Challenges in the Deep Learning-based Semantic Segmentation of Benthic Communities from Ortho-images

G. Pavoni · M. Corsini · N. Pedersen · V. Petrovic · P. Cignoni

Abstract Since the early days of the low-cost camera development, the collection of visual data has become a common practice in the underwater monitoring field. Nevertheless, video and image sequences are a trustworthy source of knowledge that remains partially untapped. Human-based image analysis is a time-consuming task that creates a bottleneck between data collection and extrapolation. Nowadays, the annotation of biologically-meaningful information from imagery can be efficiently automated or accelerated by Convolutional Neural Networks (CNN). Presenting our case studies, we offer an overview of the potentialities and difficulties of accurate automatic recognition and segmentation of benthic species. This paper focuses on the application of Deep Learning techniques to multi-view stereo reconstruction byproducts (registered images, point clouds, ortho-projections), considering the proliferation of these techniques among the marine science community. Of particular importance is the need to semantically segment imagery in order to generate demographic data vital to understand and explore the changes happening within marine communities.

Keywords Underwater monitoring · Coral reef surveys · Semantic segmentation · Automatic classification · Deep Learning

G. Pavoni (corresponding author)
Visual Computing Lab (ISTI-CNR), Italy
E-mail: gaia.pavoni@isti.cnr.it

M. Corsini
Visual Computing Lab (ISTI-CNR), Italy
E-mail: massimiliano.corsini@unimore.it

N. Pedersen
Center for Marine Biodiversity and Conservation, Scripps Institution of Oceanography, University of California San Diego, CA, United States
E-mail: nepeders@ucsd.edu

V. Petrovic
California Institute for Telecommunications and Information Technology, University of California San Diego, CA, United States
E-mail: vipetrov@eng.ucsd.edu

P. Cignoni
Visual Computing Lab (ISTI-CNR), Italy
E-mail: paolo.cignoni@isti.cnr.it

1 Moving to automated analysis: increasing the scale and the efficiency of coral-reef monitoring

Coral reefs are ecosystems of vital importance for the planet, hosting 25% of marine biodiversity. Over the past few decades, the decline of these habitats has been rapidly increasing due to factors such as thermal stress, over-fishing, or anthropogenic pollution [13]. Periodic demographic surveys allow for an understanding of how fast coral reefs are changing, by assessing the mortality, recruitment, fragmentation, or growth of colonies [19]. *In situ* surveys are still a commonly used method to gather this information but remain cumbersome and time-consuming. As such, they are limited in the scale at which they can be effectively implemented.

To derive models describing population dynamics, study the major factors impacting the health of the coral reefs or quantify the species resilience, ecologists need a highly resolved large-scale understanding of communities.

Underwater photographic surveys (speed up by the use of underwater scooters, ROVs, and autonomous vehicles) provide a rapid supply of information across larger areas. These collected images support the creation of permanent archives for future analysis, further increasing the value and utility of these datasets.

Image annotation has traditionally been performed by a human operator. Coarse community assessment of percent coverage involves the generation of several randomly-sampled points on an image and the association of each point to a known class of benthos. When the number of labeled points is sufficiently large, the presence of organisms on a reef can be statistically described. A commonly used tool for manual point-based annotations is *Coral Point Count* [24], from which coverage data on benthic organisms can be extracted. Manual point designations within photos requires a considerable amount of time. In [34], it is reported that *for a survey of about 1000 images more than 400 manual working hours are needed*. The underwater readability presents challenges leading to human error, where annotated points can fall on regions of the image that are not visible or on uncertain contours of corals. Other causes of human misclassification are the repetitiveness of the labeling task, the required experience, and above all, the inability to examine the colonies directly in the field [36-37]. Concluding, humans can be inconsistent over time and across individuals.

In [6] and [8], the authors reveal the bottleneck between the large amount of image data collected each year and the extrapolation of quantitative data by visual inspection. The National Oceanic and Atmosphere Administration (NOAA) reported that biologists analyze just 1-2% of the millions of underwater images acquired each year on coral reefs [8].

Point-based annotations are not sufficiently descriptive to quantify the growth and shrinkage, or to analyze the spatial distributions of coral colonies. These studies require the outlining of the contour for each colony (see Fig. 1), performing a per-pixel classification. The manual annotation of areas, typically demands about one hour per square meter. This task, the partitioning of an image into (disjointed) sets having semantic meaning, is commonly known as *semantic segmentation*. Nowadays, the semantic segmentation, as well as other visual recognition tasks, such as coral reefs classification [40,41], can be efficiently automated (or supported) by methods based on Convolutional Neural Networks (CNN).

This work discusses the effectiveness, challenges, and limits, of the automatic analysis of coral reef images. More precisely, we apply the semantic segmentation task on ortho-projections of point clouds. Ortho-projections and ortho-mosaics are traditional data products used by researchers to extract ecological data across broad spatial scales. We focus on a dataset composed by human-labeled ortho-projections, provided by the Center for Marine Biodiversity and Conservation, Scripps Institution of Oceanography, UC San Diego. The contribution includes:

1. A complete overview of the challenges in the fully automatic semantic segmentation of corals. Section 3 describes the issues introduced by the corals' morphology.
2. An analysis of the most suitable multi-view stereo related by-product; Section 4 motivates the choice of ortho-projections as opposed to working on digital images or 3D point clouds.
3. A methodology to deal with the semantic segmentation of large images; Section 6.
4. Strategies to improve network performances and generalization.

Additionally, Section 2 collocate this work in the literature; Section 5 discuss the case study, while Sections 9 and 10 disclose the semantic segmentation problem in the full three-dimensional context and anticipate the future directions of our study.

2 Related Works

The recognition of marine organisms, as well as other areas of image analysis, has been automated in recent years by the introduction of machine learning approaches. Most benthic annotation tools are point-based: this couples well with. patch-based classification models, as they are directly applicable to point-annotated datasets. Each input image for the training dataset is generated by cropping a square area around each annotated point. A patch-based CNN results in a single class prediction per patch. This approach leads to the problem of patch size: patches must be large enough to describe the structure of the marine organism, but small enough not to incorporate other classes. The classification of points lying on shape profiles is problematic across every patch size.

In [5], the authors proposed a machine learning approach to automatically point-label eight classes of benthic organisms and the background. These classes cover the 96% of Moorea's coral reefs, with the remaining benthos containing rare coral species with insufficient coverage for automatic labeling. Features are classified using a Support Vector Machine (SVM) classifier using both color and texture. This method reported a classification accuracy of 74.3% in evaluating images gathered in the same year (2008). The authors also published the Moorea Labeled Corals dataset, the first coral benchmark dataset containing 400,000 annotated points.

In [8], CNNs are used for the automatic classification of benthic species. The authors start from an annotated dataset of registered images containing both reflectance and fluorescence information. A CNN is trained using only the reflectance, reflectance and fluorescence averaged per-pixel, only the fluorescence, and both information concatenated in a 5-channel network. The results demonstrate that the highest accuracy (90.5%) is achieved by the three-channel network that uses the average between reflectance and fluorescence. Fluorescence is effective in increasing the contrast between corals and the background, while reflectance helps distinguishing among coral species. The comparison between the machine learning-based method proposed in [5] and CNNs using RGB images states the two approaches achieve a similar accuracies of around 87.8%. These works led to the release of *CoralNet*, a Web platform for the automatic, semi-automatic, and manual point-based annotation of benthic images. Recently, in [34], authors demonstrate a higher classification performance of CoralNet when benthic classes are restricted to functional groups, and the network is trained separately for different habitats. In [16], the authors compare the predictions of a standard RGB-based patch classification network and the annotation of human experts. They report an agreement in the benthic coverage estimation of about the 97% and a high reduction of the time required. These results demonstrate the utility of machine learning methods in underwater monitoring actions.

In [26], the authors fine-tuned a pre-trained VGG network, demonstrating that, when tiles of different dimensions are cropped around each annotated point, the performance improves, and the class imbalance is reduced. Different scale representations are then resized and given as input to the network. Corresponding output vectors are fused using a max-pooling layer before proceeding with the classification. The proposed method reports an accuracy between 69.2% and 82.8% in the various experiments performed.

As previously stated, despite the number of sampled points, a patch-based approach is not dense enough to detect changes and carry out the spatial analyses of populations, and a per-pixel classification is needed.

The manual classification of pixels is an extremely time-consuming process. In particular, when applied to corals, it requires precision in following intricate colony shapes and internal lesions (see Fig. 1). Manual semantic segmentation task by a trained biologist takes about an hour per square meter with majority of
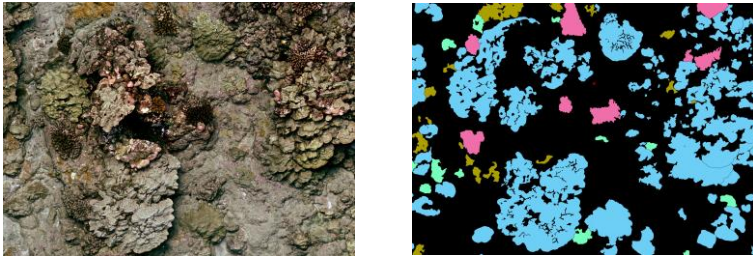
Fig. 1: An example of manually segmented coral colonies. A benthic image on the left, the manually labeled areas on the right.

the time dedicated to drawing the outline of colonies. To our knowledge, there is no available human-labeled benchmark dataset dedicated to the semantic segmentation of corals, so all the studies in the field first attempt to create an appropriate training dataset. In [3], the authors try to address the lack of densely-labeled datasets by proposing a method to propagate point annotations based on the manipulation of the fluorescence image channel. They then use the resulting labeled masks to fine-tune a SegNet [21] model. A few months ago, the same authors, in conjunction with the publication of [4], released the first extensive dataset of mask-labeled benthic images. This densely labeled comes from the propagation of sparse annotation following a multi-level superpixel approach.

In [23], the authors introduce a custom annotation tool based on SLIC and graph cuts to generate the ground truth segmented dataset. They then train the patch-based CNN architectures Resnet152 and the semantic segmentation CNN Deeplab V2, scoring an accuracy on ten classes of 90.3% and 67.70%, respectively. In comparison with ground-truth annotations, predicted areas show smoother borders and some misclassifications.

One of the main limitations to the results in terms of accuracy in studies [3, 21, 4, 23], is due to the lack of per-pixel labeled images (see Fig. 1). Approximate labels on coral contours induce uncertainty in predictions. In recent years, there has been a growth in assisted per-pixel annotation tools able to speed up the preparation of training datasets. Among them, we report TagLab[1], an AI-assisted open-source tool, designed for both large image labeling and analysis.

3 Challenges in segmenting coral species

Recognizing an object within an RGB image means determining some of its distinctive color and texture features. Effects like the color-shift due to light absorption, the scattering caused by water turbidity, the lens distortion and chromatic aberrations, can alter objects' appearance and confuse both users' interpretation and recognition algorithms. Corals, in large-area imagery, frequently display blurred outlines or appear as belonging to another class when framed under different distances. For all these reasons, species are more easily recognizable *in situ*, rather than from images. Ninio et. al. [28] report class-dependent accuracies ranging from 96% for hard corals to 80.6% for algae when discriminating species from photographs compared to in situ observations.

Automatic classification methods demonstrate lower performance on complex morphologies. Typically, CNNs for visual recognition are trained to recognize human-made objects, pedestrians, animals, and other classes that are not visually similar to marine organisms.

This affects the accuracy in predicting those natural structures that are not adequately represented in the training dataset. Furthermore, a high intra-specific variability exists between specimens belonging to the same class. Fig. 2 shows multiple colonies with variable texture, color, and morphologies. Nevertheless, all these colonies, marked in yellow and green, belong to the same genus (*Montipora*). On the other hand, very similar individuals may belong to different classes, giving rise to false negative and false positive predictions.

[1] https://github.com/cnr-isti-vclab/TagLab

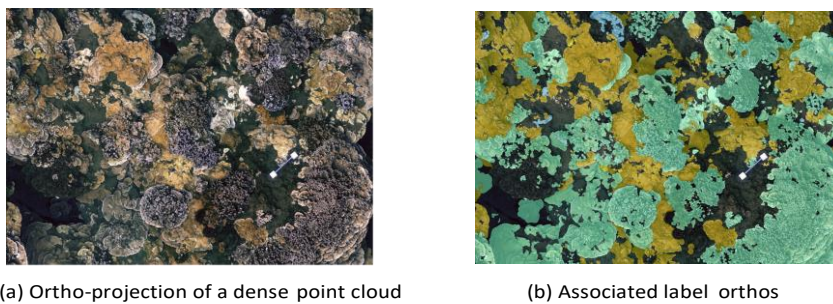(a) Ortho-projection of a dense point cloud   (b) Associated label orthos

Fig. 2: Morphological variability among *Montipora* species. All colonies portrayed in this image (except one on the upper side, marked in light blue), belong to the same genus, despite different appearances.

Finally, an additional obstacle encountered in the automatic identification of corals is the class imbalance, where the different classes in the training dataset are not equally prevalent. On the upper edge of Fig. 2, a single colony of *Porites* is marked in blue. In this area, *Porites* is an under-represented class, a hard case for the accurate automatic detection due to the lack of positive examples.

4 Imagery and derivatives for the automatic recognition of species

Ideally, to evaluate benthic changes over multi-temporal surveys, it would be desirable to exploit volumetric information. However, the automatic recognition of complex 3D shapes is complicated by occlusions and by the lack of fine geometric details essential for characterization. Classification in those cases must also consider the quality of the reconstructed geometry or the cloud sample density. The use of two-dimensional data is not only convenient, but also aligns well with historical data collected in situ or through imagery (e.g. percent cover, diameter, planar area). However, the automatic recognition of colonies from images, if not contextualized in the surrounding environment, has a limited relevance for the estimation of growth or death phenomena. Fortunately, the application of multi-view stereo reconstruction generates different types of inter-related by-products such as calibrated images, reconstructed point clouds, and ortho-projections. Furthermore, image derivatives have different properties from a machine learning perspective.

Images are a structured, high-resolution information source, arranged in a regular grid of pixels. Unregistered images do not contain information related to their context, the surrounding habitat, and the pixels in general have an unknown scale factor. When images are registered in a photogrammetric network with high feature overlap, the consistent (fully) manual per-pixel labeling is unfeasible.

Point clouds provide three-dimensional information. Color inconsistencies (such as caustics) are typically attenuated by color blending or corrected using depth information. However, they are unstructured, non-uniform data with missing information (e.g. holes) due to occlusions. The resolution of a point cloud is usually lower than the original images, and their manual annotation requires the manipulation of corals in 3D space, making labeling slow and prone to errors. The adaptation of CNN architectures to 3D data requires the transformation of the unordered point clouds into a regular voxel grid, introducing other issues such as quantization and a large memory footprint. The study of deep architectures for the direct semantic segmentation of raw point clouds is promising [11]. To our knowledge, exist a few labeled point clouds of benthic landscapes, which makes a successful automatic 3D recognition an arduous task.

Ortho-mosaics / ortho-projections are the ideal compromise between the readability of a single image and the three-dimensional information. Where depth and scale are known, ortho-mosaics can be generated from the same defined projection vector correcting perspective deformations, allowing for multi-temporal alignment and change detection of organisms through time. Colors can be corrected coherently with depth. The fixed scale of ortho-mosaics helps in preserving the physical size of morphological features, a discriminant factor in

species recognition. Last but not least, ortho-mosaics are a commonly used tool for spatial and demographic analysis of populations. The use of ortho-mosaics reduces the number of features to learn but may introduce some other problems. Image orientation errors result in local blurring or contour ghosting effects. The stitching and projection process in ortho-mosaic creation might produce local image warping in correspondence to depth discontinuities. For this reason, we propose the use of an orthorectified projection (i.e. ortho-projection) of the colored point cloud on the seabed plane to address the semantic segmentation. The use of ortho-projections avoids dealing with all stitching artifacts associated with ortho-mosaics. This geometric accuracy comes at a cost of a reduced resolution, ortho-projections exhibit a grainy appearance.

All these representations of a three-dimensional scene have their advantages and disadvantages. In order to use all the available information, the ideal approach would be to exploit them in an integrated, multi-modal fashion (see Section 10).

5 Case-study: the semantic segmentation of ortho-projections coming from the 100 Island Challenge

The 100 Island Challenge is a large-scale experiment, conducted by the Scripps Institution of Oceanography (UC San Diego) across islands in the Pacific, Caribbean, and Indian Oceans. Islands of interest have been selected to embrace a combination of human activity, oceanographic, and geomorphological conditions. Researchers want to assess which of these factors influences the structure and growth of benthic communities [14]. The spatial and demographic analysis of coral populations is conducted on top-down ortho-projections of dense point clouds to a horizontal surface plane. Surveys are conducted on the forereef at 10m depth for each island and repeated after 2-3 years to detect changes in reef populations, which are used to predict future community trajectories.

At each island, roughly 6-8 10m x 10m survey plots are collected using a large area imaging approach. For each plot, 2000-3000 superimposed images are taken using a NIKON D7000 with an 18mm focal length in a bird's eye view following two crossed lawnmower patterns, taken roughly 1-2 meters above the benthos. On the seabed are positioned scale bars and control points to form a *network* of 15 stable reference points.

The 3D model reconstruction was performed using Agisoft Metashape [2]. Self-calibration was used to estimate the camera network. The point cloud was generated at "high" resolution, with depth filtering set to "mild". To create the top-down projection of the dense cloud, ecologists used the custom visualization platform *Viscore* [31]. Ortho-projections were annotated by manually drawing colony borders using Adobe Photoshop, resulting in a fine-scale segmentation ortho. This workflow is time-intensive, requiring about one hour per square meter. Figure 3 shows the plot HAW and its associated label ortho. HAW, as well as a large number of coral reefs in the Indo-Pacific, show coverage with three predominant genera including *Porites* (light blue), *Montipora* (light green and olive), *Pocillopora* (pink), plus other rare coral taxa. The remainder of the benthos, predominantly algae and sand were classified as belonging to a generic fourth class, Background, filled in black.

The 100 Island Challenge team conducted all underwater surveys, the 3D models' generation, the generation of ortho-projections, and the manual annotations used in the following study.

The ultimate purpose of this research would be the automatic semantic segmentation of all reconstructed plots collected within the 100 Island Challenge project. Generalizing classifications to all ortho-projections created, starting from images coming from multiple cameras with varying parameters of reconstruction, different water conditions, and diverse species assemblages, is a difficult task. Moreover, there are over 800 reef-building coral species in the world, and their appearance varies across geographic regions. The dataset used contains eleven plots from seven islands, the labeled area covers about 100 square meters per ortho. Learning to distinguish all coral species requires a broader dataset, therefore, in this study, we focus the automatic recognition on three commonly occurring genera found on Indo-Pacific reefs: *Porites, Montipora,* and *Pocillopora*.

Section 6, presents the DL approach adopted in this large-scale project while section 7, reports the classification results.

6 Proposed Method

In CNN architectures, the low-level features maps store the local information to achieve an accurate pixel-wise outlining of the scene. Conversely, a correct classification results from considering the contextual information. The contextual information is added by increasing the receptive field of neurons, through the introduction of pooling layers. However, pooling layers progressively downsizes the feature maps, resulting in coarser predictions. This effect is undesirable, especially in the presence of very minute details, as in the case of corals. The Deeplab V3+ network, introduced by Le Chen et al. in 2018 [9], is one of the state of the art architectures in terms of accuracy for semantic segmentation. Authors proposed an "encoder-decoder" architecture, which uses ResNet-101 as a feature extractor and natively adopts sparse convolutions instead of pooling layers. A *sparse convolution*, a convolution having a nucleus dilated by the presence of zeros, leads to sparse activation of neurons able to capture multi-scale details without increasing the number of parameters involved or downgrading the spatial resolution. More precisely, the DeepLab V3+ adopts separable sparse convolutions, composed of atrous convolutions with point convolution among the three channels of RGB images. The contributions from convolutions at different degrees of expansion are then fused into the same layer using an efficient interpolation scheme called Atrous Pyramid Pooling. The accuracy of DeepLab V3+ on Pascal VOC 2012 [15] (a vast image dataset typically used to assess the performance on a neural network on tasks such as object recognition and segmentation) is about 89.0% mIoU (mean Intersection over Union).

Training a model from scratch takes time, computational resources, and a huge amount of data. *Transfer learning* [39] is commonly adopted in the automatic recognition field to adapt networks that have learned to solve a specific task to deal with new (similar) problems. In using a pre-trained model, the final classifier is typically replaced with a custom one, specific to the new task. When re-training pre-trained models, there are several choices: only train the classifier, train some layers and freeze the others, or leave everything unfrozen. As previously noted, coral classes have a weak visual similarity to the PASCAL VOC 2012 dataset used to train the original DeepLab V3+. In this circumstance, a simple fine-tuning of the encoder layers of the pre-trained model lead to poor results. Hence, we decided to leave all the parameters unfrozen, allowing for small updates of weights. To guarantee only minor adjustments, the learning rate was set lower than the one used during the original training. This compromise allows the network to learn dissimilar features from those contained in the Pascal VOC datasets but takes into account the limited amount of data available. Additionally, the low learning rate prevents the risk of losing previous knowledge and helps in reducing the probability of overfitting.

6.1 Dataset preparation and training

Since the plots differ in scale (see Table 6), all ortho-projections are first resampled at the same scale using Lanczos resampling. Labels are resampled utilizing nearest neighborhood sampling to preserve the values of the color codes. This rescaling operation makes the morphological features of coral classes uniform in size and resolution. The chosen common scale is 1.11 pixels/mm which is a good compromise to avoid too severe undersampling on higher-res ortho-projections and too much interpolation on lower-res ones.

Co-registered products of rescaled plots (both the ortho-projection, stored as an RGB image and the label ortho) is clipped into overlapping tiles. Then, 65% of consecutive tiles are used for network training, 20% are used for validation and the remaining 15% for the test. RGB input tiles are pre-processed on-line by subtracting the average value per channel and cropping the central area to 513 × 513 pixels after the data augmentation step (see figure 4).

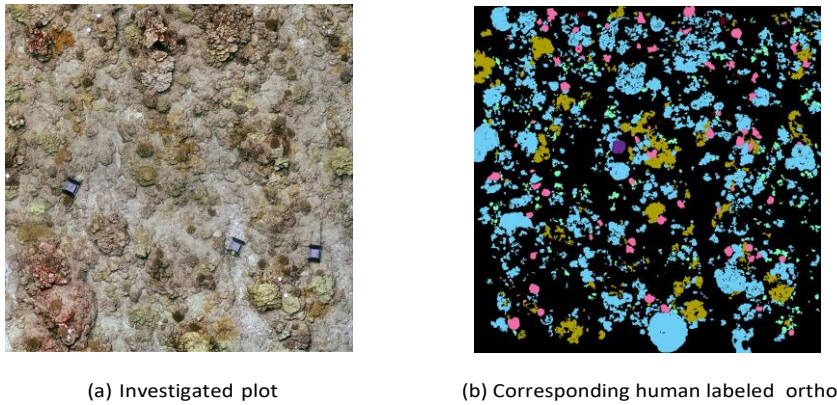(a) Investigated plot   (b) Corresponding human labeled ortho

Fig. 3: *HAW* plot (10 × 10 m). On the left the ortho-projection. On the right the corresponding manual annotations: *Porites* (light blue), *Montipora* (light green and olive) and *Pocillopora* (pink).
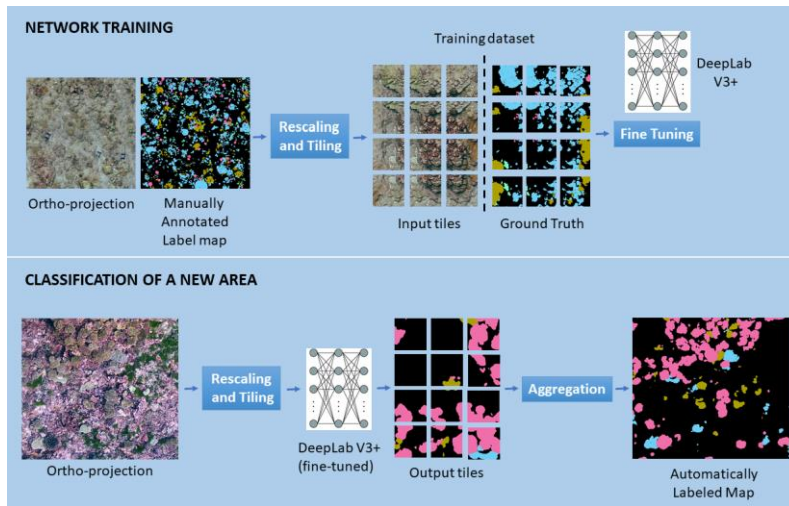

Fig. 4: Proposed method. We first rescale the input orthos (ortho-projection and label ortho) to homogenize the morphological class features. Then, they are both subdivided into tiles to create training, validation, and test datasets. The trained network is used to classify an unseen ortho-projection automatically. The new ortho is clipped into tiles too; then each tile classification scores are aggregated in the final label ortho using the procedure described in Section 6.2.

On-line data augmentation includes horizontal and vertical flips, random rotation (10-degree range), random translation (50 pixels maximum in x- and y-coordinates), and a step of *color augmentation*. We tested two different strategies to make the representation of classes less sensitive to the color absorption related to depth: the on-line color augmentation and the Local Contrast Normalization (LCN) [20]. According to the results obtained, the on-line color augmentation improved the classification results (see Section 7). We simulate the typical pixel-level alterations related to the formation of submarine images by randomly adding an RGB-shift, a random contrast change, CLAHE [35], a hue variation, a light intensity variation, and a blurring effect. The *Albumentations* library [1] offers a valid implementation of on-line color augmentation transformations. The DeepLab V3+ was trained for a variable number of instances (between 60 and 150, to avoid overfitting) using an SGD+momentum optimizer with adaptive learning rate decay, an initial learning rate of 10^{-5}, and a decay rate of 10^{-4}. Typically, class imbalance is solved by oversampling rare classes, undersampling predominant classes, or weighting the loss function. In this study, we use the cross-entropy loss function weighted by inverse frequency weights.

6.2 Infer predictions on large ortho-projections

Modern fully convolutional architectures, such as the DeepLab V3+, accept image inputs of any resolution. The only limit is available GPU memory. Despite the high amount of memory in modern graphics cards used for Deep Learning applications, ortho-projections are too large to be processed entirely. A typical remote sensing approach to overcome this problem is to subdivide the ortho into several overlapping sub-orthos, classify them, and later aggregate classification results. Border regions of tiles are usually discarded because partially framed corals might induce misclassification errors.

Multiple experiments were conducted to assess the best tile aggregation procedure. Orthos were processed after being subdivided in tiles of 1025 × 1025 pixels, with an overlap of 75% each. Predictions were then re-assembled using three approaches: without aggregating the classification scores, aggregating the classification scores using the average values, and aggregating the scores with a Bayesian approach [29]. Bayesian fusion considers the prior distribution of the specimen of interest.

7 Results

In supervised learning methods, datasets are generally divided into three subsets: a training set, a validation set (to select the best network and the optimal hyperparameters), and a test set. Performances on the test dataset (as a portion of the entire dataset) are commonly used to evaluate the network generalization capabilities. This applies to both single image datasets and orthos. However, the present species, their relative frequency on the seabed, as well as the image quality, are variable in reef monitoring. Ideally, the classifier must perform adequately on totally unseen orthos, even when belonging to a different geographical area.

"Standard" Testing approach	Testing on unseen maps
Single plot dataset - Training, validation, and test sets composed of a partition of the same plot.	*Multi-class classifier* - Limited to selected predominant classes (Montipora, Pocillopora, Porite) - Same geographical area, multiple plots - Unseen tiles both in the validation and test sets.
Multi-plot dataset - Training, validation, and test sets composed by cropping a partition of 4 different plots and mixing tiles.	*Binary classifier* - A single coral family - Different geographical areas, multiple plots - Unseen tiles both in the validation and test sets.

Fig. 5: Summary of experiments to assess the performance of the Deeplab V3+ fine-tuned following the proposed approach.

To reach this goal, we conducted several experiments. More precisely:

1. Following the standard dataset partition:
   – We first trained and tested Deeplab V3+ on single plots, to determine its performance on dense ortho-projections. As not every point cloud has the same point density and noise; this test highlights how much data quality affects learning.
   – We trained the network on a dataset composed by a mix of tiles from different plots-
2. We change the dataset partition method performing a "stress test", evaluating the networks on totally unseen orthos from different geographical areas:
   – We tested a multiple-class classifier to distinguish the three genera of interest across plots from the same island. In a similar context, the environmental conditions and 3D reconstruction parameters are variable, while the colony morphology remains coherent. For the 100 Island Challenge project, typically eight plots per island are surveyed. Therefore, inferring the classification on all plots from a few annotated plots would allow for a considerable speed-up of the manual segmentation procedure.
   – We evaluated the performance of a binary classifier for *Pocillopora*. Given the high abundance of this genus, the binary classifier is trained such that it is independent by the geographical region. To assess its performance, we tested it on totally unseen orthos coming from different geographical areas.

A scheme of these experiments is given in Figure 5. The plots used and the results obtained will be described in the next sections. Figure 3 and Figure 10 display some of the eleven orthos included in the available dataset while related information are reported in Table 6. All ortho names with the suffix *MIL* belong to the same island.

Name	Scale	Resolution (pixels)	Training tiles number
V OS	1.1507 pixel /1 mm	11081 × 11061	247
ST A	0.979 pixel / 1 mm	12628 × 13520	391
KHA	1.580 pixel / 1 mm	14647 × 13367	209
MIL-M1	0.8097 pixel /1 mm	7069 × 7433	204
MIL-M3	0.997 pixel / 1 mm	9691 × 9520	260
MIL-M4	1.1779 pixel /1 mm	10707 × 10707	216
MIL-M5	1.5385 pixel /1 mm	15661 × 14761	285
MIL-M6	1.0438 pixel /1 mm	10709 × 10428	294
FLI	1.229 pixels / 1 mm	10674 × 10320	197
HAW	1.237 pixels / 1 mm	13415 × 13582	330
MAI	1.450 pixels / 1 mm	13207 × 12659	190

Fig. 6: All orthos cover an area of approximately 10 ×10 meters. The point density of the ortho-projections is slightly different between each ortho. 65% of the tiles of each ortho comprise the training datasets. The number of tiles available for training depends on the ortho scale and resolution.

7.1 Training and testing on the same dataset

7.1.1 Semantic segmentation of a single ortho

The first experiment involves *HAW* and *MIL-M6* to verify how much the performance varies between plots. *HAW* shows a relatively uniform appearance, characterized by low chromatic variation and a flat and sandy seafloor area. In this plot, *Porites, Montipora, Pocillopora*, and Background per-pixel frequencies are respectively the 20.78%, the 7.05%, the 2.57% and the 69.60%.

Network predictions on the test set report an accuracy of 0.935 and a mean Intersection over Union (mIoU) of 0.883 (Table 1). Figure 7 shows the DeepLab V3+ results for taxonomic classification, demonstrating the ability to distinguish dead colonies within the same class. Dead colonies and dead portions of living colonies are correctly classified as belonging to the background class. Most mis- classified pixels fall on the boundaries of the automatically segmented regions, which display a slightly smoother outline to those of the ground truth classes.

The same experiment was repeated on *MIL-M6*. As visible in Figure 10, *MIL-M6* has a steeper slope and large structures where the color changes due to depth. While the Background class of *HAW* contains mostly sand, while *MIL-M6* contains additional algal and coral species with high coral coverage (see Figure 10). *MIL-M6* tiles are slightly noisier as visible in Figure 8. The Deeplab V3+ trained and tested on *MIL-M6*, reports an accuracy of 0.860 and a mIoU of 0.795 (Table 1).

In the *MIL-M6* plot, the progressive absorption of the color with depth is visible, therefore we tested the effect of both the LCN and color augmentation in tiles during pre-processing. LCN slightly degraded the performance of the automatic classifier, reaching an accuracy of 0.835 and an mIoU of 0.760 (Table 1). Con- versely, the CNN trained and tested on *MIL-M6* exploiting color augmentation reached an accuracy of 0.880, and an mIoU of 0.817 (Table 1). On-line color augmentation demonstrated an improved generalization, so we chose to apply it in all the following experiments.
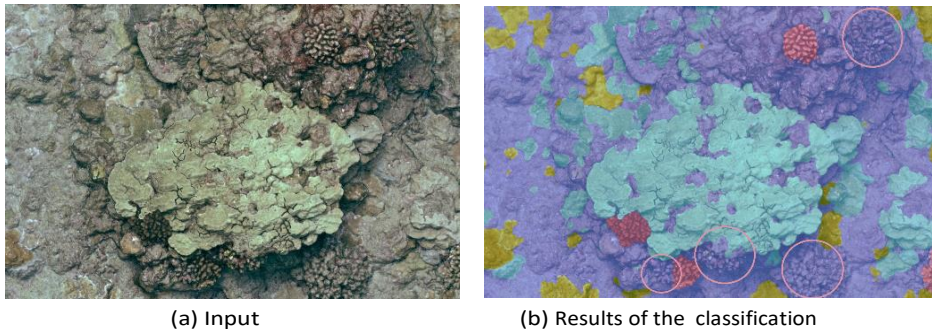
(a) Input

(b) Results of the classification

Fig. 7: Network predictions in a crop of the test area of *HAW*. The figure displays *Porites* in light blue, *Pocillopora* in pink, *Montipora* in olive green. Dead *Pocillopora* (circled) is correctly classified as Background, as well as some dead portions of *Porites*.

Experiment	Accuracy	mIoU
HAW	0.935	0.883
MIL-M6	0.860	0.795
MIL-M6 + LCN	0.835	0.760
MIL-M6 + online col. aug.	0.88	0.817

Table 1: Results of the experiment on a single plot.

(a) *HAW* training input tile (detail)
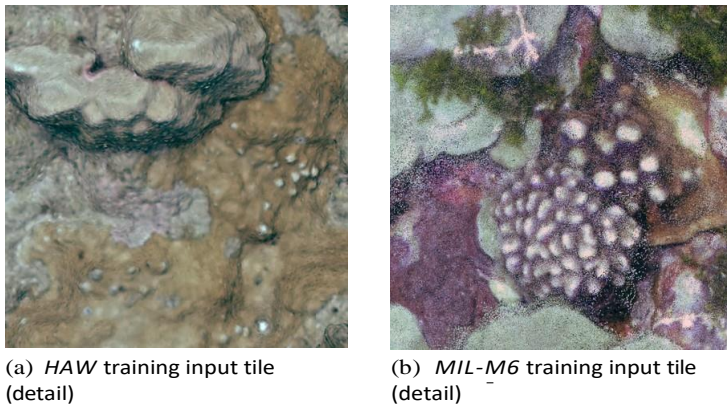
(b) *MIL-M6* training input tile (detail)

Fig. 8: Above a comparison with a tile of *HAW* and *MIL-M6* (see Figure 3 and 10). *MIL-M6* is on average noisier leading to more visual artifacts.

7.1.2 The semantic segmentation of heterogeneous datasets

For the second experiment, the Deeplab V3+ was trained and tested on a dataset mixing four plots variable in color, resolution, and geography. In particular, *MAI* portrays a flat seabed (from a distant viewpoint) with a few small colonies while *MIL-M3* can be characterized as a low-resolution point cloud distinguished by sensible color variation among colonies due to a spur and groove formation. Finally, *FLI* displays a sloped seabed with a wide color variation due to depth changes. The four ortho-projections (including *HAW*) were cropped into a training, validation, and test area, taking care to maintain consistent species variability. Overall frequencies related to the four classes on the entire four orthos training set are respectively 13.99%, 13.66%, 7.55%, and 64.8%. The purpose of this experiment is seeing if different ortho-projections with variable point cloud resolutions are mixable in the same dataset without impacting the predictions. Predictions reached an accuracy on the test area of 0.921 and a mIoU of 0.858 after 90 epochs. The performance on *MIL-M3* test tiles were the lowest, reporting an accuracy of 0.821 and an mIoU of 0.698. *MIL-M3* has a lower image quality, along with a decreased representation of tiles in the training dataset. Additionally we can observe the presence of a distinctive *Pocillopora* species (*Pocillopora zelli*) morphologically different from other individuals of the same genus (see Fig. 9).
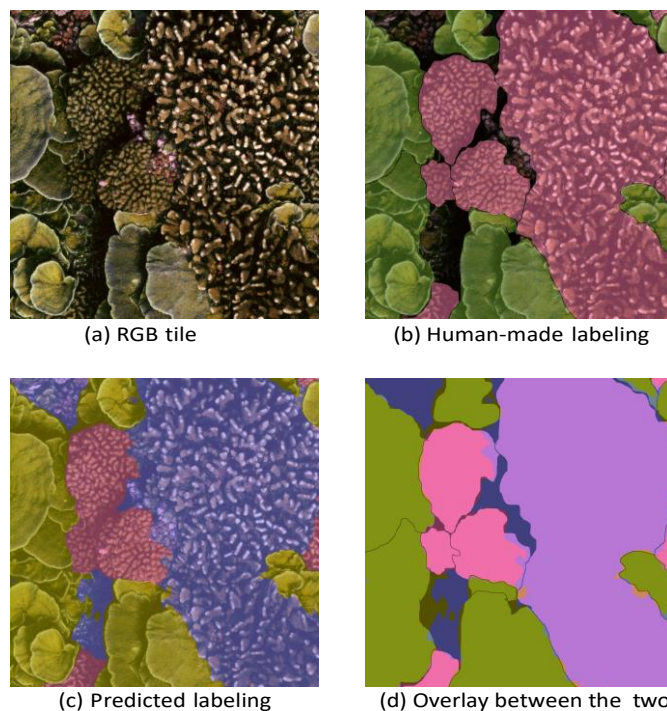
(a) RGB tile   (b) Human-made labeling

(c) Predicted labeling   (d) Overlay between the two

Fig. 9: A *Pocillopora* colony (pink color) misclassified as Background (purple), likely due to large morphological variation between species.

7.2 Evaluation of performance on totally unseen orthos

7.2.1 A 4-class classifier for a single geographical region

During the network training, a validation step is carried out after fixed amounts of epochs to asses the network performances. In the previous experiments, the performance was measured on an unseen partition of validation tiles belonging to the same plots. At the end of the training, the weights chosen are the ones that obtained the best results on the validation set. In this investigation, the validation of input tiles from totally unseen orthos compared to those belonging to the training orthos helps in choosing the network that works best for an unseen underwater scenario. In other words, we select the network that best recognizes the coral taxa, regardless of the three-dimensional reconstruction or environmental conditions.

We train the DeepLab V3+ on a dataset composed of training tiles from *MIL-M5* and *MIL-M6*, but we validate the network on a mixed set, composed validation tiles from *MIL-M5* and *MIL-M6* in addition to tiles from *MIL-M3* (see Figure 10). All the *MIL* orthos have been collected on the same island.

The resulting network was first tested on the original test set composed of *MIL-M5* and *MIL-M6*, reaching an accuracy of 0.915 and a mIoU of 0.847 (high values, as expected). For the two new test orthos, MIL-M1 obtained an accuracy of 0.886 and a mIoU of 0.801, while MIL-M4 had an accuracy of 0.972 and a mIoU of 0.947. Such a significant difference in predictions on new orthos can be explained by their respective resolution, the morphological complexity of MIL-M1 (see Figure 10), and the dense distribution of corals on the ortho.

We used the T-SNE algorithm [25]) to project the features extracted from the encoder into a 2D scatter plot (see Figure 11). The visual analysis of plotted high-level features reveals confuse boundaries between classes. Since the classes are not clearly separated, the classification is uncertain. As visible, the background of MIL-M1 is characterized by high morphological complexity. Results proved that predictions obtained from the training of MIL-M5 and MIL-M6 could be inferred on other plots belonging to the same island MIL-M1, and MIL-M4. Finally, CNN was also tested on FLI, a nearby island in the Pacific Ocean containing similar species, reporting an accuracy of 0.831, and a mIoU of 0.711. Results are reported in Table 2.

(a) *MIL-M1*

(b) *MIL-M3*

(c) *MIL-M4*

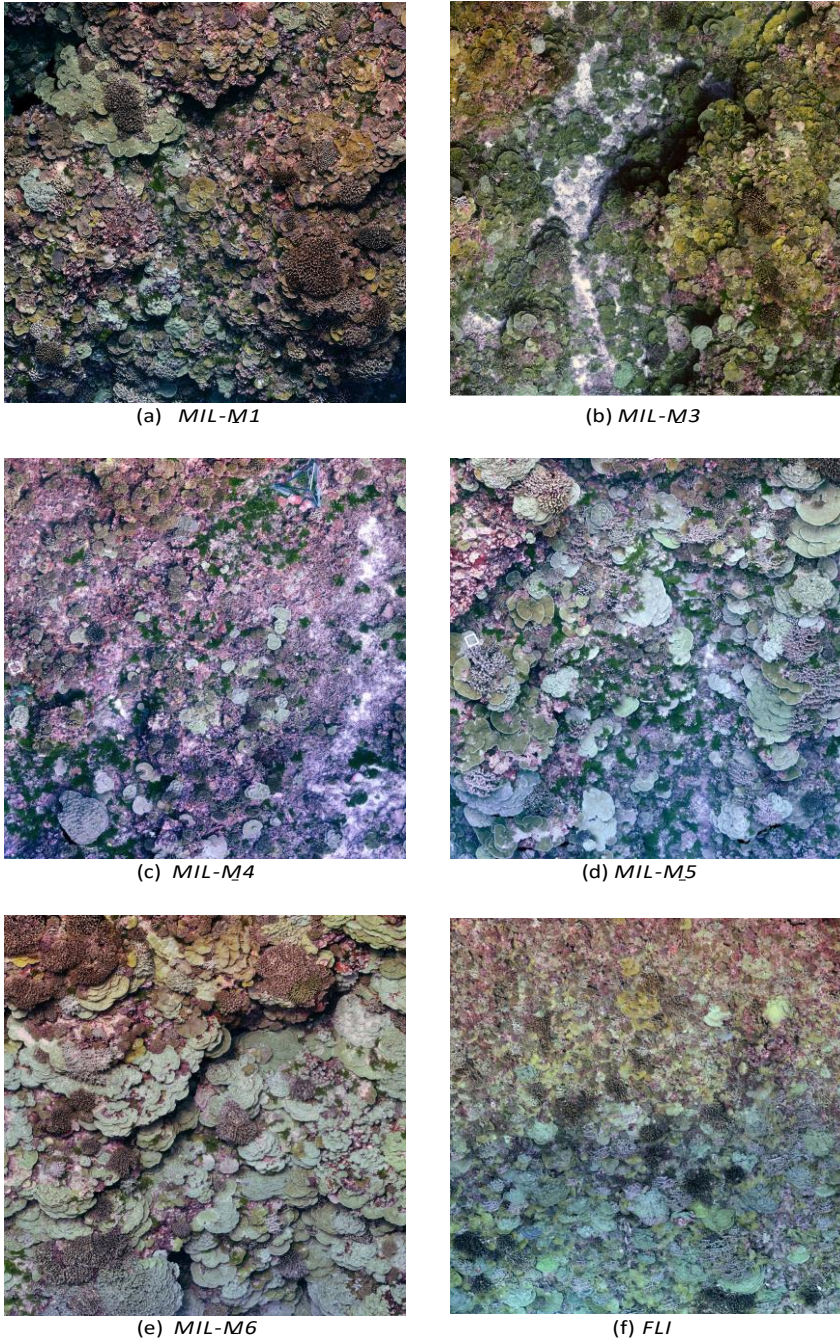(d) *MIL-M5*

(e) *MIL-M6*

(f) *FLI*

Fig. 10: Examples of plots included in the dataset. DeepLab V3+ was trained on *MIL-M5* and *MIL-M6*, validated on *MIL-M3*, and tested on *MIL-M1* and *MIL-M4*.
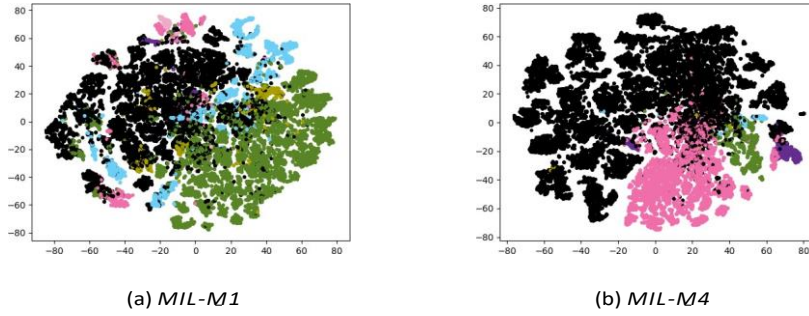
(a) *MIL-M1*   (b) *MIL-M4*

Fig. 11: Features visualization using the T-SNE algorithm [25]. Features of *MIL-M4* are better clustered compared to the ones of *MIL-M1*. An higher morphological complexity and a different coral distribution characterize the background class of *MIL-M1* (black color) which blends with coral classes.

Test set	Accuracy	mIoU
MIL-M5 + MIL-M6	0.915	0.847
MIL-M1	0.886	0.801
MIL-M4	0.972	0.947
FLI	0.831	0.771

Table 2: Performance of the CNN trained on *MIL-M5* and *MIL-M6*, validated with tiles coming from *MIL-M3*, and tested on the plots reported. Performance on *MIL* plots are overall acceptable. *FLI* is a nearby island with the same species.

7.2.2 Binary classifiers for orthos from different geographical origins

Large-scale recognition of corals is accomplished with a global photo coverage of the seafloor, not from high-resolution images of single colonies. Starting by a few meters from the seabed, some coral taxa can hardly be distinguishable, while others have more recognizable shapes, such as *Pocillopora*. Furthermore, *Pocillopora*, together with *Porites* and *Montipora*, are commonly occurring taxa, often covering a significant proportion of the benthos. For this reason, the creation of large datasets where all species are adequately represented can be accelerated by using a classifier that automatically labels common and easily recognizable species, leaving the problematic cases to biologists. With this goal, we train Deeplab V3+ to classify pixels of *Pocillopora* only, i.e., two classes: *Pocillopora* and Background.

A *Pocillopora* binary classifier was initially trained and tested starting from the four-plot dataset previously described. In order to reduce the class imbalance problem we apply a straightforward undersampling strategy: the tiles that do not contain *Pocillopora* pixels are removed. The *Pocillopora* pixels in the global dataset were about 9.33%, after the removing 16.73%.

In classifying *Pocillopora* alone the DeepLab V3+ reached an accuracy of 0.956 and a mIoU of 0.921 on a test dataset of tiles containing *Pocillopora* pixels, but an accuracy of 0.982 and mIoU 0.966 when calculated on the entire bottom where there were more *Pocillopora* pixels than 'Background'. Figure 12 compares manual segmentation (b), four-class automatic semantic segmentation (c), and automatic semantic segmentation of *Pocillopora* (d). The boxes highlight the most significant differences. In box D, *Porites* is separated in several segmentations by the human operator but is considered a single colony by the automatic algorithm. In the box E, a piece of *Pocillopora* is missing in the manual annotation but is correctly identified in both the automatic ones. In the box G, a small portion of *Pocillopora* is not detected in the automatic 4-classes segmentation, but it is correctly identified in the automatic binary one. The same situation is found in box A and F, although it is not clear which is the preferable solution. Box B contains a *Pocillopora* colony with much more rounded profiles than necessary. This effect is slightly less

pronounced in the 2-class segmentation. Finally, in the bottom right figure, on the lower edge of the tile, some pixels of *Pocillopora* are mistakenly classified. This is a common problem with pixels falling on the edges of the tiles: networks may fail in classifying

(a) RGB tile   (b) Human-made labeling

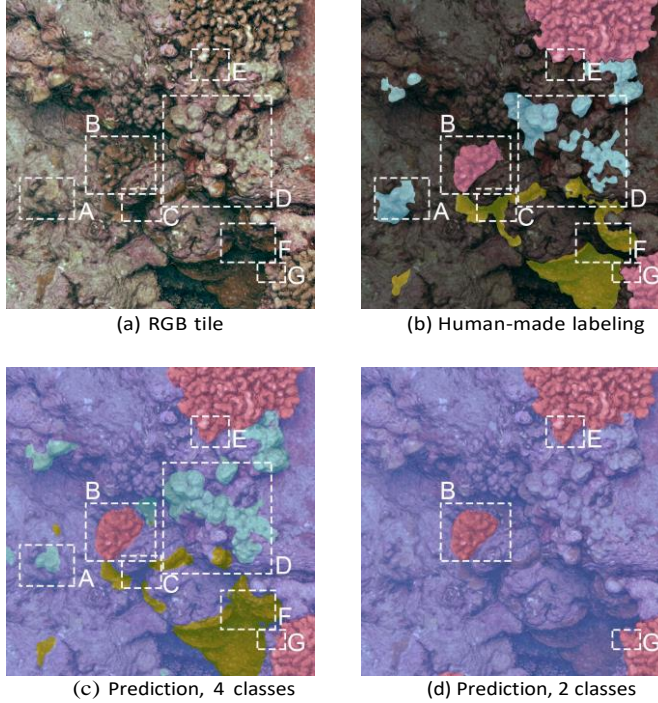(c) Prediction, 4 classes   (d) Prediction, 2 classes

Fig. 12: Results of automatic segmentation using DeepLab V3+ trained to classify respectively 4 and 2 classes (*Pocillopora* and Background). The area displayed has dimensions of 1026 × 1026 pixels.

partially portrayed objects. The experiment was repeated, adding tiles to the validation dataset belonging to different orthos, as in the previous experiment. The intent was training a binary classifier that works reasonably well independently of the geographical area. *HAW*, *FLI*, *MAI*, *MIL-M3*, and *MIL-M5* were used for training, *MIL-M1*, *MIL-M4*, and *VOS* for the validation, while *MIL-M6*, was employed for the test. The networks revealed an accuracy and a mIoU of 0.970 and 0.949 on its standard test. Performance on validation datasets were quite similar, 0.952 accuracy and 0.915 mIoU on *MIL-M1*, and 0.980 and 0.964 on *MIL-M4*. Finally, predictions on unseen orthos never ranged from the 0.940 accuracy and 0.891 mIoU on *MIL-M6* to 0.970 and 0.949 on *STA*. Orthos for which the binary classifier worked best were those with a homogeneous background class and without species morphologically similar to *Pocillopora* (see Table 3). Even if the accuracy values reached are high, predictions still require some corrections by the human expert to achieve the data quality necessary to detect changes in the colonies. Binary classifiers are a powerful resource to automate the segmentation of common classes, and, according to these experiments, they are applicable in any geographical area.

   Table 4 summarizes the results. From a qualitative point of view, the aggregation of the scores "blends" the abrupt classification discontinuity between adjacent tiles, as shown in figure 13.

Test set	Accuracy	mIoU
HAW + FLI+ MAI + MIL_M3+MIL_M5	0.970	0.949
MIL_M1	0.9526	0.915
MIL_M4	0.980	0.964
MIL_M6	0.940	0.891
STA	0.970	0.949

Table 3: Performance of the binary classifier trained on *HAW*, *FLI*, *MAI*, *MIL-M3* and *MIL-M5*, validated on *MIL-M1*, *MIL-M4*, and VOS, and tested on the following sets.

Method	Accuracy	mIoU
No fusion	0.942	0.894
Average	0.942	0.895
Bayesian fusion	0.944	0.897

Table 4: Classification results aggregating scores of overlapping tilesusing different methods.

8 Aggregation

As described in Section 6, the segmented overlapping tiles are aggregated in a single segmented ortho. The overall performance of the different strategies calculated on the plot *HAW* are reported in Table 4. Since the Bayesian fusion requires *a priori* knowledge of the distribution of the coral taxa, the average fusion is sometimes preferred. Even if numerical results are close to the tile re-arrangement without fusion, these strategies have a positive impact on the quality of the segmentation, as shown in Figure 13.

(a) Input Tile   (b) Ground truth label   (c) No aggregation   (d) Bayesian fusion
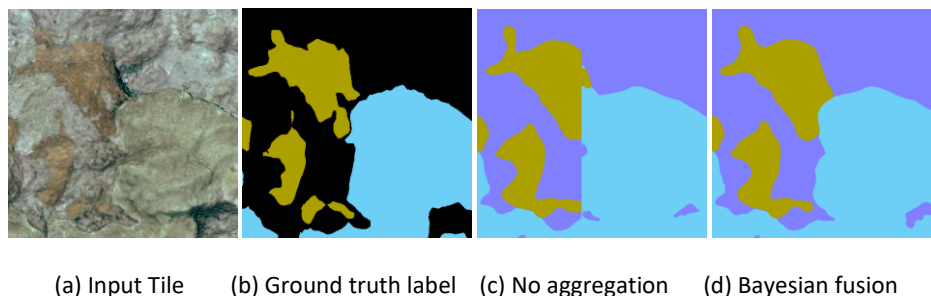
Fig. 13: Score aggregation across multiple overlapping tiles. Without aggregation predictions on tiles, the sides can be inconsistent (as shown in (c)), resulting in visible area 'cuts' after the tiles merging. The Bayesian fusion (d), as well as the average aggregation, mitigate the problem, producing a final labeling closer to the ground truth (b).

9 Discussion and Conclusions

The collection of underwater imagery for monitoring and research has rapidly expanded, leading to the accumulation of millions of images from coral reefs alone. This abundance presents challenges in annotating and extracting information from the images, as the ability of humans to manually annotate imagery is far outpaced by the rate of collection. To date, the development and use of semi-automatic and automatic annotation tools which have been adopted by coral reef scientists has accelerated the pace of image annotation [7]. However, these tools are limited in the type of data that can be extracted, as they only assist with point-based annotations, used for percent coverage estimation. While point based coverage have a widespread utility in describing current community assessments, they arguably lack the necessary information to detect change, understand the mechanisms driving that change, and make predictions about the future trajectory of these communities [32]. Demographic approaches are needed for an informative change detection, which requires data on size and abundance of colonies over time [18]. In the context of image analysis, this means semantically segmenting colonies of corals to a high degree of accuracy. The spatial information preserved when colonies are segmented across a broader landscape provides additional information about the spatial distribution inside a community, which can provide insights into biological and physical mechanisms which structure a population [14], especially when compared across life stages [30], as well as characterize the current successional stage [27]. Manual segmentation is a laborious process, even more than point-based annotation. Given the value of

ecological information provided by segmentation data, there is an urgent need for this process to become accelerated using automatic recognition.

Results obtained trained the DeepLab V3+, following the standard partition of dataset are remarkable (see Figure 7). However, from a practical perspective, the networks must infer predictions on new orthos, meaning they need to generalize. The enormous diversity of coral species presents a major issue, further complicated by the variability in appearance within and across geographic regions. Therefore, trained semantic segmentation networks were used on unseen orthos in the validation and test sets. Following this stress test procedure, we discuss:

Multi-class classifiers, which perform better on orthos from a constrained geographic region. On unseen test orthos they reach an accuracy in the range of [0.886 − 0.972] and a mIoU between [0.801 − 0.947].

Binary-classifiers trained to recognize the common species across various regions of the globe. These classifiers generate predictions on unseen orthos with accuracies between [0.940 − 0.970] and a mIoU between [0.891 − 0.949].

These results, compared with state of the art, score a better accuracy because of network improvements, training strategies, the introduction of orthos as working domain, and the use of accurate, manually-segmented, per-pixel labels for the network training. To date, the main challenges lie in:

1. the improvement of the generalization in order to deal with the extremely varied appearance of coral species.
2. the improvement in outlining specimen borders. As can be seen in Figure 12, the bulk of the uncertainty on the classification of coral colonies falls on the contours' pixels.

We expect that the performances can be improved when training on larger, taxonomically-rich datasets, as they might contain sufficient data for the network to learning rare coral taxa and for differentiating between species of the same genus, dealing misclassification caused by intra-species or intra-category morphological variability (see Figure 9). This method can be extended to the recognition of other species such as mollusks or encrusting algae. However, translucent or non-rigid, floating species, such as Macroalgae or seagrass, cannot be shaped by image-based 3D reconstruction methods, and require a direct analysis from images.

The automation of segmentation through Deep Learning techniqu ideally reduces the annotation times from the current 1h per square meter to the ten minutes of processing for a single $10m \times 10m$ plot (which can be further improved with more computational resources). However, this process might involve a manual editing step to correct errors.

Deep learning approaches assume that training and test sets contain objects following a similar distribution. Unfortunately, natural distributions of animal populations do not always meet this criterion. As it can be seen from the generalization tests of the 4-class and binary classifiers, the lowest accuracies occurred on orthos presenting a colony distribution most dissimilar from those in those orthos used for training. *Domain adaptation* is a learning task that takes into account the different features distributions between training and test sets. In this study, we reduce the domain adaptation problem by acting on the validation set and by selecting the best-performing network on new orthos.

10 Future directions

The 3D information is crucial to evaluate the volumetric change caused by the structural growth or erosion of coral reefs.

A significant number of studies focus on deep learning methodologies for 3D reconstruction and 3D semantic segmentation. However, according to a recent survey [17], current architectures are not able to produce accurate results over high-resolution reconstructions of complex objects, such as in the case with corals. Semantic segmentation of point clouds is a more mature field, where several solutions have been proposed, but many of those focus on indoor scenes. Moreover, most methods rely on volumetric representation [33], [10], affecting the resolution of the final segmentation. Other studies exploit a combination of

RGB-D representations of the scene and oriented images to solve the 3D semantic segmentation task. Dai et al. [12] presents a 2D/3D architecture (3DMV) for the 3D semantic segmentation of indoor scenes starting from RGBD scans. The 3DMV network joins feature maps extracted from RGB images together with 3D features of geometry producing per-voxel predictions.

An exciting aspect of working with ortho-projections derived from 3D image-based reconstructions is the opportunity to adopt a multi-modal approach. The manual annotation of ortho-projections is faster than the direct annotation on the point cloud, and avoids possible inconsistencies commonly found when annotating overlapping images. Starting from the labeled orthos, per-pixel class information can be propagated consistently through the reconstructed 3D geometry, and then back to the original images.

Section 9 highlights the promising performance of the proposed approach; however, despite the high accuracy values, in some cases, predictions might require some manual editing to achieve the data quality necessary to detect colonies changes. At the same time, we discussed how these excellent results are partially due to the high accuracy of the available per-pixel label. These annotations, which takes about one hour per square meter when performed in Photoshop by expert ecologists, are too much time-consuming to generate a benchmark dataset. For these two reasons, we developed TagLab[1]. This assisted annotation software, following a human in the loop approach, speed up per-pixel manual annotation, allowing at the same time the edit of automatic predictions. TagLab integrates the segmentation networks described in this work as well as another agnostic segmentation network explicitly fine-tuned for coral segmentation. TagLab also provide tools for image analysis and comparison of multi-temporal surveys. Future studies will be devoted to assessing how much the assisted annotation speeds up the experts' work.

The class imbalance and the smoother appearance of the predicted labels are two issues that we will face in future studies, starting from exploring different loss functions. The Boundary Loss [22], that has proved to be an effective solution to improve the contours prediction in imbalanced datasets. The Focal-Tversky [38], which mitigates the class imbalance problem without the need of pre-calculating class weights.

ACKNOWLEDGMENTS

References

1. E. Khvedchenya V. I. Iglovikov A. Buslaev, A. Parinov and A. A. Kalinin. Albumentations: fast and flexible image augmentations. *ArXiv e-prints*, 2018.
2. Agisoft Metashape, http://www.agisoft.com/ .
3. I. Alonso, A. Cambra, A. Muoz, T. Treibitz, and A. C. Murillo. Coral-segmentation: Training dense labeling models with sparse ground truth. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2874–2882, Oct 2017.
4. Iñigo Alonso, Matan Yuval, Gal Eyal, Tali Treibitz, and Ana C. Murillo. Coralseg: Learning coral segmentation from sparse annotations. *J. Field Robotics*, 36(8):1456–1477, 2019.
5. O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman. Automated annotation of coral reef survey images. In *CVPR*, pages 1170–1177, 2012.
6. Oscar Beijbom, Peter J. Edmunds, Chris Roelfsema, Jennifer Smith, David I. Kline, Benjamin P. Neal, Matthew J. Dunlap, Vincent Moriarty, Tung-Yung Fan, Chih-Jui Tan, Stephen Chan, Tali Treibitz, Anthony Gamst, B. Greg Mitchell, and David Kriegman. Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation. *PLOS ONE*, 10(7):1–22, 07 2015.
7. Oscar Beijbom, Peter J. Edmunds, Chris Roelfsema, Jennifer Smith, David I. Kline, Benjamin P. Neal, Matthew J. Dunlap, Vincent Moriarty, Tung-Yung Fan, Chih-Jui Tan, Stephen Chan, Tali Treibitz, Anthony Gamst, B. Greg Mitchell, and David Kriegman. Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation. *PLOS ONE*, 10, 07 2015.
8. Oscar Beijbom, Tali Treibitz, David Kline, Gal Eyal, Adi Khen, Benjamin Neal, Yossi Loya, B. Mitchell, and David Kriegman. Improving automated annotation of benthic survey images using wide-band fluorescence. *Scientific Reports*, 6:23166, 03 2016.
9. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam.

      Encoder-decoder with atrous separable convolution for semantic image segmentation.
10. Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d- r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016.
11. Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016.
12. Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 458–474, Cham, 2018. Springer International Publishing.
13. Glenn De'ath, Katharina E. Fabricius, Hugh Sweatman, and Marji Puotinen. The 27–year decline of coral cover on the great barrier reef and its causes. *Proceedings of the National Academy of Sciences*, 109(44):17995–17999, 2012.
14. Clinton Edwards, Yoan Eynaud, Gareth J. Williams, Nicole E. Pedersen, Brian J. Zgliczynski, Arthur C. R. Gleason, Jennifer E. Smith, and Stuart Sandin. Large-area imaging reveals biologically driven non-random spatial patterns of corals at a remote reef. *Coral Reefs*, 36:1291–1305, 2017.
15. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303– 338, June 2010.
16. Manuel Gonzalez-Rivero, Beijbom, Rodriguez-Ramirez, EP Bryant, Anjani Ganase, Yeray Gonzalez-Marrero, Ana Herrera-Reveles, V Kennedy, Kim, Sebastian Lopez-Marcano, Markey, P Neal, Osborne, Reyes-Nivia, M Sampayo, Stolberg, Taylor, Julie Vercelloni, Mathew Wyatt, and Hoegh-Guldberg. Monitoring of coral reefs using artificial intelli- gence: A feasible and cost-effective approach. *Remote Sensing*, 12:489, 02 2020.
17. Xianfeng Han, Hamid Laga, and Mohammed Bennamoun. Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence*, 2019.
18. Terence P. Hughes. Population dynamics based on individual size rather than age: A general model with a reef coral example. *The American Naturalist*, 123(6):778–795, 1984.
19. Terry P Hughes, James T Kerry, Mariana lvarez Noriega, Jorge G lvarez Romero, Kristen D Anderson, Andrew H Baird, Russell C Babcock, Maria Beger, David R Bellwood, Ray Berkelmans, Tom C Bridge, Ian R Butler, Maria Byrne, Neal E Cantin, Steeve Comeau, Sean R Connolly, Graeme S Cumming, Steven J Dalton, Guillermo Diaz-Pulido, C Mark Eakin, Will F Figueira, James P Gilmour, Hugo B Harrison, Scott F Heron, Andrew S Hoey, Jean-Paul A Hobbs, Mia O Hoogenboom, Emma V Kennedy, Chao-Yang Kuo, Janice M Lough, Ryan J Lowe, Gang Liu, Malcolm T McCulloch, Hamish A Malcolm, Michael J McWilliam, John M Pandolfi, Rachel J Pears, Morgan S Pratchett, Verena Schoepf, Tristan Simpson, William J Skirving, Brigitte Sommer, Gergely Torda, David R Wachenfeld, Bette L Willis, and Shaun K Wilson. Global warming and recurrent mass bleaching of corals. *Nature*, 543(7645):373377, March 2017.
20. K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Com- puter Vision*, pages 2146–2153, Sep. 2009.
21. Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *CoRR*, abs/1511.02680, 2015.
22. Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Éric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *International Conference on Medical Imaging with Deep Learning – Full Paper Track*, London, United Kingdom, 08–10 Jul 2019.
23. Andrew King, Suchendra Bhandarkar, and Brian Hopkinson. A comparison of deep learn- ing methods for semantic segmentation of coral reef survey images. pages 1475–14758, 06 2018.
24. Kevin E. Kohler and Shaun M. Gill. Coral point count with excel extensions (cpce): A visual basic program for the determination of coral and substrate coverage using random point count methodology. *Computers Geosciences*, 32(9):1259 – 1269, 2006.
25. Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
26. A. Mahmood, M. Bennamoun, S. An, F. Sohel, F. Boussaid, R. Hovey, G. Kendrick, and R. B. Fisher. Automatic annotation of coral reefs using deep learning. In *OCEANS 2016 MTS/IEEE Monterey*, pages 1–5, Sept 2016.
27. Dylan E. McNamara, Nick Cortale, Clinton Edwards, Yoan Eynaud, and Stuart A. Sandin. Insights into coral reef benthic dynamics from nonlinear spatial forecasting. *Journal of The Royal Society Interface*, 16(153):20190047, 2019.
28. R Ninio, Steven Delean, K Osborne, and H Sweatman. Estimating cover of benthic organ- isms from underwater video images: Variability associated with multiple observers. *Marine Ecology-progress Series - MAR ECOL-PROGR SER*, 265:107–116, 12 2003.
29. G. Pavoni, M. Corsini, M. Callieri, M. Palma, and R. Scopigno. Semantic segmentation of benthic communities from ortho-mosaic orthos. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W10:151– 158, 2019.
30. Nicole E. Pedersen, Clinton B. Edwards, Yoan Eynaud, Arthur C. R. Gleason, Jennifer E. Smith, and Stuart A. Sandin. The influence of habitat and adults on the spatial distribu- tion of juvenile corals. *Ecography*, 42(10):1703–1713, 2019.
31. Vid Petrovic, David Vanoni, Ashley Richter, Thomas Levy, and Falko Kuester. Visualiz- ing high resolution three-dimensional and two-dimensional data of cultural heritage sites. *Mediterranean Archaeology and Archaeometry*, 14:93–100, 01 2014.
32. B. Riegl P.J. Edmunds. Urgent need for coral demography in a world where corals are disappearing. *Marine Ecology Progress Series*, 2020.
33. Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Seg- cloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision*

(3DV), pages 537–547. IEEE, 2017.

34. Ivor D. Williams, Courtney S. Couch, Oscar Beijbom, Thomas A. Oliver, Bernardo Vargas-Angel, Brett D. Schumacher, and Russell E. Brainard. Leveraging automated image analysis tools to transform our capacity to assess status and trends of coral reefs. *Frontiers in Marine Science*, 6:222, 2019.

35. Karel Zuiderveld. Graphics gems iv. chapter Contrast Limited Adaptive Histogram Equalization, pages 474–485. Academic Press Professional, Inc., San Diego, CA, USA, 1994.

36. Culverhouse, P.F., Williams, R., Reguera, B., Herry, V. and González-Gil, S., 2003. Do experts make mistakes? A comparison of human and machine indentification of dinoflagellates. *Marine ecology progress series*, *247*, pp.17-25.

37. Durden, J.M., Bett, B.J., Schoening, T., Morris, K.J., Nattkemper, T.W. and Ruhl, H.A., 2016. Comparison of image annotation data generated by multiple investigators for benthic ecology. *Marine Ecology Progress Series*, *552*, pp.61-70.

38. Nabila Abraham, Naimul Mefraz Khan. "A novel focal tversky loss function with improved attention u-net for lesion segmentation." In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 683-687. IEEE, 2019.

39. Ian Goodfellow, Yoshua Bengio, Aaron Courville," Deep *learning"*, MIT press, 2016.

40. Ani Brown Mary, Dejey Dharma, "A novel framework for real-time diseased coral reef image classification", *Multimedia Tools and Applications,* **78,** 11387–11425 (2019).

41. Ani Brown Mary, Dejey Dharma, "Coral reef image/video classification employing novel octa-angled pattern for triangular sub region and pulse coupled convolutional neural network (PCCNN)", *Multimedia Tools and Applications,* **77,** 31545–31579 (2018).