LIMNOLOGY
and
OCEANOGRAPHY: METHODS

Limnol. Oceanogr.: Methods 2018

doi: 10.1002/lom3.10285

Automated plankton image analysis using convolutional neural networks

Jessica Y. Luo,[1,2]*[a] Jean-Olivier Irisson,[3] Benjamin Graham,[4] Cedric Guigand,[1] Amin Sarafraz,[5] Christopher Mader,[5] Robert K. Cowen[1,2]

[1]Marine Biology and Fisheries, Rosenstiel School of Marine and Atmospheric Sciences, University of Miami, Miami, Florida
[2]Hatfield Marine Science Center, Oregon State University, Newport, Oregon
[3]Sorbonne Université, CNRS, Laboratoire d'Océanographie de Villefranche, LOV, F-06230 Villefranche-sur-Mer, France
[4]Department of Statistics, University of Warwick, Coventry, United Kingdom
[5]Center for Computational Science, University of Miami, Coral Gables, Florida

Abstract

The rise of in situ plankton imaging systems, particularly high-volume imagers such as the In Situ Ichthyoplankton Imaging System, has increased the need for fast processing and accurate classification tools that can identify a high diversity of organisms and nonliving particles of biological origin. Previous methods for automated classification have yielded moderate results that either can resolve few groups at high accuracy or many groups at relatively low accuracy. However, with the advent of new deep learning tools such as convolutional neural networks (CNNs), the automated identification of plankton images can be vastly improved. Here, we describe an image processing procedure that includes preprocessing, segmentation, classification, and postprocessing for the accurate identification of 108 classes of plankton using spatially sparse CNNs. Following a filtering process to remove images with low classification scores, a fully random evaluation of the classification showed that average precision was 84% and recall was 40% for all groups. Reliably classifying rare biological classes was difficult, so after excluding the 12 rarest taxa, classification accuracy for the remaining biological groups became > 90%. This method provides proof of concept for the effectiveness of an automated classification scheme using deep-learning methods, which can be applied to a range of plankton or biological imaging systems, with the eventual application in a variety of ecological monitoring and fisheries management contexts.

Much of plankton ecology has been focused upon questions surrounding the identity, quantity, and spatial-temporal variability of planktonic organisms in aquatic systems, which has historically been addressed by various net-based sampling systems (Wiebe and Benfield 2003). Although many advanced nets were designed to overcome limitations in horizontal or vertical resolution in sampling, the emergence of plankton imaging systems represents a significant advancement for plankton ecology. Current imaging systems are able to quantify organisms within fine spatial and temporal scales, with some systems imaging organisms undisturbed and in their natural environment [e.g., Video Plankton Recorder (VPR; Davis et al. 1992), Shadow Image Particle Profiling Evaluation Recorder (SIPPER; Samson et al. 2001), ZOOplankton VISualization and Imaging System (ZOOVIS; Benfield et al. 2003), In Situ Ichthyoplankton Imaging System (ISIIS; Cowen and Guigand 2008), Underwater Vision Profiler 5 (UVP5; Picheral et al. 2010); Note that the SIPPER samples via an intake tube, and is thus not a truly undisturbed sampler]. Research using plankton imaging systems has led to new insights into the relationships between species and their fine-scale environment (e.g., Benfield et al. 2000; Ashjian et al. 2001), with implications ranging from fine-scale aggregation dynamics (Luo et al. 2014), N_2 fixation (Davis and McGillicuddy 2006), predator-prey interactions (Greer et al. 2013), carbon export (Petrik et al. 2013), and global plankton biomass estimates (Biard et al. 2016).

Though in situ plankton imaging systems were developed with a goal of reducing processing time (very lengthy for physical net samples, which requires sorting and expert identification), in reality, analyzing plankton images currently still requires extensive and time-consuming manual classification and expert taxonomic knowledge. The tradeoff is: human operators' time vs. classification accuracy vs. taxonomic

*Correspondence: jluo@ucar.edu

[a]Present address: National Center for Atmospheric Research, Boulder, Colorado

Additional Supporting Information may be found in the online version of this article.

resolution. Manual processing time is typically not reported in papers, but as an example, the manual analysis of 50+ taxa of gelatinous zooplankton within 5500 m^3 of water, imaged in 750,000 frames (13.5 inch square frames with 66 μm pixel resolution; each frame with up to 50 organisms) required the equivalent of three full man-years (Luo et al. 2014). Classification of preprocessed image segments is slightly faster; Faillettaz et al. (2016) reported a manual classification rate of 10,000 images d^{-1} into 10–15 biotic and abiotic classes, but a single multiday cruise can easily generate upward of 50 million image segments. In general, automated classification efforts currently consist of identifying small numbers of classes [five to seven classes (Davis et al. 2004; Hu and Davis 2006), and three classes (Bi et al. 2015)], but even so, few reach an acceptable benchmark of classification accuracy, commonly set at 67–83% (Culverhouse et al. 2003; Hu and Davis 2005). Alternatively, computer-assisted classification is generally used to achieve higher accuracies, which consists of a computer generated set of automated classifications followed by fully validating all images manually (Gorsky et al. 2010; Ohman et al. 2012). Consequently, it is still very difficult and time-consuming to extract high-accuracy data on many types of plankton, particularly in highly diverse areas, which limits the utility of many underwater imaging systems.

The issues with classification accuracy and speed have been pronounced with ISIIS (Cowen and Guigand 2008), which is a high resolution, large volume imager designed for sampling mesozooplankton. It typically images at a rate of 150–185 L s^{-1}, depending on tow speed. Compared with other plankton imaging systems (VPR: 10–17 mL s^{-1}, ZOOVIS: 3.6 L s^{-1}, SIPPER: 9.2 L s^{-1}, UVP5: 8–20 L s^{-1}), ISIIS records at 10–1000 times the sampling volume, which has allowed for studies on rare organisms such as larval fish (Cowen et al. 2013) or large gelatinous zooplankton (McClatchie et al. 2012; Luo et al. 2014). Particularly in subtropical zones such as the northern Gulf of Mexico, ISIIS can simultaneously record in-focus, clear images of hundreds of species, ranging from protists, diatom chains, and copepods, to shrimps, larval fish, and medusae. Therefore, in order to properly classify organisms within ISIIS images, a classifier that could handle not just a few (< 10) classes, but many classes (e.g., 30–150) is necessary.

Methods for the automated analysis of zooplankton images had early beginnings in statistical approaches, e.g., discriminant analysis (Jeffries et al. 1980, 1984), but quickly progressed to using machine learning techniques such as artificial neural networks (ANNs; Simpson et al. 1992; Culverhouse et al. 1996). Culverhouse et al. (1996) designed their ANN system explicitly for dinoflagellates, and were able to identify species with ca. 72% accuracy, which was comparable to human classification (Culverhouse et al. 2003). For a slightly broader range of taxonomic classes (five to seven classes of phytoplankton and zooplankton), an ANN-type network was combined with a support vector machine (SVM) in a dual classification method for VPR images, resulting in classification precision rates between 23% and 95% when tested on the original training set (Hu and Davis 2005, 2006). For images from the UVP5, an extensive review of different classifiers (including ANN and SVM classifiers) resulted in the adoption of a Random Forest (RF) algorithm, which consistently performed the best, even superseding the SVM classifier (ZooProcess with PkID; Gorsky et al. 2010; Gasparini and Antajan 2013). However, even with the success of the RF algorithm, most UVP5 images are still fully validated by human operators, though there have been some recent efforts toward decreasing the amount of manual labor required through the use of filtering methods (Faillettaz et al. 2016). In recent years, SVMs have continued to be used, for systems such as the SIPPER (active learning with an SVM reduces human labeling efforts; Luo et al. 2005) and ZOOVIS (SVM classifier using three classes with > 80% precision; Bi et al. 2015), while other groups have continued on with RF methods, sometimes with many more classes (47 classes, Laney and Sosik 2014; 114 classes, Schmid et al. 2016). Nonetheless, for all of these classification algorithms, a highly specific set of premeasured features was crucial for successfully training the classifier; this set of extracted features could not dynamically change, nor be automatically determined by the classifier itself. Furthermore, while the accepted benchmark for plankton classification accuracy (67–83%, Culverhouse et al. 2003) had been met in many classes by different classifiers, for biological questions particularly surrounding rare or cryptic species, a high amount of error is often untenable.

Here, we present the results of a process to develop an automated classification algorithm for ISIIS images using convolutional neural networks (CNNs), a relatively new class of methods that has revolutionized the computer vision field in recent years (Krizhevsky et al. 2012; LeCun et al. 2015), and which falls within the general category known as *deep learning*. As opposed to conventional machine-learning techniques such as ANNs, RF, and SVMs, deep-learning tools do not require extensive domain expertise (e.g., plankton imaging) and the careful engineering of feature extractors for classification. Instead, they are able to process natural data in their raw form, and automatically discover the representations that are best suited for classification. We describe a whole image processing "pipeline," which includes preprocessing, segmentation, classification, and postprocessing (Fig. 2). Then, as a proof of concept, we apply it to a set of ISIIS images collected in the northern Gulf of Mexico. While the described method is highly tuned to images collected by a particular instrument, CNNs in general (as well as the machine learning competition we ran to generate this solution) are highly versatile, and can be applied to many types of images within the biological sciences.

Methods

Description of instrument

ISIIS (Cowen and Guigand 2008) utilizes shadowgraph imaging with a line-scan camera to capture silhouette images of particles in a sampled parcel of water. This backlighting technique, with early application by Arnold and Nuttall-Smith (1974) and Ortner et al. (1979, 1981), allows for the fine taxonomic resolution of transparent organisms (e.g., gelatinous zooplankton) and the coarse taxonomic resolution of small, opaque organisms (e.g., copepods). The camera used is a 2048-pixel line-scan camera that images over a 13 × 13-cm field of view and 50-cm depth of field, with a resultant 66-μm pixel resolution. The output of the imaging is recorded as a continuous image that is parsed into square frames (2048 × 2048 pixels) at 17 frames s^{-1}. While sampling, we target a ship speed of 2.5 m s^{-1}, which results in an ISIIS sampling rate of 169 L s^{-1}. However, in practice, this sampling rate can vary from 150 L s^{-1} to 185 L s^{-1} with corresponding ship speeds of 2.25–2.75 m s^{-1}. The recorded data are ported to the surface via a fiber-optic wire, time-stamped, and saved onto a ship-based computer or raid array.

Field sampling

During July–August 2011, ISIIS was deployed over eight, 6-h transects during two oceanographic cruises onboard the NOAA ship *McArthur II* in the northern Gulf of Mexico (Fig. 1). The sampling plan was designed to capture images of species present during the day and night, at various locations and depths, and over multiple months. ISIIS sampled in tow-yo undulations from the surface to 130 m depth at the offshore sites, and from the surface to 40–60 m depth at the inshore sites.

Image preprocessing

ISIIS uses a line-scanning camera with a single row of pixels, each with its own unique light sensitivity characteristics; consequently, raw ISIIS images have a slight nonuniformity in gray-level across the image, despite the uniform distribution of incoming light. Furthermore, any dust or particles on the lens appears as vertical lines in the raw, square frame (Fig. 3a). These lines and image nonuniformities are corrected in a radiometric calibration called "flat-fielding" in which we calculate a calibration frame (Fig. 3b) that is subtracted from the raw frame. The calibration is calculated per frame; since the objects of interest occupied only a small amount of the frame (based on initial tests, we assumed it to be < 20%), we ignored those outliers and calculated a column-averaged frame for calibration. Thus, the resultant frame after flat-fielding is devoid of vertical nonuniformities that could bias the segmentation and classification (Fig. 3c).

Next, in order to equalize the image histogram, we normalized the contrast within each frame using the OpenCV 2.4 "equalizeHist" command (https://docs.opencv.org/2.4/modules/imgproc/doc/histograms.html). The histogram normalized frame allowed for the better detection of regions of interest (ROIs) for segmentation (Supporting Information Fig. S1).

Finally, due to the fact that sampling included coastal waters with high turbidity (from the Mississippi River plume), we calculated a signal-to-noise (SNR) ratio for each frame in order to filter out the highly noisy frames captured in turbid waters. The SNR was computed by first calculating a

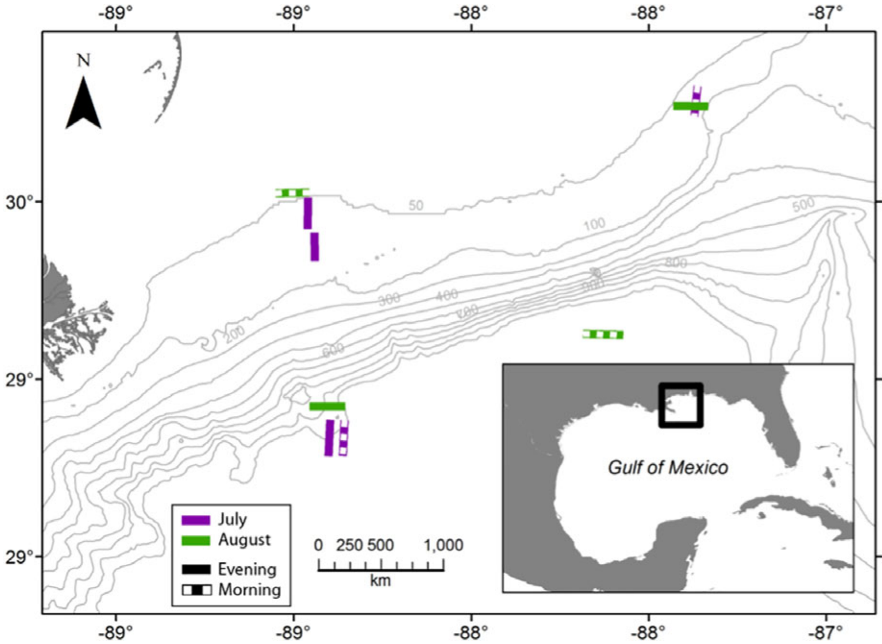

Fig. 1. Sampling sites in the northern Gulf of Mexico, spanning 2 months (July and August 2011), in nearshore and offshore sites, occurring during evening (solid lines) or morning (dashed lines) times. Each transect was sampled over 6 h, and consisted of tow-yo undulations, from the surface to a maximum of 50 m for the inshore sites and 130 m for the offshore sites.

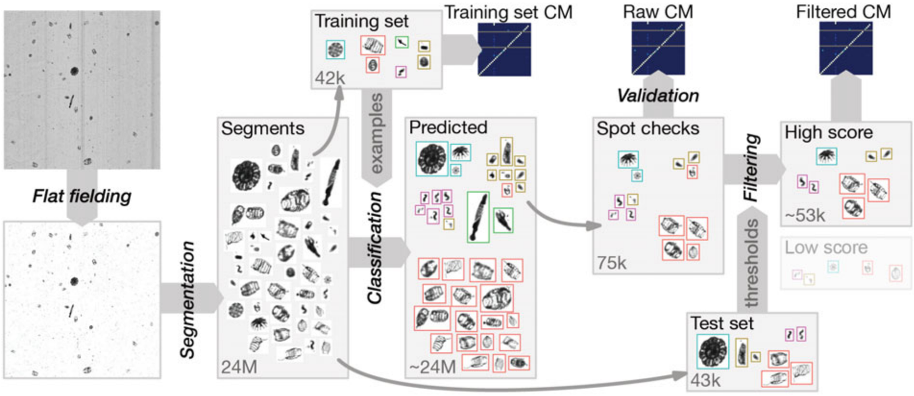

Fig. 2. Flowchart overview of the processing steps. Raw frames were first flat-fielded and corrected, then segmented into smaller image segments. A training set was generated to train the image classifier, which was a convolutional neural net (CNN). The full dataset of ~ 24M images were classified. Afterward, a random subset of 75k images were spot-checked (manually validated) to estimate the accuracy of the classifier. Separately, a 43k-test set was also validated and used to set probability thresholds, which separated the classified dataset into low-probability (discarded into "unknown") and high-probability images (retained). CMs were generated to evaluate classifier performance at each step.

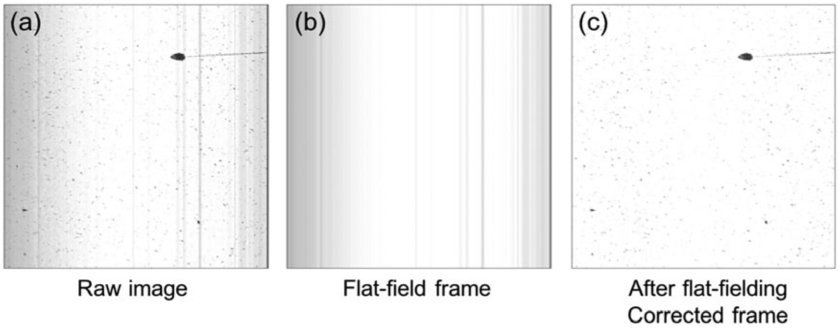

Fig. 3. An example of the flat-fielding process, showing the raw image (**a**), the flat-field frame that was removed from the raw image (**b**), and the corrected frame (**c**).

cleaned-up frame, or the "signal-frame," which was simply done by applying a 3 × 3 median filter to the histogram normalized frame. The difference between the histogram normalized frame and the signal frame was then considered the "noise-frame." SNR was then calculated as the log of the ratio between vector norm (l^2-norm) values of the two images:

$$\mathrm{SNR} = 20\log_{10}\left(\frac{|F_{\mathrm{signal}}|}{|F_{\mathrm{noise}}|}\right) \tag{1}$$

where F_{signal} is the signal-frame and F_{noise} is the noise-frame; and the vector norm values was calculated using the OpenCV 2.4 "norm" function, and is defined as:

$$|\mathbb{X}| = \sqrt{\sum\nolimits_{k=1}^{n} |x_k|^2}, \mathrm{where}\, \mathbb{X} = [x_1, x_2, x_3, \ldots, x_n] \tag{2}$$

After calculating the SNR on a set of representative images, we found a clear difference in SNR values between frames captured from turbid vs. not turbid waters (Supporting Information Fig. S2). Thus, we used a threshold cutoff of SNR = 25 to discard extremely noisy images, which were approximately 26% of all frames originally captured. Note that through earlier efforts, we have found that images from highly turbid waters often require manual identification of images, so we sought to limit this study to images captured from more typically oceanic waters. This exclusion of noisy images should be considered an effect of the sampling environment, rather than the image processing method, as images collected in oceanic waters rarely had high SNRs.

Segmentation

Preprocessed frames were then segmented using the ISIIS image segmentation software (Tsechpenakis et al. 2007, 2008; Iyer 2012; http://cs.iupui.edu/~gavriil/vital/MVISIIS). The segmentation software uses an unsupervised machine learning technique, K-harmonic means clustering, to detect ROIs from the raw images. Iyer (2012) tested six different clustering methods for segmentation (K-means, iterative K-means, fuzzy C-means, Isodata, Spectral algorithm, and K-harmonic means), and chose the K-harmonic means method because it achieved the highest accuracy rates (95%) at relatively fast speeds and was easily implemented for parallel processing. In our implementation, we found that the segmentation process was further improved after the addition of the image histogram equalization step (*see* "Image preprocessing" section). Finally, the segmented images were given a unique name that refers to its time-stamp and location within the original frame. This naming convention allows for each image to be quickly associated with neighboring images, the original frame, shipboard GPS, and the environmental data recorded by the instrument.

Automated image classification

Convolutional neural networks

Segmented images were classified using convolutional (or deep) neural networks (CNNs), a method that is able to process images directly and automatically discover the characteristics within the images that are best suited for classification. Deep-neural networks make use of the fact that natural images can be analyzed in a hierarchical fashion, with lower-level features organizing to form higher-level features (e.g., pixels to edges, edges to body parts, and body parts into organisms), and have been used in numerous applications from speech and face recognition (Lawrence et al. 1997; Hinton et al. 2012) to predictions of galaxy morphology (Dieleman et al. 2015). The four key ideas that characterize CNNs (local connections, shared weights, pooling, and the use of many layers) facilitate minimal preprocessing and require no prior knowledge in designing features for classification; this represents a significant advance compared with traditional machine learning methods such as ANNs and SVMs (LeCun et al. 2015).

Spatially sparse convolutional neural networks (SparseConvNets) were initially designed for the recognition of Chinese handwriting. SparseConvNets recognize that the background of an image often occupies many pixels and not processing them allows the CNN approach to be applied more efficiently, with less computational cost (Graham 2014). Plankton images can also be considered "sparse images," as the majority of the image, even in segmented images, is background. The actual particle or organism occupies a relatively small percentage of the pixels in the image. Thus, not processing the background (white) pixels results in a much faster classification process.

An application of SparseConvNets with Fractional Max-Pooling (Graham 2015) was initially developed as part of the Poisson Process team for the 2015 National Data Science Bowl competition (3-month machine learning competition to classify ca. 60,000 ISIIS plankton images within 121 categories; dataset available at Cowen et al. 2015, see competition solution at: www.kaggle.com/c/datasciencebowl/forums/t/13158/poisson-process-competition-report-and-code/). For the competition, a number of similar models were used to generate an ensemble solution. We chose the best single model from team Poisson Process and made small modifications to improve overall speed with little apparent changes in accuracy.

As opposed to ANNs, which process images as vectors, CNNs process images as three-dimensional arrays. Images are represented in a computer as three-dimensional arrays with size $N \times N \times C$, where the first two dimensions ($N \times N$) are spatial dimensions, representing the number of pixels in the image, and the last dimension (C) is the number of color-channels. In our case, the input image has dimensions of $N \times N \times 1$, as there is only one color channel in monochrome images (RGB color images have $C = 3$). However, the C dimension does not necessarily have to represent only true colors, but rather can be generalized and expanded to represent abstract "features" of an image. On a basic level, convolutional networks work by going through an iterative process of collecting features and appending them as two-dimensional slices to the C dimension. These additional abstract color-channels are "value-added" images, as they represent increasingly higher-level features, as the algorithm progresses from the bottom of the network to the top. Examples of features that would be detected at the bottom of the network include edges or combinations of edges, and at the top of the network, these features would be something biologically relevant, such as tails or antennae. These collections of features are constructed by a numerical optimization technique which involves iteratively showing the network training images from which it can learn discriminative features useful for classification.

Our network is constructed as a sequence of two alternating types of layers, termed convolutional and pooling layers. Convolutional layers form the main building block for CNNs, as they detect local combinations of features. Pooling layers operate by merging semantically similar features into one (for a general description, please see LeCun et al. 2015). The network has 13 convolutional layers in total, separated by 12 pooling layers. The n-th convolutional layer looks at overlapping 2×2 pixel regions of the image below, producing an output image with $32*n$ color-channels. The number of color-channels increases as we rise through the network in the expectation that we will produce an increasingly rich description of the contents of the image. The interleaved pooling layers reduce the spatial size of the input image, but leave the number of color-channels unchanged. We use fractional max-pooling with scaling factor of $1/\sqrt{2}$. The scaling is multiplicative, so the image shrinks exponentially as we climb the network. This reduction in resolution offsets the increase in the number of color-channels, ensuring computational feasibility. After the last convolutional layer, we calculate the average of each color-channel over the spatial dimensions. We then perform multinomial logistic regression on the set of color-channel features to predict the class of the image. We used the SparseConvNet (https://github.com/btgraham/

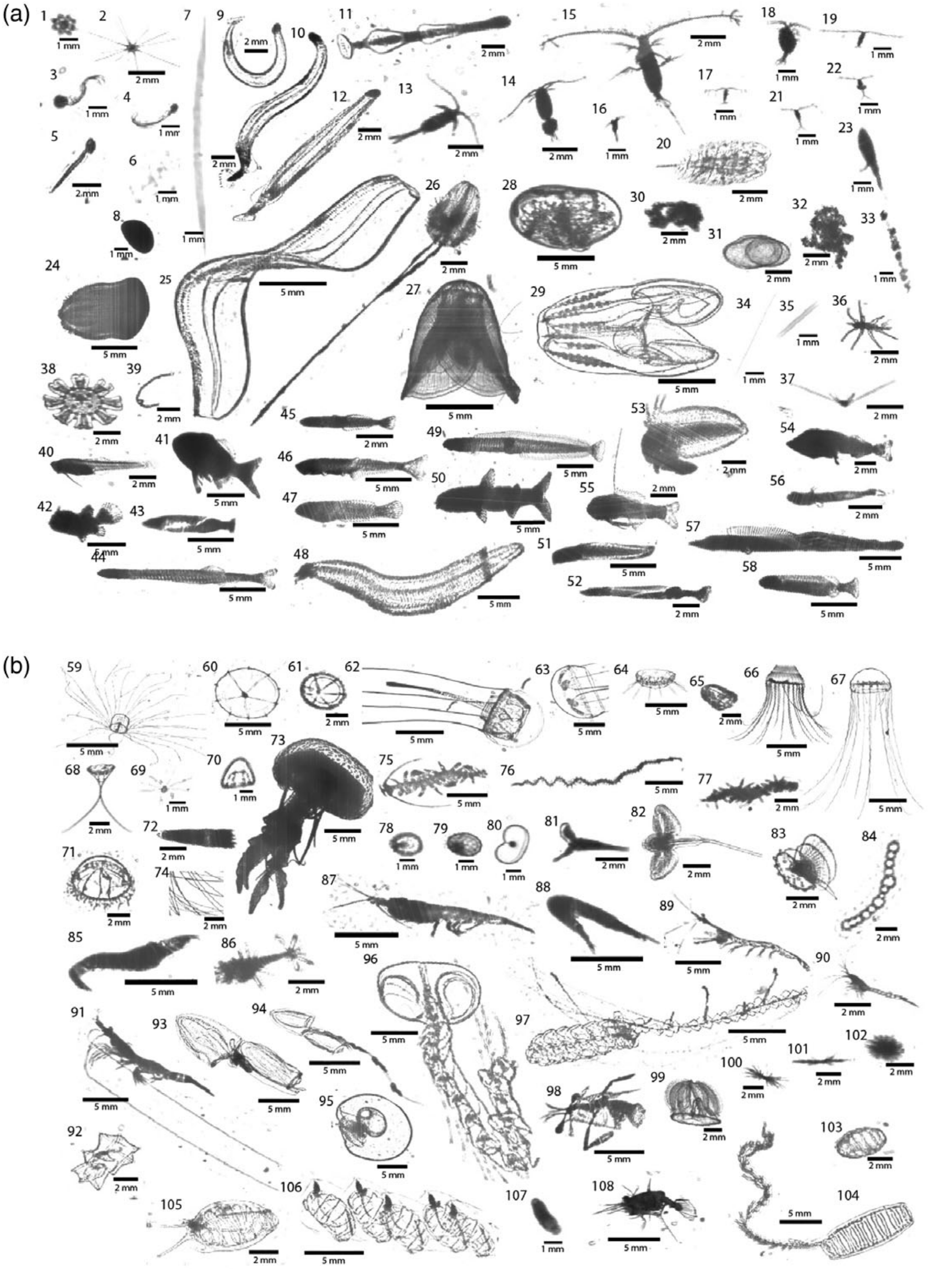

Fig. 4. (**a**, **b**) Example images from each of the 108 classes within the learning set. Classes [*and corresponding groups*] are: (1) acantharia protist 1, (2) acantharia protist 2 [*protist*]; (3) appendicularian sinusoidal tail, (4) appendicularian slight curve, (5) appendicularian straight [*appendicularian*]; (6) artifact 1, (7) artifact 2, (8) bubbles [*artifact*]; (9) chaetognath c-curved, (10) chaetognath curved, (11) chaetognath dark, (12) chaetognath straight [*chaetognath*]; (13) copepod calanoid, (14) copepod calanoid eggs, (15) copepod calanoid eucalanus, (16) copepod calanoid flatheads, (17) copepod calanoid frilly antennae, (18) copepod calanoid large, (19) copepod calanoid small long-antennae [*copepod calanoid*]; (20) copepod cyclopoid copilia [*copepod copilia*]; (21) copepod cyclopoid oithona, (22) copepod cyclopoid oithona eggs [*copepod oithona*]; (23) copepod escape [*copepod calanoid*]; (24) ctenophore beroida [*ctenophore beroida*]; (25) ctenophore cestida [*ctenophore cestida*]; (26) ctenophore cydippid [*ctenophore cydippid*]; (27) ctenophore lobata mnemiopsis, (28) ctenophore lobata ocyropsis, (29) ctenophore lobata type 1 [*ctenophore lobata*]; (30) detritus sparse blob, (31) detritus casings, (32) detritus dark, (33) detritus filamentous [*detritus*]; (34) diatom chain string, (35) diatom chain tube [*diatom chain*]; (36) echinoderm

SparseConvNet) software package, which takes advantage of the sparsity of the images to reduce the computational burden, to train the CNN.

Network training

Images (n=42,564) were manually sorted in 108 classes to serve as a training set. We used 100+ classes to accurately represent the taxonomic diversity in the data (Fig. 4). Initially, we started with a training set that was a subset of the data (not shown), which represented the actual proportions of objects in each class, but refined the training set by adding in rare classes. Since we were most interested in rarer groups (e.g., larval fish, jellies, etc.), they were inflated to provide a greater number of representative samples for the training set. The total number of images in each class of the training set is provided in Supporting Information Table S1.

In addition, at each "epoch" (i.e., training cycle), SparseConvNet picks examples from the training set and performs data augmentation (randomly rotates, skews, and scales each image), hence creating subtle variations of the original shapes and simulating new training examples. This procedure is fairly common in CNNs and helps to generalize a model based on a limited set of examples.

We trained for 150 epochs, as this represented the point at which the error rate plateaued at a minimum value (14.9–15.1%). Using a g2.2xlarge instance on the Amazon elastic computing cloud (one NVIDIA GPU with 1536 CUDA cores), training 150 epochs took ca. 24 h.

Model predictions

To make the prediction robust, each of the 23.4M images in the full data set was passed through the fitted network 24 times, each time with different data augmentation parameters. The probabilities for each object to belong to each class, predicted by the model, were averaged over the 24 predictions, the maximum was found, and the corresponding class was considered as the predicted class. Total prediction time was 165 machine hours (though < 36 actual hours, as we used five GPU instances in parallel).

Classification groupings

The 108 original classes in the training set were mapped onto 37 broader groups, which represented taxonomic or functional groupings that were more relevant for ecological analyses (Fig. 4; Supporting Information Table S2). For example, many of the original classes were created for automated image classification purposes, with the distinctions between classes only morphological (e.g., straight vs. curved appendicularians) or due to an imaging or segmentation artifact (e.g., cropped bells and tentacles). Others were created to distinguish between different forms within a diverse class (e.g., detritus) that would otherwise pollute many other classes. Last, some taxonomic classes were grouped together for filtering purposes and further analyses (e.g., the fish groups).

Model performance

Confusion matrices (CMs) are a tool for quantifying classifier accuracy (e.g., Hu and Davis 2005; Bi et al. 2015). The calculated CM statistics included three values and three rates, calculated separately for each class *i*: values were numbers of true positives (TP_i), false positives (FP_i, type I error), and false negatives (FN_i, type II error). The rates calculated were precision (P_i, Eq. 3), recall (R_i, Eq. 4), and the F1-score, which is the harmonic mean of the precision and recall rates ($F1_i$, Eq. 5). For a given class, precision quantifies the "purity" of the prediction and recall quantifies the "completeness" of the prediction.

$$\mathrm{P_i} = \mathrm{TP_i}/(\mathrm{TP_i} + \mathrm{FP_i}) \tag{3}$$

$$\mathrm{R_i} = \mathrm{TP_i}/(\mathrm{TP_i} + \mathrm{FN_i}) \tag{4}$$

$$\mathrm{Fl_i} = 2 \times \mathrm{P_i} \times \mathrm{R_i}/(\mathrm{P_i} + \mathrm{R_i}) \tag{5}$$

A self-prediction of the training set and the associated CM represent the theoretical maximum of the classifier performance. It was computed to give a benchmark for determining which classes had naturally high variability and which ones were relatively homogenous (Supporting Information Table S1).

brachiolaria, (37) echinoderm pluteus [*echinoderm*]; (38) ephyra [*ephyra*]; (39) fecal pellets [*detritus*]; (40) fish bregmacerotidae, (41) fish carangidae, (42) fish ceratioidei, (43) fish echeneidae, (44) fish engraulidae, (45) fish gobiidae, (46) fish gonostomatidae, (47) fish labroidei, (48) fish leptocephali, (49) fish microdesmidae, (50) fish myctophidae, (51) fish ophidiidae, (52) fish phosichthyidae, (53) fish pleuronectiformes, (54) fish scombridae, (55) fish serranidae, (56) fish synodontidae, (57) fish trichiuridae, (58) fish xyrichtys [*fish*]; (59) hydromedusae rhopalonematidae [*hydro rhopalonematidae*], (60) hydromedusae eucheilota, (61) hydromedusae haliscera [*hydro other*]; (62) hydromedusa liriope tetraphylla, (63) hydromedusa liriope cut-off-bell [*hydro liriope*]; (64) hydromedusae narcomedusae other [*hydro narcomedusae*]; (65) hydromedusae rhopalonema 2 [*hydro rhopalonematidae*], (66) hydromedusae solmaris rhodoloma, (67) hydromedusae solmaris spp, (68) hydromedusae solmundella, (69) hydromedusae tiny solmaris [*hydro narcomedusae*]; (70) hydromedusae type 1 small bell, (71) hydromedusae type 2, (72) hydromedusae type 3 [*hydro other*]; (73) medusa pelagia noctiluca [*medusa pelagia*]; (74) medusa tentacles [*hydro other*]; (75) polychaete type 1, (76) polychaete type 2, (77) polychaete type 3 [*polychaete worm*]; (78) protist noctiluca, (79) protist radiolarian, (80) protist radiolarian clear [*protist*]; (81) pteropod type 1, (82) pteropod type 2, (83) pteropod type 3 [*pteropod*]; (84) radiolarian chain [*radiolarian chain*]; (85) shrimp caridean, (86) shrimp caridean small [*shrimp other*]; (87) shrimp euphausiid, (88) shrimp euphausiid escape posture [*shrimp euphausiid*]; (89) shrimp lucifer [*shrimp lucifer*]; (90) shrimp mysid [*shrimp other*]; (91) shrimp sergestidae [*shrimp sergestidae*]; (92) siphonophore ablyidae [*siph ablyidae*]; (93) siphonophore calycophoran pointy head no-stem, (94) siphonophore calycophoran pointy head with-stem, (95) siphonophore calycophoran round head [*siph calycophoran*]; (96) siphonophore lilyopsis rosea [*siph lilyopsis rosea*]; (97) siphonophore physonect [*siph physonect*]; (98) stomatopods [*stomatopods*]; (99) tornaria acorn worm larvae [*tornaria*]; (100) trichodesmium bow-tie, (101) trichodesmium tuft, (102) trichodesmium puff [*trichdodesmium*]; (103) tunicate doliolid, (104) tunicate doliolid budding, (105) tunicate salp, (106) tunicate salp chains [*tunicates*]; (107) unknown dark blob [*detritus*]; (108) zoea [*zoea*].

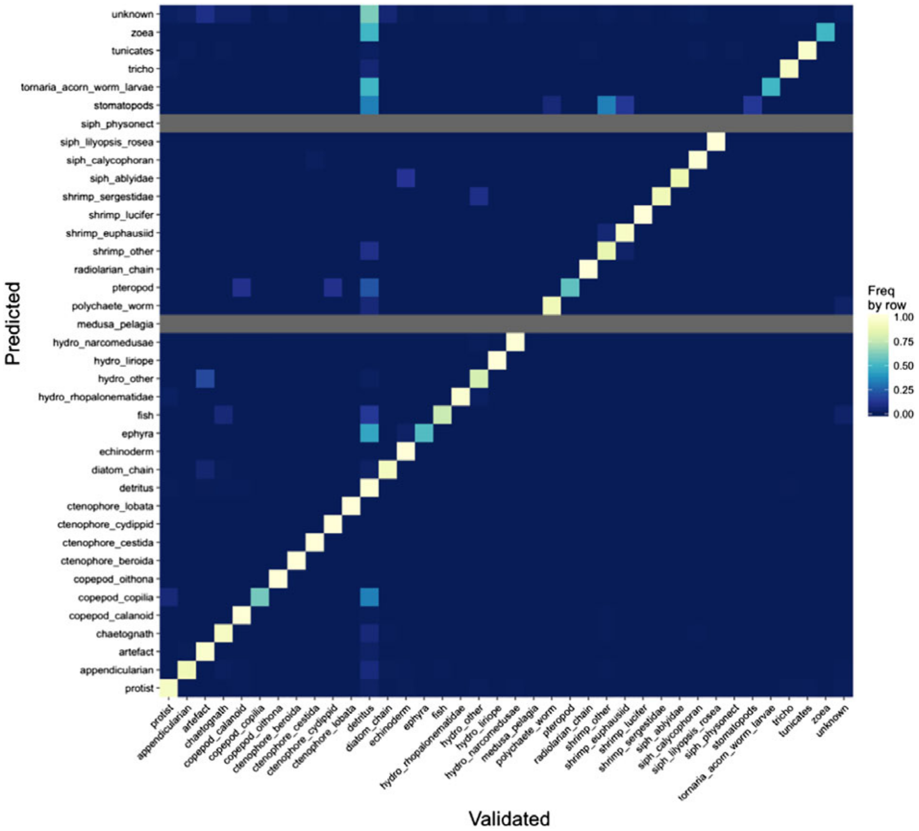

Fig. 5. Confusion matrix on the 75,000 random images, classified into 108 classes, and then grouped into 38 groups (including unknowns). Low-probability images were moved into Unknowns. Rows show computer-predicted classes, and columns show human-validated classes. Color indicates proportion of images sorted from computer-predicted classes into manually verified classes, scaled by row. Gray rows indicate rare categories where no (high probability) images were randomly selected for validation.

To evaluate the classification success on the full dataset, we performed spot-checks: 75,000 predicted images were picked randomly (0.30% of the total dataset) and their identification was manually validated. The corresponding CM is shown in Supporting Information Table S2 (see column "Without probability filtering").

Probability filtering

In the full dataset, images of organisms spanned the range in terms of quality: small (e.g., early life stage) to large (e.g., adult), blurry to clear, oriented toward, away, or to the side of the camera, etc. Thus, the images that were more difficult to predict, or less archetypal often were associated with a low prediction score. The prediction score is an output of any classification algorithm: for each candidate image, the algorithm computes a score (often a probability) associated with every category in the training set. Classification is then just a matter of picking the maximum score. However, for difficult to identify objects that could fit in many classes, even the maximum score can be low, reflecting a low confidence in the classification. Therefore, we used this score to filter classified images into "high" vs. "low" likelihood of correct classification using a threshold value set for each class. Faillettaz et al. (2016) first demonstrated this approach, showing that the removal of "low-confidence images" (in their case, over 70% of their original dataset) still allowed for the prediction of true spatial distributions of many taxa.

In the present study, we determined the appropriate threshold values for each class by predicting a new, independent, 43,000-member test set. All images in this test set were manually identified, which allowed us to detect prediction errors. For each class, we set the threshold value to be the classification score *above which* 95% of images were correctly classified into the corresponding *group* (as opposed to the *class* itself). Groups were used because many of the 108 classes were separated based on morphological distinctions with little ecological relevance (e.g., "chaetognaths *curved*" vs "chaetognaths *straight*"). The 95% level, which resulted in 29.6% of images discarded (though individual classes varied, Supporting Information Table S3), was chosen as a compromise between improving classification accuracy and retaining enough images for ecological analyses. As a comparison, if the thresholds were set at the 90% level, then only 19% of images would

Table 1. Comparison of precision, recall, and F1-ratio for the prediction of the training set compared to the full dataset (from the random 75,000 spot-checks, after applying the filtering thresholds), calculated at the group level. Gray rows indicate rare groups with < 25 images in the spot-checked set.

	Training set			Full dataset		
Class	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
appendicularian	97.5	99.2	98.3	90	39.1	54.5
artifact	94.6	96.8	95.7	96.5	90.5	93.4
chaetognath	98.7	99	98.8	92.7	46.9	62.3
copepod_calanoid	98.3	99.2	98.7	98.8	62.4	76.5
copepod_copilia	97.7	97.7	97.7	60	75	66.7
copepod_oithona	97.5	99.8	98.6	100	57.6	73.1
ctenophore_beroida	98.8	89.4	93.9	100	20	33.3
ctenophore_cestida	100	99.3	99.6	100	33.3	50
ctenophore_cydippid	100	83.3	90.9	100	3.1	6
ctenophore_lobata	99.3	98.7	99	100	14.3	25
detritus	97.4	92.6	94.9	98.2	55.3	70.8
diatom_chain	97.6	98	97.8	92.3	78.6	84.9
echinoderm	98.2	98.6	98.4	100	16.5	28.3
ephyra	100	100	100	52.8	47.5	50
fish	99.5	99.8	99.6	76.3	38.5	51.2
hydro_liriope	93.5	96.1	94.8	100	21.4	35.3
hydro_narcomedusae	97.4	95.5	96.4	98.8	23	37.3
hydro_other	88.8	92.3	90.5	79.7	15	25.2
hydro_rhopalonematidae	95.4	97	96.2	96.7	33.1	49.3
medusa_pelagia	98.8	95.3	97	NA	0	NA
polychaete_worm	99	99	99	90	45.8	60.7
protist	99.1	99.6	99.3	93.9	53.7	68.3
pteropod	99	94.6	96.8	55.6	14.7	23.3
radiolarian_chain	100	100	100	100	77.8	87.5
shrimp_euphausiid	95.8	98.8	97.3	94.1	69.6	80
shrimp_lucifer	93.1	94.7	93.9	100	32.5	49.1
shrimp_other	97.4	91.7	94.5	86.2	12.3	21.5
shrimp_sergestidae	92.1	89.4	90.7	90	50	64.3
siph_ablyidae	96.7	97.8	97.2	87.5	21.9	35
siph_calycophoran	95.9	95.6	95.7	98.6	24.6	39.4
siph_lilyopsis_rosea	92	89.6	90.8	100	44.4	61.5
siph_physonect	95	80	86.9	NA	0	NA
stomatopods	95.3	95.3	95.3	13.3	66.7	22.2
tornaria	98.4	98.4	98.4	50	25	33.3
trichodesmium	92.3	95.5	93.9	93.2	29.5	44.8
tunicates	97.8	99.2	98.5	95.5	50.9	66.4
zoea	97.9	100	98.9	1.5	92.7	3

be cut, but at 99% level, then 63% of all images would be discarded.

The discarded, "low-confidence images," were put into the "unknown" category. Since this affected some of the 75,000 randomly selected images used to compute the confusion matrix, a postfiltering confusion matrix was then recalculated. The differences between the CM stats before vs. after probability filtering are shown in Supporting Information Table S2, and the section "With probability filtering" gives the CM for the final processed dataset, which can be used for future ecological studies.

Results

A total of 2.4 million raw ISIIS frames (nearly 40 h of imaging, 10 TB data) were collected from eight transects in the northern Gulf of Mexico. The raw ISIIS frames were segmented into 23.4 million images (27 GB), and classified into all

108 categories (Supporting Information Table S2). After filtering, 64.3% of the images were retained and 35.7% discarded. Aside from the detritus and artifact images, there were 1.62 million images of phytoplankton and protists and 1.37 million images of mesozooplankton.

Training set prediction: Accuracy benchmarks

Overall, the F1-score (the harmonic mean between precision and recall) for all classes was 88.1, with 67% with a F1-score over 90, and 83% with a F1 score over 80. The hardest classes to predict were the fish classes, with a mean F1 score of 70.7 (e.g., myctophid fishes were often confused for other types of fishes), and the easiest classes to predict included the protists (mean F1 of 98.0), cyclopoid copepods (mean F1 of 96.1), and chaetognath classes (mean F1 of 93.4) (Supporting Information Table S1). At the group level, the F1-scores increase, such that the lowest was 85.6 (physonect siphonophores), and 70% had F1-scores of 95 or above (Table 1).

Image filtering and CMs

In total, filtering removed 30% of all images, though this percentage differed by category. Out of 108 categories, 26 were well-predicted (over 60% retained after filtering), including the diatom chains, chaetognaths, dark detritus, protists, doliolids, and three calanoid copepod classes. Many of these 26 classes also fell within the top quarter in terms of numerical abundance (10 classes, containing 18.1 out of 23.4 million images). In particular, the main artifact class (imaging artifacts, as opposed to detritus), which comprised over 8.2 million images, or 35% of the total, were well-predicted, and over 93% were retained. However, the well-predicted classes were not only the common classes, since some of the rare but morphologically monotypic (e.g., the cestid ctenophores and goby fishes) also performed very well. In contrast, 22 classes were very heavily filtered, where less than 10% were retained. These classes included six (out of 14) hydromedusae, five (out of 19) fish, two polychaete worms, two siphonophores, one copepod and one shrimp, and tended to be the less common but morphologically diverse classes (Fig. 5, Supporting Information Table S3).

Results from the 75,000 random spot-checks showed that filtering improved the mean classification precision rate at the group level by 33% points, from 53% to 84% precision (Supporting Information Table S2). If only the biological groups were considered (thus excluding artifacts, detritus, and unknown, which was nearly 80% of the dataset), this increase was just slightly greater, from 51% to 87% precision. Twelve of the biological groups had less than 25 randomly drawn images (Supporting Information Table S2, also marked in gray in Table 1); these groups were very rare, each representing less than 0.12% of the total biological data. Excluding the rare biological groups, the precision rate after filtering was 90.7%.

Naturally, using the filtering thresholds decreased the total recall rate, by 23 percentage points, from 63% to 40%. For the nonrare biological groups (*n* = 23), the decrease was greater (31 points), but final recall rate was similar (39%). However, the F1-score, which is the harmonic mean of the precision rate and recall rate, only increased slightly, from 49% to 51%.

The final classification comparison was conducted between the (postfiltering) classifier and the training set, which represents the difference between a full dataset classification and the theoretical maximum for a classifier (Table 1). On the full dataset, classification precision was actually close to or even exceeded that of the training set, which was possible because of the application of the filtering thresholds. Despite the corresponding decrease in the recall rate, a comparison of F1-scores showed that a few of the biological groups had a less than 10-point difference (*Oithona* copepods, sergestid shrimp, and ablyid siphonophores). Groups that were less common, or had a lot of natural variability, such as other shrimp, pteropods, and cydippid ctenophores, showed a much greater difference, of 70–80 points, which was largely due to low recall rates postfiltering. However, the average difference in F1-scores for the (nonrare) biological groups was 40%, representing a moderate difference between the final classifier and the training set.

Discussion

We demonstrate the successful application of an image processing procedure, using a deep learning CNN, to classify a ~ 40 h, 10 TB in situ plankton imaging dataset containing 25 million image segments into 108 classes. After applying a filtering threshold on the classification probabilities, and grouping the classes into 37 taxonomically and functionally meaningful groups, the average classifier precision on nonrare biological groups (*n* = 23) was 90.7%, which is higher than any previous attempt on high-sampling volume, in situ plankton images.

Since Culverhouse et al. (2003) published a finding that trained personnel are only able to achieve 67–83% self-consistency on an expert plankton classification task, that range has existed as a sort of de facto benchmark within the plankton imaging field in which computer classification can be considered to be as good as human classification (e.g., Hu and Davis 2005, 2006). In reality, Culverhouse et al.'s (2003) findings were specific to a *difficult* identification task, in which morphologically variable dinoflagellate species (genus *Dinophysis*) were being distinguished from each other. For in situ plankton images, it is not very difficult for a human to distinguish between broad plankton community-based groups (e.g., calanoid copepods, shrimps, and larval fish), but rather, the difficulty only lies when distinguishing within certain taxa (e.g., between larval mesopelagic fishes, or between small decapod shrimps). Of course, classification difficulty may vary due to environmental conditions and ecosystem composition. Nonetheless, we suggest that this benchmark should be revisited, and raised to at least 90%. In manual sorting for the present dataset as well as others with 120+ classes (e.g., Cowen et al. 2015), the proportion of unknowns, in which an expert operator is unable to sort the image, ranged from < 1% for 30–35 classes to ca. 5% for 120–130 classes. In the present case, application of a method incorporating deep machine learning and filter thresholding resulted in over 90% precision on all the nonrare biological groups; this approaches the point in which we may consider an automated classifier to be as good as a human operator in sorting common plankton groups at higher taxonomic levels.

Application of the Faillettaz et al. (2016) filtering method gives us the ability to select for the highest probability images, and subsequently manipulate the precision levels (and by association, the recall rates) in the final classifier. Without filtering, the raw SparseConvNet classification precision and recall for the dataset would have been 53% and 63%, respectively. The filtering step modified the classification statistics (resulting in 84% precision and 40% recall for all groups), but allowed us to ensure the best description of biological patterns, which was important given the scientific goals of the image analysis procedure. Applying a classification filter would probably increase the overall performance of other previously published classification schema, and would likely temper the difference between our results and those earlier studies. Secondarily, we also note that it is not sufficient to judge a classifier by the class precision alone; the recall rate must also be incorporated. We therefore propose a more widespread adoption of the F1-score, which is the harmonic mean of the precision and recall.

CNNs represent a significant advance over traditional machine learning methods, because they are designed to learn and automatically extract feature descriptors (LeCun et al. 2015). Aside from the construction of the neural network architecture, the single most important factor determining the success of the classifier was the training set. Fernandes et al. (2009) had proposed a computer-assisted method for determining the optimal number of classes (settled on 30), using a Tree-Augmented Naïve Bayes classifier. In our case, since deep-learning methods are capable of classifying many more classes, we manually defined 108 classes and then grouped them into 37 groups after classification, but future efforts with CNNs should utilize some amount of computer assistance in determining the identity and quantity of classes.

Deep-learning methods require large amounts of training data, and our 42,000 item training set for 108 categories was likely on the low end; significant amounts of data augmentation was necessary. However, this is still an order of magnitude greater than the training sets used by traditional machine learning plankton image classifiers: Hu and Davis (2006) used 200 images per class for seven classes, Bi et al. (2015) used 210 images total for three classes, and Faillettaz et al. (2016) used 5979 images for 14 classes. Our choice of using a "natural" training set, where rare classes were augmented but not to the quantity of the most common classes, was a decision following our broader research objectives of describing mesozooplankton (including larval fish) distributions. These organisms are relatively rare, especially compared to protists and diatom chains, and thus needed special attention within the training set. Augmenting rare groups in the training set is naturally a time-consuming process. However, if the scientific objective of the image analysis system was to classify the detritus and common phytoplankton, then a more representative training set would achieve higher accuracies (Chang et al. 2012). Furthermore, to the extent possible, it was necessary to include the range of images, from the best (clearest, sharpest) image to the worst (most ambiguous, blurry) image, and to divide classes not only by taxonomy, but also morphological differences. Still, there were classes that did not perform very well (e.g., "shrimp other"), but were too difficult to separate further.

The development of our classifier (an application of the spatially sparse CNN, Graham 2014, 2015) was achieved following the 2015 National Data Science Bowl, a Kaggle.com machine learning competition. For the competition, we used the same ISIIS imaging system as in the present paper, but data from a different sampling region (Straits of Florida; competition data available at Cowen et al. 2015). While crowd-sourcing and machine learning competitions are not within the scope of the present paper (but discussed in Robinson et al. 2017), there were some key lessons we learned through the process that determined the successful application of the present classifier. First, in many competition settings, teams submit results that are an average of multiple models, also known as ensembles, which are computationally expensive and not necessarily the most realistic for real-world use. Therefore, it was critical to identify the single best model, which may or may not be part of the best ensemble (it was not in our case). Second, further development of the classification scheme was necessary after the competition ended. Essential to our success was the inclusion of a bio-computing specialist who could bridge the gap between the biologists and the computer scientists. Finally, the design of the competition dataset was also highly important, as it determined the types of solutions that emerged. We found that it was essential that the dataset had all the qualities of a good training set (ratio of images within rare vs. common classes, inclusion of high- and low-quality images, and separation of classes by taxonomy and morphology). These three key points facilitated the successful transfer of an image classifier between the competition and the present context.

As plankton datasets, both physical (e.g., the Continuous Plankton Recorder archive) and digital (the growing ISIIS collection), get larger and more comprehensive, it is critical to note that the amount of samples to sort at a particular taxonomic resolution will always depend on the scientific question *and* the time available for analyses. For some questions, such as the spatio-temporal variability in ichthyoplankton distributions (Richardson et al. 2010) or the niche shift of sibling species (Beaugrand et al. 2002), manually sorting physical samples to the genus or species level is necessary, but in those cases, only a relatively small number of organisms can realistically be sorted. For other questions, such as the fine-scale distribution of broad taxonomic groups, the complete analyses of samples collected by high throughput imaging systems is most adapted. In that case, manual sorting would be time prohibitive, especially with increasing numbers of classes (we estimate that sorting into 40–50 classes, which can be done at 5000 images d^{-1}, occurs at roughly half the rate of sorting into 10–15 classes). Therefore, computer-assisted or fully automated classification becomes more expedient. Even the necessity of creating a training set, with associated independent test set, for each sampling region is time consuming (by our estimates, ca. 2–3 months). The future development of a master, global-level training set with regional filters could facilitate a more rapid image classification process. This could eventually lead to a minimal amount of manual identification work for each additional dataset (i.e., for spot-checks for the final confusion matrix, which would yield a class-specific correction factor for densities). The combination of the speed of classification, use of 100+ classes, high precision, and only needing to do small amounts of manual sorting would significantly increase the utility of plankton imaging systems, as we will be able to classify millions to billions of in situ plankton images quickly and accurately.

It is evident that with the recent interest in plankton (e.g., from the *Tara* Oceans project, Bork et al. 2015) that there are many additional questions and areas for exploration regarding the base of the marine food chain. Imaging systems are inherently complementary to net-based sampling; physical samples are always going to be necessary for ecological questions requiring fine taxonomic resolution and the analysis of hard structures (e.g., otoliths), isotopes, or genetics. However, large-volume imaging systems can be particularly useful for addressing questions regarding rare, gelatinous, or large organisms in the context of predator-prey dynamics, horizontal and vertical aggregations, and fine-scale relationships to the environment. Plankton imaging systems can also provide important validation data for regional and global ocean ecosystem models, which suffer from insufficient data for constraining patterns and processes. The development of whole, integrated pipelines for plankton image analysis can enhance the utility of automated classification tools, and can eventually lead to the goal of real-time image processing done at sea. Combined with some net-sampling for taxonomic validation, plankton imaging systems can be an incredibly powerful tool, with applications in ocean monitoring and fisheries management, as well as in addressing many of the fundamental questions still existing within plankton ecology.

Data availability statement

All manually classified images from the full training set and test sets (43K probability filtering set and 75K random spot-check set), as well as text files containing predicted and validated classes for all test sets will be available on Zenodo.org (doi: 10.5281/zenodo.836492). Executable files for the segmentation code will also be on Zenodo.org. Source code for SparseConvNet is available at https://github.com/btgraham/SparseConvNet, as well as on Zenodo.org.

References

Arnold, G. P., and P. B. N. Nuttall-Smith. 1974. Shadow cinematography of fish larvae. Mar. Biol. **28**: 51–53. doi:10.1007/BF00389116

Ashjian, C. J., C. S. Davis, S. M. Gallager, and P. Alatalo. 2001. Distribution of plankton, particles, and hydrographic features across Georges Bank described using the Video Plankton Recorder. Deep-Sea Res. Part II Top. Stud. Oceanogr. **48**: 245–282. doi:10.1016/S0967-0645(00)00121-1

Beaugrand, G., P. C. Reid, F. Ibañez, J. A. Lindley, and M. Edwards. 2002. Reorganization of North Atlantic marine copepod biodiversity and climate. Science **296**: 1692–1694. doi:10.1126/science.1071329

Benfield, M. C., C. S. Davis, and S. M. Gallager. 2000. Estimating the in-situ orientation of *Calanus finmarchicus* on Georges Bank using the Video Plankton Recorder. Plankton Biol. Ecol. **47**: 69–72.

Benfield, M. C., C. J. Schwehm, R. G. Fredericks, G. Squyres, S. F. Keenan, and M. V. Trevorrow. 2003. Measurements of zooplankton distributions with a high-resolution digital camera system, p. 17–30. *In* L. Seuront and P. G. Strutton [eds.] Handbook of scaling methods in aquatic ecology: Measurement, analysis, simulation. CRC Press.

Bi, H., Z. Guo, M. Benfield, C. Fan, M. Ford, S. Shahrestani, and J. Sieracki. 2015. A semi-automated image analysis procedure for in situ plankton imaging systems. PLoS One **10**: e0127121. doi:10.1371/journal.pone.0127121

Biard, T. and others 2016. *In situ* imaging reveals the biomass of giant protists in the global ocean. Nature **532**: 504–507. doi:10.1038/nature17652

Bork, P., C. Bowler, C. de Vargas, G. Gorsky, E. Karsenti, and P. Wincker. 2015. *Tara* Oceans studies plankton at planetary scale. Science **348**: 873–873. doi:10.1126/science.aac5605

Chang, C.-Y., P.-C. Ho, A. R. Sastri, Y.-C. Lee, G.-C. Gong, and C.-H. Hsieh. 2012. Methods of training set construction: Towards improving performance for automated mesozooplankton image classification systems. Cont. Shelf Res. **36**: 19–28. doi:10.1016/j.csr.2012.01.005

Cowen, R. K., and C. M. Guigand. 2008. *In Situ* Ichthyoplankton Imaging System (ISIIS): System design and preliminary results. Limnol. Oceanogr.: Methods **6**: 126–132. doi:10.4319/lom.2008.6.126

Cowen, R. K., A. T. Greer, C. M. Guigand, J. A. Hare, D. E. Richardson, and H. J. Walsh. 2013. Evaluation of the In Situ Ichthyoplankton Imaging System (ISIIS): Comparison with the traditional (bongo net) sampler. Fish. Bull. **111**: 1–12. doi:10.7755/FB.111.1.1

Cowen, R. K., S. Sponaugle, K. L. Robinson, and J. Luo. 2015. PlanktonSet 1.0: Plankton imagery data collected from

F.G. Walton Smith in Straits of Florida from 2014-06-03 to 2014-06-06 and used in the 2015 National Data Science Bowl (NODC Accession 0127422). NOAA National Centers for Environmental Information. Dataset. Available from https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.nodc:0127422 (Accessed July 2018).

Culverhouse, P. F., R. Williams, B. Reguera, R. E. Ellis, and T. Parisini. 1996. Automatic categorization of 23 species of dinoflagellate by artificial neural network. Mar. Ecol. Prog. Ser. **139**: 281–287. doi:10.3354/meps139281

Culverhouse, P. F., R. Williams, B. Reguera, V. Herry, and S. González-Gil. 2003. Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. Mar. Ecol. Prog. Ser. **312**: 297–309. doi:10.3354/meps247017

Davis, C. S., S. M. Gallager, M. S. Berman, L. R. Haury, and J. R. Strickler. 1992. The Video Plankton Recorder (VPR): Design and initial results. Arch. Hydrobiol. Beih. Ergeb. Limnol. **36**: 67–81.

Davis, C. S., Q. Hu, S. M. Gallager, X. Tang, and C. J. Ashjian. 2004. Real-time observation of taxa-specific plankton distributions: An optical sampling method. Mar. Ecol. Prog. Ser. **284**: 77–96. doi:10.3354/meps284077

Davis, C. S., and D. J. McGillicuddy. 2006. Transatlantic abundance of the N2-fixing colonial cyanobacterium *Trichodesmium*. Science **312**: 1517–1520. doi:10.1126/science.1123570

Dieleman, S., K. W. Willett, and J. Dambre. 2015. Rotation-invariant convolutional neural networks for galaxy morphology prediction. Mon. Not. R. Astron. Soc. **450**: 1441–1459. doi:10.1093/mnras/stv632

Faillettaz, R., M. Picheral, J. Y. Luo, C. Guigand, R. K. Cowen, and J.-O. Irisson. 2016. Imperfect automatic image classification successfully describes plankton distribution patterns. Meth. Oceanogr. **15–16**: 60–77. doi:10.1016/j.mio.2016.04.003

Fernandes, J. A., X. Irigoien, G. Boyra, J. A. Lozano, and I. Inza. 2009. Optimizing the number of classes in automated zooplankton classification. J. Plankton Res. **31**: 19–29. doi:10.1093/plankt/fbn098

Gasparini, S., and E. Antajan. 2013. PLANKTON IDENTIFIER: A software for automatic recognition of planktonic organisms. Available from http://www.obs-vlfr.fr/~gaspari/Plankton_Identifier/index.php (Accessed August 2016).

Gorsky, G. and others 2010. Digital zooplankton image analysis using the ZooScan integrated system. J. Plankton Res. **32**: 285–303. doi:10.1093/plankt/fbp124

Graham, B. 2014. Spatially-sparse convolutional neural networks. eprint arXiv:14096.070v1.

Graham, B. 2015. Fractional max-pooling. eprint arXiv:1412.6071v4.

Greer, A. T., R. K. Cowen, C. M. Guigand, M. A. McManus, J. C. Sevadjian, and A. H. V. Timmerman. 2013. Relationships between phytoplankton thin layers and the fine-scale vertical distributions of two trophic levels of zooplankton. J. Plankton Res. **35**: 939–956. doi:10.1093/plankt/fbt056

Hinton, G. and others 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag. **29**: 82–97. doi:10.1109/MSP.2012.2205597

Hu, Q., and C. Davis. 2005. Automatic plankton image recognition with co-occurrence matrices and support vector machine. Mar. Ecol. Prog. Ser. **295**: 21–31. doi:10.3354/meps295021

Hu, Q., and C. Davis. 2006. Accurate automatic quantification of taxa-specific plankton abundance using dual classification with correction. Mar. Ecol. Prog. Ser. **306**: 51–61. doi:10.3354/meps306051

Iyer, N. 2012. Machine vision assisted *in situ* ichthyoplankton imaging system, p. 61. M.S. thesis. Purdue Univ.

Jeffries, H. P., K. Sherman, R. Maurer, and C. Katsinis. 1980. Computer processing of zooplankton samples, p. 303–316. *In* V. Kennedy [ed.] Estuarine perspectives. Academic Press.

Jeffries, H. P., M. S. Berman, A. D. Poularikas, C. Katsinis, I. Melas, K. Sherman, and L. Bivins. 1984. Automated sizing, counting and identification of zooplankton by pattern recognition. Mar. Biol. **78**: 329–334. doi:10.1007/BF00393019

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. Imagenet classification with deep convolutional neural networks, p.1097–1105. Adv. Neural Inf. Process. Syst. 25

Laney, S. R., and H. M. Sosik. 2014. Phytoplankton assemblage structure in and around a massive under-ice bloom in the Chukchi Sea. Deep-Sea Res. Part II Top. Stud. Oceanogr. **105**: 30–41. doi:10.1016/j.dsr2.2014.03.012

Lawrence, S., C. L. Giles, A. C. Tsoi, and A. D. Back. 1997. Face recognition: A convolutional neural-network approach. IEEE Trans. Neural Netw. **8**: 98–113. doi:10.1109/72.554195

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. Nature **521**: 436–444. doi:10.1038/nature14539

Luo, J. Y., B. Grassian, D. Tang, J.-O. Irisson, A. T. Greer, C. M. Guigand, S. McClatchie, and R. K. Cowen. 2014. Environmental drivers of the fine-scale distribution of a gelatinous zooplankton community across a mesoscale front. Mar. Ecol. Prog. Ser. **510**: 129–149. doi:10.3354/meps10908

Luo, T., K. Kramer, D. B. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins. 2005. Active learning to recognize multiple types of plankton. J. Mach. Learn. Res. **6**: 589–613. doi:10.1109/ICPR.2004.1334570

McClatchie, S. and others 2012. Resolution of fine biological structure including small narcomedusae across a front in the Southern California Bight. J. Geophys. Res. Oceans **117**: C04020. doi:10.1029/2011JC007565

Ohman, M. D., J. R. Powell, M. Picheral, and D. W. Jensen. 2012. Mesozooplankton and particulate matter responses to a deep-water frontal system in the southern California Current System. J. Plankton Res. **34**: 815–827. doi:10.1093/plankt/fbs028

Ortner, P. B., S. R. Cummings, and R. P. Aftring. 1979. Silhouette photography of oceanic zooplankton. Nature **277**: 50–51. doi:10.1038/277050a0

Ortner, P. B., L. C. Hill, and H. E. Edgerton. 1981. In-situ silhouette photography of Gulf Stream zooplankton. Deep-

Sea Res. Part A Oceanogr. Res. Pap. **28**: 1569–1576. doi: 10.1016/0198-0149(81)90098-4

Petrik, C. M., G. A. Jackson, and D. M. Checkley Jr. 2013. Aggregates and their distributions determined from LOPC observations made using an autonomous profiling float. Deep-Sea Res. Part I Oceanogr. Res. Pap. **74**: 64–81. doi:10.1016/j.dsr.2012.12.009

Picheral, M., L. Guidi, L. Stemmann, D. Karl, G. Iddaoud, and G. Gorsky. 2010. The underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. Limnol. Oceanogr.: Methods **8**: 462–473. doi:10.4319/lom.2010.8.462

Richardson, D. E., J. K. Llopiz, C. M. Guigand, and R. K. Cowen. 2010. Larval assemblages of large and medium-sized pelagic species in the Straits of Florida. Prog. Oceanogr. **86**: 8–20. doi:10.1016/j.pocean.2010.04.005

Robinson, K. L., J. Y. Luo, S. Sponaugle, C. Guigand, and R. K. Cowen. 2017. A Tale of two crowds: Public Engagement in Plankton Classification. Front. Mar. Sci. 4. doi: 10.3389/fmars.2017.00082

Samson, S., T. Hopkins, A. Remsen, L. Langebrake, T. Sutton, and J. Patten. 2001. A system for high resolution zooplankton imaging. IEEE J. Ocean. Eng. **26**: 671–676. doi:10.1109/48.972110

Schmid, M. S., C. Aubry, J. Grigor, and L. Fortier. 2016. The LOKI underwater imaging system and an automatic identification model for the detection of zooplankton taxa in the Arctic Ocean. Meth. Oceanogr. **15–16**: 129–160. doi:10.1016/j.mio.2016.03.003

Simpson, R., R. Williams, R. Ellis, and P. F. Culverhouse. 1992. Biological pattern recognition by neural networks. Mar. Ecol. Prog. Ser. **79**: 303–308.

Tsechpenakis, G., C. Guigand, and R. K. Cowen. 2007. Image analysis techniques to accompany a new *in situ* ichthyoplankton imaging system, p. 1–6. *In* OCEANS 2007. Aberdeen, UK. doi: 10.1109/OCEANSE.2007.4302271

Tsechpenakis, G., C. Guigand, and R. K. Cowen. 2008. Machine vision-assisted *in situ* ichthyoplankton imaging system. Sea Technol. **49**: 15–20.

Wiebe, P. H., and M. C. Benfield. 2003. From the Hensen net toward four-dimensional biological oceanography. Prog. Oceanogr. **56**: 7–136. doi:10.1016/S0079-6611(02)00140-4

Acknowledgments

We thank Booz Allen Hamilton for sponsoring and for Kaggle.com for hosting the 2015 National Data Science Bowl competition. Many thanks to Robin Faillettaz, Christian Briseño-Avena, and Kelly Robinson for helpful discussions, and to Chase Cobb and Clare Hansen for assistance in image analysis. Funding for this project came from NOAA NRDA Contract AB133C-11-CQ-0050 to RKC, and a National Science Foundation Graduate Research Fellowship and University of Miami Maytag Fellowship to JYL. RKC and JYL were also supported on NSF OCE Grant 1419987 (to RKC and Su Sponaugle) during the preparation of this manuscript.

Conflict of Interest

None declared.

Submitted 3 February 2017
Revised 13 September 2018
Accepted 17 September 2018

Associate editor: Malinda Sutor