Enhancing marine debris identification with convolutional neural networks

Victoria Wahlig[1], Ricardo A. Gonzales[2]

[1] Falmouth High School, Falmouth, Maine

[2] Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom

SUMMARY

Oceanic debris, predominantly plastic, inflicts catastrophic damage on marine ecosystems, threatens aquatic life through entanglement, strangulation, and starvation, and increasingly accumulates as microplastics within the tissues of marine species consumed by humans. Assessing and mitigating this crisis is complicated by the wide-ranging distribution of waste by ocean currents to remote and deep-sea locations. In this study, we developed a deep learning model to discern and identify components of images captured by an underwater remotely operated vehicle (ROV). We hypothesized that the model would achieve an 80% accuracy. We utilized image segmentation, a process that groups or "masks" all pixels associated with a specific object in the image, to recognize and delineate image components (trash, animal, plant, ROV). We trained our model using the established TrashCan 1.0 dataset, which comprises images captured by ROVs in the Sea of Japan. Our model, a convolutional neural network employing U-Net architecture, formulated feature maps for each object within the images, enabling prediction of object classes. During the testing phase, we compared the model-generated object masks against reference masks to establish the Dice Similarity Coefficient (DSC) for each object class (trash = 0.81 ± 0.38, animal = 0.85 ± 0.35, plant = 0.88 ± 0.31, ROV = 0.86 ± 0.29, overall average = 0.84 ± 0.36). The consistent categorization of image components by our model demonstrates its potential as an effective ROV-borne tool capable of identifying and targeting hitherto inaccessible ocean trash. Our study lays a robust foundation for future research and additional applications.

INTRODUCTION

The issue of plastic pollution in the oceans is of considerable global concern (1). Estimates suggest that 8 million metric tons of plastic waste enter marine environments every year, contributing to an existing pile of approximately 200 million metric tons (2). This form of pollution has widespread ecological repercussions, negatively impacting ocean ecosystems, marine life, and even human health via the food chain (3–7). While traditional methods for addressing plastic pollution, such as manual cleanups and landfill disposal, are promising, these methods are not sufficient to tackle the scale of the problem (8). With ocean currents spreading waste over vast areas, including remote and deep-sea locations, the need for innovative, scalable solutions is imperative.

Current waste management methods, particularly in the marine context, fall short in various ways. Manual clean-ups, though commendable in their grassroots engagement, have limitations regarding the amount of area that can be covered and the depth of ocean that can be reached. Technologies like artificial coastlines designed to collect ocean-borne waste are not without flaws either; these are expensive to deploy and maintain, and often result in the unintentional capture of marine life (9). The limitations of existing methods signify the critical need for innovative strategies, specifically those that are capable of targeting waste in a more efficient and ecologically sensitive manner, extending beyond the surface and shallow waters to the deeper ocean.

The use of Remotely Operated Vehicles (ROVs) offers a compelling alternative to current practices. In recent years, some projects, such as SeaClear's fleet of ROVs, have shown promise in coastal waste collection (10). However, these are not without their complexities, involving a network of scouting drones, multiple ROVs, and central control systems (11). The current operational reach of such ROVs is confined to shallow, near-shore waters. Therefore, ongoing research and development are essential for enhancing their capabilities to make them effective in dealing with waste in deeper and more remote areas of the ocean.

Integration of sophisticated imaging technology and machine learning algorithms could significantly enhance the capabilities of ROVs (12). Current underwater vision systems have reached a level of clarity and reliability that makes them a viable input source for machine learning algorithms. These images could be used as input to deep learning models to enable ROVs to perform specific tasks. Utilizing deep learning algorithms, ROVs could be programmed to recognize and classify objects in images captured by these advanced vision systems, with the ability to customize visual input and processing to fit the needs of the deep learning model enhancing trash detection reliability and consistency. Additionally, the incorporation of image segmentation technology into these algorithms can aid ROVs in precisely isolating specific objects without disturbing the surrounding ecosystem. This process assigns different components of an image to specific classes (e.g., trash, plant, animal, background), creating color-specific "masks" for each component, thereby enabling the ROV to act based on the information presented by the segmented image (**Figure 1**). This could allow for more efficient removal of identified trash while logging data about identified marine life, improving the overall effectiveness of ocean cleanup efforts.

In this study, we developed and evaluated a deep learning model capable of identifying components of an ROV-acquired
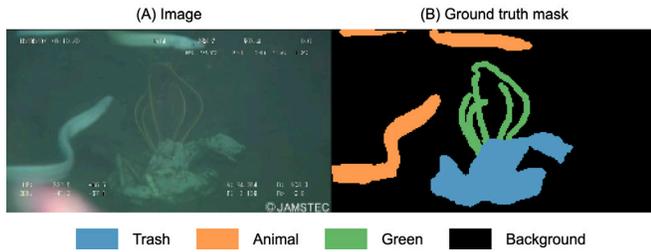
Figure 1: Image Segmentation in Remotely Operated Vehicle (ROV) Photography. (A) An original ROV-captured image (B) with its corresponding manually contoured segmentation mask in their respective colors.

underwater image. We trained the model, which was based on a convolutional neural network utilizing U-Net architecture, using a dataset of ROV-acquired images labeled with colored mask overlays (13). We then tested the model's efficacy on a separate dataset to assess its ability to autonomously identify and categorize image components. The overarching aim of this project was to showcase the potential of outfitting underwater ROVs with convolutional neural networks as a means of identifying and addressing the global issue of oceanic plastic pollution. With a current semantic segmentation algorithm achieving an accuracy of 40% (14), we hypothesized that our model would outperform this by achieving at least 80% accuracy in segmenting distinct underwater objects. Our model employed different, more tailored parameters than the existing algorithm, allowing it to achieve a higher accuracy. The study resulted in a high accuracy across the categories of ROVs, plants, animals, and trash. All categories surpassed the 80% accuracy benchmark, demonstrating the correctness and consistency of our model in a segmentation task.

RESULTS

In this study, we created a deep learning algorithm to detect objects in an image captured by an underwater ROV, with the goal of finding and isolating trash for eventual removal. We utilized a convolutional neural network model developed using the U-Net architecture, which was trained to segment images from the TrashCan dataset (images taken by ROVs) into four distinct classes: trash (with 8 subclasses), animal (with 7 subclasses), plants, and ROV appendages (15). This dataset was composed of 7,212 images of different classes: trash, plants, animals, and ROVs. The annotation of the images to create masks was done in a previous research study (15). This process had the potential for human error and biases, which our trained model would eliminate (as it creates masks based off of clearly visible patterns). However, these annotations are significant enough to establish a valid comparison later, in the statistical analysis. The preprocessing of data ensured uniformity across the dataset, eliminating inconsistencies in file format and aiding in the identification of numerical errors or incorrect file sorting. We used the Dice Similarity Coefficient (DSC) to measure of how well the real mask (created "by hand") matches the predicted mask generated by the model (16). The DSC compares the overlap between the real and predicted masks, using the pixels shared by the two. The higher the DSC, the more accurate

Trash								
fabric	gear	metal	paper	plastic	rubber	wood	others	mean
0.88 ± 0.32	0.83 ± 0.37	0.81 ± 0.38	0.95 ± 0.20	0.65 ± 0.45	0.97 ± 0.16	0.94 ± 0.23	0.48 ± 0.47	**0.81 ± 0.38**

| Animals							Plants	ROV
fish	starfish	shells	crab	eel	others	mean	mean	mean
0.77 ± 0.42	0.85 ± 0.35	0.95 ± 0.21	0.85 ± 0.35	0.86 ± 0.33	0.84 ± 0.36	**0.85 ± 0.35**	**0.88 ± 0.31**	**0.86 ± 0.29**

Table 1: Mean Dice Similarity Coefficients (DSCs) for Individual and Overall Classes. Segmentation performance measured in DSCs (mean ± standard deviation) for each identified class: trash (8 subclasses), animals (7 subclasses), plants, and remotely operated vehicle (ROV), computed from 7,212 real and predicted mask pairs. The overall model across all classes has an average DSC of 0.84 ± 0.36, indicating substantial overlap between the hand-drawn and predicted masks. This data highlights the model's proficiency in recognizing various underwater objects within the images.

the model's segmentation.

The DSC values for each class and subclass, as well as a collective average for all samples, were calculated with both the mean and standard deviation of these values accounted for (**Table 1**). Overall, the model yielded a high segmentation performance of 0.84 ± 0.36 across all samples, and 0.81 ± 0.38, 0.85 ± 0.35, 0.88 ± 0.31, and 0.86 ± 0.29, for trash, animals, plants, and ROV, respectively (**Figure 2**). The distribution of segmentation performance across all samples showed that most values were located between 0.8 and 1.0, indicating high model accuracy as there was considerable overlap between the ground truth and predicted masks (**Figure 3**). This was also demonstrated for each individual aggregated class (**Figure 4**). Although the standard deviation values show some spread in the DSC values, they are within a range that still supports the notion of high performance,
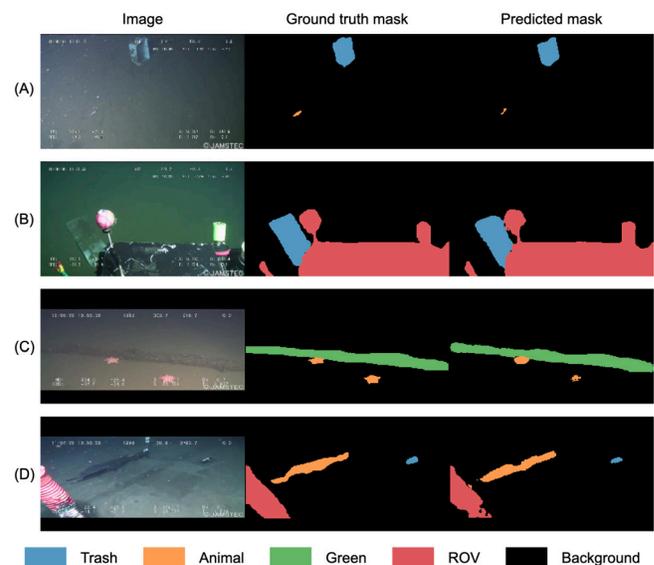
Figure 2: Class-specific Visualized Examples and Model Predictions. Four examples (A-D) with each row consisting of a photograph, a manually contoured ground truth mask, and the corresponding predicted mask generated by our model.
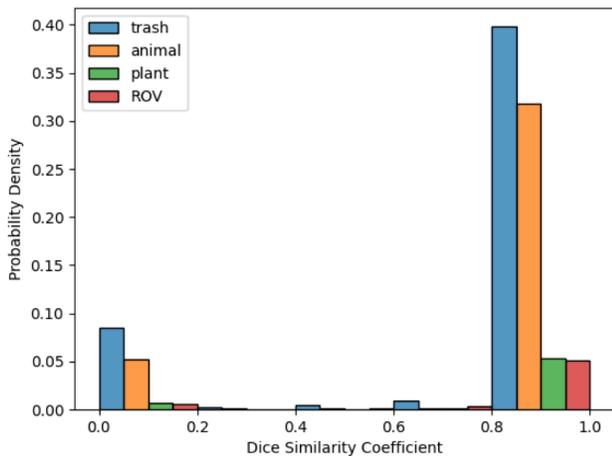
Figure 3: Dice Similarity Coefficients Distribution for Aggregated Classes. Histogram of the distribution of Dice Similarity Coefficients (DSCs) across four main categories: trash, animal, plant, and remotely operated vehicle (ROV). These categories represent an aggregation of the original 16 classes. Each class, represented by a unique color, is subdivided into 5 bins within the DSC range of 0 to 1. The vertical axis represents the probability density.

especially when considered in the context of the overall mean. Overall, the model demonstrated high performance, accurately segmenting objects across all classes with an accuracy greater than 80%.

DISCUSSION

We aimed to train a deep learning model to identify and categorize components in underwater images captured by ROVs, with a focus on detecting plastic pollution. Employing the U-Net architecture of convolutional neural networks, the model demonstrated strong potential in recognizing various components in a dataset of ROV-acquired images. Our study contributes progress towards improved accuracy of underwater object detection and shows the potential benefits of using deep learning algorithms in environmental conservation efforts. When implemented on an ROV (equipped with a camera to provide a reading of the surrounding environment), it may be able to accurately identify trash in a variety of locations. Given the dataset diversity in terms of trash and animal types, it can be used in many ocean ecosystems across the world. As deep underwater environments are all hidden from sunlight and have extreme pressure and low temperature, their ecosystems resemble each other, allowing our AI algorithm to be applied nearly anywhere.

The study revealed a pronounced probability density for higher DSCs across categories like ROVs, plants, and animals, contrasting with the category of trash, which exhibited more variability. The broad range of possible appearances for trash could account for this difference, indicating potential areas for model improvement. We found that sand strewn across objects posed a challenge for the model, suggesting that future iterations might benefit from techniques forcing the model to focus more on shape and other features besides texture. The high accuracy level of over 80%, along with a significant concentration of DSCs towards the higher end

of the scale, suggests that our model is, in most instances, predicting on par with an expert level performance (17). The model also had an inference time of around one second per image, allowing a model-equipped ROV to quickly detect objects in gathered image frames. These findings support the effectiveness and potential of the proposed model in the segmentation and identification of various components within underwater images.

In contrast to the preceding research that employed a mask R-CNN architecture on the TrashCan dataset, our approach leverages an adapted U-Net model (13, 15, 18). Despite the variation in architectural choices, a comparative evaluation revealed a notable enhancement in segmentation performance by our model, successfully categorizing all 16 classes, an advancement from an Average Precision of 0.30 attained in the previous work (14). Although there were minor differences in validation protocols between these models, the overarching similarity in evaluation conditions substantiates a credible comparison. The divergence in Average Precision (AP) reported in the earlier project (AP of 0.30) and the Dice Similarity Coefficient (DSC) achieved in our work (DSC of 0.84) underscores the potential superior performance of our adapted U-Net model in the underwater image segmentation task. The comparative assessment, albeit indirect due to the different metrics, suggests that the architectural modifications and other optimizations in our model could offer a more robust solution for this specific segmentation challenge. This growth in accuracy will strengthen operating ROVs, allowing them to be more effective and precise in trash removal.

For future research, addressing the challenges of texture detection and data collection could improve model performance. As the model is trained to find patterns in data, differences in texture – such as sand, common in underwater images – can throw off the consistency of patterns from one object to a similar one. A proposed solution involves artificially overlaying sand onto objects in training images,
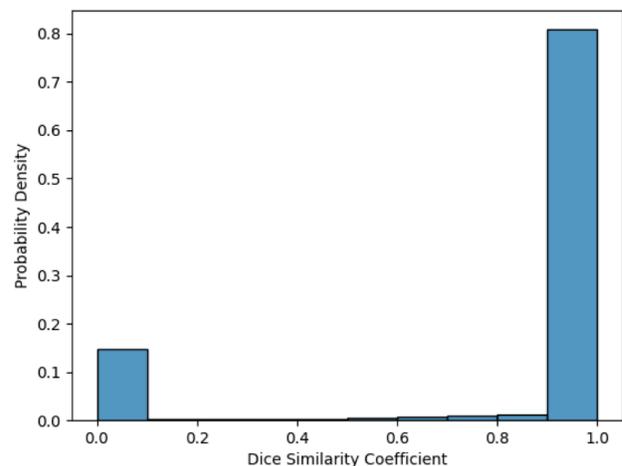
Figure 4: Overall Dice Similarity Coefficients Distribution. Histogram of the probability density distribution of Dice Similarity Coefficients (DSCs) for the entire dataset, without distinction of classes. All instances across the 16 original classes are grouped together, forming a unified category. The DSC range from 0 to 1 is divided into 10 bins, each representing a distinct level of model accuracy.
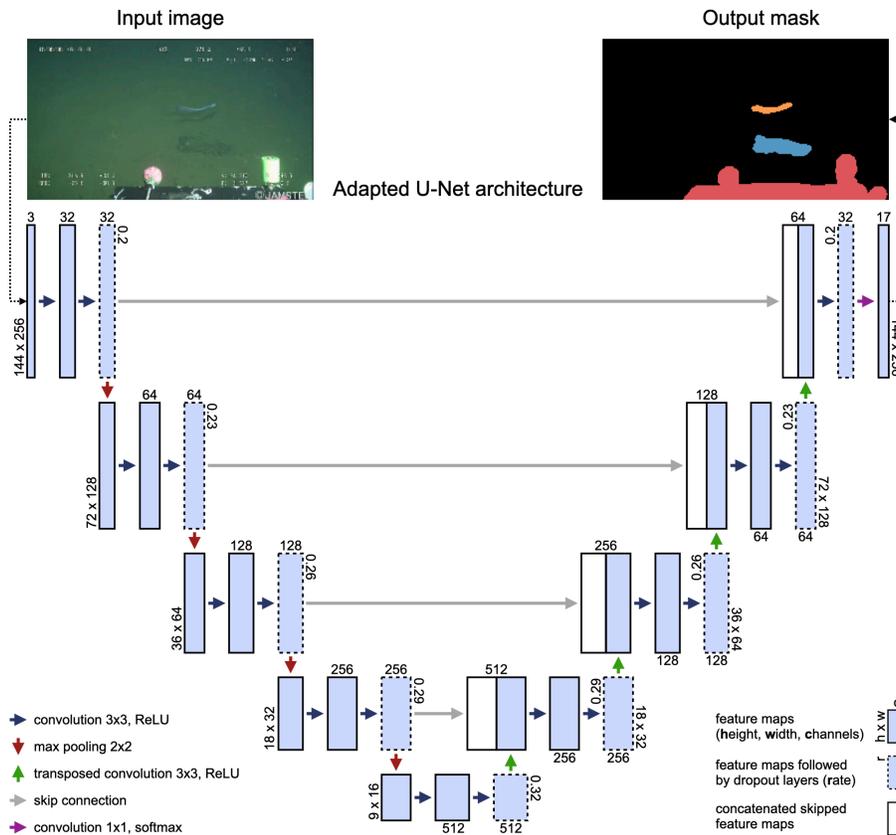
Figure 5: Adapted U-Net model architecture. An encoder-decoder structure with symmetric skip connections, based on U-Net (11). Each encoder stage features 3x3 convolutions, rectified linear unit (ReLU) activations, dropout layers, and 2x2 max pooling. Conversely, the decoder has 3x3 transposed convolutions, ReLU activations, dropout layers, and concatenates with skip connections from the encoder. The final stage uses a 1x1 convolution with a softmax activation for multi-class segmentation. The model processes the input image to generate an output mask, where trash is depicted in blue, animal in orange, remotely operated vehicle in red, and background in black.

thereby compelling the model to focus on shape and object features. The limitations of time required for implementation, training, and data collection, and the paucity of underwater ROV initiatives are areas that future work could also address. Encouraging more research into leveraging deep learning for environmental preservation can stimulate more people to produce optimally efficient models and contribute to broader initiatives against plastic pollution.

MATERIALS AND METHODS
Imaging Dataset
The deep learning model in this study was trained and tested using the dataset TrashCan 1.0 (15). Dataset images came from videos sourced from ROVs deployed in the Sea of Japan. These ROVs were operated by the Japan Agency of Marine Earth Science and Technology, an agency that has gathered data since 1982. This dataset was composed of 7,212 images of different classes, including marine debris, ROVs (e.g., appendages, projections), plants and animals, and trash. Each image was manually ("by hand") annotated with color-specific masks in a previous research study (15). Trash was further sub-categorized into TrashCan-Material (e.g., plastic, paper, rubber, wood) and TrashCan-Instance (e.g., bag, bottle, can, clothing, wrapper). TrashCan 1.0 included testing and validation sets that were used in different

phases of model creation.

Data Preprocessing
The dataset was initially uploaded into Google Drive. After mounting Google Drive and creating directories for file paths, file data was transitioned into a Google Colaboratory document (19). All files shared the same format. For colorful RGB images, which include a color channel for red, green, and blue, the shape was defined as (144, 256, 3) - representing the height, width, and number of color channels. The corresponding masks had a similar shape definition (144, 256, 17). Masks were one-hot encoded, which meant that the input mask had 17 channels (the sixteen classes, plus one for the background). If a mask or image did not fit these specifications, it was resized using bilinear interpolation for the input image and nearest neighbor interpolation for the mask. The file IDs for images and masks were combined and collated into a list. The paired masks and images were then assigned category IDs. These sixteen classes included ROV appendages, plants, animals (fish, starfish, shell, crab, eel, etc.), and trash (fabric, fishing gear, metal, paper, plastic, rubber, wood, etc.). While the model was trained with all sixteen classes, these four larger categories were used to simplify interpretation of results. Each mask and image was then saved as a PNG file. This pre-processing step organized

the development data (which comprised of training and testing sets) into distinct subsections, from which the model later learned.

Image Segmentation

Image segmentation represents a sophisticated approach to image processing, wherein components of input images can be differentiated into various object classes such as animal, plant, ROV, and trash (20). Every pixel, the minute areas of illumination composing an object on a display screen, is assigned a specific identifying color based on the object it contributes to. This pixel transformation facilitates the creation of a "mask" that, when superimposed on the original image, defines the type and boundaries of objects within the image, thereby segmenting the image. This project employed a model that was initially trained using hand-segmented images and masks, but with sufficient training, gained the capability to independently generate these masks. Image segmentation, though more complex, offers more detail than the alternative method of image detection, which relies on bounding boxes to mark the general area of an object's presence rather than its precise location and form. This additional detail is particularly desirable in scientific research fields, especially biology and medicine, where greater precision is required. The project utilized semantic segmentation, a method that groups all objects of a category together during image segmentation.

Model Hyperparameters

Many hyperparameters were adjusted to tune the model for optimal accuracy. The ideal model is characterized by the best-fitted parameters. Training occurs in a series of epochs, defined as complete passes of the entire dataset through the model. Multiple iterations allow data samples to be observed and model parameters to be updated, with the precise number varying depending on the specific model. A fixed number of epochs, 175 in this case, are used to prevent endless running of the model, which may not contribute to improved accuracy (21). Additionally, each batch fed into the model contains a certain quantity of training images, denoted as the batch size, which in this instance was set at 16. The Adam optimizer was chosen for this model due to its efficiency and implementation effectiveness, as it updates model parameters throughout the training process, thereby enhancing accuracy (22). The categorical crossentropy loss function was applied in this model. These hyperparameters dictated the setup and operational process of the model throughout its training phase.

Model Architecture

The model employed is a convolutional neural network, designed with U-Net architecture (13) (**Figure 5**). As a blueprint for the model, the architecture allowed model weights or parameters to progressively become more accurate and precise during the training phase. The convolutional neural network consisted of convolutional layers, acting like filters that traverse the input and concentrate on local regions to generate predictions, a strategy more effective than analyzing the entire image at once. Within each layer, numerous convolutional kernels, acting as individual filters, were present. Each of these kernels was an integer matrix applied to a pixel set of identical size, with each pixel multiplied by its corresponding value in the matrix. The results formed a grid,

creating a feature map. These maps highlighted detected features in an image, isolating important elements that compose an object, and ultimately aiding the model to predict the object's class (23).

The model leveraged U-Net architecture, an encoder-decoder framework with residual connections, incorporating rectified linear activation function (ReLU) and max pooling layers. ReLU is a piecewise linear function commonly used in neural networks due to its effectiveness and accuracy. ReLU passes positive inputs directly while nullifying non-positive ones, determining inter-neuronal transfers, introducing non-linearity into the network, and aiding in learning complex data patterns for improved performance. Max pooling was utilized to prevent overfitting, abstracting data, and simplifying parameters. The max pooling layer identified maximum values from segments of a feature map, indicating the most dominant features in a sample. Each convolutional block was succeeded by a dropout layer to further prevent overfitting. Our model, tailored for an input shape of (144, 256, 3), had a depth of 4, an increasing dropout rate of 0.2, and handled 17 classes (including the image background).

Implementation Details

Preprocessing and model training was conducted in Google Colaboratory due to its beneficial integration with Google Drive, offering a streamlined method to organize and store files pertaining to each data type (image or mask) and dataset (training and testing) (19). Furthermore, the Tensorflow framework was employed (24). The duration required for training a model varies, influenced by factors such as parameters, architecture, and data type. In this case, the model ran for approximately 3 hours to complete 175 epochs, a period that did not include the preprocessing time of about 30 minutes or the time taken to load libraries and define functions, which is roughly 5 minutes per run. The time needed for inference was approximately one second per image. The code used for data preprocessing, model training, and evaluation is available on GitHub at the following repository:
https://github.com/victoria-wahlig/trashcan-segmentation-project.

Statistical Analysis

Predicted masks were visualized from each class. The Dice Similarity Coefficient (DSC), also known as the Sørensen-Dice index, was calculated for each individual class, as well as for the average of all resulting masks. The DSC is a measure of how well the real mask (created manually or "by hand") matches the predicted (model-generated) mask (16). The real mask serves as a baseline to which the model's predicted mask can be compared. The formula is

$$DSC = 2\,|A \cap B|/(|A| + |B|)$$

where A and B are the real and predicted masks, respectively. This formula estimates the overlap between A and B, which ideally will be 1. An accurately predicted mask should considerably overlay onto the actual mask, with almost no pixels visible that are not shared by the two masks. To integrate multi-class segmentation into our analysis, the DSC was calculated individually for each layer of each image and then averaged for each class. The DSC served as a consistent and accurate metric for evaluating segmentation

performance across all images and classes.

REFERENCES
1. Industry Agenda. The new plastics economy rethinking the future of plastics. *The World Economic Forum*: Geneva, Switzerland, 2016
2. Ritchie, Hannah, and Roser, Max. "Plastic pollution." *Our World in Data* (2018).
3. Lebreton, Laurent. "Evidence That the Great Pacific Garbage Patch Is Rapidly Accumulating Plastic." *Scientific Reports*, vol. 8, no. 1, 22 Mar. 2018, https://doi.org/10.1038/s41598-018-22939-w.
4. Yadav, Vinay, et al. "Framework for Quantifying Environmental Losses of Plastics from Landfills." Resources, *Conservation and Recycling*, vol. 161, Oct. 2020, p. 104914, https://doi.org/10.1016/j.resconrec.2020.104914.
5. Gall, Sarah C., and Thompson, Richard C. "The Impact of Debris on Marine Life." *Marine Pollution Bulletin*, vol. 92, no. 1-2, Mar. 2015, pp. 170–179, https://doi.org/10.1016/j.marpolbul.2014.12.041.
6. Kvale, Karin, et al. "Zooplankton Grazing of Microplastic Can Accelerate Global Loss of Ocean Oxygen." *Nature Communications*, vol. 12, no. 1, 21 Apr. 2021, p. 2358, https://doi.org/10.1038/s41467-021-22554-w.
7. Marsden, Pete, et al. Microplastics in drinking water. *World Health Organization*, 2019
8. Ocean Conservancy. "International Coastal Cleanup." *Trash Free Seas*, 2023, https://oceanconservancy.org/trash-free-seas/.
9. Morrison, Evan, et al. "Evaluating the Ocean Cleanup, a Marine Debris Removal Project in the North Pacific Gyre, Using SWOT Analysis." *Case Studies in the Environment*, 2019, https://doi.org/10.1525/cse.2018.001875.
10. Smith, Melissa. "Could Autonomous Robots Combat Marine Debris?" *SportDiver*, 2020, https://www.sportdiver.com/could-autonomous-robots-combat-marine-debris.
11. Huvenne, Veerle A.I., et al. "ROVs and AUVs." *Submarine Geomorphology*, July 2017, pp. 93–108, https://doi.org/10.1007/978-3-319-57852-1_7.
12. Nian, Rui, et al. "ROV-Based Underwater Vision System for Intelligent Fish Ethology Research." *International Journal of Advanced Robotic Systems*, vol. 10, no. 9, Sept. 2013, p. 326, https://doi.org/10.5772/56800.
13. Ronneberger, Olaf, et al. "U-Net: Convolutional Networks for Biomedical Image Segmentation." *Lecture Notes in Computer Science*, vol. 9351, 2015, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.
14. Hong, Jungseok, et al. "TrashCan: A Semantically-Segmented Dataset towards Visual Detection of Marine Debris." *ArXiv.org*, July 2020, https://doi.org/10.48550/arXiv.2007.08097.
15. Hong, Jungseok, et al. "TrashCan 1.0 an Instance-Segmentation Labeled Dataset of Trash Observations." *Data Repository for the University of Minnesota*, July 2020, https://doi.org/10.13020/g1gx-y834.
16. Dice, Lee R. "Measures of the Amount of Ecologic Association between Species." *Ecology*, vol. 26, no. 3, July 1945, pp. 297–302, https://doi.org/10.2307/1932409.
17. Isensee, Fabian, et al. "NnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation." *Nature Methods*, vol. 18, no. 2, Dec. 2020, pp. 203–211, https://doi.org/10.1038/s41592-020-01008-z.
18. He, Kaiming, et al. "Mask R-CNN." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, pp. 1–1, https://doi.org/10.1109/tpami.2018.2844175.
19. Bisong, Ekaba. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*. New York, Apress, 2019.
20. Minaee, Shervin, et al. "Image Segmentation Using Deep Learning: A Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, 2021, pp. 1–1, https://doi.org/10.1109/tpami.2021.3059968.
21. Yang, Li, and Abdallah Shami. "On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice." *Neurocomputing*, vol. 415, Nov. 2020, pp. 295–316, https://doi.org/10.1016/j.neucom.2020.07.061.
22. Kingma, Diederik P., and Ba, Jimmy. "Adam: A method for stochastic optimization." *3rd International Conference on Learning Representations* (ICLR 2015), 2015, https://doi.org/10.48550/arXiv.1412.6980.
23. Liu, Qing, et al. "A Review of Image Recognition with Deep Convolutional Neural Network." *Intelligent Computing Theories and Application*, 2017, pp. 69–80, https://doi.org/10.1007/978-3-319-63309-1_7.
24. Abadi, Martin, et al. "TensorFlow: A System for Large-Scale Machine Learning." *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi.