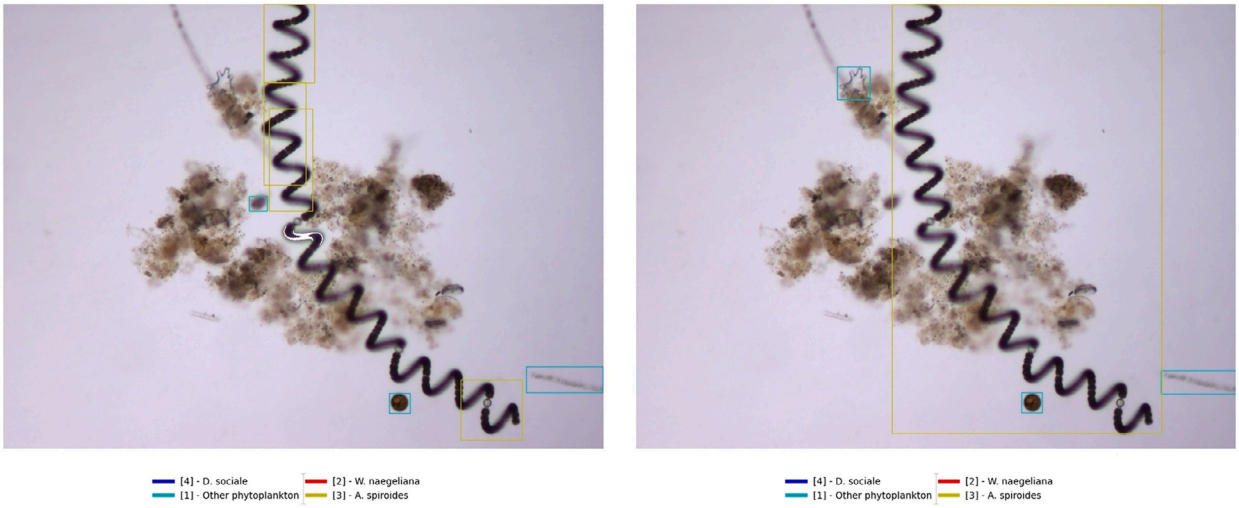
Fig. 8 depicts an example of A. spiroides error for RetinaNet, where the model tends to produce multiple detections (Fig. 8a) for a single A. spiroides specimen (Fig. 8b). The reason behind this behavior is attributed to the structural characteristics of the method. The classification and regression heads of RetinaNet process all the possible ROIs in the image. In this way, multiple subareas of a single specimen are processed by the heads, along with the areas covering the whole specimen. Since the specimens usually have homogeneous texture and morphological appearance, consequently producing similar feature maps, the classification head has difficulties discarding these subareas as background, while, on the other hand, full specimen regions may contain large portions of background (or other specimens) in combination with the target specimen features. Thus, RetinaNet classifies multiple small background samples as A. spiroides, as well as a large A. spiroides specimen as background, producing confusions in both directions. These errors occur along all the folds in RetinaNet and are inherent to the method. In contrast, in Faster R-CNN the A. spiroides errors are located in one specific fold, revealing that they can be situational. In this context, also taking into account the overall performance superiority in all the classes, Faster R-CNN is a more adequate option for the phytoplankton detection and recognition problem.

4.4. Time complexity

Table 5 shows the time costs for both Faster R-CNN [34] and RetinaNet [33]. We can observe that RetinaNet is around 15% and 20% faster than Faster R-CNN. However, these improvements are not sufficient if we consider the remarkable performance advantages of Faster R-CNN. Therefore, we conclude that, despite the longer execution times, Faster R-CNN is the better option due to its superior performance.

4.5. Comparison with previous work

We present a comparison between the proposed work and Rivas-Villar et al. [19], the only state-of-the-art work on the same dataset as ours, which is required in order to provide a fair and significant comparison. Their work addresses the same problem under the same conditions, and consists of a fully automatic approach, based on classical image processing techniques, that performs the detection and recognition tasks in separate modules that, in contrast to our work, need to be individually adjusted. Both methods are adjusted over the same data and center the performance analysis on the same target classes. The evaluation of both methods is performed over the holdout split used by Rivas-Villar et al. [19]. Table 6 shows the class distribution for this holdout split. In addition, both methods are evaluated using the metrics and criteria described in Section 3.5. In order to do that, as these metrics are



(a) Prediction. (b) Ground Truth.

Fig. 8. Example of an A. spiroides error in RetinaNet, related to the lack of a ROI filtering mechanism.

Table 5
Time costs in seconds for Faster R-CNN and RetinaNet on a RTX 3080.

	Training	Validation/ Inference
Iteration Faster R-CNN	0.064	0.044
Iteration RetinaNet	0.051	0.037
Epoch Faster R-CNN	14.85	2.55
Epoch RetinaNet	11.83	2.15
150 epochs Faster R-CNN	2227.50	382.50
150 epochs RetinaNet	1774.50	322.50

Table 6
Number of specimens per class for the holdout split used in the comparison with Rivas-Villar et al. [19].

	Train	Test
W. naegeliana	175	58
A. spiroides	48	10
D. sociale	314	58
Other	666	128

not reported in [19], we computed the performance of the integrated detection and classification system of Rivas-Villar et al. for this work.

Table 7 shows the precisions at the maximum recall levels for both methods. It is observed that the proposed work outperforms Rivas-Villar et al. [19] in recall by a margin between 25% and 35% in all the target species. Regarding the precision, the proposed work even improves it despite achieving higher recalls, which means considering more predictions for the precision calculation. However, our work obtains an inferior precision in A. spiroides. The reason behind this is that their method introduces a considerable amount of background noise in the detection and dedicates an ad hoc step to discard it. Thus, their method is less prone to produce false negatives than false positives, affecting the recall instead of the precision. In contrast, our work achieves a superior recall, usually implying producing more predictions that can potentially penalize the precision with the presence of false positives. In spite of this, the recall improvement is more relevant and remarks the superiority of the proposed work in A. spiroides with respect to Rivas-Villar et al. [19].

Fig. 9 shows the confusion matrices for both methods. We observe the same performance improvements in all the target species as in Table 7. For the proposed work (Fig. 9a), the confusions between species only occur between W. naegeliana and other non-target species. Thus, the main confusion cause is related to specimens from the target species labeled as background. These confusions with

Table 7
Comparison of precisions at maximum recall levels between our work and Rivas-Villar et al. [19] work. Best values are highlighted in bold.

	Ours		Rivas-Villar et al.	
	Precision (%)	Max. Recall (%)	Precision (%)	Max. Recall (%)
W. naegeliana	**89.47**	**82.50**	82.35	56.90
A. spiroides	80.00	**80.00**	**100.00**	50.00
D. sociale	**84.09**	**92.50**	79.31	57.50
Other	61.54	**62.5**	**69.00**	53.91

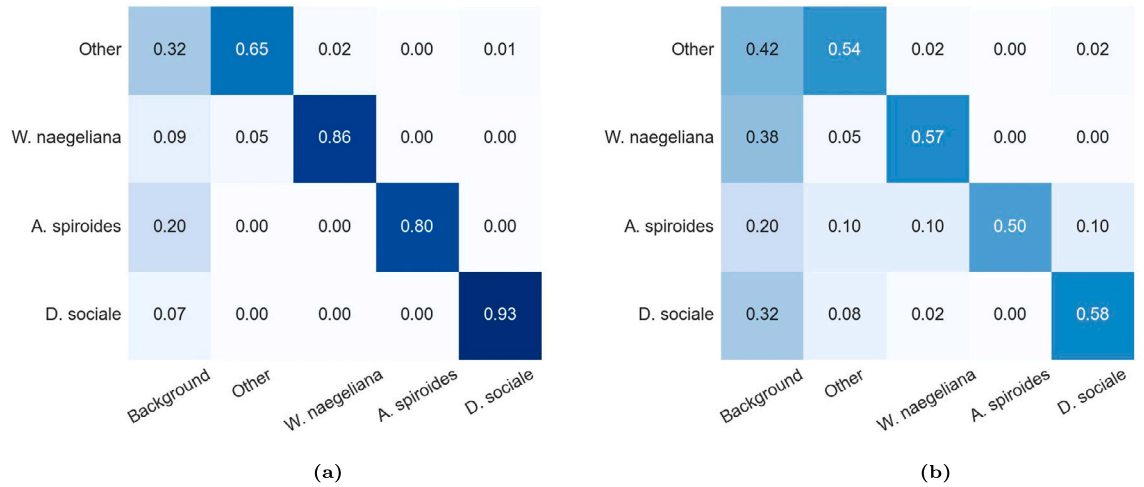


Fig. 9. Confusion matrices for (a) our method and (b) Rivas-Villar et al. [19], using the same holdout split. The rows represent the ground truths, and the columns the predictions.

Table 8
Number of false positives, i.e. background samples classified as a target class, for our method and Rivas-Villar et al. [19].

	Other	W. naegeliana	A. spiroides	D. sociale
Ours	44	3	3	6
Rivas-Villar et al.	24	2	0	2

background also occur in their method to a larger extent (Fig. 9b). However, their method also produces confusions between A. spiroides and the other target species, showing an inferior discriminative capability in the classification.

Table 8 shows the number of background samples classified as a target class, corresponding with the unnormalized first row of the confusion matrices. The proposed method classifies more background samples as a target class. In the case of W. naegeliana and D. sociale, these confusions do not impact the results given the abundant amount of specimens. In contrast, the 3 false detections in A. spiroides have a considerable impact on the performance metrics, since there are only 10 specimens in the test set, and are the reason for the 80% precision in Faster R-CNN. However, the achievement of a higher recall level is more relevant and a priority in our work. Therefore, the overall performance of Faster R-CNN is also superior in A. spiroides despite the effects of these confusions. The reason behind their superior false positive performance is that they dedicate a full step, which needs to be adjusted ad hoc, to discard the non-phytoplankton samples. In this way, their method produces more false negatives than false positives. Moreover, this complex pipeline complicates the adjustment, penalizing the flexibility of the method. In contrast, our method incorporates the complete detection and recognition process in a single end-to-end trainable module and obtains an overall superior performance in all the classes.

Fig. 10 shows a comparison between both methods in the response to W. naegeliana colonies. We observe that their method (Fig. 10c) groups the specimens into a single detection. In contrast, the proposed method (Fig. 10b) is able to detect them as single specimens, providing a more accurate solution in relation to the ground truth (Fig. 10a). In this context, the proposed method performs a more exact specimen-wise detection, in spite of the specimens being close together or even overlapping. Therefore, our proposal provides a higher quality specimen location and a more accurate specimen counting, even in complex cases.

In addition to the performance improvements, as well as the more exact detection that our method provides in comparison to Rivas-Villar et al. [19], there are several advantages implicit to our proposed method. Our method automatizes the feature extraction process and is trainable end-to-end, easing the adjustments that are needed for the detection and recognition tasks, which are complex in the multi-specimen domain. Moreover, this removes the ad hoc adjustments that are present in their work, increasing the level

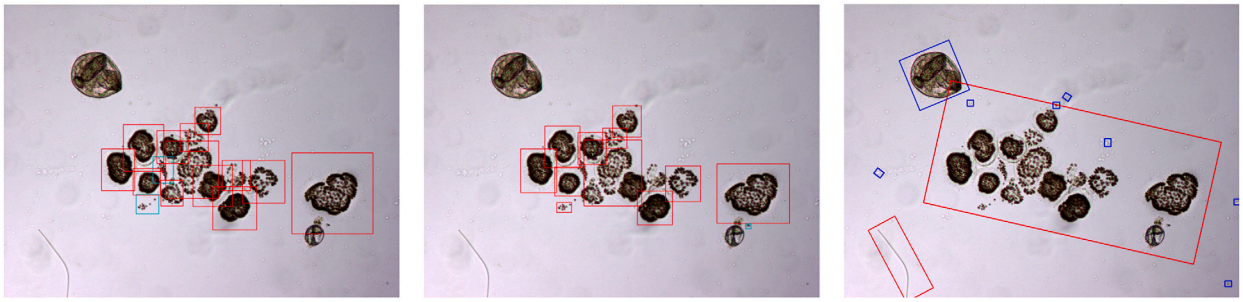


Fig. 10. Colony differences between our work and Rivas-Villar et al. [19]. We show (a) the ground truth, (b) our prediction and (c) Rivas-Villar et al. [19] prediction. In our work and the ground truth, the bounding box color represents the object class. In the example of their work, the color represents if the detections were merged into a colony (red) or if it was directly detected as a single specimen (blue).

of abstraction from the domain. This results in a simplified structure, avoiding the complexity of the automatic pipeline proposed by Rivas-Villar et al. and easing the extrapolation and scalability of the method, which would require ad hoc readjustments in their case. Thus, we propose a more flexible and less domain-dependent method that even provides several performance improvements for the phytoplankton detection and recognition problem.

5. Conclusion

Some phytoplankton species are able to generate toxins, which can have a negative impact on health if they are concentrated in a high amount. The analysis of the water is currently performed manually by the experts, and most of the proposed automatic methods are classical approaches with ad hoc procedures and domain dependencies. In this paper, we proposed to test Faster R-CNN and RetinaNet, two widely used object detectors with similar overall architectures but different working paradigms. We used a dataset of images that were acquired using a simplified and systematic protocol with fixed focal point and magnification, reducing the expert workload. However, this introduces specimens that are out of focus, increasing the complexity of the problem. Thus, the goal of this work is to improve the flexibility and domain abstraction of the method while keeping the acquisition process simple.

The experiments revealed that Faster R-CNN was remarkably superior to RetinaNet in all aspects, mainly due to architectural and functional differences. The RPN from Faster R-CNN is specifically dedicated to selecting the candidate ROIs, whereas RetinaNet has to process all the possible ones. As a result, RetinaNet has problems on distinguishing complete specimens from subareas due to their homogeneous appearances, generating multiple smaller incorrect detections. These problems affect the results, as Faster R-CNN achieves ∼32% more recall than RetinaNet and ∼28% more precision in a critical species such as A. spiroides. Overall, the achieved Faster R-CNN performance is superior to RetinaNet, which suggests that two-stage detectors are more adequate for this problem. In addition, we provide a considerable performance improvement with respect to the reference state-of-the-art previous work, which is a classical approach over the same dataset and target species. Overall, the obtained results are remarkable considering the limited dataset size in comparison to typical large scale deep learning datasets. However, we acknowledge that the set of analyzed species is still small and limited. In future work, we plan to expand the dataset by including new water samples in the dataset from other lakes in order to increase the number of target species, following the proposed simplified systematic acquisition protocol. This, also considering the remarkable performance of the method herein proposed, would increase the variety of environments in which our system could successfully operate. Furthermore, another limitation of our study is that we only tested ResNet-50 as the backbone for both Faster R-CNN and RetinaNet, since it is widely used in the state of the art. Although the results are satisfactory, ideally, more backbones could be tested, specially ResNets of bigger and smaller sizes, which would allow us to find the ideal backbone size for this dataset.

In conclusion, we successfully built a deep learning methodology based on Faster R-CNN, a two-stage detector, and found that one-stage detectors are not suitable for microscopy phytoplankton images in comparison. Also, the nature of our dataset minimizes the image acquisition work from the experts but increases the complexity of the task that the automatic system has to deal with. However, the proposed approach is able to adapt to the complex dataset and obtain remarkable results with respect to previous work while improving the domain abstraction and flexibility. Overall, the proposed system is able to perform toxic phytoplankton detection and classification despite the complexities that such tasks entail.

CRediT authorship contribution statement

Jorge Figueroa: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **David Rivas-Villar:** Writing – review & editing, Resources, Investigation, Data curation. **José Rouco:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Jorge Novo:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Data availability

The dataset used in this work (FMPD) has been published (http://varpa.org/research/phytoplankton#FMPD).

Acknowledgements

This research was funded by Ministerio de Innovación, Government of Spain [research projects PDC2022-133132-I00, PID2019-108435RB-I00 and TED2021-131201B-I00]; Consellería de Cultura, Educación e Universidade, Xunta de Galicia through the pre-doctoral grant contracts ref. ED481A 2021/147 and ED481A 2023/059, and Grupos de Referencia Competitiva, grant ref. ED431C 2020/24; CITIC, Centro de Investigación de Galicia ref. ED431G 2019/01, receives financial support from Consellería de Educación, Universidade e Formación Profesional, Xunta de Galicia, through the ERDF (80%) and Secretaría Xeral de Universidades (20%).

References

[1] H.W. Paerl, V.J. Paul, Climate change: links to global expansion of harmful cyanobacteria, Water Res. 46 (5) (2012) 1349–1363, https://doi.org/10.1016/j.watres.2011.08.002, cyanobacteria: impacts of climate change on occurrence, toxicity and water quality management.
[2] B. Whitton, M. Potts, The Ecology of Cyanobacteria: Their Diversity in Time and Space, Springer, 2002.
[3] I. Chorus, M. Welker, Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management, CRC Press, 2021.
[4] A. Quesada, D. Sanchis, D. Carrasco, Cyanobacteria in Spanish reservoirs. How frequently are they toxic?, Limnetica 23 (2004), https://doi.org/10.23818/limn.23.09.
[5] M.R. First, L.A. Drake, Performance of the human “counting machine”: evaluation of manual microscopy for enumerating plankton, J. Plankton Res. 34 (12) (2012) 1028–1041, https://doi.org/10.1093/plankt/fbs068.
[6] K. Vuorio, L. Lepistoe, A.-L. Holopainen, Intercalibrations of freshwater phytoplankton analyses, Boreal Environ. Res. 12 (2007) 561–569 [cited 13/12/2022], http://www.borenv.net/BER/archive/pdfs/ber12/ber12-561.pdf.
[7] P. Culverhouse, R. Williams, B. Reguera, V. Herry, S. González-Gil, Do experts make mistakes? A comparison of human and machine identification of dinoflagellates, Mar. Ecol. Prog. Ser. 247 (2003) 17–25, https://doi.org/10.3354/meps247017.
[8] C.S. Davis, S.M. Gallager, M. Marra, W. Kenneth Stewart, Rapid visualization of plankton abundance and taxonomic composition using the video plankton recorder, Deep-Sea Res., Part 2, Top. Stud. Oceanogr. 43 (7) (1996) 1947–1970, https://doi.org/10.1016/S0967-0645(96)00051-3.
[9] A. Remsen, T. Hopkins, S. Samson, What you see is not what you catch: a comparison of concurrently collected net, optical plankton counter, and shadowed image particle profiling evaluation recorder data from the northeast Gulf of Mexico, Deep-Sea Res., Part 1, Oceanogr. Res. Pap. 51 (2004) 129–151, https://doi.org/10.1016/j.dsr.2003.09.008.
[10] Y. Nagashima, Y. Matsumoto, H. Kondo, H. Yamazaki, S. Gallager, Development of a realtime plankton image archiver for auvs, in: 2014 IEEE/OES Autonomous Underwater Vehicles (AUV), 2014, pp. 1–6.
[11] H.M. Sosik, R.J. Olson, Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry, Limnol. Oceanogr., Methods 5 (6) (2007) 204–216, https://doi.org/10.4319/lom.2007.5.204.
[12] N. Barteneva, I. Vorobjev, D. Basiji, A. Lau, T.T.W. Wong, H.C. Shum, K. Wong, K. Tsia, M. Hildebrand, A. Davis, R. Abbriano, H. Pugsley, J. Traller, S. Smith, R. Shrestha, O. Cook, E. Sanchez-Alvarez, K. Manandhar-Shrestha, B. Alderete, Imaging Flow Cytometry: Methods and Protocols, Methods in Molecular Biology, vol. 1389, Springer, 2016.
[13] E. Álvarez, A. López-Urrutia, E. Nogueira, Improvement of plankton biovolume estimates derived from image-based automatic sampling devices: application to flowcam, J. Plankton Res. 34 (6) (2012) 454–469, https://doi.org/10.1093/plankt/fbs017.
[14] M.G. Camoying, A.T. Yñiguez, FlowCam optimization: attaining good quality images for higher taxonomic classification resolution of natural phytoplankton samples, Limnol. Oceanogr., Methods 14 (5) (2016) 305–314, https://doi.org/10.1002/lom3.10090.
[15] E. Álvarez, M. Moyano, A. López-Urrutia, E. Nogueira, R. Scharek, Routine determination of plankton community composition and size structure: a comparison between FlowCAM and light microscopy, J. Plankton Res. 36 (1) (2013) 170–184, https://doi.org/10.1093/plankt/fbt069.
[16] E. Álvarez, A. López-Urrutia, E. Nogueira, S. Fraga, How to effectively sample the plankton size spectrum? A case study using FlowCAM, J. Plankton Res. 33 (7) (2011) 1119–1133, https://doi.org/10.1093/plankt/fbr012.
[17] K. Rodenacker, B. Hense, U. Jütting, P. Gais, Automatic analysis of aqueous specimens for phytoplankton structure recognition and population estimation, Microsc. Res. Tech. 69 (9) (2006) 708–720, https://doi.org/10.1002/jemt.20338.
[18] K. Schulze, U. Tillich, T. Dandekar, M. Frohme, Planktovision - an automated analysis system for the identification of phytoplankton, BMC Bioinform. 14 (2013) 115, https://doi.org/10.1186/1471-2105-14-115.
[19] D. Rivas-Villar, J. Rouco, R. Carballeira, M.G. Penedo, J. Novo, Fully automatic detection and classification of phytoplankton specimens in digital microscopy images, Comput. Methods Programs Biomed. 200 (2021) 105923, https://doi.org/10.1016/j.cmpb.2020.105923.
[20] D. Rivas-Villar, J. Rouco, M.G. Penedo, R. Carballeira, J. Novo, Automatic detection of freshwater phytoplankton specimens in conventional microscopy images, Sensors 20 (22) (2020), https://doi.org/10.3390/s20226704.
[21] Q. Hu, C. Davis, Automatic plankton image recognition with co-occurrence matrices and support vector machine, Mar. Ecol. Prog. Ser. 295 (2005) 21–31, https://doi.org/10.3354/meps295021.
[22] J.-y. Zhao, H. Guo, X.-b. Sun, A research on the recognition of chironomid larvae based on svm, in: 2009 Pacific-Asia Conference on Circuits, Communications and Systems, 2009, pp. 610–613.
[23] I. Corrêa, P. Drews, M. Silva de Souza, V.M. Tavano, Supervised microalgae classification in imbalanced dataset, in: 2016 5th Brazilian Conference on Intelligent Systems (BRACIS), 2016, pp. 49–54.
[24] T. Luo, K. Kramer, D. Goldgof, L. Hall, S. Samson, A. Remsen, T. Hopkins, Recognizing plankton images from the shadow image particle profiling evaluation recorder, IEEE Trans. Syst. Man Cybern., Part B, Cybern. 34 (4) (2004) 1753–1762, https://doi.org/10.1109/TSMCB.2004.830340.
[25] P. Culverhouse, R. Simpson, R. Ellis, J. Lindley, R. Williams, T. Parisini, B. Reguera, I. Bravo, R. Zoppoli, G. Earnshaw, H. McCall, G. Smith, Automatic classification of field-collected dinoflagellates by artificial neural network, Mar. Ecol. Prog. Ser. 139 (1996) 281–287, https://doi.org/10.3354/meps139281.
[26] L. Boddy, C. Morris, M. Wilkins, L. Al-Haddad, G. Tarran, R. Jonker, P. Burkill, Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data, Mar. Ecol. Prog. Ser. 195 (2000) 47–59, https://doi.org/10.3354/meps195047.

[27] I. Correa, P. Drews, S. Botelho, M.S. de Souza, V.M. Tavano, Deep learning for microalgae classification, in: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, pp. 20–25.
[28] E.C. Orenstein, O. Beijbom, Transfer learning and deep feature extraction for planktonic image data sets, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, pp. 1082–1088.
[29] H. Lee, M. Park, J. Kim, Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning, in: 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 3713–3717.
[30] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, https://doi.org/10.1016/j.media.2017.07.005.
[31] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais Prabhat, Deep learning and process understanding for data-driven Earth system science, Nature 566 (7743) (2019) 195–204, https://doi.org/10.1038/s41586-019-0912-1.
[32] G. Ciaparrone, F. Luque Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, F. Herrera, Deep learning in video multi-object tracking: a survey, Neurocomputing 381 (2020) 61–88, https://doi.org/10.1016/j.neucom.2019.11.023.
[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2020) 318–327, https://doi.org/10.1109/TPAMI.2018.2858826.
[34] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149, https://doi.org/10.1109/TPAMI.2016.2577031.
[35] H. Zheng, N. Wang, Z. Yu, Z. Gu, B. Zheng, Robust and automatic cell detection and segmentation from microscopic images of non-setae phytoplankton species, IET Image Process. 11 (11) (2017) 1077–1085, https://doi.org/10.1049/iet-ipr.2017.0127.
[36] A. Verikas, A. Gelzinis, M. Bacauskiene, I. Olenina, E. Vaiciukynas, An integrated approach to analysis of phytoplankton images, IEEE J. Ocean. Eng. 40 (2) (2015) 315–326, https://doi.org/10.1109/JOE.2014.2317955.
[37] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2014, pp. 580–587.
[38] R. Girshick, Fast R-CNN, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440–1448.
[39] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.
[40] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: single shot multibox detector, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 21–37.
[41] K.T. Le, Z. Yuan, A. Syed, D. Ratelle, E.C. Orenstein, M.L. Carter, S. Strang, K.M. Kenitz, P. Morgado, P.J.S. Franks, N. Vasconcelos, J.S. Jaffe, Benchmarking and automating the image recognition capability of an in situ plankton imaging system, Front. Mar. Sci. 9 (2022), https://doi.org/10.3389/fmars.2022.869088.
[42] K. Hacıefendioğlu, H.B. Başağa, O.T. Baki, A. Bayram, Deep learning-driven automatic detection of mucilage event in the Sea of Marmara, Turkey, Neural Comput. Appl. 35 (9) (2023) 7063–7079, https://doi.org/10.1007/s00521-022-08097-1.
[43] M. Yang, W. Wang, Q. Gao, C. Zhao, C. Li, X. Yang, J. Li, X. Li, J. Cui, L. Zhang, Y. Ji, S. Geng, Automatic identification of harmful algae based on multiple convolutional neural networks and transfer learning, Environ. Sci. Pollut. Res. 30 (6) (2023) 15311–15324, https://doi.org/10.1007/s11356-022-23280-6.
[44] S.-S. Baek, J. Pyo, Y. Pachepsky, Y. Park, M. Ligaray, C.-Y. Ahn, Y.-H. Kim, J. Ahn Chun, K. Hwa Cho, Identification and enumeration of cyanobacteria species using a deep neural network, Ecol. Indic. 115 (2020) 106395, https://doi.org/10.1016/j.ecolind.2020.106395.
[45] P. Qian, Z. Zhao, H. Liu, Y. Wang, Y. Peng, S. Hu, J. Zhang, Y. Deng, Z. Zeng, Multi-target deep learning for algal detection and classification, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), 2020, pp. 1954–1957.
[46] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Netw. 5 (2) (1994) 157–166, https://doi.org/10.1109/72.279181.
[47] R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks, in: S. Dasgupta, D. McAllester (Eds.), Proceedings of the 30th International Conference on Machine Learning, PMLR, Atlanta, Georgia, USA, in: Proceedings of Machine Learning Research, vol. 28, 2013, pp. 1310–1318 [cited 21/04/2023], http://proceedings.mlr.press/v28/pascanu13.
[48] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936–944.
[49] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755.
[50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[52] S. Ruder, An overview of gradient descent optimization algorithms, arXiv:1609.04747, 2017.
[53] I. Sutskever, J. Martens, G. Dahl, G. Hinton, On the importance of initialization and momentum in deep learning, in: S. Dasgupta, D. McAllester (Eds.), Proceedings of the 30th International Conference on Machine Learning, PMLR, Atlanta, Georgia, USA, in: Proceedings of Machine Learning Research, vol. 28, 2013, pp. 1139–1147 [cited 21/04/2023], https://proceedings.mlr.press/v28/sutskever13.html.
[54] A. Krogh, J.A. Hertz, A simple weight decay can improve generalization, in: NIPS, 1991.
[55] N. Ismoilov, S.-B. Jang, A comparison of regularization techniques in deep neural networks, Symmetry 10 (2018) 648, https://doi.org/10.3390/sym10110648.