Coral-Segmentation:
Training Dense Labeling Models with Sparse Ground Truth

Iñigo Alonso[1] Ana Cambra[1] Adolfo Muñoz[1] Tali Treibitz[2] Ana C. Murillo[1]
[1]DIIS-i3A. Universidad de Zaragoza, Spain.
[2]Charney School of Marine Sciences. University of Haifa, Israel.

Abstract

Biological datasets, such as our case of study, coral segmentation, often present scarce and sparse annotated image labels. Transfer learning techniques allow us to adapt existing deep learning models to new domains, even with small amounts of training data. Therefore, one of the main challenges to train dense segmentation models is to obtain the required dense labeled training data. This work presents a novel pipeline to address this pitfall and demonstrates the advantages of applying it to coral imagery segmentation. We fine tune state-of-the-art encoder-decoder CNN models for semantic segmentation thanks to a new proposed augmented labeling strategy. Our experiments run on a recent coral dataset [4], proving that this augmented ground truth allows us to effectively learn coral segmentation, as well as provide a relevant score of the segmentation quality based on it. Our approach provides a segmentation of comparable or better quality than the baseline presented with the dataset and a more flexible end-to-end pipeline.

1. Introduction

Semantic image segmentation, or dense image labeling, assigns a category label to each image pixel. This problem has been widely studied in the past and, as many other applications, it has achieved extraordinary results with deep learning based approaches [22]. However, there are many domains where obtaining large amounts of good quality dense labeled segmentation data, which is required to train such approaches, is highly costly and tedious to obtain.

Tasks to monitor different aspects of wildlife can highly benefit of automatic semantic segmentation approaches, from animal recognition in videos [18] to coral identification in underwater survey imagery [4]. Unfortunately, datasets of this kind often only provide a weakly labeled ground truth. This is the case in our work, which is focused on quantifying coral abundance. Coral reefs have a high ecological and economical value [6]. Sadly, in the past decades a variety of anthropogenic stressors caused a severe decline in coral coverage around the world [7]. This rapid change rate requires creating automatic methods for quick evaluation of reef health, that is currently done manually.

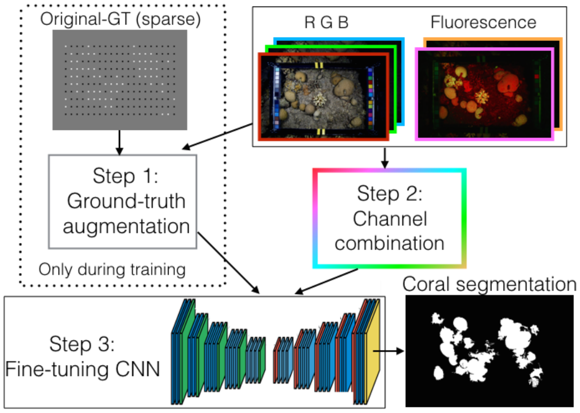

Figure 1. Coral segmentation pipeline based on CNN segmentation model. Step 1: sparse ground truth available is augmented to facilitate training. Step 2: input multimodal data is combined to use the more discriminative channels. Step 3: fine-tuning.

Recent work on this topic proposed a system to classify patches from underwater imagery into several classes of common corals and other textures that occur frequently in underwater scenarios [4]. This work highlights the benefits of using fluorescence data to more easily discriminate among coral and non coral regions. Following their conclusions, we explore the use of RGB combined with fluorescence channels, but we target an end-to-end dense coral segmentation per image, as opposed to training per-patch classification. This problem can be formulated as an image segmentation into *coral/no-coral*. Our work (summarized in Fig. 1) addresses two challenges to achieve this goal:

- Lack of large amounts of accurate labeled data. The available datasets do not have detailed segmentation ground truth, but only a few sparse labeled points.
- How to use multimodal input (RGB + fluorescence) with a state-of-the-art image segmentation model.

The main contribution of our work is an effective approach to fine-tune state-of-the-art encoder-decoder CNN models for semantic segmentation with a combination of multi-modal data when only very sparse ground truth labels are available. We first study and propose different strategies to augment the sparse coral labeled data available into dense labels. This enables us to fine-tune existing CNN models even if there is not a large amount of labeled data. We also perform an exhaustive evaluation of different ways to combine the fluorescence and RGB information.

Our experimental results demonstrate how the proposed simple augmentation of ground truth labels provides valuable and effective additional information to train an end-to-end coral segmentation model. Our approach presents several advantages with respect to prior work based on individual patch classification, such as a better fit to the coral regions contours and a decoupled dependency on the existence of multimodal data. This is an important property of our pipeline, that it allows us to take advantage of the multi-modal data only during training to augment the labeled data, but still train a model that does not require those input channels, i.e., accepts only RGB input. This is relevant because often the fluorescence information is not available. This pipeline can also be applied to other multi-modal information such as other multiespectral data the same way we applied it to fluorescence information.

Another significant insight from the experiments on this work is the effective and meaningful segmentation results evaluation that can be obtained with the presented performance scores based on the augmented ground truth.

2. Related Work

We next discuss the most relevant topics to the presented approach are state-of-the-art methods on semantic segmentation and strategies to deal with a lack of the required training data. Besides, we also comment on related works about the particularities of automatic semantic segmentation of underwater imagery from coral reefs.

Semantic image segmentation. Superpixel segmentation approaches, such as SLIC [1] or SEEDs [17] algorithms, typically provide an over-segmentation of the input image, and have been the basis for earlier works on semantic image segmentation based on superpixel classification and superpixel based label propagation. On the other hand, successful encoder-decoder CNN based segmentation approaches [2, 12] have achieved state-of-the-art results on semantic segmentation problems lately. The recent survey on image segmentation by Zhu et al [22] provides a more detailed discussion of solutions for this long studied problem. Our approach takes both recent CNN based end-to-end semantic segmentation models and superpixel segmentation algorithms as important ingredients. Besides, it is designed to implicitly consider the context information around each superpixel. Many prior work highlights the importance of modeling the context information for different visual classification tasks, and so do many previous approaches on the particular problem of semantic image segmentation. For example, Yong et al [20] presented an approach where semantic context modeling helps a visual recognition task for novelty detection in wildlife scenes, or Mostajabi et al [14] highlighted the improvements obtained in superpixel classification by using superpixel context.

Working with biological imagery, it is very common to find weakly labeled datasets. This presents a lot of challenges and opportunities to develop weakly labeled training methods. For example, Venkitasubramanian at al [18] propose how to train animal recognition system in videos with weak supervision, thanks to the use of multimodal data. This lack of enough training data is specially crucial in semantic segmentation approaches, because acquiring accurate segmentation is a tedious task, often unfeasible.

Lack of training data. The lack of (good) labeled training data is a common issue when building and training deep learning based systems. We can find multiple strategies to overcome this problem, briefly discussed next.

Data augmentation, i.e., generating additional data by altering the original labeled data, is a very common solution. Many works have used variations of this strategy, including for example the well know *Alexnet* model [11], that was trained augmenting the training data by applying image translations and horizontal reflections and altering the intensities of the RGB values. A more recent solution to augment the training set, or to actually completely generate an artificial data set, is to generate *synthetic data* [8, 15]. This strategy provides perfect ground truth labels of plenty of concepts, as long as the image rendering or simulation platform support that information. This type of methods do not always transfer properly from data to real data, in part because for many problems is hard to simulate the right amount of variability needed for the training data. Other recent work proposing how to deal with the fact of *no labeled data* [16] at all, describes how to adapt an existing model when there is no training data available for the new domain.

Other common strategy to deal with lack of good training data is to build approaches that can learn from *weakly labeled data*, which is much easier to obtain. Lu et at recently presented a survey on different approaches to train semantic segmentation from noisy and weakly labeled data [13], which discusses these problems and presents many related solutions. This work covers the augmentation of weak labeling focusing on detecting the noisy labels. They propose a pipeline which allows to segment the images with only image-level labels introducing a intermediate labelling variable so that they can learn which are noisy labels.

Sometimes weak label means per-image label as opposed to per-pixel, e.g., in the work from Kolesnikov and Lampert [10], that proposes a new composite loss function that allows us to train CNN models for image segmentation using weakly labeled data consisting of per-image class labels. Other times, like in our case, weak label means that the labeling is very sparse, as opposed to having a dense per pixel labeling. Vernaza et al [19] propose how to simultaneously learn a label-propagator and the image segmentation model. This approach propagates the ground truth labels from a few traces, to estimate the main object boundaries in the image and provide a label for each pixel. This work is maybe the closest related to our approach in the sense that they also demonstrate benefits when training CNN based segmentation using the propagated sparse available labels. Differently from this work, we do not have continuous traces as labels, but a sparse grid of points equally spread over the image, as detailed in next section, and we do not learn how to propagate the available labels. Instead, we take advantage of the fluorescence data available to augment the labeled data. Our work is inspired by the discussed prior work, but none of the existing examples demonstrates how to train a dense semantic segmentation model with such sparse and isolated labeled points as those available for the coral datasets.

Coral imagery segmentation. Obtaining good quality images from coral natural scenarios and their annotations is a challenging task, as well as automatically recognizing the corals on those images [4], [5].

As previously mentioned, our work studies and proposes how to face the challenges to enable latest results on semantic segmentation using CNNs to the segmentation of coral imagery. Prior work has demonstrated how the use of multi-modal data can facilitate this problem, in particular combining RGB images with fluorescence images [4]. This work has shown that CNN based approaches provide a higher performance than other methods evaluated in earlier works, such us SVM approaches [3] concerning multi-modal data in coral segmentation. We build upon these conclusions, but instead of building a per-patch classifier, we work on an end-to-end segmentation model based on fine-tuning state-of-the-art models from other domains, such as [2], as described in the next section.

3. Proposed Segmentation Approach

This section details the proposed approach to achieve dense semantic segmentation using sparse ground truth.

3.1. Problem statement

The main challenge considered in this work is how to learn a good semantic image segmentation given a very sparse ground truth to learn the model.

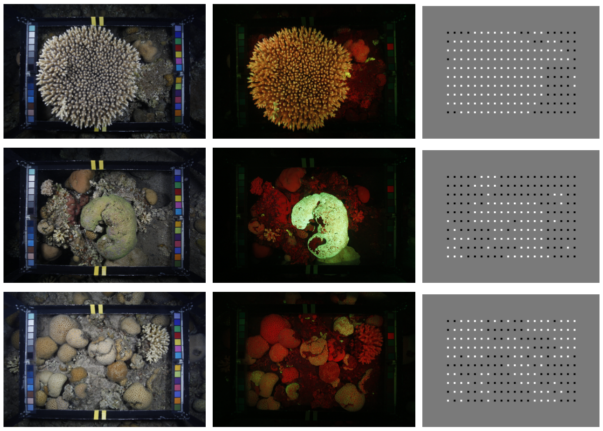
(a) RGB (b) Fluorescence (c) Original-GT

Figure 2. Three examples of the input data available in the dataset. Each row contains corresponding (a) original RGB image, (b) fluorescence image and (c) available sparse ground truth labels. These are single pixel labels, enlarged for visualization purposes. White pixels are coral. Black pixe0ls are non-coral.

The input for our particular problem is a set of multi-modal image channels (in particular, RGB and fluorescence images) and a sparse set of labels. The challenges from using the multi-modal inputs are not only about how to combine them but also that the different sensor images can be misaligned. As far as the ground truth is concerned, the main challenge is to find how to augment a sparse ground truth into a dense one. Fig. 2 shows some examples of the input data, highlighting the very sparse labeled set of points in the images. The images have $1078x976$ resolution but the ground truth has only 200 pixels labeled per image. Taking into account that the dataset has 142 training images, we only have 28400 training pixels (much smaller than the amount of pixels we have to classify in a single image).

The expected output for the semantic segmentation is a matrix where each pixel of the input image is classified (in our case into coral or no-coral classes).

3.2. Learning the coral segmentation model

Our proposed segmentation approach consists of the three steps detailed next and summarized in Fig. 1.

3.2.1 Ground truth augmentation

The most relevant challenge is the very sparse ground truth, because typically to train a CNN for semantic segmentation dense ground truth is needed. We evaluate three strategies to obtain this dense labeling, as shown in Fig. 3.

Patches-GT. This strategy is the more straightforward. We expand the labeled ground truth pixels into labeled patches around those pixels. This strategy assumes that the surrounding pixels of a labeled one are the same kind.

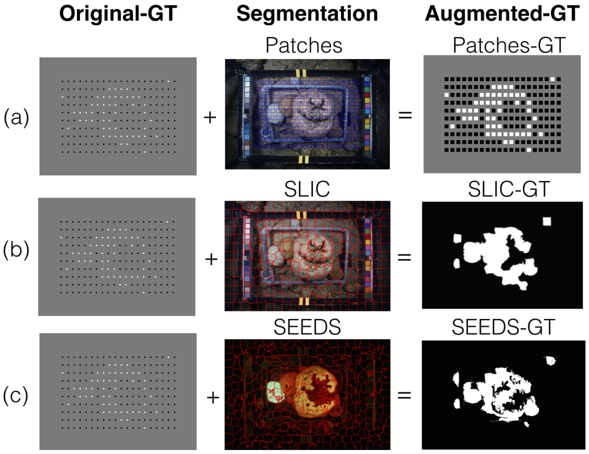

Figure 3. Ground truth augmentation methods that we considered. (a) small patches around original-GT labeled pixels; (b) SLIC and (c) SEEDS superpixels, computed on RGB or fluorescence images, used to expand the original-GT. SLIC and SEEDs can be augmented using either RGB or fluorescence image superpixels but fluorescence yields a much better segmentation.

Several patch sizes were tested and 25x25 pixel patches gave the best results (using 1078 x 976 images) providing 125000 labeled pixels per image instead of 200.

Superpixels (SLIC-GT, SEEDS-GT). We apply these superpixel segmentation methods to the images. This allows us to match the original labeled pixels to each segmentation. This method gives a better and more accurate solution. The outcome augmentations of SLIC [1] and SEEDS [17] superpixels are similar. Visually, SEEDS-GT fits better to the shape of the coral. These methods can fail specially when the corals are too small or the have holes. The Fig. 4 shows some cases of failure of the SEEDS-GT. Nevertheless, these approaches seem pretty similar to the RGB images. These superpixel augmentation can be obtained from any of the multi-modal images (see Fig. 3).

This step is independent from the segmentation prediction. Therefore, this augmentation can be obtained with fluorescence images and the segmentation output from the RGB images. The experimental results from the next Sec. 4 analyze the differences of using with different augmented ground truths in our pipeline.

3.2.2 Input channel combination

This step combines the available input channels. We evaluate several combinations of the available multi-modal data (as summarized in Fig. 5).

Using 3-channel input combination. First, since the base CNN model we use for fine-tuning has a three channel input [2], the intuitive approach is to select three out of the available channels. The combinations considered are based on previous studies on the different channels [4]. This study concludes that the two first fluorescence channels are more discriminating than the RGB channels and that within the RGB channels, the red channel is the most important.

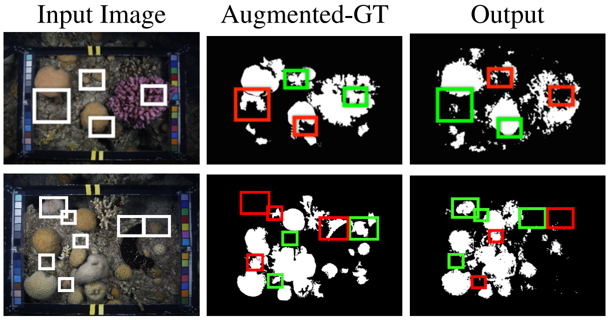

Figure 4. Even though the Augmented-GT (SEEDS-GT in these images) used to train our system is noisy, these examples show that the final segmentation obtained with our trained model detects regions that are missed in the augmented GT computed for those examples. We have manually highlighted incorrect predictions with red squares and good predictions with green squares.

Other input combination. Another insight from prior work we consider is that the different modalities available may not be perfectly registered. Therefore, this may impact the training if joining the inputs in earlier layers, as opposed to later ones. Then, other strategies we have evaluated use all the input channels available. They are based on combining the output of two different CNNs (one trained with fluorescence and other with RGB channels). This has been implemented in two ways: training two CNNs separately and then combining their outputs, or training them together.

3.2.3 Fine-tuning existing segmentation CNN model

The final step consists of training the model with the augmented-GT. The state-of-the-art image segmentation systems use CNN based models, which offer excellent accuracy. Our goal is to adapt existing semantic segmentation models to our target classes. In particular, we fine-tune SegNet [2] model with the coral images.

Segnet is a well-known encoder-decoder CNN for semantic segmentation, trained on urban scenes. It has a symmetrical structure in terms of convolutions and deconvolutions which allows to learn significantly well. Other approaches use only one deconvolution layer at the end of the network, as proposed in [12]. For example, good results on ImageNet scene segmentation challenges [21] were achieved applying this technique to the RESNET-50 model [9]. However, it performed worse (5% less accuracy) than using SegNet for our problem, maybe due to the larger number of deconvolutions applied in Segnet.

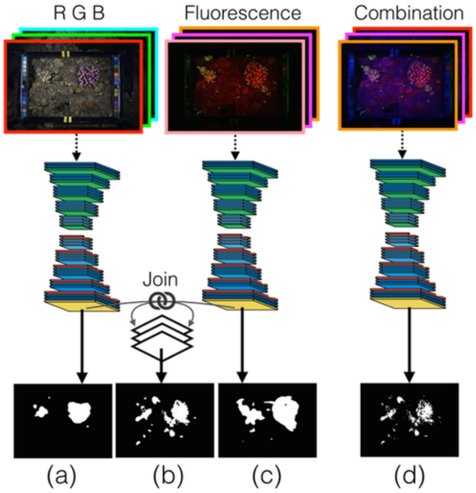

Figure 5. Different strategies to combine available image channels to train an end-to-end segmentation model: (a) fine-tuning with a 3-channel input using RGB data only; (c) using fluorescence data only; (d) combination of both ($Fluor_1 + Fluor_2 + Red$); (b) joining two of the fine-tuned models.

We keep the original SegNet for finetuning with three input channel combinations, while we performed slight modifications to its original network design for the experiments where we join two net structures. We also use the median frequency balancing [2] in the loss function (1). We use the cross-entropy loss [12] as the objective function for training the network. Adding the median frequency balancing (ϕ) to this function looks like this:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} \phi_{y^{(i)}} \left[y^{(i)} \log \hat{y}^{(i)} + (1-y^{(i)}) \log(1-\hat{y}^{(i)})\right]\right], \tag{1}$$

where m is the number of labeled pixels, $y^{(i)}$ is the label, $\hat{y}^{(i)}$ is the CNN predicted output. This gives a better performance on our data-set. Every class is weighted in the loss function with the ratio of the median of class frequencies computed on the entire training set divided by the class frequency. This implies the classes with low number of labeled pixels will have a higher weight. Thus, the CNN is not affected by the differences on the number of class samples.

4. Experiments

The following experiments analyze different aspects and variations of our approach for coral segmentation and compare the results obtained with prior work on the same data.

4.1. Set-up

Data-set. All the following experiments are run on the Eilat Fluorescence Corals dataset [4]. The dataset consists of 212 coral annotated multimodal image-pairs: RGB and fluorescence images. There are 200 labeled pixels per image, assigning to each of them a label from coral and non-coral classes[1]. Note that this ground truth is very sparse, since images have 1078 x 976 resolution. The data is split into a training-set of 142 randomly selected image-pairs, and a test-set with the remaining 70 image-pairs.

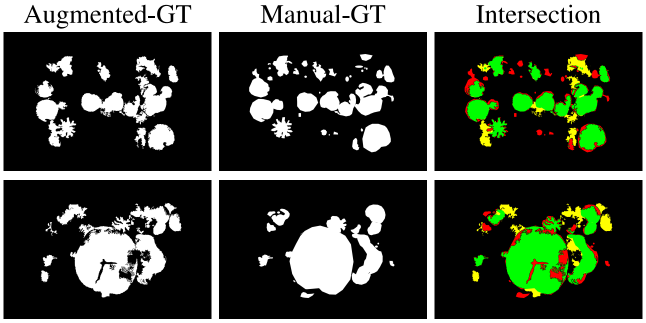

Figure 6. Examples of Augmented-GT and Manual-GT. The intersection of both shows green/red/yellow pixels when labeled as coral in both/only manual-GT/only augmented-GT respectively.

Evaluation. We use standard accuracy, recall and precision scores for the evaluation of the results computed according to different strategies:

Original-GT based sparse scores. The scores computed based on the original ground truth (Original-GT) are not fully representative, as it will be shown next. Intuitively, 200 pixels labeled out of around a million per image are not a dense ground truth for dense image labeling.

Superpixel-GT and Manual-GT based dense scores. The augmented ground truth we generate based on superpixels (Superpixel-GT) is an approximated but dense labeling, which as shown next gives a reliable evaluation. The fact of having very sparse ground truth is a challenge not only to train but also to evaluate in a meaningful way the dense labeling results. The representativity of this augmented ground truth can be seen in multiple visual results. Besides, we include comparisons using a few (7% of the testing data) detailed manual segmentations (Manual-GT) performed by an expert. This helps to further validate the Augmented-GT and the segmentation results. The average accuracy of the values in the augmented-GT with respect to the Manual-GT is of 93% (for the 5 images with Manual-GT available). Fig. 6 shows examples comparing these two segmentations.

4.2. Ground truth augmentation

Our work copes with the challenge of having a very sparse ground truth available to train a dense image labeling/segmentation model. The following results evaluate the use of different augmented ground truth. Some examples are shown in Fig. 7. All of them use the same model to be fine-tuned (SegNet [2]) and the same three input channels

[1] http://datadryad.org/resource/doi:10.5061/dryad.t4362

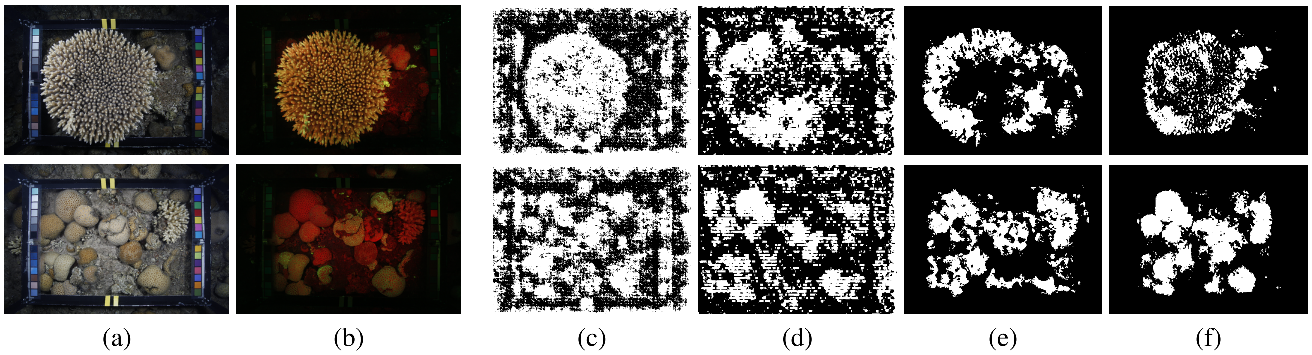

Figure 7. Coral segmentation using different augmented ground truth strategies. Two examples of corresponding RGB (a) and fluorescence (b) images and the coral segmentation obtained using a model trained with the sparse Original-GT (c) and with several augmented-GT: Patches-GT (d), SEED-GT (e) and SLIC-GT (f). Superpixel ground truths yield more accurate results.

Table 1. Coral segmentation (average pixel classification accuracy). Training and evaluation with different ground truth (GT).

Evaluation:	Original-GT	Patches-GT	SLIC-GT	SEEDS-GT
Training:	**(sparse)**		**(dense)**	**(dense)**
Original-GT	0.56	0.53	0.43	0.42
Patches-GT	0.77	0.80	0.67	0.67
SLIC-GT	**0.81**	0.80	**0.89**	**0.90**
SEEDS-GT	0.78	0.77	0.85	0.86

(two fluorescence channels and Red channel from RGB image). Note how noisy the results are when training with a sparse ground truth. The models trained with Original-GT and Patches-GT also give inaccurate predictions on the edges due to the lack of labeling on those regions, i.e., the patches-GT provides a segmentation with squared artifacts.

Table 1 summarizes these experiments. Each row shows the results for a different training option. Each column shows the accuracy computed over different sets of pixels (e.g., the evaluation with Original-GT means we compute the accuracy considering only the 200 labeled pixels per image). We can observe that the superpixel based approaches present better quantitative and qualitative results. These results illustrate how the proposed augmented ground truth is more suitable for training and more representative for the evaluation, as we analyze further in the following subsection 4.4 experiments. Out of the box superpixel segmentation gives much better results when computed on the fluorescence images, rather than on the RGB images, as it can be seen in Fig. 8. This is expected, since the fluorescence values are much higher on living beings in the scene images.

Our proposed pipeline allows us to take advantage of this multimodal input for the ground truth augmentation but still train the segmentation model with only one data modality. Even though the augmented ground truth based on superpixels is approximated, the model can still learn the coral regions very robustly. It even segments coral regions that were not included correctly as coral ground truth (as it can be seen in the examples in Fig. 4).

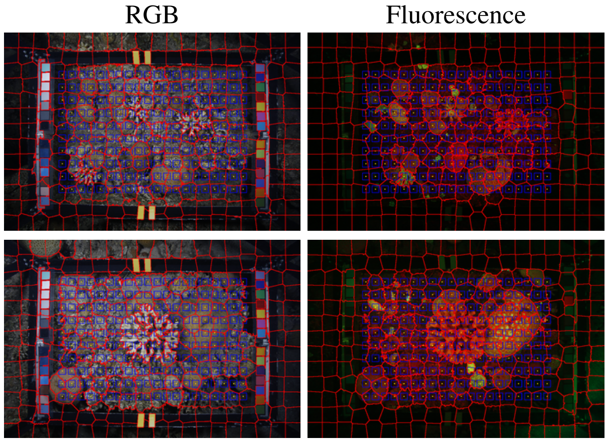

Figure 8. Superpixel segmentation of the image (red boundaries). Segmentation on fluorescence images fits better the coral regions.

4.3. Input channel combinations

These experiments evaluate different ways to combine the available input channels, i.e., RGB and fluorescence image channels, as explained in Sec. 3. The best results were obtained finetuning directly the original SegNet model, using the augmented ground truth. In particular, we consider SEEDS-GT and SLIC-GT, since they performed clearly better than the other options considered in previous subsection.

A summary of the results of the different three input channel combinations experiments is shown in Table 2. Fig. 9 shows visual examples of these experiments. Every combination has been trained with varying hyperparameters to get the best possible model. The configuration which gives better results uses the median frequency balancing, training $50k$ iterations with a learning rate of $2x10^{-4}$.

Additional experiments were carried out using all input channels as described in previous section:

- Training a fine-tuned CNN for each modality and joining the output of their probabilities for each class.

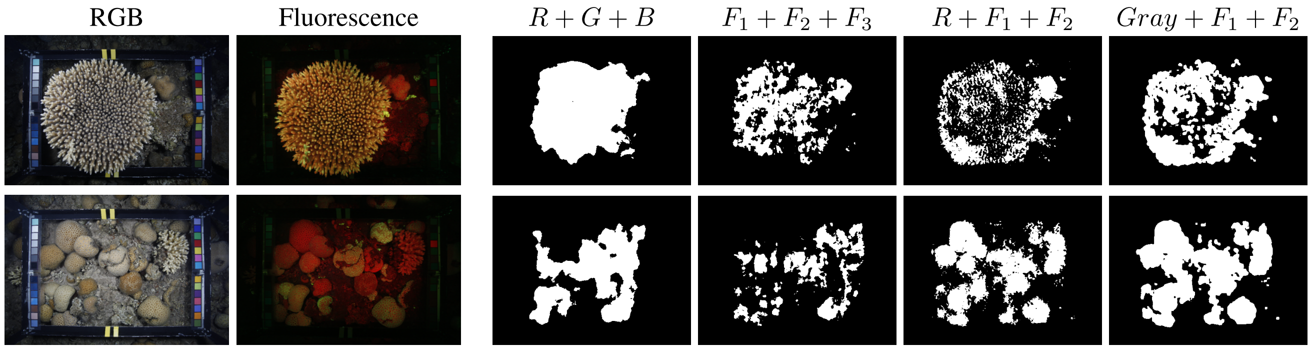

Figure 9. Coral segmentation using the proposed augmented-GT and different input channel combinations. The results of the different combinations are detailed in Table 2. The $Gray+F_1+F_2$ combination yields the best qualitative results.

Table 2. Coral segmentation (classification results per pixel) with different 3-channel combinations as input to finetune SegNet model.

3-Channel combinations	Average Accuracy	Coral Recall	No-coral Recall	Coral Precision	No-coral Precision
	Evaluation: **Original-GT** based **sparse scores**				
RGB only	0.76	0.43	0.89	0.60	0.80
Fluor only	0.79	0.52	0.90	0.61	0.83
$R+F_1+F_2$	0.80	0.63	0.87	0.64	0.86
$Gray+F_1+F_2$	**0.81**	**0.74**	0.84	0.65	0.89
	Evaluation: **Superpixel-GT** based **dense scores**.				
RGB only	0.87	0.43	0.94	0.64	0.89
Fluor. only	0.89	0.44	0.96	0.67	0.91
$R+F_1+F_2$	0.90	0.52	0.96	0.66	0.92
$Gray+F_1+F_2$	**0.91**	**0.61**	0.96	0.66	0.95

R, G, B: RGB channels
F_1, F_2: Fluorescence channels 1, 2 respectively
$Gray$: The average of the RGB channels

Table 3. Coral segmentation (classification results per pixel).

	Average Accuracy	Coral Recall	No-coral Recall	Coral Precision	No-coral Precision
	Evaluation: **Original-GT** based **sparse scores**				
Superpixel based (Ours)	0.81	0.74	0.84	0.65	0.89
Patch based$^+$	**0.94**	0.87	0.96	0.87	0.96
	Evaluation: **Superpixel-GT** based **dense scores**				
Superpixel based (Ours)	**0.91**	0.61	0.96	0.66	0.95
Patch based$^+$	0.90	0.59	0.94	0.63	0.95
	* Evaluation: **Manual-GT** based **dense scores**				
Superpixel based (Ours)	**0.92**	**0.79**	0.93	0.69	0.97
Patch based$^+$	0.90	0.60	0.95	0.66	0.94

*Computed only over the 5 images with Manual-GT available
$^+$Simulated result using [4] assuming 94% of patches correctly classified

- Fine-tuning a new CNN joining the two fine-tuned SegNet models after their last convolutional layer.

We discarded to train a model with larger input size because prior work showed better results with latter join of the data, probably because the images are not perfectly registered. We then combined two CNN models, one trained for RGB, and other for fluorescence data. None of them explored improved the performance, probably because of too large of a network model and not enough data to train it.

Although using only RGB information does not achieve the highest performance, it presents a promising direction. Our approach can use the fluorescence information only for the ground truth augmentation, and still train a model that takes as input RGB only data.

As expected from the results in prior work running patch classification [4], the best input combination contains fluorescence and RGB channels. Using models trained with a combination of both types of input data modalities ($Gray+F_1+F_2$ or $R+F_1+F_2$) provides the highest average accuracy and recall of the coral class (which is the most significant for the application of interest).

4.4. Patch vs. Superpixel based segmentation

The following results demonstrate the differences and advantages of the presented approach with respect to the baseline presented with the studied dataset, a patched-based classification approach. Table 3 shows comparable overall accuracy, recall and precision for both methods. Interestingly, our approach outperforms the patch-based method when evaluating on the Manual-GT, according to the dense scores, while the sparse scores benefit the per-patch approach. A more qualitative analysis of this comparison is shown in Fig. 10. We can see that superpixel-based approach produces more coral-like shapes in the segmentation and follows better the object contours. An important drawback of the patch-based approach is an implicit lack of per pixel precision, which does not happen in the presented end-to-end pipeline. Additional segmentation examples of the final pipeline configuration are shown in Fig. 11.

Another advantage of our superpixel based approach is that it provides a more flexible pipeline, where we can take advantage of valuable multimodal data only during training (i.e., using it only for the data augmentation).

Moreover, using a metric based on sparse data labels, when the output is dense, can be less representative than us-

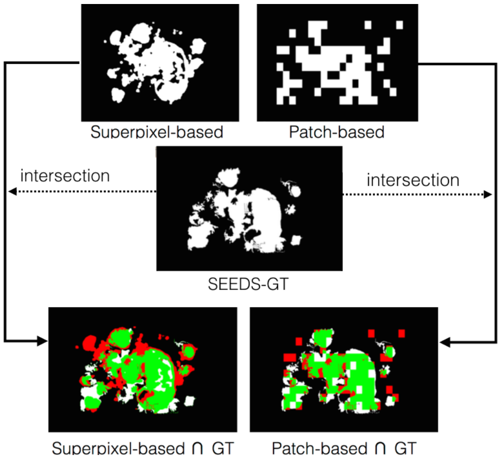

Figure 10. Coral segmentation with patch-based or superpixel-based (ours) approaches compared to augmented ground truth (SEEDS-GT). The patch-based shows the simulated output of results in [4]. Superpixel-based shows results using fine-tuned SegNet using $Gray + F_1 + F_2$ input channels. The intersection images show coral pixels correctly (green) and incorrectly (red) labeled.

ing scores based on an approximate but dense ground truth, as the one we use. The sparse scores are evaluating just 0.0002% of the pixels per image. Our results show the scores based on the augmented ground truth serve as a good quality evaluation for the segmentation. The last rows in Table 3 show how the scores using the available Manual-GT are closer to those using Superpixel-GT than to scores obtained using the Original-GT. This verifies the good representativity of the augmented ground truth, as shown in previous Fig. 6.

Although the augmented ground truth has some noise, i.e., incorrect labeling of both positive and negative pixels, our results show that the segmentation model is still learned effectively due to the huge increase in the number of training data (labeled pixels).

5. Conclusions

We have presented a novel pipeline which makes up for the lack of labeled data for semantic segmentation training. This has an important impact on semantic segmentation scenarios where the available datasets present sparse and scarce labels on the annotated images. We demonstrate that this augmented ground truth allows us to effectively learn the coral segmentation when finetuning a state-of-the-art CNN for semantic segmentation. Our results show the benefits of using the proposed augmentation of sparse image labels. We have analyzed the influence of variations in the labeling augmentation and the experiments show the superpixel based methods work better than other more direct options. Besides, we also show how the augmented ground truth can serve as a more significant way to evaluate the dense segmentation with dense scores.

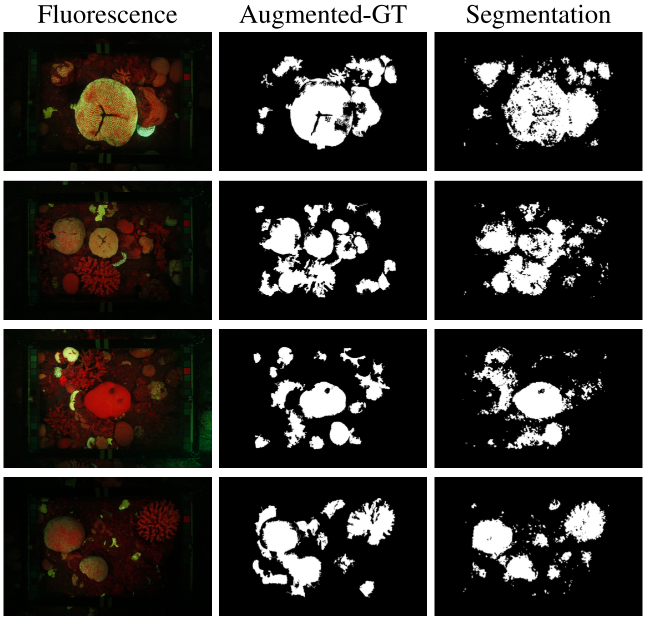

Figure 11. Coral segmentation results obtained from a model trained with the augmented (dense) ground truth and $Gray + F_1 + F_2$ input channels. Each row depicts the fluorescence image, augmented ground truth with SEEDS-GT, and the respective coral segmentation result.

Following previous results which highlight the benefits of using fluorescence information to recognize corals in images, we study different ways of taking advantage of this kind of multi-modal inputs. We have shown how useful the multi-modal input is as well in the proposed end-to-end dense labeling. Our flexible pipeline allows us to relax the requirements of the multi-modal input, fluorescence in our case. Since fluorescence data is not always available, a nice property of our pipeline is that we can still benefit partially of that type of input for the augmentation (during training), and still train a segmentation model that does not require it.

As future steps, we plan to explore other state-of-the-art CNN architectures for semantic segmentation, as well as studying more sophisticated multi-modal combinations and labeling augmentation methods.

Acknowledgments

This research has been partially funded by the European Union (CHIST-ERA IGLU), Spanish Government (projects DPI2015-65962-R, DPI2015-69376-R) and Aragon regional government (Grupo DGA T04-FSE). TT was supported by The Leona M. and Harry B. Helmsley Charitable Trust. The authors would like to give special thanks to Adi Zweifler for her help as an expert for hand-labeling some coral scene images and NVIDIA Corporation for the donation a Titan Xp GPU used for this research.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 2, 4

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015. 2, 3, 4, 5

[3] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman. Automated annotation of coral reef survey images. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1170–1177. IEEE, 2012. 3

[4] O. Beijbom, T. Treibitz, D. I. Kline, G. Eyal, A. Khen, B. Neal, Y. Loya, B. G. Mitchell, and D. Kriegman. Improving automated annotation of benthic survey images using wide-band fluorescence. *Scientific reports*, 6, 2016. 1, 3, 4, 5, 7, 8

[5] J.-N. Blanchet, S. Déry, J.-A. Landry, and K. Osborne. Automated annotation of corals in natural scene images using multiple texture representations. *PeerJ Preprints*, 4:e2026v2. 3

[6] H. S. Cesar. Coral reefs: their functions, threats and economic value. *Collected essays on the economics of coral reefs*, pages 14–39, 2000. 1

[7] P.-Y. Chen, C.-C. Chen, L. Chu, and B. McCarl. Evaluating the economic damage of climate change on global coral reefs. *Global Environmental Change*, 30:12–20, 2015. 1

[8] A. Gupta, A. Vedaldi, and A. Zisserman. Synthetic data for text localisation in natural images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[10] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision (ECCV)*. Springer, 2016. 3

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012. 2

[12] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 2, 4, 5

[13] Z. Lu, Z. Fu, T. Xiang, P. Han, L. Wang, and X. Gao. Learning from weak and noisy labels for semantic segmentation. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 39(3):486–500, 2017. 2

[14] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[15] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[16] B. Sun and K. Saenko. Deep CORAL: correlation alignment for deep domain adaptation. *CoRR*, abs/1607.01719, 2016. 2

[17] M. Van den Bergh, X. Boix, G. Roig, B. de Capitani, and L. Van Gool. SEEDS: Superpixels extracted via energy-driven sampling. In *European conference on computer vision*, pages 13–26. Springer, 2012. 2, 4

[18] A. N. Venkitasubramanian, T. Tuytelaars, and M.-F. Moens. Wildlife recognition in nature documentaries with weak supervision from subtitles and external data. *Pattern Recogn. Lett.*, 81(C):63–70, Oct. 2016. 1, 2

[19] P. Vernaza and M. Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[20] S.-P. Yong, J. D. Deng, and M. K. Purvis. Novelty detection in wildlife scenes through semantic context modelling. *Pattern Recognition*, 45(9):3439–3450, 2012. 2

[21] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ADE20K dataset. *CoRR*, abs/1608.05442, 2016. 4

[22] H. Zhu, F. Meng, J. Cai, and S. Lu. Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. *Journal of Visual Communication and Image Representation*, 34:12–27, 2016. 1, 2