LIMNOLOGY
and
OCEANOGRAPHY: METHODS

*Limnol. Oceanogr.: Methods* 22, 2024, 47–64

doi: 10.1002/lom3.10591

# Taming the data deluge: A novel end-to-end deep learning system for classifying marine biological and environmental images

**Hongsheng Bi,[1]* Yunhao Cheng,[2] Xuemin Cheng,[3] Mark C. Benfield,[4] David G. Kimmel,[5] Haiyong Zheng,[2] Sabrina Groves,[1] Kezhen Ying[6]**

[1]University of Maryland Center for Environmental Science, Solomons, Maryland, USA
[2]Ocean University of China, Qingdao, Shandong, People's Republic of China
[3]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, Guangdong, People's Republic of China
[4]Louisiana State University, Baton Rouge, Louisiana, USA
[5]Alaska Fisheries Science Center, Seattle, Washington, USA
[6]Photobio Tech LTD, Shenzhen, Guangdong, People's Republic of China

## Abstract

Underwater imaging enables nondestructive plankton sampling at frequencies, durations, and resolutions unattainable by traditional methods. These systems necessitate automated processes to identify organisms efficiently. Early underwater image processing used a standard approach: binarizing images to segment targets, then integrating deep learning models for classification. While intuitive, this infrastructure has limitations in handling high concentrations of biotic and abiotic particles, rapid changes in dominant taxa, and highly variable target sizes. To address these challenges, we introduce a new framework that starts with a scene classifier to capture large within-image variation, such as disparities in the layout of particles and dominant taxa. After scene classification, scene-specific Mask regional convolutional neural network (Mask R-CNN) models are trained to separate target objects into different groups. The procedure allows information to be extracted from different image types, while minimizing potential bias for commonly occurring features. Using in situ coastal plankton images, we compared the scene-specific models to the Mask R-CNN model encompassing all scene categories as a single full model. Results showed that the scene-specific approach outperformed the full model by achieving a 20% accuracy improvement in complex noisy images. The full model yielded counts that were up to 78% lower than those enumerated by the scene-specific model for some small-sized plankton groups. We further tested the framework on images from a benthic video camera and an imaging sonar system with good results. The integration of scene classification, which groups similar images together, can improve the accuracy of detection and classification for complex marine biological images.

Imaging systems are increasingly being used to study aquatic organisms and their interactions with the environment at different spatial and temporal scales (Solan et al. 2003; Wiebe and Benfield 2003; Smith and Rumohr 2013; Durden et al. 2016; Shortis et al. 2016; Irisson et al. 2022). In light of the rapid developments and diverse applications of various imaging systems, automated image processing encounters several challenges. A major hurdle is the customization and system-specific nature of most automated image processing procedures (MacLeod et al. 2010; Durden et al. 2016; Luo et al. 2018; Orenstein et al. 2020), which limits their broader applicability. In addition, environment-driven separation protocols further compound this issue as they often necessitate different levels of disassociation of target and nontarget objects. For instance, separating objects in images from turbid coastal water is more challenging than those from clear offshore water. Further challenges include segmentation method, unbalanced taxonomic samples, size class differences, particle saturation, overcrowding, and overlap.

The challenges of developing a general image processing framework arise from differences in image properties associated with different imaging technologies and different environments. For example, processing plankton images primarily

*Correspondence: hongshengbi@gmail.com

**Author Contribution Statement:** H.B., Y.C., M.B., and H.Z. were responsible for the conceptualization of this study. Data acquisition was conducted by H.B. and K.Y.. Y.C., H.B., X.C. and K.Y. developed the program, conducted the experiment, and drafted the initial manuscript. Reviewing and editing of the final manuscript were performed by H.B., D. K., M.B., and S.G. All authors contributed to the development of the methodology.

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

means separating target and nontarget objects in the water column, whereas processing benthic images requires separation of foreground objects from both the background and nontarget objects. In previous cases, these tasks have been distinguished and processed by convolutional neural networks (CNNs) (Cheng et al. 2019; González et al. 2019; Piechaud et al. 2019; Wang et al. 2020). A commonly applied architecture in aquatic biological images starts with potential targets, denoted by the region of interest (RoI), and the segmented RoIs are standardized before feature description and classification by a pretrained machine learning model (Bi et al. 2015; Cheng et al. 2019; Irisson et al. 2022; Mittal et al. 2022). However, accurate segmentation and class unbalance remain two major issues, both of which undermine the performance and accuracy of the common framework by producing excessive false positives and biased results.

The goal of accurate segmentation of RoIs from underwater images is to separate individual organisms and ensure that each RoI only has one potential target for the subsequent classification. While accurate segmentation is relatively easy for images collected in water with low particle density, it remains a challenge in water with higher particle density (Fig. 1). Using plankton images collected from coastal waters as an example, images are often crowded either with planktonic organisms or other particles (Fig. 1a–c). When images are saturated with other particles, it is difficult to separate target organisms from nontarget particles (Fig. 1a). When organisms have a complex structure, over-segmentation often occurs (Bi et al. 2015). Furthermore, overlap among different organisms makes the task of segmenting individuals even more challenging (Fig. 1a–e). The issue of crowded images and overlap among different organisms arises from two different aspects. Patchy distribution is a common feature of marine organisms (Levinton 1995), and when an imaging system is towed through a patch of marine organisms, we would expect overcrowded images and overlap among individual organisms. As technology advances, modern imaging systems are often equipped with increased field of view and depth of field, which further exacerbate the issue of overcrowded and overlapping objects.

Image segmentation, dividing an image into distinct regions or segments based on certain characteristics, is a long-standing problem in computer vision and most existing techniques are not suitable for noisy environments (Pal and Pal 1993; Song and Yan 2017). Recent efforts on developing segmentation techniques, specifically for crowded underwater images, partially alleviate this issue (Cheng et al. 2020; Song et al. 2022); but given the complexity and uncertainty in underwater images, unsupervised deep learning approaches like region-based CNN (R-CNN) offer a more promising solution (Minaee et al. 2021). The R-CNN models combine a region proposal network (RPN) to locate RoIs, a CNN model to describe features of RoIs generated from RPN proposals, and a classification layer to predict final bounding boxes and classes. Mask R-CNN is a combination of Faster R-CNN and a fully convolution network which outperforms existing models in instance-level segmentation and recognition by delineating individual objects within an image (Ren et al. 2015; He et al. 2017). Faster R-CNN is an extension of R-CNN, in which the RPN shares convolutional features with the subsequent detection network, allowing for end-to-end training and eliminating the need for the selective search algorithm (Girshick 2015; Ren et al. 2017). In the present study, we test the feasibility of using Mask R-CNN as the first step in processing noisy marine biological and environmental images.

In marine environments, skewed frequency distribution across taxonomic groups is common and such imbalanced class distributions have detrimental impacts on classification performance because the model often oversamples the rare groups and under samples the abundant groups (Buda et al. 2018; Zhang et al. 2020; Sharma et al. 2022). While a more balanced class distribution fits our preference for unbiased results for both common and rare taxonomic groups, deep learning models require tens of thousands of labeled images to achieve high accuracy, which is almost impossible for rare taxonomic groups. Meanwhile, the large size variation that spans up to several orders of magnitude, also complicates the imbalance issue. Large organisms occupy the greatest numbers of pixels, and morphological features can persist throughout the convolution process, whereas small organisms occupy small amounts of pixels and morphological features likely disappear after a few iterations of convolution. Therefore, small organisms need more labeled images than large organisms to reach the same level of accuracy. This training imbalance can either exacerbate or alleviate the unbalance class issue, depending on the steps that precede convolution.

There is no simple solution to the common issue of unbalanced classes in deep learning, but proper data-level operations could be helpful in addressing this issue (Sharma et al. 2022). We propose to alleviate the issue of unbalanced class at the data level by incorporating a scene classification model. Scene classification uses the layout of organisms within the scene, in addition to the ambient context, to group similar images. Using underwater plankton images as an example, images with high concentration of other particles (Fig. 1a) are often collected around locations or times with strong physical mixing. Single-species dominated images are often collected within a bloom period or a patch of the organisms (Fig. 1c–e). Scene classification reduces the diversity and uncertainty within each category, allowing a more targeted classification model to be trained. The inclusion of a scene classification model improves the alignment between observed data and the library compared to the full model. This integration offers certain advantages, such as a more balanced model for images captured in clear water (Fig. 1f), in contrast to the full model. However, during bloom periods, the scene classification model may create a more skewed model compared to the full model. In addition, it can aid in addressing rare

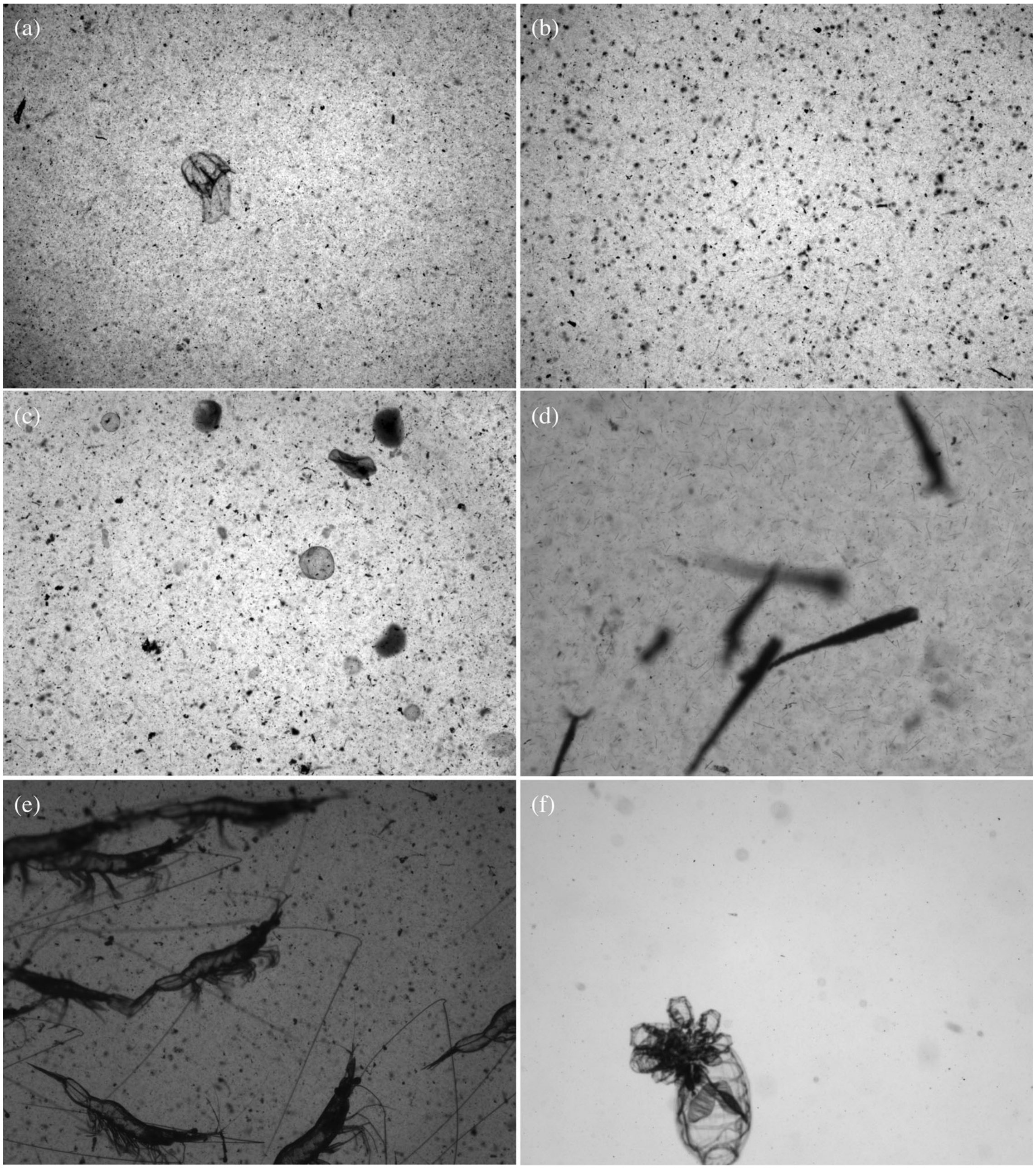


**Fig. 1.** Example images for selected scene categories: (**a**) high-concentration scene with large amounts of particles, (**b**) *Noctiluca* scene with each black dot indicating one *Noctiluca*, (**c**) *Phaeocystis* scene showing the semi-transparent round-shaped colonies, (**d**) Pteropoda scene showing clustered *Creseis acicula* and line-shaped *Lyngbya* algae, (**f**) shrimp scene showing clustered individuals, and (g) low-density scene showing a clean image with one budding *Thaliacea*.

species by reducing the number of taxa within each scene, thus facilitating their identification.

In the present study, we propose a general framework to process underwater images (Fig. 2). The procedure starts with scene classification to separate images based on their context and layout of dominant organisms. This process is followed by a separate Mask R-CNN model, trained for each scene, to facilitate foreground object detection and recognition. To assess the impact of integrating scene classification on accuracy, we conducted a comparative analysis of the full model and the scene-specific model using the same testing dataset. The full model represents the common architecture and is accomplished through a Mask R-CNN model for automated recognition of underwater organisms without scene classification. The scene-specific model starts with a scene classification model followed by a dedicated model for each scene for

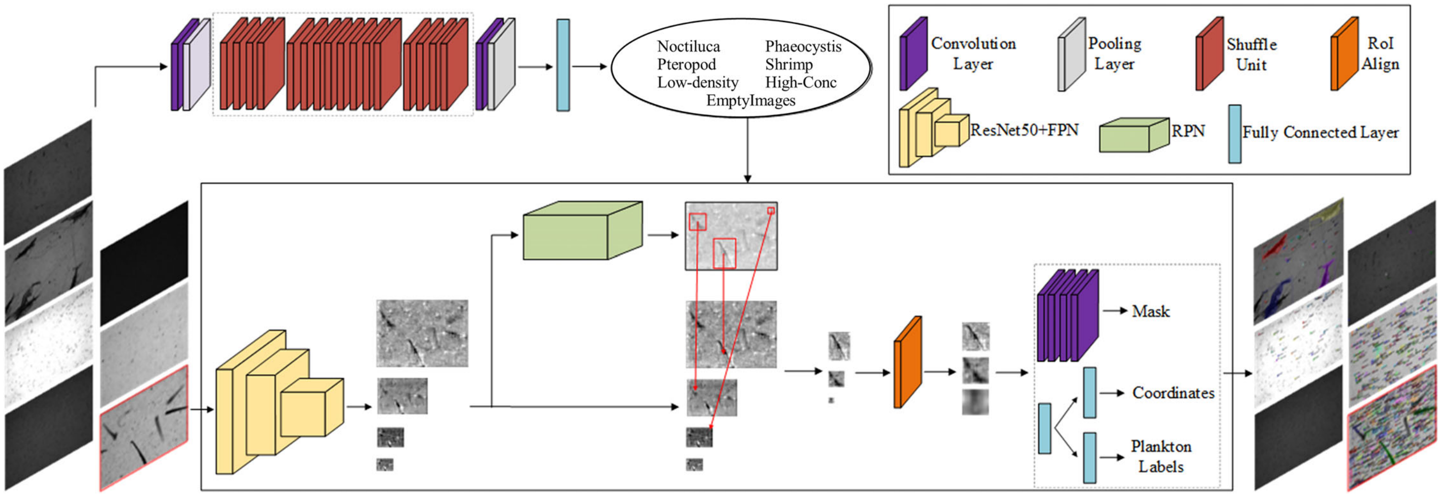


**Fig. 2.** Diagram of the proposed procedure. The procedure starts with a scene classification using a lightweight ShuffleNet. For example, six scenes and an additional scene for empty images were selected based on image contents such as dominant taxa group or concentration of particulates. A separate Mask R-CNN model was trained for each scene category and potential objects were first detected through a region proposal network and then classified by a residual neural network (lower box). The final output included an image with predicted results and segmented objects for each taxa group.

recognition. We present metrics, including counts, precision, and the number of missing individuals, for each approach using underwater plankton images from coastal waters collected by PlanktonScope (Song et al. 2020; Liu et al. 2021). We then expand the comparison to test for general applicability using images collected by different underwater imaging systems.

## Methods

### Full model

Mask R-CNN combines a Faster CNN for object identification and a fully convolutional network for recognition (He et al. 2017). Faster CNN includes two networks, a CNN and RPN. First, the network detects region proposals that are defined as regions in the feature map which contain the object. In the second stage, the network predicts bounding boxes and object class for each of the proposed regions obtained in the first stage. Each proposed region can be of different size and the size of these proposed regions is then fixed by the RoI pooling method, which helps to preserve spatial information. Mask R-CNN extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition. It also replaces the RoI pooling with RoIAlign which makes object detection more efficient and accurate while simultaneously generating a high-quality segmentation mask for each instance. The output from the RoIAlign layer is analyzed by a residual neural network (ResNet) for object classification. Finally, the procedure segments recognized objects based on the corresponding mask generated during object detection (He et al. 2017).

The Mask R-CNN model was implemented in Python using the Detectron 2 package (Wu et al. 2019). In the present study, 979 underwater plankton images were labeled using the open-source annotation tool LabelMe (Wada 2016). LabelMe creates and manages annotations in a JavaScript Object Notation (JSON) file format, which allows for easy data exchange between different programming languages and platforms. The JSON generated by the LabelMe software contains information about the annotated images, including the objects present in the image, their bounding boxes, and any associated attributes or labels. It is a widely used standard for storing image annotations, facilitating the development of deep learning models in object detection, image segmentation, and recognition. Within these images, we identified 17 taxonomic groups and labeled 32,227 individual organisms ranging from small algae, $< 100\ \mu$m in diameter, to small pelagic fish, $\sim 4$ cm. The first three most abundant taxonomic groups are the line-shaped algae *Lyngbya*, near-transparent ellipsoid *Noctiluca*, and copepods (Table 1). We trained a full model, using the labeled images, and assessed the accuracy of foreground object identification and classification to evaluate the model performance.

For all training datasets, including the full model, scene classification model, and each scene-specific recognition model, the data were divided into training and validation sets in a 70 : 30 ratio. All models were trained using a batch size of 4 and underwent 20,000 iterations during the training process. The training was performed on NVIDIA RTX graphics cards, specifically the RTX 2000, 3000, and 4000 series.

### Scene-specific model

The scene-specific model begins by classifying underwater images into different scenes and is followed by a scene-specific Mask R-CNN model for objection detection and classification (Fig. 2). Scene classification depends on the ambient context,

**Table 1.** Number of images, identified taxonomic groups, and labeled individuals for each taxonomic group in the full model and scene-specific models.

| | | | Different scenes | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Full model | High concentration | Noctiluca | Phaeocystis | Pteropod | Shrimp | Low density |
| Full frame images | Number of images | 979 | 151 | 99 | 81 | 324 | 124 | 281 |
| Number of labeled individuals | Appendicularia | 489 | 29 | 13 | 147 | 207 | 4 | 239 |
| | Chaetognatha | 209 | 22 | 42 | 45 | 12 | 9 | 81 |
| | Copepoda | 3466 | 144 | 353 | 531 | 1039 | 119 | 1952 |
| | *Creseis* | 1225 | 0 | 0 | 0 | 1225 | 9 | 188 |
| | Echinodermata | 451 | 5 | 388 | 34 | 16 | 0 | 12 |
| | Fish | 10 | 0 | 0 | 0 | 0 | 10 | 0 |
| | Larval fish | 1 | 0 | 0 | 0 | 0 | 0 | 3 |
| | *Lyngbya* | 20,212 | 1635 | 1593 | 815 | 18,370 | 205 | 10,107 |
| | Medusae | 246 | 73 | 28 | 38 | 50 | 6 | 109 |
| | *Noctiluca* | 4950 | 189 | 3661 | 4 | 6 | 782 | 595 |
| | *Phaeocystis* | 521 | 143 | 86 | 259 | 40 | 0 | 0 |
| | Polychaete | 21 | 0 | 0 | 0 | 0 | 0 | 21 |
| | *Scylla* larvae | 14 | 0 | 0 | 0 | 0 | 0 | 14 |
| | Shrimp | 262 | 76 | 7 | 0 | 2 | 154 | 7 |
| | *Skeletonema* | 3 | 2 | 0 | 0 | 1 | 0 | 0 |
| | Spiral diatom | 111 | 13 | 16 | 0 | 24 | 0 | 15 |
| | Thaliacea | 108 | 9 | 54 | 5 | 13 | 5 | 57 |

shape, and layout of biological targets within the image. This step is accomplished through an efficient lightweight CNN model, namely ShuffleNet, which utilizes pointwise group convolution and channel shuffle to reduce computation cost while maintaining accuracy (Ma et al. 2018; Zhang et al. 2018). We use underwater plankton images to illustrate how we set up and train the scene classification model.

In the present study, underwater plankton images were manually sorted into six scene categories (Fig. 1):

1. images with evenly distributed, abundant particles (Fig. 1a), which are common in estuaries and nearshore waters due to riverine and resuspended bottom particles;
2. randomly distributed, *Noctiluca*-dominated images (Fig. 1b);
3. randomly distributed *Phaeocystis* colonies, which exhibit round clustering (Fig. 1c);
4. *Creseis acicula* Pteropoda clusters (Fig. 1d) and the co-occurring, evenly distributed, line-shaped *Lyngbya* algae;
5. clustered, small shrimp (Fig. 1e) with co-occurring, randomly distributed *Noctiluca*; and
6. clear water, low concentration plankton (Fig. 1f).

For the plankton scene classification model, we included 546, 1060, 768, 508, 336, and 766 full frame images in the high-concentration, *Noctiluca*, *Phaeocystis*, *Creseis*, shrimp, and clear water scenes, respectively. The number of images, taxonomic groups, and labeled targets for each scene are summarized in Table 1. For the training dataset, the data were split into a training set and a validation set in a 70 : 30 ratio.

In the second phase, we first classified the 979 labeled images into 6 different scene groups, and then a separate Mask R-CNN model was trained using the labeled images for each scene category (Fig. 2). Note that we manually selected a small subset of labeled images to be included in different scenes. Out of the 979 total images utilized in the current study (corresponding to the sum of the diagonal elements in Table 2), 84 images were cross-referenced (represented by the sum of the nondiagonal elements in Table 2). The goal was to represent the occurrence of various species more accurately across different scenes, while the total amount of information for both the full model and the scene-specific model remains unchanged. The inclusion of some duplicate images also ensured the scene-specific model receives sufficient data for effective training. For instance, in the case of line-shaped algae, *Lyngbya*, which often co-occurs with *Creseis*, it was essential to provide ample examples for the model to learn from. Labeling *Lyngbya* is an extremely time-consuming process, often taking several weeks to label a single image due to its small size and frequent occurrences in large numbers. Similarly, *Noctiluca* often co-occurred with shrimp, thus we included some labeled images from the *Noctiluca* scene to improve the representation of *Noctiluca* in the shrimp scene. Likewise, we observed that *Noctiluca* frequently co-occurred with shrimp. To enhance the representation of *Noctiluca* in the shrimp scene, we intentionally included some labeled images from the *Noctiluca* scene.

**Table 2.** The number of images in each scene is denoted by the diagonal elements, while the nondiagonal elements represent images that are cross-referenced across different scenes.

| | High conc. | Noctiluca | Phaeocystis | Pteropoda | Shrimp | Low den. |
|---|---|---|---|---|---|---|
| High conc. | 102 | 1 | 4 | 2 | 21 | 21 |
| Noctiluca | 1 | 85 | 3 | 3 | 2 | 5 |
| Phaeocystis | 0 | 1 | 76 | 0 | 0 | 4 |
| Pteropoda | 0 | 0 | 0 | 322 | 0 | 2 |
| Shrimp | 0 | 8 | 0 | 3 | 113 | 0 |
| Low den. | 0 | 0 | 0 | 0 | 3 | 281 |

### Comparison between full model and scene-specific model for plankton images

The performance of full model and scene-specific models was evaluated using accuracy metrics and a confusion matrix. During the process of model training, two accuracy metrics were calculated for the Mask R-CNN model: the accuracy for foreground object detection and the accuracy for object classification. For the scene classification model, a single accuracy metric was calculated. The general form of accuracy is $(\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN})$, where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. A confusion matrix was also constructed for the scene classification model. The confusion matrix includes the actual and predicted classes obtained by a classification system. Each row represents an actual class example; each column represents the state of a predicted class.

The performance of both the full model and the scene classification model was evaluated and compared using a testing dataset comprised of 100 underwater plankton images. These images were collected in coastal waters, specifically in Guangdong, People's Republic of China, and the Columbia River Plume, USA, utilizing the PlanktonScope, an in situ shadowgraph imaging system (Song et al. 2020; Liu et al. 2021). Twenty nonempty full-frame images were randomly selected for each scene, and then images were processed using both the full- and scene-specific models. Both models output the predicted results, segmented objects for each class, and the counts for each class. We then manually examined the segmented objects, moved the misclassified objects to the correct class, and performed a recount. Given that Mask R-CNN does not produce information for the nontarget classes, and it is almost impossible to perform objective visual counts for small semi-transparent organisms like *Noctiluca* and *Lyngbya*, we used precision as a metric to evaluate model performance instead of accuracy. Precision measures the accuracy of positive predictions made by the model which reflect the model's ability to correctly identify true-positive instances among all the positive predictions it makes. Precision is calculated as: $\text{Precision} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, where TP is true positives and FP is false positives. A high precision value indicates that the model has a low false-positive rate, meaning that when it predicts a positive class, it is highly likely to be correct. On the other hand, a low precision value suggests that the model is making a significant number of false-positive predictions, which is often a problem in plankton image recognition.

### Examples of processing benthic images and sonar images

Benthic videos were collected in the Arctic using a camera system manufactured by A.G.O. Environmental Electronics Ltd, Victoria, B.C., Canada (Cooper et al. 2019). The system includes two positioning lasers, an undersea video camera, with onboard monitoring and recording on a ship-based video camcorder, and hand deployment using a 200-m electronic cable. A key challenge in processing benthic images is separating organisms from a complex background because benthic organisms often blend into their habitats and suboptimal image quality only exacerbates this problem. To illustrate the difficulties encountered when processing benthic images using traditional methods and highlight the benefits of the current framework, we conducted a comparison between the results of RoI extraction using thresholding binarization and our approach. We separated benthic images into four scenes: images with aggregated organisms, complex background including dead shells and debris, isolated organisms, empty images without organisms. A scene classifier was trained with 120 images in each scene category. For demonstration purposes, we only included relatively few labeled images, 4–30, for each scene to train the scene-specific model. A set of 20 randomly selected benthic images were used to test the performance of the model.

Adaptive resolution imaging sonar (ARIS) systems are ideal tools to study pelagic organisms ranging from jellyfish to small fish (Lankowicz et al. 2020; Shahrestani et al. 2020). Sonar images were collected in Chesapeake Bay using the ARIS 1800, a shipboard-mounted imaging system with a downward-looking angle of 30°–45° (Sound Metrics Corp). These cameras have improved resolution and produce near-video quality images of the water column up to a 35 m range from the camera lens, and can resolve objects down to 3 mm. However, sonar images have much lower resolution than optical cameras

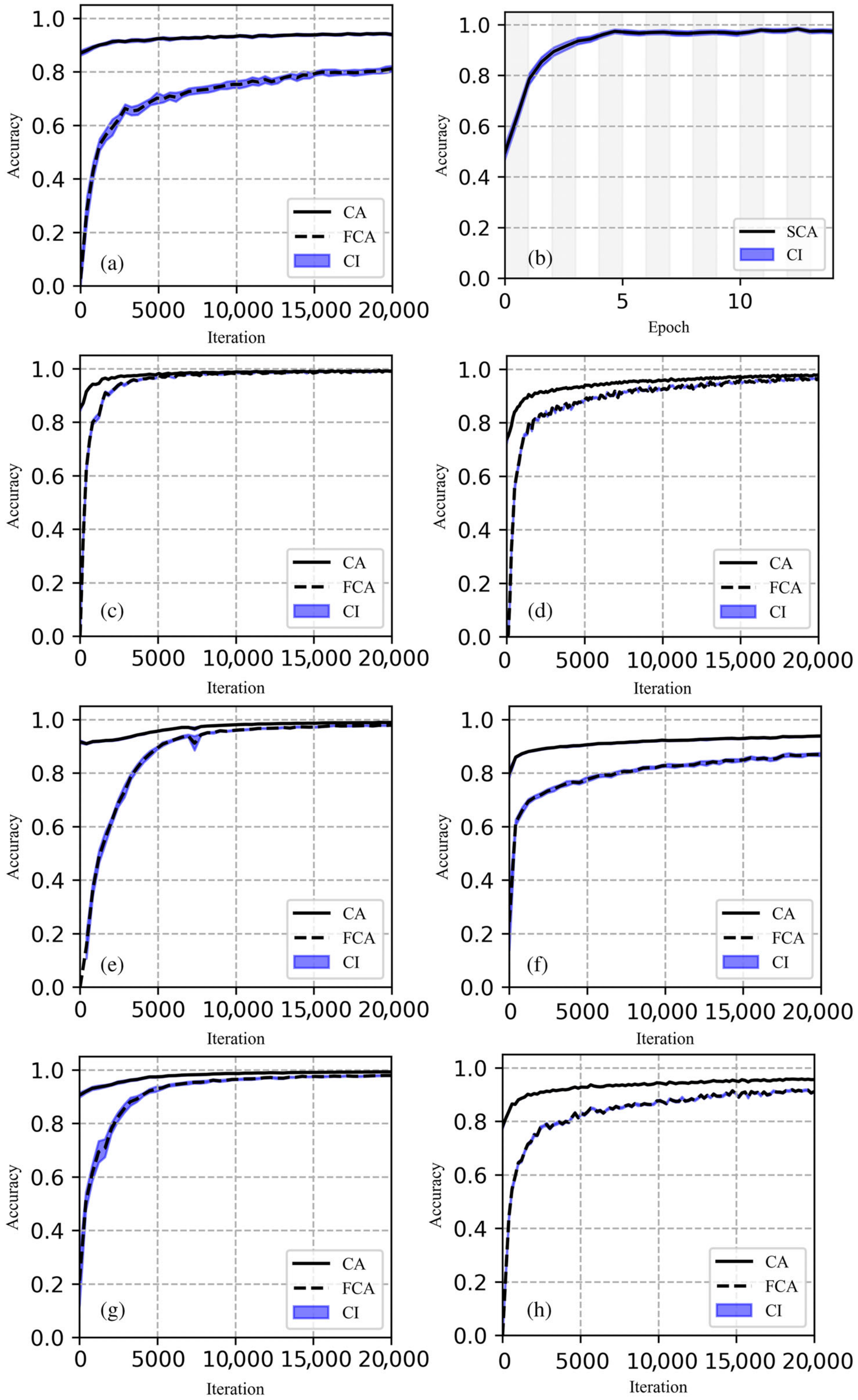


**Fig. 3.** Classification accuracy (CA), foreground classification accuracy (FCA), confidence interval (CI), and scene classification accuracy (SCA) for full model (**a**), scene classification model (**b**), high concentration model (**c**), *Nocticula* model (**d**), *Phaeocystis* model (**e**), Pteropoda model (**f**), Shrimp model (**g**), and low-density model (**h**).

and individual objects including small pelagic fish have few distinct morphological features. In most cases, visual identification relies on auxiliary information, such as schooling shape, size, and other spatial characteristics. For demonstration purposes, we separated sonar images into eight scenes: (1) empty images with only beam patterns; (2) sea floor; (3) fish schools without sea floor; (4) fish school with sea floor; (5) jellyfish without sea floor; (6) jellyfish with sea floor; (7) mysid shrimp swarms with sea floor; and (8) mysid shrimp swarms without sea floor. Note that the sea floor often has a strong signal and is an important feature to separate scenes because it both determines the layout of an image, and also affects image contrast. We also used a set of 20 randomly selected sonar images to test the performance of the proposed framework.

## Results

### Comparison between full model and scene-specific model for plankton images

The full model achieved 73% accuracy in foreground object detection and 87% accuracy in object classification using the labeled underwater plankton dataset (Fig. 3a). In the proposed framework, the accuracy of the initial step, scene classification, reached 98% using the selected full-frame plankton images (Fig. 3b) and the six scene-specific models generally reached higher accuracy than the full model (Fig. 3c–h). The confusion matrix (Fig. 4) suggests that the scene classification model can separate full frames into the corresponding scene with 94–99% accuracy for most scene categories. The frequent co-occurrence of shrimp and *Noctiluca* likely caused the relatively low accuracy in the shrimp scene. Among the scene-specific models, the high-concentration and shrimp-specific models had the highest accuracies, with 98% and 96% foreground classification accuracy and 97% and 96% classification accuracy respectively (Fig. 3c,g). Pteropoda and low-density specific models had the lowest accuracies with 80% and 81% foreground classification accuracy and 89% and 92% classification accuracy. The increased scene-specific model performance suggests that scene classification reduces the variation, or uncertainty, among images.

Both the full model and scene-specific model performed well on large organisms (Figs. 5, 6; Table 3). Full frame images for Figs. 5 and 6 were provided as Supporting Information Appendices 1–12. In the high-concentration scene, both procedures effectively identified and segmented targets from very noisy and low contrast backgrounds) with the scene-specific approach performing slightly better than the full model. In the shrimp scene, both procedures identified and segmented shrimp correctly with the full model yielding more false positives for Chaetognatha, Copepoda, Medusae, and Shrimp groups, while the full model found more *Noctiluca* than the shrimp scene-specific model (Fig. 5c,d; Table 3). This disparity was caused by the larger sample of labeled *Noctiluca* in the full model library (4950 individuals) than in the shrimp scene-specific library (782 individuals; Table 1). In the low-density scene, both procedures performed well (Fig. 5e,f) except that the low-density scene-specific model tended to miss targets due to misclassification (Table 3).

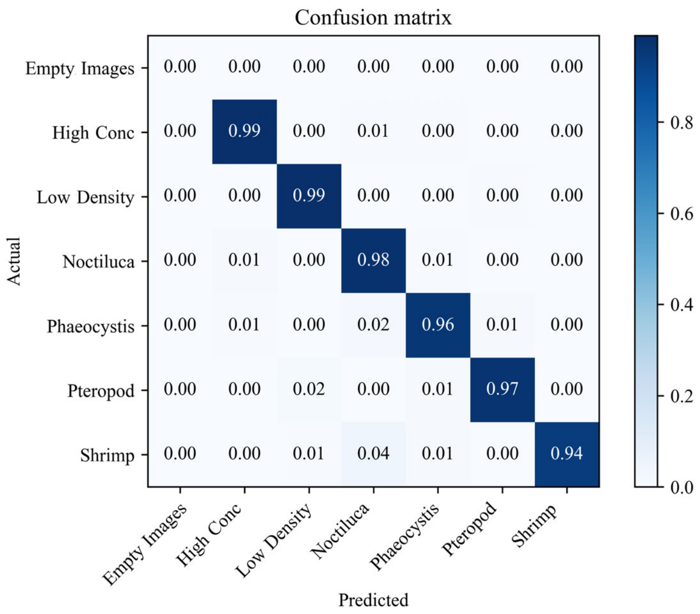
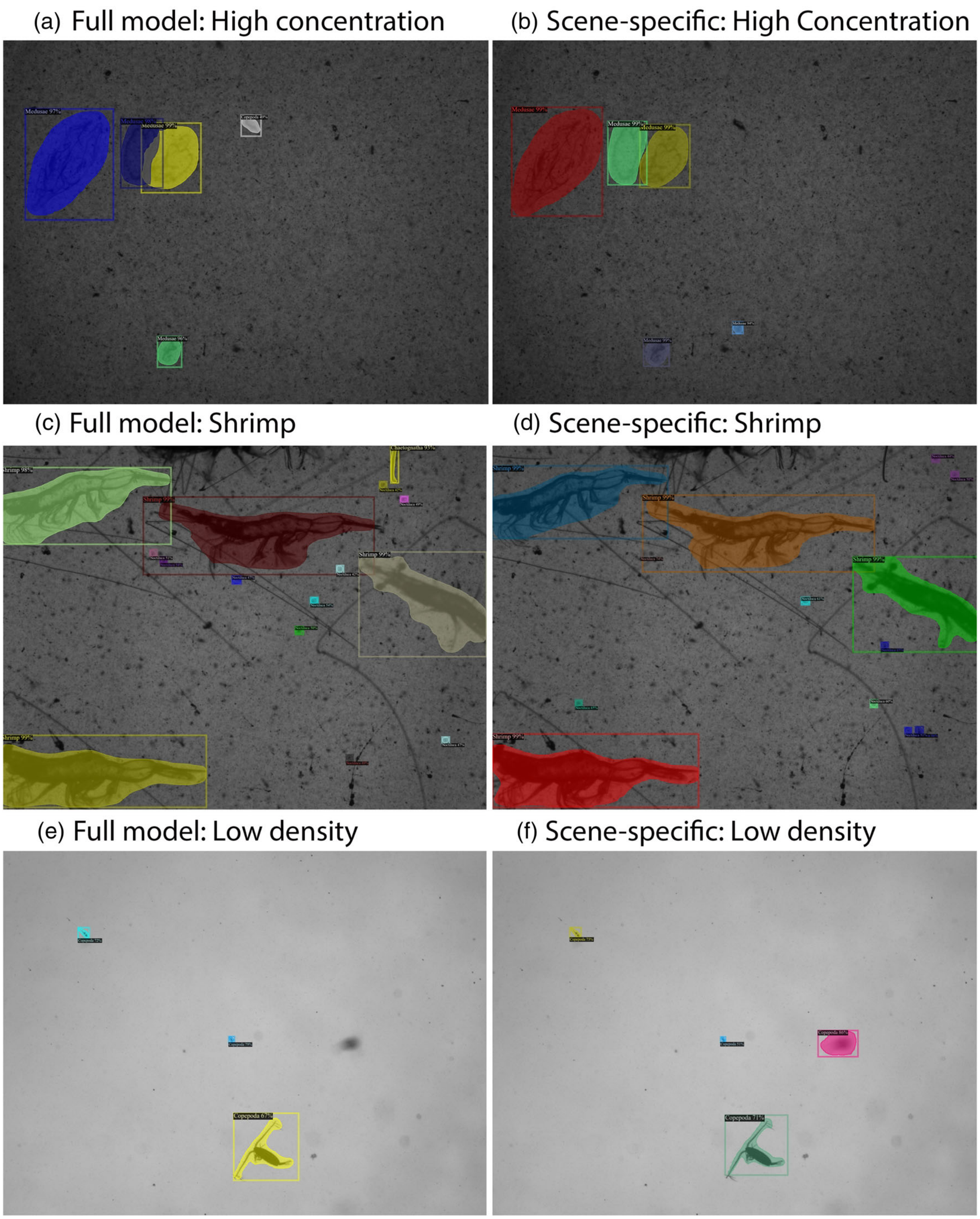


**Fig. 4.** Normalized confusion matrix for the scene classification model with each row representing an actual class example and each column representing the state of a predicted class.

When image complexity increased, the scene-specific models generally outperformed the full model with improved precision and reduced missing targets (Fig. 6; Table 3). As an example, in the 20 testing images depicting the *Noctiluca* scene (Fig. 6a,b), the full model and scene-specific model identified 5348 and 1488 *Noctiluca* individuals, respectively. However, visual counts resulted in 1595 individuals being observed. This discrepancy in the number of detected *Noctiluca* individuals contradicts the fact that the full model library contains approximately 26% more labeled *Noctiluca* individuals compared to the library specific to the *Noctiluca* scene (Fig. 6a,b). In the 20 testing images representing the *Phaeocystis* scene, the full model and scene-specific model detected 363 and 666 *Phaeocystis* colonies, respectively. However, visual counts revealed a total of 672 *Phaeocystis* colonies (Fig. 6c,d). Within the 20 testing images of the *Pteropoda* scene, the 2 dominant groups were *Creseis* and *Lyngbya*. The full model identified 68 *Creseis* and 691 *Lyngbya*. Conversely, the scene-specific model also identified 68 *Creseis*, matching the count from the full model, but detected 948 *Lyngbya*, which is 37% more than the count from the full model (Fig. 6e,f).

The accuracy of the scene-specific model for other small organisms, like copepods, is also much higher than the full model (Table 3). The full inclusion model approach also tends to merge multiple RoIs for small organisms, such as *Lyngbya* clusters, which were often falsely recognized as a single *Lyngbya* individuals (Fig. 6e,f). The occurrence of merged RoIs increased as we increased the number of large organisms, such


(a) Full model: High concentration

(b) Scene-specific: High Concentration

(c) Full model: Shrimp

(d) Scene-specific: Shrimp

(e) Full model: Low density

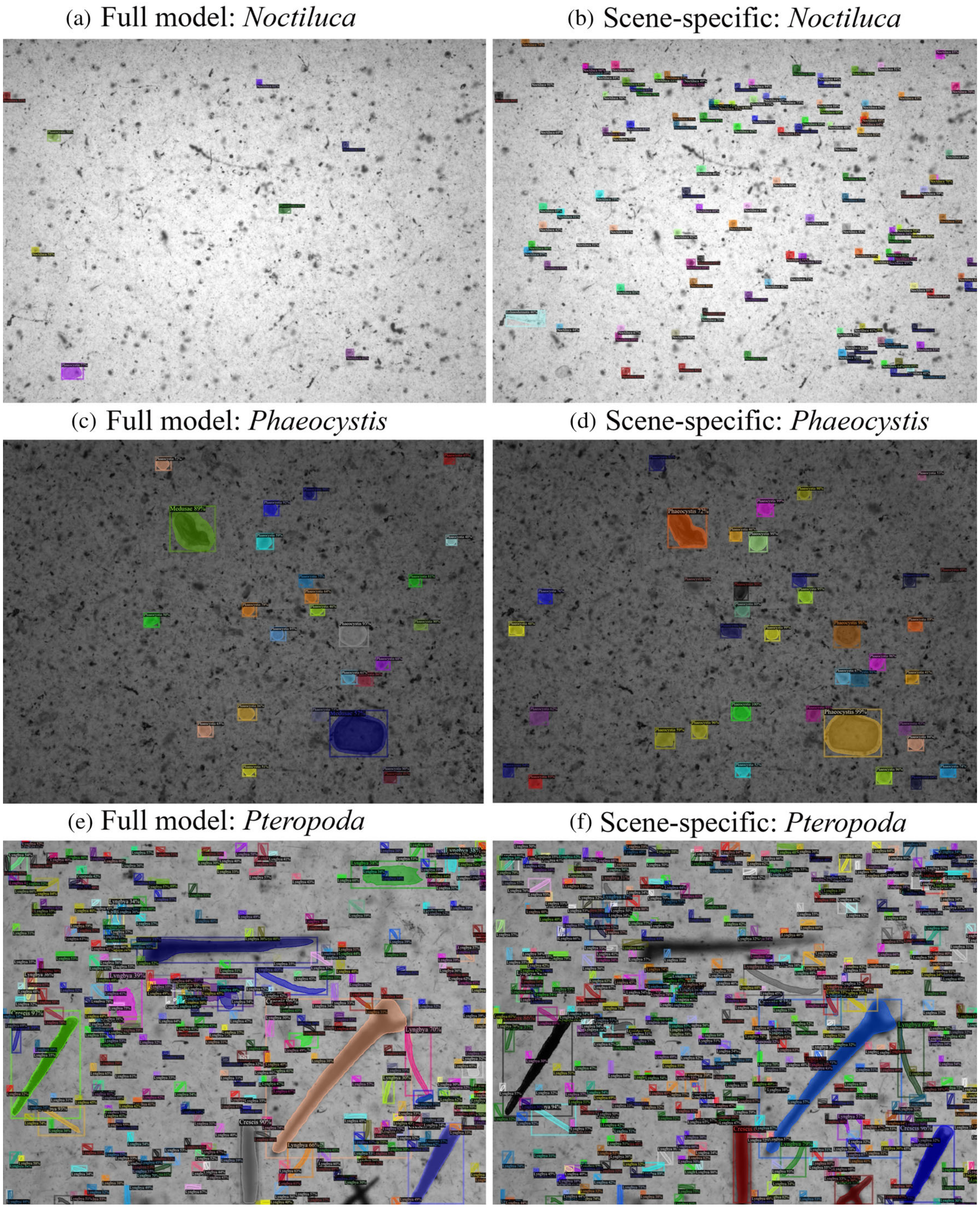
(f) Scene-specific: Low density

**Fig. 5.** Examples of processed underwater plankton images from three different scene categories using the full model and scene-specific models.

as mysid shrimp, in the library. However, the scene-specific models effectively avoided the merging RoIs issue.

### Examples of processing benthic and sonar images

Effectively extracting RoIs is a necessary, but challenging, first step in image recognition. Using benthic images as an example, it was difficult to extract intact RoIs from the background using traditional binarization methods, like adaptive thresholding, because the resulting RoIs were disconnected and fragmented into small pieces (Fig. 7a,b). For images with aggregated benthic organisms (Fig. 7c), the traditional binarization approach was even less effective and failed to yield any



(a) Full model: *Noctiluca*

(b) Scene-specific: *Noctiluca*

(c) Full model: *Phaeocystis*

(d) Scene-specific: *Phaeocystis*

(e) Full model: *Pteropoda*

(f) Scene-specific: *Pteropoda*

**Fig. 6.** Example of processed underwater plankton images with high complexity.

meaningful results. By comparison, the proposed approach performed well on low-contrast benthic images (Fig. 7c,d). The number of images, scene categories, identified taxonomic groups, and labeled individuals for the full model and scene specific models are provided in Table 4. For the full model, the foreground classification accuracy and classification accuracy reached 99% and 95%, respectively. For the scene-specific models, the accuracy of scene classification model reached 98%, the foreground classification accuracies ranged 97–99%, and classification accuracies ranged 96–97% (Table 4). When we applied the trained models to the testing benthic images, the full model missed some individuals of the dominant taxonomic groups, $\sim$ 13% of sand

**Table 3.** Results from the processed underwater plankton images using the full model, scene-specific models, and visual counts.*,† The following abbreviations are used: Prec for precision, Chaeto for Chaetognatha, Append for Appendicularia, and Echino for Echinodermata. The shaded and white portions indicate different scene categories alternately. The "/" symbol denotes an unavailable value when the denominator is either zero or unavailable. Note that achieving 100% precision with a limited number of samples may not be statistically representative of the model's overall performance.

| | | Full model | | | Scene specific model | | | |
|---|---|---|---|---|---|---|---|---|
| Scene | Label | Count | Prec. (%) | Miss. | Count | Prec. (%) | Miss. | Visual count |
| High conc. | Chaeto. | 1 | 100 | 0 | 1 | 100 | 0 | 1 |
| | Copepoda | 4 | 25 | 0 | 1 | 100 | 0 | 1 |
| | Medusae | 45 | 89 | 11 | 29 | 100 | 0 | 29 |
| | *Noctiluca* | 1 | 100 | 0 | 1 | 100 | 0 | 1 |
| | *Phaeocystis* | 19 | 100 | −13 | 16 | 100 | −16 | 32 |
| | Shrimp | 3 | 100 | 0 | 3 | 100 | 0 | 3 |
| Noctiluca | Append. | 2 | 0 | 0 | 0 | / | 0 | 0 |
| | Chaeto. | 11 | 100 | −2 | 13 | 100 | 0 | 13 |
| | Copepoda | 37 | 48 | −11 | 37 | 76 | −1 | 29 |
| | Echino. | 19 | 100 | −18 | 45 | 82 | 0 | 37 |
| | Medusae | 8 | 13 | 0 | 0 | / | −1 | 1 |
| | *Noctiluca* | 348 | 100 | −1247 | 1488 | 100 | −107 | 1595 |
| | *Lyngbya* | 15 | 100 | / | 0 | / | −15 | 15 |
| | *Phaeocystis* | 10 | 0 | 0 | 0 | / | 0 | 0 |
| | Shrimp | 5 | 20 | 0 | 0 | / | 0 | 1 |
| Phaeocy. | Copepoda | 7 | 86 | −14 | 34 | 62 | 0 | 21 |
| | Cresis | 2 | 0 | 0 | 0 | / | 0 | 0 |
| | Echino. | 1 | 100 | 0 | 2 | 50 | 0 | 1 |
| | *Lyngbya* | 52 | 83 | 0 | 0 | 0 | 0 | 41 |
| | Medusae | 12 | 0 | 0 | 0 | / | 0 | 0 |
| | *Noctiluca* | 16 | 69 | 0 | 0 | / | −11 | 11 |
| | *Phaeocystis* | 363 | 100 | −309 | 666 | 100 | −6 | 672 |
| | Shrimp | 2 | 50 | 0 | 3 | 33 | 0 | 1 |
| | Thaliacea | 7 | 0 | 0 | 0 | - | 0 | 0 |
| Pteropoda | Append. | 7 | 43 | −4 | 9 | 56 | −2 | 8 |
| | Copepoda | 105 | 53 | −45 | 128 | 74 | −4 | 99 |
| | *Creseis* | 68 | 100 | 0 | 68 | 100 | 0 | 68 |
| | Echino. | 1 | 0 | 0 | 0 | | 0 | 0 |
| | *Lyngbya* | 691 | 100 | 257 | 948 | 100 | 0 | 948 |
| | Medusae | 1 | 0 | 0 | 0 | / | 0 | 0 |
| | *Noctiluca* | 6 | 17 | 0 | 0 | / | −1 | 1 |
| | *Phaeocystis* | 6 | 0 | 0 | 0 | / | 0 | 0 |
| | Shrimp | 1 | 0 | 0 | 0 | / | 0 | 0 |
| Shrimp | Chaeto. | 1 | 0 | 0 | 0 | / | 0 | 0 |
| | Copepoda | 3 | 0 | 0 | 0 | / | 0 | 0 |
| | Echino. | 1 | 50 | 0 | 0 | / | −1 | 1 |
| | Fish | 2 | 100 | 0 | 2 | 100 | 0 | 2 |
| | Medusae | 1 | 0 | 0 | 0 | / | 0 | 0 |
| | *Noctiluca* | 160 | 100 | 0 | 80 | 100 | −80 | 160 |
| | Shrimp | 51 | 100 | 4 | 47 | 100 | 0 | 47 |
| Low den. | Append. | 2 | 0 | −3 | 3 | 33 | −2 | 2 |
| | Chaeto. | 1 | 100 | 0 | 2 | 50 | 0 | 1 |
| | Copepoda | 32 | 81 | 0 | 28 | 86 | −1 | 25 |

*(Continues)*

**Table 3.** Continued

| | | Full model | | | Scene specific model | | | |
|---|---|---|---|---|---|---|---|---|
| Scene | Label | Count | Prec. (%) | Miss. | Count | Prec. (%) | Miss. | Visual count |
| | *Creseis* | 0 | / | 0 | 2 | 0 | 0 | 0 |
| | *Lyngbya* | 0 | / | 0 | 8 | 75 | 0 | 6 |
| | Medusae | 0 | / | −2 | 0 | / | −2 | 2 |
| | *Noctiluca* | 8 | 0 | 0 | 0 | / | 0 | 0 |
| | Thaliacea | 4 | 75 | 0 | 1 | 100 | −2 | 3 |

*Negative missing values indicate individuals that were not detected and misclassified; positive missing values indicate possible over-segmentation and dual labeled objects in which an object was recognized as more than one taxonomic group based on a 40% probability threshold.

†For small semi-transparent organisms like *Noctiluca* and *Lyngbya*, visual counts were performed based on the model output in which identified organisms were labeled.

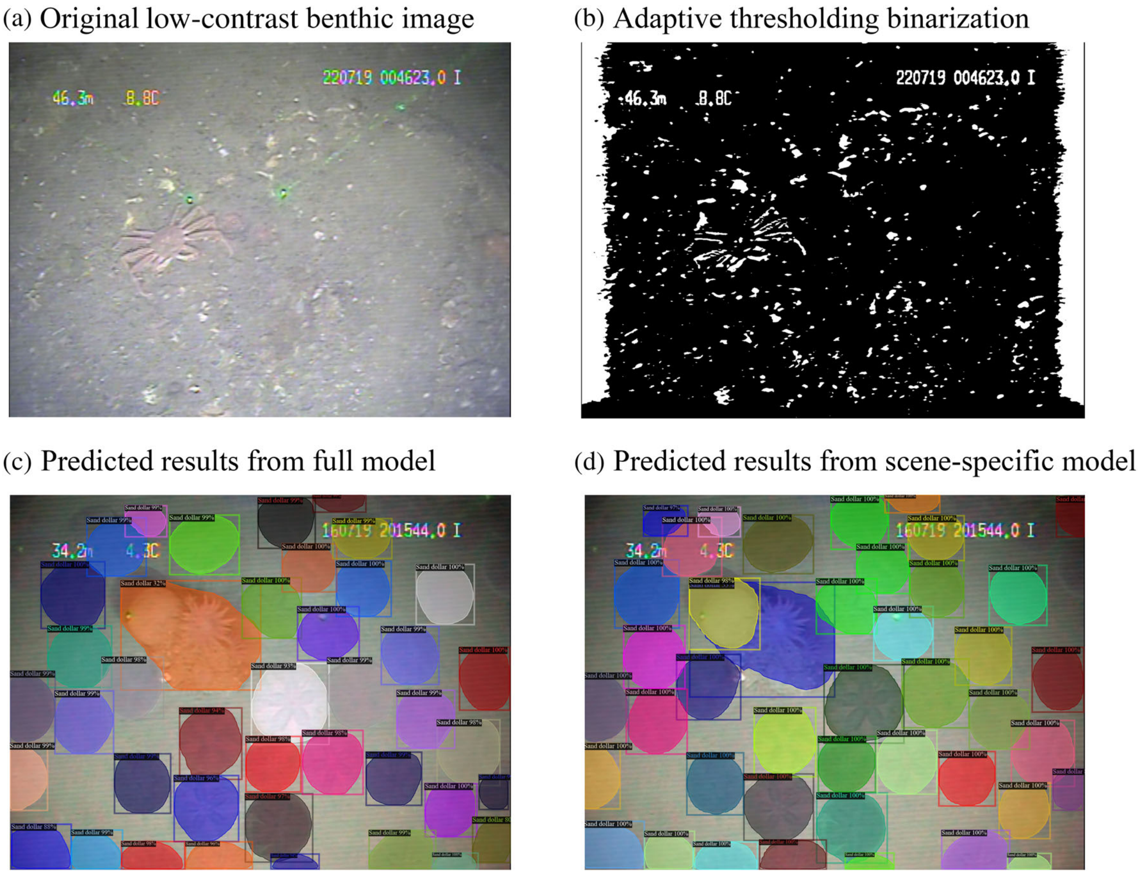

(a) Original low-contrast benthic image
(b) Adaptive thresholding binarization
(c) Predicted results from full model
(d) Predicted results from scene-specific model

**Fig. 7.** Benthic video images: (**a**) an original benthic video image with a crab, (**b**) binarized image using an adaptive thresholding approach to segment potential regions of interest, (**c**) predicted results for an aggregated scene image using the full model, and (**d**) predicted results for the same aggregated scene image using scene-specific model. Note in image (**c**), there are a few sand dollars that were missed and a merge of sand dollars.

dollar, ~ 60% of brittle star, and ~ 67% of sea anemone (Table 4). For testing images, the scene-specific model occasionally misclassified sea anemones as shells, and vice versa. Given the small number of labeled organisms in the library, we expect that the performance of both models to improve rapidly as the number of labeled organisms increases.

Sonar images are often in relatively low resolution and lack details at individual level (Fig. 8a,d). In many cases, soft-bodied jellyfish were obscured, with only the bell portion of their bodies visible. The number of images, scene categories, identified taxonomic groups, and labeled individuals are provided in Table 5. For the full model, the foreground

**Table 4.** Number of benthic images, identified taxonomic groups, number of labeled individuals and accuracies for the full model and scene-specific models. The white portion in the table displays information on model training, while the shaded portion shows the testing results.

| | | Model | | | | Testing results | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Full | | | Scene | | | Visual |
| | | Full | Aggregated | Complex | Isolated | Count | Precision (%) | Missing | Count | Precision (%) | Missing | Count |
| Full frame images | Number of images | 31 | 6 | 10 | 15 | 20 | 100 | | 20 | | | 20 |
| Number of labeled individuals | Anemone | 7 | 0 | 0 | 7 | 2 | 100 | 4 | 6 | 83 | 1 | 6 |
| | Brittle Star | 11 | 0 | 0 | 11 | 8 | 100 | 12 | 21 | 95 | 1 | 20 |
| | Crab | 8 | 1 | 0 | 7 | 5 | 100 | 1 | 2 | 100 | 4 | 6 |
| | Coral | 6 | 1 | 0 | 5 | 0 | / | / | 0 | / | / | 0 |
| | Fish | 1 | 0 | 0 | 1 | 0 | / | 1 | 0 | / | 1 | 1 |
| | Sand dollar | 212 | 212 | 0 | 0 | 376 | 100 | 54 | 430 | 100 | 0 | 430 |
| | Shell | 241 | 0 | 241 | 0 | 4 | 100 | 8 | 12 | 91 | 1 | 12 |
| | Starfish | 2 | 0 | 0 | 2 | 1 | 100 | 0 | 1 | 100 | 1 | 2 |
| | Worm | 1 | 0 | 0 | 1 | 0 | / | / | 0 | / | / | 0 |
| Accuracy | Foreground accuracy (%) | 98 | 98 | 97 | 99 | | | | | | | |
| | Classification accuracy (%) | 95 | 97 | 96 | 96 | | | | | | | |

classification accuracy and classification accuracy reached 94% and 97%, respectively. For the scene-specific models, scene classification accuracy was 96%, the foreground classification accuracies ranged from 92% to 100%, and classification accuracies ranged from 90% to 99% (Table 5). When both models were applied to the testing images, the full model often had trouble distinguishing mysid swarms from small pelagic fish schools (Fig. 8b,e); it tended to confuse small forage fish schools and mysid schools (Fig. 8e). The scene-specific model successfully identified mysid swarms (Fig. 8c) and enumerated small forage fish within the small pelagic fish school (Fig. 8f). In summary, the full model overestimated the number of small fish, jellyfish and mysid swarms, while results from the scene-specific model were consistent with visual counts (Table 5).

## Discussion

The new approach takes advantage of recent progress in artificial intelligence and is designed to address several challenges related to the distribution patterns of marine organisms in different environments. The inclusion of RPN significantly increased our ability to separate each RoI from its background. The RPN first generates a set of region proposals for each object, classifies each proposed region as foreground or background, and finally it produces the best fit region proposal for each object. The RPN model also identified nontarget objects, and therefore could effectively reduce the number of objects that needed to be classified by the following CNN model. This initial step leads to subsequent increases in accuracy, reduced false positives, and reduced computational demands. Results suggest that the RPN approach is an ideal candidate as a common approach for object detection in different types of images, from marine biome.

When compared to a single, unbiased Mask R-CNN model for all images, the advantages of the proposed framework are threefold. First, a single unbiased classifier often oversamples rare groups resulting in more false positives and undersamples abundant groups leading to more false negatives. A scene-specific classification model for each scene takes into account the inherently patchy distribution of marine organisms meaning that dominant taxon could be different in different images. A scene-specific model for each scene also allows a better match in data distribution between samples and libraries, and therefore increases accuracy in both object extraction and recognition. Second, the scene specific approach provides a better method for dealing with large size differences among objects, for example, an algal cell vs. an adult krill. A single unbiased classifier underestimates or completely misses small organisms. Third, the scene specific model reduces the variation and uncertainty among images by separating images with similar layout into the same scene which subsequently improves the model performance.

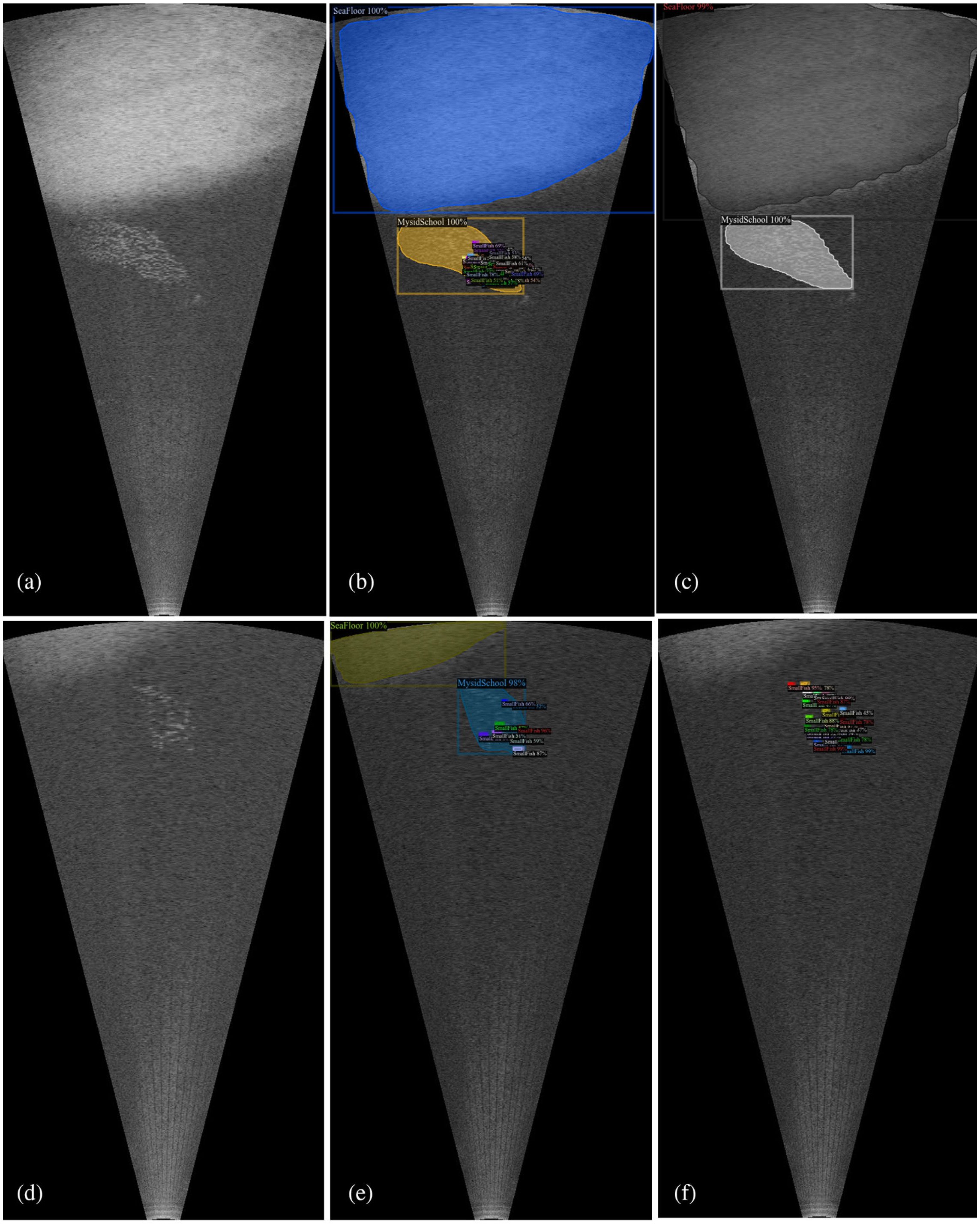


**Fig. 8.** Sonar images: (**a**) a sonar image showing seafloor and a swarm of mysids in the middle water column, and (**b**) predicted results from the full model, (**c**) predicted results from the scene-specific result; (**d**) seafloor and a near bottom school of small forage fish, (**e**) predicted results from the full model, and (**f**) predicted results from the scene-specific model.

The choice of Mask R-CNN as the backbone approach because it stands out as an accurate and versatile approach, making it ideal for tasks requiring precise instance segmentation (He et al. 2017). In comparison to popular one-shot detectors like You Only Look Once (YOLO), where object detection is performed by convolving the full frame image (Redmon et al. 2016), Mask R-CNN employs a common convolution framework for both full frame images and region

**Table 5.** Number of sonar images, identified taxonomic groups, labeled individuals and accuracies for the full model and scene-specific models. The white portion in the table displays information on model training, while the shaded portion shows the testing results.

| | Full frame images | Scenes | | | | Accuracy | |
|---|---|---|---|---|---|---|---|
| | Number | Sea floor | Small fish | Mysid swarm | Jellyfish | Foreground accuracy (%) | Classification accuracy (%) |
| Full model | 249 | 154 | 6346 | 159 | 130 | 94 | 97 |
| EmptyFloor | 30 | 30 | 0 | 0 | 1 | 100 | 99 |
| FishFloor | 29 | 29 | 1757 | 0 | 0 | 94 | 90 |
| FishNoFloor | 16 | 0 | 2949 | 0 | 8 | 92 | 91 |
| JellyfishFloor | 30 | 30 | 683 | 9 | 40 | 98 | 96 |
| JellyfishNoFloor | 29 | 0 | 613 | 2 | 55 | 100 | 98 |
| MysidFloor | 61 | 61 | 154 | 74 | 0 | 100 | 99 |
| MysidNoFloor | 51 | 0 | 0 | 74 | 26 | 100 | 99 |
| Full model | 20 | 17 | 634 | 5 | 6 | | |
| Scene model | 20 | 4 | 487 | 4 | 4 | | |
| Visual counts | 20 | 17 | 503 | 4 | 4 | | |

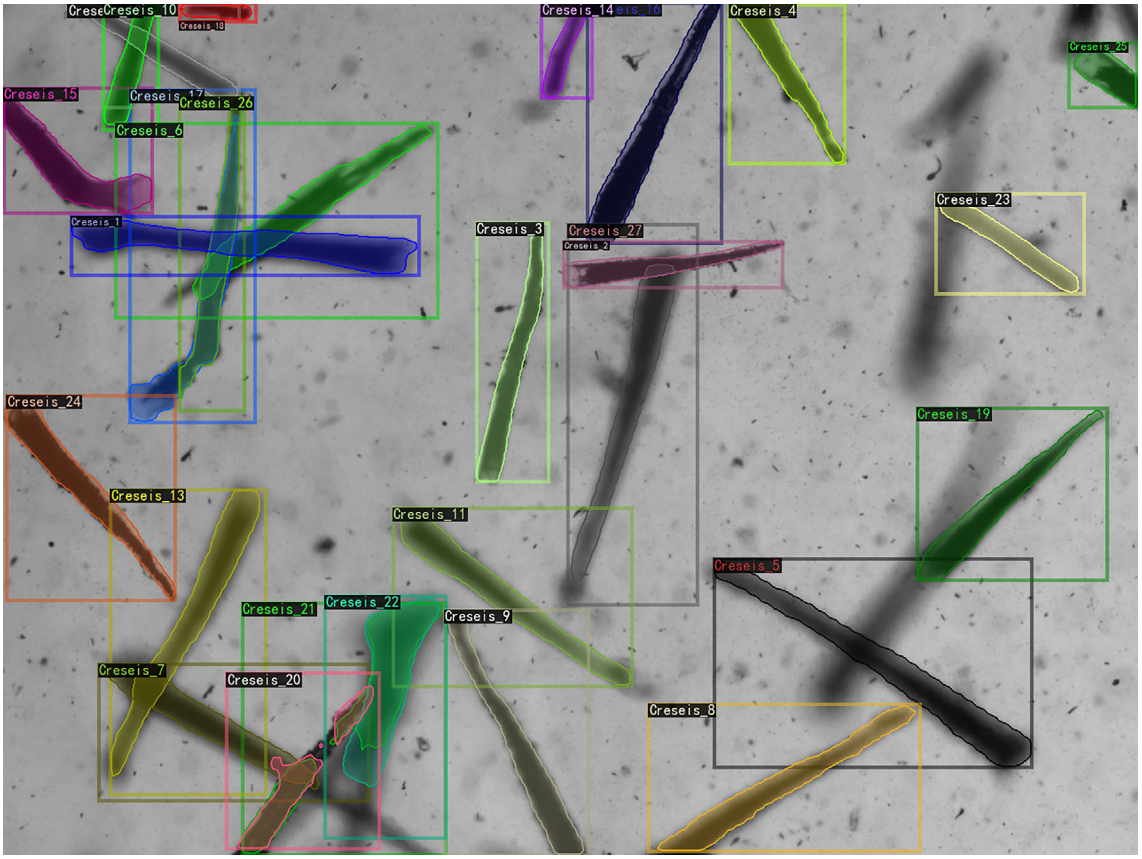


**Fig. 9.** A plankton image dominated by *Creseis acicula* to illustrate the overlapping issue. In total, there were 30 individuals by visual examination and 5 were either partially presented or out of focus. The scene-specific model recognized 26 individuals with one over-segmented individual.

proposals. One-shot detectors are often faster, making them suitable for real-time applications and objects with well-defined edges and features (Prasetyo et al. 2020; Sumit et al. 2020). However, they may sacrifice fine-grained details in object localization and segmentation (Huang et al. 2022; Muñoz-Benavent et al. 2022; Xu et al. 2022). For underwater plankton images, which frequently feature complex noisy backgrounds and overlapping individuals, a robust two-stage approach like Mask R-CNN proves to be an ideal choice. Nonetheless, it is worth noting that our preliminary experiments using YOLO showed similar results: over 100 times faster processing speed than Mask R-CNN, excellent results for small fish, shrimp, and even small copepods, but relatively weaker results for *Noctiluca*, which has faint edges. We envision that both approaches have great value for processing marine environment images with different purposes, such as real-time applications vs. post-processing. Depending on the specific requirements of the task, one can choose the most suitable algorithm to achieve their objectives effectively.

The Mask R-CNN approach also addresses the long-standing issue of broken and overlapping objects. Traditional binarization approaches often generate broken objects resulting in over segmentation which leads to more errors during classification as illustrated in the benthic images. The RPN used in the new approach has a clear advantage over other segmentation approaches, such as thresholding binarization. With advances in optics leading to significant increases of camera depth of field in order to increase the imaging volume, the likelihood of overlapping objects becomes a more pronounced issue. The new approach preserves the advantages of depth of field by computationally reducing the degree of overlap between objects (Fig. 9). In a test image featuring numerous overlapping *Creseis acicula*, a total of 31 individuals were present, among which 5 individuals were either out of focus or only partially visible. The program exhibited high accuracy in recognizing most individuals, and the only instance of over-segmentation occurred with a single individual, resulting in a final count of 27 individuals. Furthermore, the new approach leads to a notable reduction in false positives by integrating a simple binary classifier within the RoI segmentation step. This refinement results in fewer proposed regions being fed to the CNN model, contributing to improved accuracy and efficiency.

The integration of scene classification can significantly reduce the number of labeled images required for training a model for image recognition. The level of variation refers to the diversity and complexity of the images in the dataset, such as differences in lighting conditions, backgrounds, orientations, object sizes, and poses. When images in the dataset exhibit low variation, meaning they share similar characteristics and patterns, a smaller number of labeled images may be sufficient for training the model. The model can generalize well to unseen data and achieve good results as shown by the present study, as it can learn from a relatively small representative set of images. Conversely, high variation among images demands a larger number of labeled examples for successful model training. More labeled images are required to ensure that the model can learn a wide range of features and patterns, making it more robust and capable of handling different scenarios. The number of labeled images is intertwined with the complexity of the intended task. For example, training a model to recognize large organisms will require fewer labeled individuals than training a model for recognition and segmentation tasks involving organisms like *Noctiluca* and small jellyfish with weak features and edges. On the computational side, scene classification was performed using a lightweight network, ShuffleNet, which requires less than 1% of the time needed by the Mask-RCNN to process a single image. The model training time for the ShuffleNet model is significantly shorter compared to that of the Mask-RCNN model.

The proposed framework can process different types of images as illustrated in the present study: microscopic, complex plankton images, benthic video images, and sonar images. Furthermore, this unified framework fundamentally addresses the issue of customized system-specific algorithms and provides an opportunity to compare different systems within the same image processing framework. The future development of our imaging processing framework will unlock the potential for real-time observation of plankton using PlanktonScope. A notable example of our success in this area is the deployment of a PlanktonScope system along the Yangjiang coast since 2021, positioned in front of a nuclear power plant cooling water intake. This system has been instrumental in providing real-time data on plankton density, with a specific focus on mysid shrimp, which is crucial as mysid swarms could potentially clog cooling water intake (Bi et al. 2022). Furthermore, we have also deployed another PlanktonScope system in Chesapeake Bay since February 2023. This deployment aims to monitor plankton dynamics in the mid-stem of the Bay, contributing valuable insights to the ecosystem dynamics. The availability of real-time plankton monitoring data serves as invaluable ecosystem indicators, supplying essential information for decision-makers and managers to make informed choices. In conclusion, our proposed framework stands to significantly streamline the deployment of imaging systems in ecological studies by facilitating swift bulk image processing and extraction of pertinent ecological information.

## References

Bi, H., and others. 2015. A semi-automated image analysis procedure for in situ plankton imaging systems. PloS One **10**: e0127121. doi:10.1371/journal.pone.0127121

Bi, H., and others. 2022. Temporal characteristics of plankton indicators in coastal waters: High-frequency data from PlanktonScope. J. Sea Res. **189**: 102283. doi:10.1016/j.seares.2022.102283

Buda, M., A. Maki, and M. A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. **106**: 249–259. doi:10.1016/j.neunet.2018.07.011

Cheng, K., X. Cheng, Y. Wang, H. Bi, and M. C. Benfield. 2019. Enhanced convolutional neural network for plankton identification and enumeration. PloS One **14**: e0219570. doi:10.1371/journal.pone.0219570

Cheng, X., K. Cheng, and H. Bi. 2020. Dynamic downscaling segmentation for Noisy, low-contrast in situ underwater plankton images. IEEE Access **8**: 111012–111026. doi:10.1109/ACCESS.2020.3001613

Cooper, L. W., and others. 2019. A video seafloor survey of epibenthic communities in the Pacific Arctic including Distributed Biological Observatory stations in the northern Bering and Chukchi seas. Deep-Sea Res. II Top. Stud. Oceanogr. **162**: 164–179. doi:10.1016/j.dsr2.2019.05.003

Durden, J. M., and others. 2016. Perspectives in visual imaging for marine biology and ecology: From acquisition to understanding. Oceanogr. Mar. Biol. Annu. Rev. **54**: 1–72. doi:10.1201/9781315368597

Girshick, R. 2015. Fast R-CNN, p. 1440–1448. *In* Proceedings of the IEEE International Conference on Computer Vision. doi:10.1109/ICCV.2015.169

González, P., A. Castaño, E. E. Peacock, J. Díez, J. J. Del Coz, and H. M. Sosik. 2019. Automatic plankton quantification using deep features. J. Plankton Res. **41**: 449–463. doi:10.1093/plankt/fbz023

He, K., G. Gkioxari, P. Dollár, and R. Girshick. 2017. Mask R-CNN, p. 2961–2969. *In* Proceedings of the IEEE International Conference on Computer Vision. doi:10.1109/ICCV.2017.322

Huang, H., X. A. Feng, J. Jiang, P. Chen, and S. Zhou. 2022. Mask RCNN algorithm for nuclei detection on breast cancer histopathological images. Int. J. Imaging Syst. Technol. **32**: 209–217. doi:10.1002/ima.22618

Irisson, J.-O., S.-D. Ayata, D. J. Lindsay, L. Karp-Boss, and L. Stemmann. 2022. Machine learning for the study of plankton and marine snow from images. Ann. Rev. Mar. Sci. **14**: 277–301. doi:10.1146/annurev-marine-041921-013023

Lankowicz, K. M., H. Bi, D. Liang, and C. Fan. 2020. Sonar imaging surveys fill data gaps in forage fish populations in shallow estuarine tributaries. Fish. Res. **226**: 105520. doi:10.1016/j.fishres.2020.105520

Levinton, J. S. 1995. Marine biology: Function, biodiversity, ecology. Oxford Univ. Press.

Liu, D., K. Ying, Z. Cai, H. Huang, and H. Bi. 2021. Outburst of *Creseis acicula* in southwest Daya Bay in July 2020. Ocean. Limnol. Sin. **52**: 1438–1447.

Luo, J. Y., and others. 2018. Automated plankton image analysis using convolutional neural networks. Limnol. Oceanogr. Methods **16**: 814–827. doi:10.1002/lom3.10285

Ma, N., X. Zhang, H.-T. Zheng, and J. Sun. 2018. Shufflenet v2: Practical guidelines for efficient CNN architecture design, p. 116–131. *In* Proceedings of the European Conference on Computer Vision (ECCV). Springer, Cham. doi:10.1007/978-3-030-01264-9_8

MacLeod, N., M. Benfield, and P. Culverhouse. 2010. Time to automate identification. Nature **467**: 154–155. doi:10.1038/467154a

Minaee, S., Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos. 2021. Image segmentation using deep learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **44**: 1. doi:10.1109/TPAMI.2021.3059968

Mittal, S., S. Srivastava, and J. P. Jayanth. 2022. A survey of deep learning techniques for underwater image classification. IEEE Trans. Neural Netw. Learn. Syst. **34**: 1–15. doi:10.1109/TNNLS.2022.3143887

Muñoz-Benavent, P., and others. 2022. Impact evaluation of deep learning on image segmentation for automatic bluefin tuna sizing. Aquac. Eng. **99**: 102299. doi:10.1016/j.aquaeng.2022.102299

Orenstein, E. C., and others. 2020. The Scripps plankton camera system: A framework and platform for in situ microscopy. Limnol. Oceanogr. Methods **18**: 681–695. doi:10.1002/lom3.10394

Pal, N. R., and S. K. Pal. 1993. A review on image segmentation techniques. Pattern Recogn. **26**: 1277–1294. doi:10.1016/0031-3203(93)90135-J

Piechaud, N., C. Hunt, P. F. Culverhouse, N. L. Foster, and K. L. Howell. 2019. Automated identification of benthic epifauna with computer vision. Mar. Ecol. Prog. Ser. **615**: 15–30. doi:10.3354/meps12925

Prasetyo, E., N. Suciati, and C. Fatichah. 2020. A comparison of YOLO and Mask R-CNN for segmenting head and tail of fish, p. 1–6. *In* 2020 4$^{th}$ International Conference on Informatics and Computational Sciences (ICICoS). IEEE, doi:10.1109/ICICoS51170.2020.9299024

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. 2016. You only look once: Unified, real-time object detection, p. 779–788. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, doi:10.1109/CVPR.2016.91

Ren, S., K. He, R. Girshick, and J. Sun. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**: 1137–1149. doi:10.1109/TPAMI.2016.2577031

Shahrestani, S., H. Bi, D. Liang, K. Lankowicz, and C. Fan. 2020. Multi-scale spatial dynamics of the Chesapeake Bay nettle, *Chrysaora chesapeakei*. Ecosphere **11**: e03128. doi:10.1002/ecs2.3128

Sharma, S., A. Gosain, and S. Jain. 2022. A review of the oversampling techniques in class imbalance problem, p. 459–472. *In* International Conference on Innovative Computing and Communications. Springer.

Shortis, M., E. H. Abdo, and A. A. Dave. 2016. A review of underwater stereo-image measurement for marine biology and ecology applications, p. 269–304. *In* R. N. Gibson, R. J. A. Atkinson, and J. D. M. Gordon [eds.], Oceanography and marine biology. CRC Press.

Smith, C. J., and H. Rumohr. 2013. Imaging techniques, p. 97–124. *In* A. Eleftheriou and A. McIntyre [eds.], Methods for the study of marine benthos. Blackwell Science.

Solan, M., and others. 2003. Towards a greater understanding of pattern, scale and process in marine benthic systems: A picture is worth a thousand worms. J. Exp. Mar. Biol. Ecol. **285–286**: 313–338.

Song, Y., and H. Yan. 2017. Image segmentation techniques overview, p. 103–107. *In* 2017 Asia Modelling Symposium (AMS). IEEE. doi:10.1109/AMS.2017.24

Song, J., and others. 2020. Early warning of Noctiluca scintillans blooms using in-situ plankton imaging system: An example from Dapeng Bay, P.R. China. Ecol. Indic. **112**: 106123. doi:10.1016/j.ecolind.2020.106123

Song, J., W. Jiao, Z. Cai, and H. Bi. 2022. A two-stage adaptive thresholding segmentation for noisy low-contrast images. Ecol. Inform. **69**: 101632. doi:10.1016/j.ecoinf.2022.101632

Sumit, S. S., J. Watada, A. Roy, and D. Rambli. 2020. Object detection deep learning methods, YOLO shows supremum to Mask R-CNN. J. Phys. Conf. Ser. **1592**: 042086. doi:10.1088/1742-6596/1529/4/042086

Wada, K. 2016. Labelme: Image polygonal annotation with python.

Wang, N., J. Yu, B. Yang, H. Zheng, and B. Zheng. 2020. Vision-based in situ monitoring of plankton size spectra via a convolutional neural network. IEEE J. Ocean. Eng. **45**: 511–520. doi:10.1109/JOE.2018.2881387

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

Wu, Y., A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. 2019. Detectron2.

Xu, X., and others. 2022. Crack detection and comparison study based on faster R-CNN and mask R-CNN. Sensors **22**: 1215. doi:10.3390/s22031215

Zhang, H., L. Huang, C. Q. Wu, and Z. Li. 2020. An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset. Comput. Netw. **177**: 107315. doi:10.1016/j.comnet.2020.107315

Zhang, X., X. Zhou, M. Lin, and J. Sun. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices, p. 6848–6856. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: doi:10.1109/CVPR.2018.00716

## Acknowledgments

We thank Dr. Cooper for providing benthic videos. PlanktonScope images were collected by Bi and Ying in Shenzhen coastal waters, and sonar images were collected by Bi, Lankowicz, and Shahrestani in Chesapeake Bay. X.C., Y.C., and K.Y were supported by the National Key Research and Development Program of China under Grant 2017YFC1403600 and the Shenzhen Science and Technology Innovation Program under grant KCXFZ20211020163557022.

*Submitted 28 January 2023*
*Revised 13 October 2023*
*Accepted 25 October 2023*

*Associate editor: Clare E. Reimers*