SEMANTIC SEGMENTATION OF BENTHIC COMMUNITIES FROM ORTHO-MOSAIC MAPS

G. Pavoni[1,2]*, M. Corsini[2,3], M. Callieri [2], M. Palma [4], R. Scopigno [2]

[1] Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Pisa, Italy
[2] Istituto di Scienze e Tecnologie dell'Informazione "A.Faedo", CNR, Pisa, Italy
[3] Dipartimento di Ingegneria "Enzo Ferrari", Università degli Studi di Modena e Reggio Emilia
[4] Dipartimento di Scienze della Vita e dell'Ambiente, Università Politecnica delle Marche, Ancona, Italy

Commission II, WG II/9

KEY WORDS: Underwater Monitoring, Ortho-mosaic Map, Semantic Segmentation, Coral Reefs

ABSTRACT:

Visual sampling techniques represent a valuable resource for a rapid, non-invasive data acquisition for underwater monitoring purposes. Long-term monitoring projects usually requires the collection of large quantities of data, and the visual analysis of a human expert operator remains, in this context, a very time consuming task. It has been estimated that only the 1-2% of the acquired images are later analyzed by scientists (Beijbom et al., 2012). Strategies for the automatic recognition of benthic communities are required to effectively exploit all the information contained in visual data. Supervised learning methods, the most promising classification techniques in this field, are commonly affected by two recurring issues: the wide diversity of marine organism, and the small amount of labeled data.
In this work, we discuss the advantages offered by the use of annotated high resolution ortho-mosaics of seabed to classify and segment the investigated specimens, and we suggest several strategies to obtain a considerable per-pixel classification performance although the use of a reduced training dataset composed by a single ortho-mosaic. The proposed methodology can be applied to a large number of different species, making the procedure of marine organism identification an highly adaptable task.

1. INTRODUCTION

In recent years, neural networks have been successfully used to recognize marine organisms. In particular, CNN have demonstrated to obtain reasonably good performance in the segmentation of benthic communities (King et al., 2018, Alonso et al., 2017). While speeding up the recognition step, the use of neural networks require the preparation of a large training dataset. The commonly used labeling methodology for underwater classification is a point-based manual annotation on all the photos of the input dataset, that has to be carried out by experienced personnel, resulting in a very time-consuming process. To our knowledge, when using point-wise annotations, all the existing approaches based on SVM, CNN or FCNN models ((Beijbom et al., 2015), (King et al., 2018), (Alonso et al., 2017), (Mahmood et al., 2016)) adopt a patch-based labeling, cropping a square area around each annotated point. The Patch-based labeling could lead to sparse training datasets (Alonso et al., 2017); furthermore, the annotated points falling close to the contours of the specimen introduce a certain amount of uncertainty in the annotation, depending on the size of the extracted patch. In (King et al., 2018) the authors compare the performance of the state of the art architectures, using different annotation types. Best results were obtained by free-hand drawing labels on the investigated specimens, however, producing a similar dataset is extremely time-consuming. Nowadays, the polygonal annotation of ortho-mosaics inside GIS tools is a raising trend among biologists that employ geo-referenced maps to study the spatial distribution of populations. In this work, we analyze efficient ways to use this new, available, data format for the training of a semantic segmentation CNN for benthic communities.

In section 2 we describe the advantages of working on polygon-annotated ortho-photo map when training CNN architectures. Then, we discuss several simple strategies, exploiting the properties of the starting data, to improve the per-pixel classification performance.

- A biologically-inspired method to partition the input map into the *training*, *validation* and *test* datasets.
- A simple but effective oversampling method, able to cope with the class imbalance problem.
- A way to aggregate network scores using the prior information about the actual coverage of the specimens on the surveyed area.

Finally, as a final step in our workflow, we employ a validation tool, to analyze the semantic segmentation of complex natural structures (Pavoni et al., 2019). This tool can be used to validate the network predictions, or to correct the human labeling inconsistency, feeding back into the network a cleaner and enhanced dataset for a possible re-training step. The improvements obtained by introducing these methods in the network training and execution are outlined in section 4.

2. METHODS

Coral reefs are populated by a huge amount of species, and since we are working with RGB images, learning to classify them coincide with learning some peculiar features in their morphology and color. However, the task of identifying this features by photographs, both for human users and for supervised learning methods, is complicated not only by the intraspecific mutability of

*Corresponding author (gaia.pavoni@isti.cnr.it)

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W10-151-2019 | © Authors 2019. CC BY 4.0 License.

marine organisms, but also by environmental factors, by perspective issues or by the flaws of underwater images. The same coral colony, framed from a different view angle or distance may appear totally different. Underwater photos are affected by color and sharpness changes (due, respectively, to the water absorption of some wavelenghts and to the turbidity), by distortions and by chromatic aberrations (caused by the interaction of the camera lens with the port of the camera housing and with the water medium). From a machine learning perspective, the use of ortho-mosaic maps reduce the amount of factors to learn by removing some of the described inconsistencies.

Ortho-mosaics maps are rectified, have a constant pixel size, and always present an azimuthal viewpoint, framing all the specimens with a coherent view direction. Additionally, color variations across photos are smoothed out by the blending process, and chromatic aberration is mostly removed. These factors reduce the non-biological variability of the corals images, helping the network to learn what really differentiates them from a biological-only perspective.

Ortho-mosaics incorporate information of the actual metric scale, depth and the geographical coordinates. The output of the networks can be easily used for computing specimen areas, abundance and coverage of the species. From the perspective of the coral segmentation task, the physical size of morphological features, as well as the depth of the coral colonies, are a discriminant factor in classification.

Thanks to geo-referencing, identified objects can be relocated in a three dimensional context preserving the spatial distribution of the specimen as an additional and valuable information supporting monitoring activity.

As we will show in the next, working with maps also allows us to exploit the scale and the spatial distribution of the population to solve issues commonly found in preparing training datasets, such as the partition and the class imbalance.

Finally, inspired by the polygonal annotation made by (Palma et al., 2017) for the calculation of biometric indicators, we used the manual polygonal labeling as an annotation solution for a supervised learning dataset. In this context, with respect to the point-wise photographic annotation, the use of ortho-photo mosaic seabed maps leads to faster labeling time, since drawing an approximate polygon around a specimen is faster than annotate hundreds of points on several images. Additionally, the polygonal annotation allows the use of a segmentation network instead of a patch-based classification one, providing a pixel-wise classification.

While solving many issues, ortho-photo mosaics may introduce different problems. Small registration errors of the input images may cause local blurring or ghosting in the final map, that have to be somehow recognized and removed from the input dataset. The stitching and projection process causes local image warping close to geometric discontinuities (e.g. the borders of the more protruding corals), we overcome this problems by masking these areas to exclude them from the loss computation.

2.1 Biologically-inspired dataset partition

Typically, to train a network in a reliable way, the available input data is split into three datasets: one for the actual training, one for the validation (used to tune the network's hyper-parameters) and one for the testing of the network performance (to assess the generalization capability of the trained network). In order to properly work, these three datasets must be representative of the whole data. A simple random partition works well in datasets which are intrinsically uniform, such as the ones for the automatic recognition of pedestrian and cars. However, in our case, we are dealing with a continuous space (the reef map), where the organisms follow a non-uniform population distribution (Edwards et al., 2017). For these reasons, instead of subdividing the data into random non-overlapping parts, we chose three sub-areas of the monitored seabed by using ecology metrics describing the spatial patterns of benthic communities.

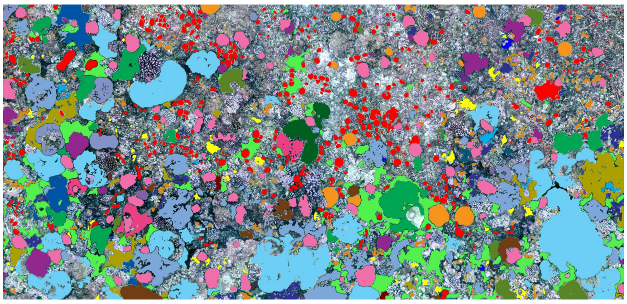

Figure 1. Coral class distribution. Courtesy of Scripps Institution of Oceanography.

Among all the landscape ecology metric commonly used (and implemented in many popular software such as FRAGSTAT) we select three that describe the colonies distribution and are invariant to the shuffle of the tiles. The *Patch Size Coefficient of Variation*:

$$PSCV = \frac{100.0 \cdot \mathrm{std}(\text{Coral's areas on the Patch})}{\mathrm{mean}(\text{Coral's areas on the Landscape})},$$

which measures the standard deviation of the size of specimens as a percentage of the mean size all over the dataset. It is commonly used to describe the landscape area variability. *Patch Richness* and *Patch Coverage* are related, respectively, to the number and density of specimens. Obviously, other metrics can be integrated in this method: for example, at the moment we are not considering metrics related to the perimeters because our polygon labeling is not that much detailed.

The selection criteria proceeds by choosing on the map a couple of non overlapping windows (our *Patches*) of the approximately dimensions of about the 15% of the entire surveyed area. Metrics are computed on each window and on the remaining area; a weighted sum of the calculated values is assigned to each of the three regions as a similarity score (S). This process is iterated an arbitrary number of times ($\sim 10,000$), and then the triplet with the best values of S is chosen as validation, test and training area respectively.

In order to combine the metrics in a weighted sum in an homogeneous way, we express them in percentage w.r.t the statistics of the labeled population. The weight used have been set empirically after some tests. Values of S close to 0.0 means that the three areas have very similar statistical characteristics, great values that the area chosen are very different w.r.t the population.

This strategy is trivial when applied on a single class, but still produces a better balanced partition with respect to a random choice. When the number of classes to segment increases, this method will give an even stronger advantage, as a manual selection would be impractical and a random process would not be viable.

The image on the figure 1 is a labeled ortho-projection of a three-dimensional reef reconstruction carried out by the Scripps Institution of Oceanography (100 Islands Challenge Team, 2019); this clearly shows how difficult it might be to manually select areas displaying an adequate class representation. For such complex

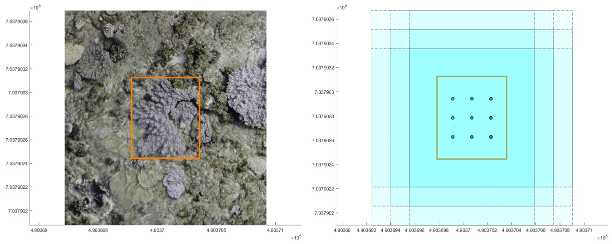

Figure 2. (Right) Bounding box of a coral in the ortho-mosaic map. (Left) Cropping Windows for the oversampling in data space; the points indicate the centers of such windows. This coral will be replicated 9 times in the final dataset.

cases, the two windows might be divided into smaller areas with separate scores, and then re-combined to reach a properly balanced set.

2.2 Corals Oversampling

The dataset used for the tests, as many other datasets of this type, only contains a small number of representatives for the species under monitoring, compared to the extension of the surveyed area. This causes a significant problem of class imbalance when training the network. To solve this issue, we propose a simple oversampling strategy in data-space, based on the actual area coverage of the specimens.

We subdivide each annotated coral into a set of overlapping subparts: the number of parts is proportional to the size the coral, while their arrangement follows the shape of the specimen. The bounding box of each coral is regularly sampled, using as a step the average size of the smaller specimens: a cutting window is applied on each of these sampled points. The size of the cutting window is determined by the size of the network training input and by the maximum translation applied during the data augmentation. Each cropping window give us a new input image tile for the input dataset. An example of the cropped tiles of a coral is shown in Fig. 2.
This sampling produces a set of tiles that follow the size and shape of the specimens, making possible to feed the network with all the coral borders, and to further apply a random-displacement and/or rotation step of augmentation at training time, further increasing the coral percentage.

This strategy is motivated by the presence of a very large number of small corals (which are therefore well represented in the dataset), and only a few of large ones. Since we are working with natural, growing structure, the idea is to give the same importance to small and large corals into the training and validation sets.

Classic feature-space oversampling are difficult to apply in this specific case, because the pre-trained CNNs typically employed for features extraction are trained on dataset that does not contain this type of visual data. Additionally, standard augmentation alone would not be able to take into account the spatial continuity of the larger individuals and not guarantee to cover all the corals border. Working with ortho-mosaic makes this processing step possible, as the ground sampling distance is known, and larger specimen are not scattered across multiple photos but are represented by a single area on the map.

Figure 3 shows a dataset sample, largest coral appears several time in different position into the cropped area.

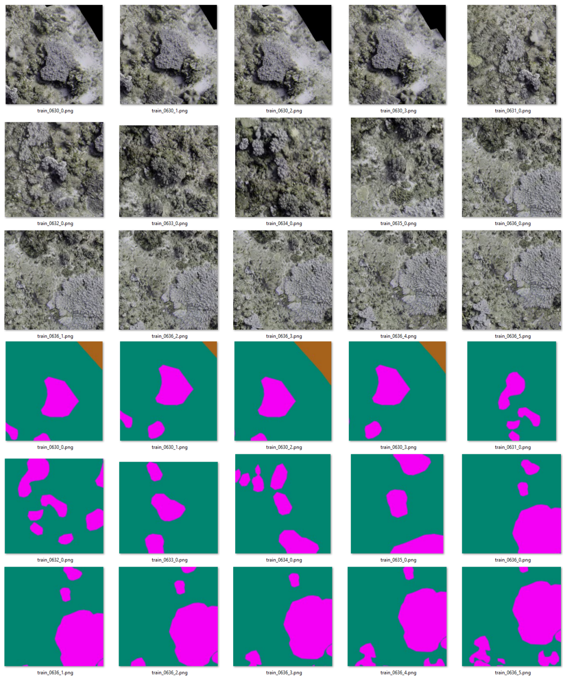

Figure 3. An excerpt of the dataset after the coral oversampling. (Top) RGB images. (Bottom) Corresponding labels.

2.3 Re-assembling the output of the CNN

Ideally, most of the CNN are translation-invariant because are based on convolutional filters and max pooling layers. However, the padding operation in the convolutional filters introduces small but significant differences. In remote sensing applications, the size of the ortho-photo map is usually too large to be processed entirely with a single pass for memory constraints. In these cases, the segmentation is applied on an overlapping sliding window to ensure the class consistency, in particular on the image borders. Since this approach produces more than one per-pixel classification score, a method to re-assemble the segmented map is required.

The standard method to obtain the final scores is to simply average the overlapping results (Liu et al., 2017, Audebert et al., 2017). Here, we propose to employ a method already used in multi-view stereo matching (Ma et al., 2017); using the Bayesian Fusion to aggregate the scores that belong to the same pixel.

Defining $S_N = \{s_1, s_2, \ldots, s_N\}$ a set of classification scores for a given pixel, generated by the sliding window in different positions, according to the Bayes rule we can write:

$$p(y|S_N) = \frac{p(S_N|y, S_{N-1})p(y|S_{N-1})}{p(S_N)} \tag{1}$$

where y is the output of the network for that pixel. By assuming that the scores are i.i.d, it is possible to write:

$$p(y|S_N) = \mu p(y) \prod_{i=1}^{N} p(s_i|y) \tag{2}$$

where μ is a constant. At this point, the final Bayesian aggregation becomes:

$$p(y = 0|S_N) \quad = \quad p(y = 0) \prod_{i=1}^{N} p(s_i|y)$$

https://doi.org/10.5194/isprs-archives-XLII-2-W10-151-2019 |

$$p(y=1|S_N) \quad = \quad p(y=1)\prod_{i=1}^{N} p(s_i|y) \tag{3}$$

Note that $p(y=0|S_N)$ and $p(y=1|S_N)$ should be normalized to obtain the probability that the pixel belongs to the species of interest or not.

An interesting aspect of this formulation is that it includes the *a priori* probabilities of the presence of the species of interest: i.e. if we know in advance that, in the surveyed area, the coverage of the species is not more than the $X\%$ of the seabed, it is possible to take into account this information, and produce more reliable scores for the pixels with an uncertain prediction (probabilities close to 0.5).

3. VALIDATION OF THE SEGMENTATION

The human error in labeling specimens adds uncertainty both at the training phase and at the testing of the network. This lack of consistency depends on the operator's experience in distinguishing exactly those types of marine organisms but also on the repetitiveness of the task. In (Beijbom et al., 2015), authors show that the expert introduce an intra-annotator error of about the $10-11\%$. With the aim to reduce such source of errors we proposed an interactive validation tool (implemented in PyQt) that allows the user to confirm, reject or copy an existing annotation with a single click.

The tool interface is shown in figure 4. In the *Comparison Panel*, the two main windows, allow the user to analogize the original human annotation (on the left) with the network predictions (on the right). The *Navigation Panel*, in the top-right, enable to register the actions performed within the map.
This tool can be used in two different ways. It is possible to work on the classification output of the network applied to an unseen area, in order to efficiently correct erroneous predictions, calculating at the same time the adjusted percentage of abundance and coverage of the species. However, it is also possible to use it to check the classification output of the network when applied on the entire input dataset: in this case, it allows to quickly correct the errors in the input annotations, by comparing the results with the original labels. At this point, a more correct labeling might be exported and used to re-train the network.

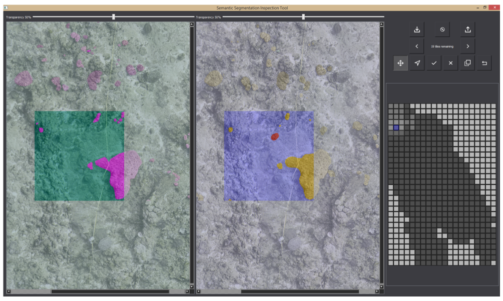

Figure 4. The semantic segmentation validation tool.

4. RESULTS

We test our strategies on a 150 × 50 meters wide ortho-mosaic of the barrier reef already investigated in (Palma et al., 2017), containing various coral species, as well as rocks and sand regions, labeled by a single biologist (see Fig. 5). One pixel of the orthophoto map covers around $1.14mm$.

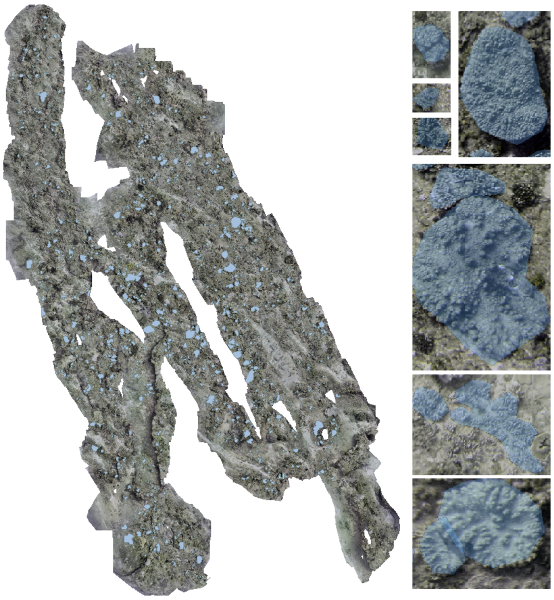

Figure 5. The *Mozambique* geo-referenced ortho-mosaic with the corresponding polygonal labels. On right, some examples of annotated corals.

The ortho-mosaic, generated with Agisoft PhotoScan, was built from color-corrected images. For this purpose, we employed a combination of the CLAHE algorithm (Zuiderveld, 1994) in the *Lab* color space with a successive small auto adjustment of the RGB components. This global color adjustment works very well in our case, the mean value of the images extracted from the ortho-mosaic for the training is always close to middle gray.

We only have a single labeled coral class, the *Soft Coral Digitate*, which shows a large intraspecific morphological variance (see Fig. 5) and covers approximately just the 6.4% of the seabed. The 'other' class contains elements that are poor in features, such as sand, but also other corals classes morphologically similar to the monitored class, which are thus excellent candidates to be false positives. Our labels are loosely-fitting polygons surroundings the corals, marked on a separate layer using QGIS, not fully coherent with their edges (see Fig. 5). However, this annotation technique is very fast and probably the most suitable in relation to the smooth appearance of our data.

4.1 Network and training parameters

We do our experiments using a standard pre-trained CNN architecture for semantic segmentation tasks, the *Bayesian SegNet* (Kendall et al., 2015). Dataset tiles are pre-processed by subtracting the mean value, no further normalization are required thanks to the previous color adjustment. The fine-tuning of the network is obtained using an initial learning rate of $5 \cdot 10^{-5}$. This learning rate is reduced by a factor 5 every 50 epochs, for a total of 150 epochs. The optimizer is Adam with a small weights L_2 regularization (0.0005). Higher values of the regularization tends to oversmooth the coral borders.

Each input images is a tile of the ortho-map of 448×448 pixels (two times the input size of the pre-trained SegNet). We used an image of such size to permit large translation during the data augmentation. In particular, each input tile is randomly flipped horizontally and/or vertically, translated in a range of ± 50 pixels, and rotated in a range of ± 10 degrees. The augmented image is cropped centrally to obtain the input size of the network

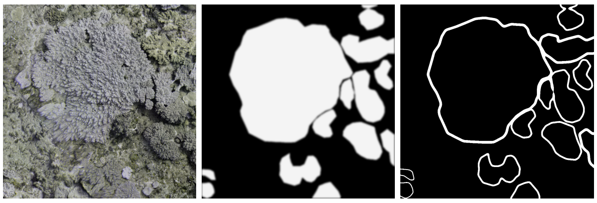

Figure 6. A dataset tile (Left), the probability map (Center), cross-entropy mask (Right)

(224 × 224). We decided for these augmentation parameters after some empirical tests. The batch size used is 32.

The architecture follows the typology of the input labeling. The Bayesian Segnet, returns homogeneous regions predictions, i.e. blobs, that are suitable to approximate the polygonal annotation. However, coral colonies borders are generally jagged: more recent architectures, such as *DeepLabV3+* (Chen et al., 2018), could be used to obtain a more precise per-pixel labeling.

Strategies to deal with the inaccurate labeling of the polygonal annotations. Polygonal annotation of corals is more accurate then the one based on points, but some labeling problems still remain. Coral contours are often too irregular to be annotated with a simple polygon, and on high-resolution orthomosaics, texture artifacts may appear in correspondence of depth discontinuities. To deal with this problems, we tested two different strategies to deal with the *uncertainty* around corals borders.

In the first strategy, probability maps are calculated by individually rasterizing each label, and applying a gaussian filter with a kernel of 20 pixels ($\sigma = 5$ pixels) to smooth out its borders. The idea is to treat the labels as probability maps, with probability to have coral equals to 1.0 when inside the corals, but with a decreasing ramp to 0.0 across the area of the border (see Fig. 6). In this case a binary cross-entropy loss is used.

In the second strategy, the rasterized labels are thresholded to carve out a thin mask covering the area across the border such that the thickness of the mask is proportional to the size of the coral (see Fig. 6). This time we decided to exclude the borders from the loss calculation. The rationale is try to learn the inner region of specimens, i.e. the "inner patterns" of the corals. This time we chose a cross-entropy loss function because we are dealing with binary maps. The adaptive masking is necessary to prevent loosing too much useful data on small corals. In our dataset, we set the minimum thickness to 7 pixels and a maximum one of about 15-16 pixels.

According to our tests we can state that the second solution is inefficient. The network is not able to segment the coral borders properly; the learned "inner patterns" does not guarantee better precision w.r.t. the whole polygonal labeling. Conversely, the solution based on treating labels as probability maps and adding an uncertain field around the borders is able to reduce the FPR to about 0.8-1.0%.
This result, combined with the Bayesian Fusion (see Table 5), gives the best performance in terms of False Positive Rate (FPR), i.e. 2.5%, w.r.t the network with the best accuracy (0.960 vs 0.962) and the best F1-score (0.641 vs 0.650). This result is very encouraging and worthwhile more investigation to better assess its advantages.

In recent works, as in (Maninis et al., 2018), the uncertainty of the sketch labels is solved by transforming them into heat maps, then added in an additional information channel to train the network. The Bayesian Segnet that we used for our tests was pre-trained with three channels, but we reserve to try this approach with a multi-channel network because it seems very promising.

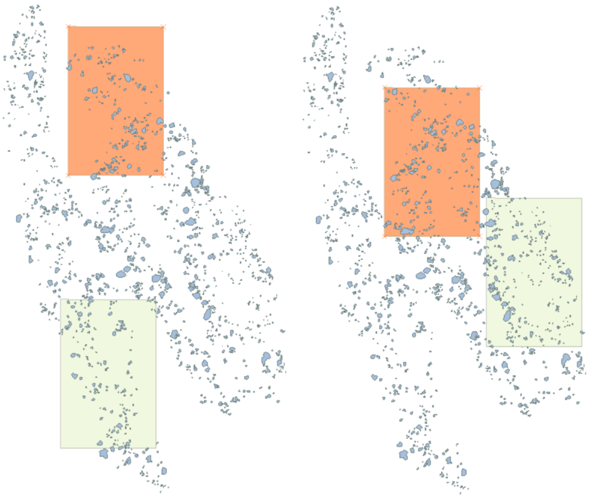

Figure 7. The *Random Window* dataset (on the left) and the *Selected Windows* dataset (on the right). Orange and Olive color indicate respectively the validation and the test Area.

Dataset		SCD Class (predicted)	Other Class (predicted)
Random Windows	SCD Class	0.511	0.489
	Other class	0.027	0.973
Selected Windows	SCD Class	0.585	0.415
	Other class	0.029	0.971
Random Windows + Weighted	SCD Class	0.872	0.128
	Other Class	0.146	0.854
Selected Windows + Weighted	SCD Class	0.881	0.119
	Other class	0.104	0.896

Table 1. Confusion matrices for the biologically-inspired selection and random selection.

4.2 Biologically-inspired dataset partition

We used our area selection method, based on the spatial analysis of the populations, to identify the best training, validation and test area. These areas are shown in Figure 7. The obtained S scores are 5.1, 5.5 and 5.8 for the test, validation and training dataset respectively. This means that the descriptive statistics of these areas are all quite close to one another, as we wanted.

From now on, we will refer to the dataset obtained by splitting the tiles into training, validation and test sets using the areas selected with our method as *Selected Windows*. Similarly, we will refer with *Random Windows* to the dataset in which the tiles are split using randomly selected areas. In this case, the scores are a bit higher (around 20.0), i.e. the random areas are still quite similar but not be as good as the ones chosen with our method.
Tables 1 and 2 summarize the obtained results. The term 'weighted' means that, to compensate the class imbalance, we also introduced a weighted cross entropy loss with the weights set as the inverse of the class frequency.

As performance metrics, we consider the overall **accuracy** and the **F1-score**, that measures the test accuracy more efficiently in

Dataset	Accuracy	F1-Score
Random Windows	0.949	0.511
Selected Windows	0.944	0.591
Random Windows + Weighted	0.856	0.388
Selected Windows + Weighted	0.895	0.535

Table 2. Comparison of accuracy and F1-Score between the biologically-inspired selection and random selection.

uneven class distribution. To better assess the amount of coral correctly classified w.r.t the *Other* class, we also report the corresponding confusion matrices. Note that due to the strong imbalance of the classes, a small classification error on the *Other* class greatly increases the amount of incorrectly classified pixels.
According to the metrics used the network trained using the Selected Windows is able to classify correctly more SCD (58.5% vs 51.1% of the Random Windows) with basically the same FPR. As a matter of fact, it has the best F1-Score. The number of tiles used for training with the Random Windows is greater than the Selected Windows (906 vs 820), however, the biologically-inspired selection still outperforms the random selection.
This result is significant, also considering that we are working on a relatively regular colonies distribution: the Mozambique dataset portrays a flat reef, without major slopes, located centrally inside the barrier and not subject to external currents. Tests conducted on other sets of Random windows with comparable scoring give similar qualitative results. The performances greatly degrades when the Windows have much higher values of S, i.e. > 40.
Regarding the use of the weighted cross-entropy loss, the imbalance of our test case is so severe (the number of pixels of the Soft Coral Digitate are about 6.4% of the entire ortho-map) that this standard technique is unable to produce good results (compare the $F1 - Score$ in table 2). This further motivates the use of our oversampling strategy to reduce the imbalance problem.

4.3 Corals Oversampling

As already stated, our dataset is heavily unbalanced (6.4% of coral pixels according to the provided labeling information). Our simple oversampling method (see Section 2.2) is able to feed the network with an amount of coral pixels between the 35% and 40% of the training dataset, making it balanced and increasing the performance considerably.

Tables 3 and 4 report the results obtained on the Random Windows and the Selected Windows dataset after our data balancing step. The oversampling step during the corals cropping tiles is about 132 pixels, which corresponds to cover the area with a step of 16 cm in physical space. The Dense Oversampling test used a step of 66 pixels, producing a higher number of cropped tiles for the large corals. Note that this solution increases the True Positive Rate (TPR) and the accuracy significantly, outperforming the class weighting. The best accuracy reach 94.5% with an F1-score of 0.674. Intuitively, a severe oversampling also introduces redundant data that does not improve the performance anymore; this can be seen by looking at results of the the dense oversampling test.
Unlike doing a random augmentation, with this data oversampling approach we are sure to cover the area of each SCD instance.

4.4 Bayesian Fusion vs Standard Fusion

Dataset		SCD Class (predicted)	Other Class (predicted)
Random Windows + Oversampling	SCD Class	0.793	0.207
	Other class	0.064	0.926
Selected Windows + Oversampling	SCD Class	0.809	0.191
	Other class	0.044	0.956
Selected Windows + Dense Overs.	SCD Class	0.826	0.174
	Other class	0.048	0.952

Table 3. Comparison between selected and random areas.

Dataset	Accuracy	F1-Score
Random W. + Oversampling	0.929	0.537
Selected W. + Oversampling	0.945	0.674
Selected W. + Dense Oversampling	0.943	0.670

Table 4. Performance of the proposed oversampling approach.

As described in Section 2.3, the classification of a large map generated by using a sliding window approach may cause problems like inconsistent classification at tiles borders. The standard method to reduce such inconsistencies is to average the scores on a set of overlapping window. Typical overlap values used are 50% or 75%. In the following, we compare the averaging of the scores (after the Softmax layer) with our Bayesian Fusion approach.

Figure 8 shows an example of output probabilities produced for a single sliding window. As clearly visible, the proposed Bayesian Fusion can produce scores that show less uncertainty with respect to those produced by simply averaging the input. This reduces the presence of ambiguous range values, i.e. around 0.4-0.6, and helps to remove the smaller wrongly-classified SCD. The performance evaluated against the ground truth of the scores aggregated using the Bayesian Fusion are slightly better than the ones of the scores aggregated with the simple average (see Table 5). Note that the overlap of 75% does reduce the performance, most probably due to the weak labeling at the corals' edge, that corresponds to the zone of high uncertainty. The role of the prior probabilities can be easily understood by taking a look at the confusion matrix reported in Table 6. Since we know that the presence of the SCD has a low probability w.r.t the Other class, the effect of the Bayesian fusion is to make the coral classification more strict, but at the same time the FPR is greatly reduced.

4.5 Using the validation tool

We compared two successive annotations performed in QGIS by the same biologist after a few months. The biologist needed approximately 25 hours to verify and correct the previous annotation using QGIS. Employing the validation tool for executing the same task, the biologist took only 9 hours (about 9.2 seconds per instance).
Figure 10 shows in light blue the polygons annotated consistently, in dark blue the false negatives (labels missing in the first session), and in red the false positives (labels incorrectly classified in the first session). More accurate results about the validation tool performance are described in (Pavoni et al., 2019).

Method	Overlap	Accuracy	F1-Score
Average	50%	0.958	0.640
Bayesian	50%	0.962	0.650
Average	75%	0.957	0.641
Bayesian	75%	0.959	0.644

Table 5. Performance of the proposed Bayesian Fusion approach vs the usual Averaging method. Prior probabilities are set 0.2 for the SCD class and 0.8 for the "other" class.

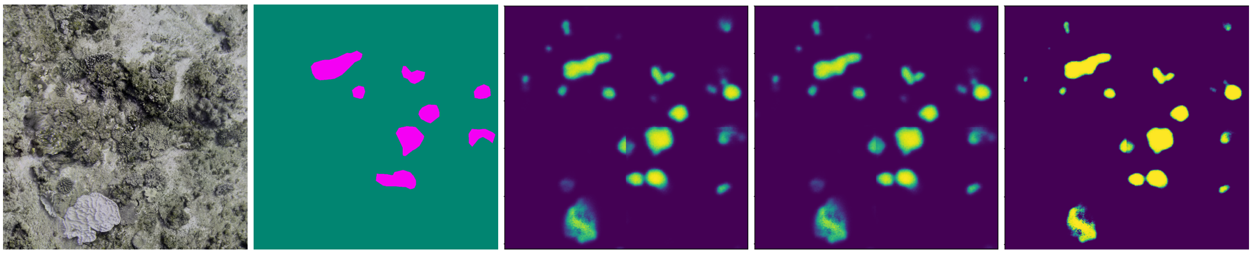

Figure 8. (From Left to Right) Input image. Manually annotated labels. Classification scores without aggregation. Average scores. Bayesian Fusion.

Method		SCD Class (predicted)	Other Class (predicted)
Average	SCD Class	0.808	0.192
	Other class	0.036	0.964
Bayesian1	SCD Class	0.772	0.228
	Other class	0.029	0.971
Bayesian2	SCD Class	0.718	0.282
	Other class	0.027	0.973

Table 6. Comparison between Averaging and Bayesian Fusion. All methods use 50% overlap. Note that Bayesian1 corresponds to the prior probabilities set to 0.2 and 0.8 respectively, while Bayesian2 corresponds to 0.1 and 0.9 (in this last case we have accuracy=0.960 and F1-score=0.639).

This time, we used the tool to validate the predictions with our best performance network, Selected Windows + Oversampling + Bayesian (50%), over the Test area. The biologist needed approximately one hour to complete the task, with the following results:

- 73 new instances, about the 15% of the Test area coral pixels, were detected. Since coral instances are easily identified when suggested by an automatic segmentation, this might justify the development of an assisted input tool.
- 40 small (about 10%) specimens predicted have not been validated because considered "uncertain", confirming the complexity of the task.
- The FPR, that was about 2.9% in the confusion matrix, decreased to 1.8% after the biologist validation.
- The TPR increased, in the same manner, from to 77.2% to 81%.

The new accuracy of the network, according with these values, became **0.967**.

4.6 Discussion and Qualitative Comparison

We report here a qualitative comparison with the work by King et al. (King et al., 2018) where the authors analyze several FCNN architectures in the task of semantic segmentation of coral reefs. In this field, making analogies with the performance of other supervised techniques is complicated by the lack of standard benchmark datasets. The input data are very similar to ours since the different classes have been annotated directly onto an ortho-photo map using a proprietary tool, but the SegNet has not been tested in the comparison. The most promising network tested in (King et al., 2018), DeepLab v2, gives an accuracy of about 0.677 in classifying ten classes. We reached a maximum accuracy of 0.950 in classifying the *Soft Coral Digitate* class.

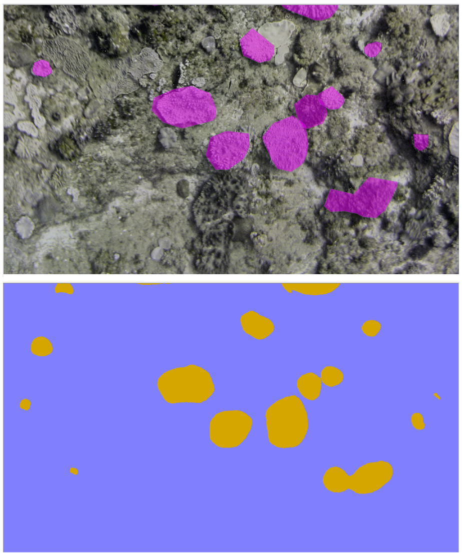

Figure 9. (Above) A close-up of the polygon labeled ortho-mosaic. (Below) The corresponding semantic segmentation obtained by our network on the Test area.

In general, we can state that the overall performance of our network is in line or better than the current state of the art: an example of a segmentation result is shown in (see Fig. 9). We point out that other similar solutions, without the described improvement strategies, requires larger dataset and more information to obtain similar quantitative performance, for example, the use of fluorescence data in addition to the RGB data by Alonso et al. (Alonso et al., 2017).

Regarding the proposed methods, the biologically-inspired dataset partition has demonstrated to work properly. We think it can be even more efficient with a higher number of classes and it can be improved choosing more specific metrics for the specimen under evaluation.
According to the results, the oversampling strategies based on size- and shape-driven cropping of the specimen has been very effective to overcome the lack of data often characterizing this type of study. The Bayesian Fusion has been able to obtain slightly improve the performance w.r.t the standard averaging method. We underline that this strategy in general, and it can be applied also in other monitoring applications based on ortho-photo maps.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a set of methods for the improvement of semantic segmentation of benthic communities using ortho-mosaic maps. The proposed strategies are automatic and exploit the characteristics of metricity and continuity of ortho-photo

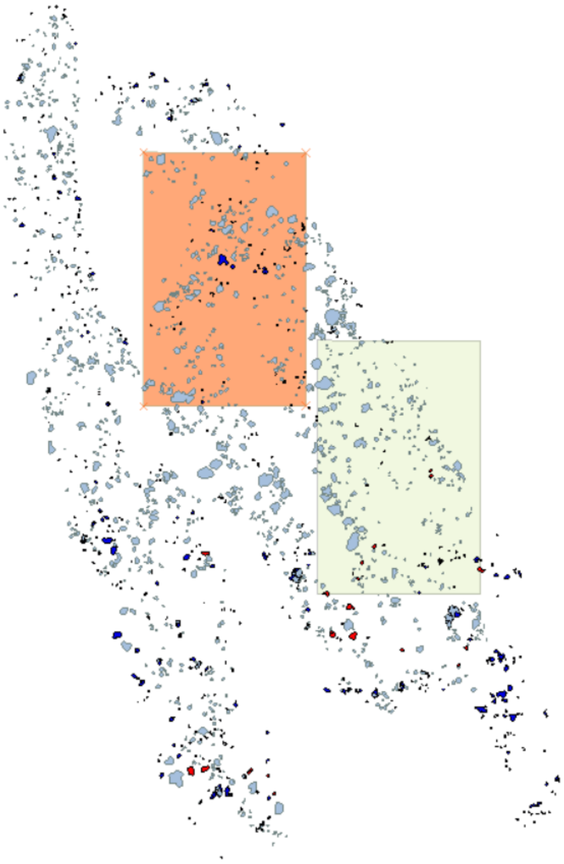

Figure 10. Human annotation errors. After a second annotation session, some of the original labels were found to be false positive (red) or false negative (dark blue).

maps. The results of these solutions are encouraging, despite the low quality of the input at our disposal and despite the presence of incorrect annotations.
The next step will be to test these methods on other datasets, containing more classes and different information layers. We are currently working on speeding up the human labeling step in order to obtain labels that better follows specimens border without compromising the advantages in terms of speed of the polygonal annotation. More accurate labeling would justify the choice of a finer-grain segmentation network. One of the greatest advantages of working with ortho-mosaics coming from 3D reconstruction of the seabed is the opportunity to adopt a multi-modal approach, combining the depth from the DEM maps with the RGB value from the textures. This would make also the evaluation criterion more robust; different species thrive at different depth ranges.

Life beneath the surface is characterized by a marvelous variety of animal and plant species. According to our experience, every team of biologists deals and needs to identify a very specific class of organisms. A customizable detection tool, that includes all the steps from the annotation to the segmentation and to the validation of the results (with a streamlined interface), is the final purpose of our research activity.

ACKNOWLEDGEMENTS

This work was supported by the Marie Sklodowska-Curie Action, Horizon 2020 within the project GreenBubbles (Grant Number 643712).

REFERENCES

100 Islands Challenge Team, 2019. 100 Islands Challenge Project. http://100islandchallenge.org. Online; accessed 22 February 2019.

Alonso, I., Cambra, A., Muoz, A., Treibitz, T. and Murillo, A. C., 2017. Coral-segmentation: Training dense labeling models with sparse ground truth. In: *ICCVW 2017*, pp. 2874–2882.

Audebert, N., Le Saux, B. and Lefèvre, S., 2017. Semantic segmentation of earth observation data using multimodal and multiscale deep networks. In: S.-H. Lai, V. Lepetit, K. Nishino and Y. Sato (eds), *ACCV 2016*, pp. 180–196.

Beijbom, O., Edmunds, P. J., Kline, D. I., Mitchell, B. G. and Kriegman, D., 2012. Automated annotation of coral reef survey images. In: *CVPR2012*, pp. 1170–1177.

Beijbom, O., Edmunds, P. J., Roelfsema, C., Smith, J., Kline, D. I., Neal, B. P., Dunlap, M. J., Moriarty, V., Fan, T.-Y., Tan, C.-J., Chan, S., Treibitz, T., Gamst, A., Mitchell, B. G. and Kriegman, D., 2015. Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation. *PLOS ONE* 10, pp. 1–22.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *ECCV*.

Edwards, C., Eynaud, Y., Williams, G. J., Pedersen, N. E., Zgliczynski, B. J., Gleason, A. C. R., Smith, J. E. and Sandin, S., 2017. Large-area imaging reveals biologically driven non-random spatial patterns of corals at a remote reef. *Coral Reefs* 36, pp. 1291–1305.

Kendall, A., Badrinarayanan, V. and Cipolla, R., 2015. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *CoRR*.

King, A., Bhandarkar, S. M. and Hopkinson, B. M., 2018. A comparison of deep learning methods for semantic segmentation of coral reef survey images. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Liu, Y., Minh Nguyen, D., Deligiannis, N., Ding, W. and Munteanu, A., 2017. Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery. *Remote Sensing*.

Ma, L., Stueckler, J., Kerl, C. and Cremers, D., 2017. Multi-view deep learning for consistent semantic mapping with rgb-d cameras. In: *IEEE International Conference on Intelligent Robots and Systems (IROS)*, Vancouver, Canada.

Mahmood, A., Bennamoun, M., An, S., Sohel, F., Boussaid, F., Hovey, R., Kendrick, G. and Fisher, R. B., 2016. Automatic annotation of coral reefs using deep learning. In: *OCEANS 2016 MTS/IEEE Monterey*, pp. 1–5.

Maninis, K.-K., Caelles, S., Pont-Tuset, J. and Van Gool, L., 2018. Deep extreme cut: From extreme points to object segmentation. In: *Computer Vision and Pattern Recognition (CVPR)*.

Palma, M. A., Casado, M. R., Pantaleo, U. and Cerrano, C., 2017. High resolution orthomosaics of african coral reefs: A tool for wide-scale benthic monitoring. *Remote Sensing* 9, pp. 705.

Pavoni, G., Corsini, M., Palma, M. and Scopigno, R., 2019. A validation tool for improving semantic segmentation of complex natural structures. In: *Eurographics 2019 - Short Papers* (in press).

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