Departamento de
Informática e Ingeniería
de Sistemas
Universidad Zaragoza
1542

Trabajo Fin de Máster

Segmentación Semántica con Modelos de Deep Learning y Etiquetados No Densos

Semantic Segmentation with Deep Learning Models and Sparse or Weak Labels

Íñigo Alonso Ruiz

Directora: Ana Cristina Murillo Arnal

Co-directora: Ana Belén Cambra Linés

Máster en Ingeniería informática
Departamento de Informática e Ingeniería de Sistemas
Escuela de Ingeniería y Arquitectura
Universidad de Zaragoza

22 de Enero de 2018

TRABAJOS DE FIN DE GRADO / FIN DE MÁSTER

Escuela de Ingeniería y Arquitectura Universidad Zaragoza

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD

(Este documento debe acompañar al Trabajo Fin de Grado (TFG)/Trabajo Fin de Máster (TFM) cuando sea depositado para su evaluación).

D./Dª. Iñigo Alonso Ruiz,

con nº de DNI 73013097B en aplicación de lo dispuesto en el art. 14 (Derechos de autor) del Acuerdo de 11 de septiembre de 2014, del Consejo de Gobierno, por el que se aprueba el Reglamento de los TFG y TFM de la Universidad de Zaragoza,

Declaro que el presente Trabajo de Fin de (Grado/Máster) Máster, (Título del Trabajo) Segmentación semántica con modelos de Deep Learning y etiquetados no densos

es de mi autoría y es original, no habiéndose utilizado fuente sin ser citada debidamente.

Zaragoza, 22 de Enero de 2018

Fdo: Íñigo Alonso Ruiz

Resumen

La segmentación semántica es un problema muy estudiado dentro del campo de la visión por computador que consiste en la clasificación de imágenes a nivel de píxel. Es decir, asignar una etiqueta o valor a cada uno de los píxeles de la imagen. Tiene aplicaciones muy variadas, que van desde interpretar el contenido de escenas urbanas para tareas de conducción automática hasta aplicaciones médicas que ayuden al médico a analizar la información del paciente para realizar un diagnóstico o operaciones.

Como en muchos otros problemas y tareas relacionados con la visión por computador, en los últimos años se han propuesto y demostrado grandes avances en los métodos para segmentación semántica gracias, en gran parte, al reciente auge de los métodos basados en aprendizaje profundo o *deep learning*.

A pesar de que en los últimos años se están realizando mejoras constantes, los modelos de *deep learning* para segmentación semántica presentan un reto que dificulta su aplicabilidad a problemas de la vida real: necesitan grandes cantidades de anotaciones para entrenar los modelos. Esto es muy costoso, sobre todo porque en este caso hay que realizarlo a nivel de píxel.

Muchos conjuntos de datos reales, por ejemplo datos adquiridos para tareas de monitorización del medio ambiente (grabaciones de entornos naturales, imágenes de satélite) generalmente presentan tan solo unos pocos píxeles etiquetados por imagen, que suelen venir de algunos *clicks* de un experto, para indicar ciertas zonas de interés en esas imágenes. Este tipo de etiquetado hace que sea muy complicado el entrenamiento de modelos densos que permitan procesar y obtener de manera automática una mayor cantidad de información de todos estos conjuntos de datos.

El objetivo de este trabajo es proponer nuevos métodos para resolver este problema. La idea principal es utilizar una segmentación inicial de la imagen multi-nivel de la imagen para propagar la poca información disponible. Este enfoque novedoso permite aumentar la anotación, y demostramos que pese a ser algo ruidosa, permite aprender de manera efectiva un modelo que obtenga la segmentación deseada. Este método es aplicable a cualquier tipo de dispersión de las anotaciones, siendo independiente del número de píxeles anotados. Las principales tareas desarrolladas en este proyecto son:

- Estudio del estado del arte en técnicas de segmentación semántica (la mayoría basadas en técnicas de *deep learning)*
- Propuesta y evaluación de métodos para aumentar (propagar) las etiquetas de las imágenes de entrenamiento cuando estas son dispersas y escasas
- Diseño y evaluación de las arquitecturas de redes neuronales más adecuadas para resolver este problema

Para validar nuestras propuestas, nos centramos en un caso de aplicación en imágenes submarinas, capturadas para monitorización de las zonas de barreras de coral. También demostramos que el método propuesto se puede aplicar a otro tipo de imágenes, como imágenes aéreas, imágenes multiespectrales y conjuntos de datos de segmentación de instancias.

Summary

Semantic segmentation is a broadly studied problem in the field of computer vision. It consists of pixel-level image classification. That is, a value or label will be assigned to every pixel in the image. Semantic segmentation has a variety of applications such as the understanding of urban scenarios for autonomous driving, or the understanding of medical images to help doctors when diagnosing or even operating. As for many other computer vision problems, in recent years, a lot of improvements have been made in semantic segmentation approaches. These improvements are in big part thanks to deep learning techniques.

In spite of this fact, deep learning models for semantic segmentation present some challenges. This challenges hinder the real life applicability. The most important one is the need of a large amount of labels to be able to train it. The labeling process implies a very high cost, specially when the labels required are pixel-level.

Many datasets, such as those captured for environment monitoring (satellite imagery or natural environment) usually present only a few labeled pixels per image. These labeled pixels are obtained by an expert specifying areas of interest of each image with a few clicks. This kind of labeling makes the training process of dense models more costly.

The goal of this project is to propose novel methods to solve this problem. The main idea is to use a multi-level superpixel segmentation of the image to propagate the available sparse labeling information. Our novel approach allows to augment the labeling. We demonstrate that this method allows to effectively learn the dense semantic segmentation in spite of the noise the labeling may have. This method can be applied to any kind of labeling sparsity, being independent on the number of labeled pixels.

The main tasks developed in this project are:

- The study of the state-of-the-art of semantic segmentation (most of them are deep learning approaches)
- Proposal and evaluation of labeling augmentation methods when the labeling is weak or sparse
- Design and evaluation of suitable neural network architectures for this problem

In order to validate our proposal, we focus the experimentation on underwater imagery, captured for coral reef monitoring. We also demonstrate the proposed approach can be applied to other kind of images, such as aerial images, multi-spectral images and instance segmentation datasets.

Contents

Chapter 1

Introduction

This section introduces the motivation and the context of the project, as well as a summary of its content and scope.

1.1 Motivation

The field of deep learning has pushed the state of the art in plenty of computer vision related applications in the last years, such as image classification, object detection and recognition, image captioning, semantic segmentation, among a long list [LBH15]. A summary of the key ideas of deep learning, as well as its relation with semantic segmentation, can be found in Appendix B.

This project works on improved techniques for the problem of semantic segmentation, or dense image labeling. This problem consists of assigning a label to each pixel in the image giving a result of the same spatial dimensions as we can see in the example of Fig. 1.1.

Typically, deep learning approaches to learn to perform semantic segmentation of an image require a lot of densely labeled examples. This semantic segmentation labels are really costly to collect, because every pixel of every image has to be labeled. The

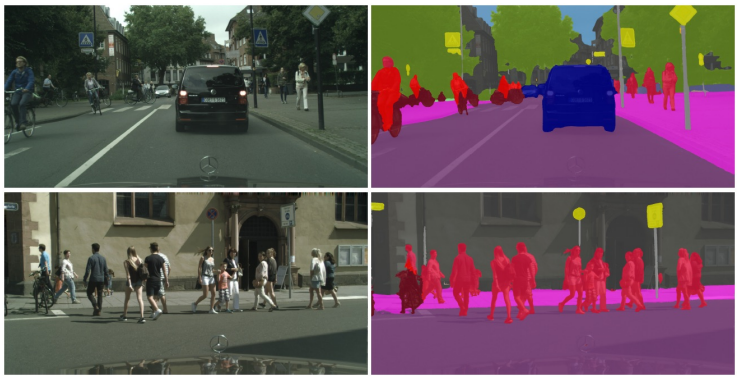


Figure 1.1: Example of a scene image (left) and its corresponding dense labeling, or *semantic segmentation*, (right) considering several class labels from urban areas, such as person, sign, road, etc. (Image from [KVK16])

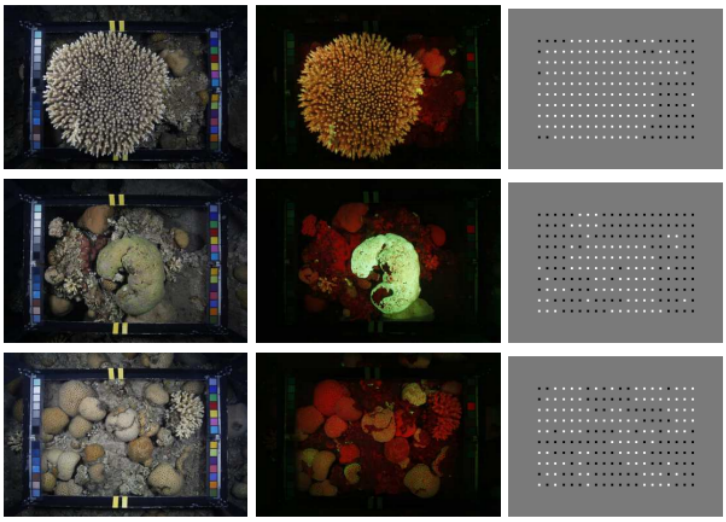


Figure 1.2: Example of a coral images (a), (b) and the sparse labeling available (c). This Original-GT is visually enhanced to be able to see the labeled pixels. There are only 200 labeled pixels per image.

main challenge considered in this work is how to learn semantic segmentation models when the available labels are not as dense as those need by existing deep learning techniques. Figure 1.2 shows how this sparse labeling typically looks like: a sparse set of labeled points, typically equally distributed across the image.

In particular, this work is focused on use cases where it is common to have a lot of data, very sparse labels, but a dense labeling of the images is needed. The main case of study of this work is imagery from coral reefs, due to three reasons:

- Biological datasets usually have this kind of labeling. Each image tends to have between 20 and 500 labeled pixels.
- It is a significant environmental problem. Coral reefs are valuable ecology regions in danger, and its careful surveillance requires creating automatic methods for quick evaluation of reef health, that is currently done manually.
- The research group has access to plenty of coral datasets through a collaboration with marine laboratories.

The challenges and the related work are further detailed in next chapter 2.

Personal motivation. Besides the technical challenges described, and the fact that I really like Machine Learning and Data Analysis, before I get my degree and start looking for a job, I wanted to try what academic/research life is like, and performing this project was a good opportunity to do so.

1.2 Context

This project has been carried out within the Robotics, Perception and Real time group (RoPeRT), a research group in the Aragón Institute of Engineering Research (I3A). In particular, the research of this project has been developed in collaboration with the two thesis supervisors, Ana Cristina Murillo and Ana Belén Cambra, and the collaborators with whom we have published the initial results of the work in this project (see Appendix A). The most recent results will be submitted to another conference in the near future.

1.3 Task and Goals

The **general goal** of this project is to advance on new techniques for semantic segmentation of images with deep learning techniques when the labeling is weak or sparse. In order to get there, this project has covered the following tasks (their temporal extent is summarized in the diagram from Fig. 1.3):

- **Task 1**. Learn about the **tools and frameworks** to use. This tools are mostly *Tensorflow, OpenCV*. I have used this tools before but for very simple tasks. This first step of the work is aimed to go deeper into their usage.
- **Task 2**. To **study related works** and current state of the art approaches: well known semantic segmentation and superpixel techniques; *deep learning* for semantic segmentation, with emphasis on approaches that work with weak labeling.
- **Task 3**. To **propose and implement** a an approach to train a dense semantic segmentation model with sparse labeling. This task has started from earlier works in the group on semantic segmentation. I will use some existing code for neural networks and superpixels segmentation. The main sub-tasks here are
  - to develop and evaluate new strategies to augment sparse training data.
  - to design or adapt existing models which are more suitable for this problem and type of data.
- **Task 4**. Carry out experiments with different real use-cases, with special focus on the challenging underwater datasets for coral region monitoring, and **evaluate** the different designed approaches.
- **Task 5**. Gather the conclusions and write the approach description and results in a technical **report**.

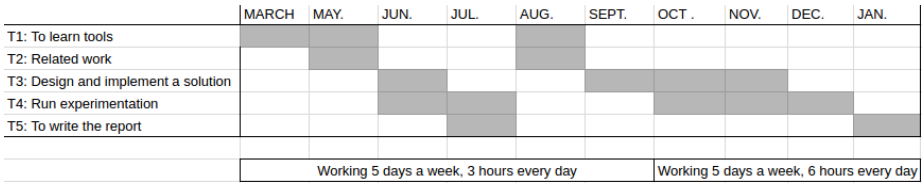


Figure 1.3: This diagram summarizes the approximated temporal distribution of the tasks in the project.

The more specific challenges considered in this work will be detailed in the next chapter, with a more detailed discussion of the related works.

1.4 Outline

The remainder of this document contains, in chapter 2, the description of the challenges to overcome, the contributions of this work and the related work; in chapter 3, the details of the proposed approach specifying the summary of our pipeline; in chapter 4 the carried out experiments and the evaluation of every step of our pipeline in several datasets; in chapter 5, we summarize the conclusions of this work.

Chapter 2

Challenges and Related Work

This chapter discusses the related work from the most relevant topics to this project, emphasizing the challenges related to each of them. We discuss existing strategies to deal with weak and sparse labeling and lack of training data and discuss state-of-the-art methods on convolutional neural networks and more specifically for semantic segmentation. Besides, we briefly mention the particularities of automatic semantic segmentation of underwater imagery from coral reefs.

2.1 Lack of training data and labeling

As briefly mentioned in the introduction, the lack of labeled training data is a common issue when building and training deep learning based systems. We find several strategies to overcome this problem in prior work, which we could organize in two big groups.

Data augmentation, i.e., generating extra data by altering the original labeled data keeping a realistic appearance, is a very common solution. Many works have followed this strategy, including for example the well know *Alexnet* model [KSH12], that was trained augmenting the training data by applying image translations and horizontal reflections, altering the intensities of the RGB values. A more recent solution to augment the training data is to generate *synthetic data* [GVZ16, RSM$^+$16]. This strategy provides perfect ground-truth labels with only the cost of constructing the simulated scenes on the simulation platform. Nevertheless, it is not a flawless solution. There are several problems due to the difficulty of generating realistic images and with all the variations and options real world has. Ros et al. [RSM$^+$16] presented a large dataset which includes lot of variations achieving impressive results on real images above all when training synthetic and real data at the same time.

Another common strategy to deal with the lack of good training labeled data is to build build approaches that can learn from **weakly labeled data**, which is much easier to obtain[XSU15, PVK17]. Lu et at carried out a survey on different approaches to train semantic segmentation from noisy and weakly labeled data [LFX$^+$17], which discusses these problems and presents many related solutions. This work covers the augmentation of weak labeling focusing on detecting the noisy labels.

Sometimes weak labeling means **per-image labels** as opposed to per-pixel. Some

works propose to modify the CNN architecture [PKD15, PC15]. Durand et al. [DMTC17] propose to train a classification neural network to learn the features of the classes and then to work with the feature maps to get an accurate segmentation results. Another work, from Kolesnikov and Lampert [KL16], 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. Another approach is to turn the sparse pixel-level labels into image-level labels cropping patches around the labeled pixels [BTK+16].

Other times, like in our case, weak label means that the **labeling is very sparse**, as opposed to having a dense per pixel labeling. Uhrig et al [USS+17] propose a new CNN architecture: Sparsity Invariant CNNs. This architecture is based on sparse convolutions. Sparse convolutions allow to learn from sparse labeling. This method has been proved to work well between 5% and 70% of sparsity. This method focuses reconstruct the depth map from a sparse LIDAR sensor information. Vernaza et al [VC17] 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. Alonso et al [ACM+17] use superpixel segmentation methods to augment the sparse labeling into a dense labeling allowing an usual encoder-decoder CNN to learn the semantic segmentation. This work is the closest to ours and is one of the baselines we use to evaluate our work.

2.2 Semantic image segmentation

Semantic segmentation is a topic with significant improvements in the recent years [GGOEO+17]. Deep learning approaches have achieved state-of-the-art results on semantic segmentation problems lately such us the Mask-RCNN [HGDG17] and Tiramisu architecture [JDV+17]. This recent survey on image segmentation by Zhu et al [ZMCL16] provides a more detailed discussion of solutions for this long studied problem. We find numerous prior work based on superpixel segmentation approaches, such as PB [ZHMB11] or SEEDs [VdBBR+12] algorithms, which have been the basis for earlier works on semantic image segmentation, based on superpixel classification and superpixel based label propagation. Focusing on the previously discussed challenge of lack of dense training examples, we can find recent work specially relevant for our work, which uses state-of-the-art CNN models to learn segmentation from partially labeled training data, introducing a new partially supervised training paradigm and weight transfer function [HDH+17].

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 instance, Yong et al [YDP12] presented an method where modeling the semantic context helped visual recognition task for novelty detection in wildlife scenes, or Mostajabi et al [MYS15] highlighted the improvements obtained in superpixel classification by using superpixel context.

Our approach joins both recent CNN based semantic segmentation models and superpixel segmentation algorithms as key ingredients. We compare several representative architectures used for this task [LSD15, JDV+17, HLWvdM17], making use of the proposed superpixel-based methods to augmented the training data. Besides, it is designed to implicitly consider the context information around each superpixel.

2.3 Coral detection and segmentation

The lack of good labeling is a common feature on lot of domains. A good example are biological datasets, which is our main case of study, more specifically, coral reefs monitoring and detection.

Automatic analysis of this type of underwater datasets is a real-life example of a domain with weakly labeled data that needs dense segmentation. For instance, CoralNet[1] is a collaboration project which focuses on coral reef analysis. It has a lot of classification datasets from all over the world (half a million images spread across 632 sources with 1420 robots assisting the annotation work of the almost 20 Million point locations). Unfortunately, this kind of datasets only provide a weakly labeled ground truth (typically a few pixels labeled by an expert). Coral reefs have a high ecological and economical value [Ces00], but in the past decades a variety of events are causing a severe decline in coral coverage around the world [CCCM15]. This rapid change rate requires creating automatic methods for quick evaluation of reef health, that is currently done manually, making the process too slow and tedious.

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 [BDLO, BTK$^+$16]. Lots of efforts have being made to share datasets from all over the world in order to improve methods on this kind of sparse labeled data such us CoralNet[2] ot The Aqua Project[3].

Recent works have presented several approaches of both data collection using multi-robot teams [SCH$^+$17] and automated image analysis of coral reefs [MLD$^+$17], with the purpose of automatizing the detection and monitoring of the coral reefs health. Beijbom et al. [BTK$^+$16] show that CNN based approaches provide higher performance than other methods, such as SVM based approaches [BEK$^+$12]. We build upon these conclusions, but instead of building a per-patch classifier, we work on a state-of-the-art end-to-end segmentation model using our labeling augmentation strategy.

2.4 Challenges and contributions.

As previously mentioned, solutions for semantic segmentation have witnessed a significant improvement in recent years thanks in big part to convolutional neural networks [GGOEO$^+$17], and a lot of applications have seen the impact of these improvements, such us autonomous driving or medical applications. Many other fields could benefit from these improvements but it is not always feasible to obtain the large amount of labeled training data required by the existing techniques. Semantic segmentation models need pixel-level annotations in order to train, but this type of labeling is very time consuming and often needs human experts. Therefore, in a lot of datasets there are only available image-level annotations or a few pixels are labeled. This brings an interesting challenge: *how to train dense models with very sparse or weak labels*. This would allow a lot of domains to enhance their processes and help them to extract more detailed information and conclusions from their data.

In this work, we work on this challenge, focusing our study in improving the automatic analysis of coral reef datasets. We carry out an in-depth analysis on the ground-

[1] https://coralnet.ucsd.edu/

[2] https://coralnet.ucsd.edu/

[3] http://cim.mcgill.ca/ mrl/data.html

truth augmentation using superpixels segmentation methods. We focus our efforts not only on getting good results on the coral dataset but to create a method which can be generalized to other domains and datasets.

Our work (summarized in next Chapter, in Fig. 3.1) 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. The purpose is to learn a good semantic segmentation given a very sparse ground-truth using only RGB images and very few images.
- The labeling augmentation method has to be flexible on the amount of labeled pixel as well as be able to work in different types of images and domains.

In this work, we are focusing on coral datasets (coral segmentation), although we also demonstrate the proposed methods on other domains.

The main contribution of this work is a labeling augmentation method based on superpixels, which allows to effectively train fully convolutional neural networks to get accurate pixel-level classification when only very sparse ground truth labels are available. We first study different CNN approaches to get dense segmentations from the sparse ground-truth and then we study how different superpixels segmentation methods perform on our multi-level superpixel algorithm. This enables us to get segmentations results very similar as if training with fully-annotated ground-truths. We also perform experiments on other different datasets to demonstrate this method also works in other domains.

Our experimental results demonstrate how the proposed augmentation of ground truth labels provides valuable and effective additional information to train an end-to-end segmentation model. Our approach presents several advantages with respect to prior work based on individual patch classification and based on more simple labeling augmentations, getting better results and being more flexible on the sparse ground-truth and images on which the method can work.

Chapter 3

Learning Semantic Segmentation From Weak Labeling

This section details the proposed pipeline to achieve dense semantic segmentation from only RGB images and weak labeling.

3.1 Summary of the pipeline

The purpose of semantic segmentation is to get the corresponding per-pixel labeling of a given image. In this section we summarize the pipeline proposed to overcome the challenges detailed in previous chapters. Figure 3.1 presents a diagram of the main steps in our approach, described next, an end-to-end CNN model for semantic segmentation, which is trained with the augmented labeling we propose. Next we detail how we have built these two parts of our system, and the options studied.

First, we study different CNN approaches applicable to this specific problem. This problem can be treated as a semantic segmentation problem or as a classification problem. We study different neural networks and approaches in order to use the one which fits better for this problem.

Then, we explain how the proposed multi-level superpixel augmentation works. The most important challenge is to find how to augment a sparse and weak labeling into a dense one to enable a convolutional neural network to properly train. This technique has to be flexible on the number of labeled pixels. We show that this method is flexible on the type of images and on the sparsity of the weak labeling carrying out experiments on datasets of different domains and scenarios.

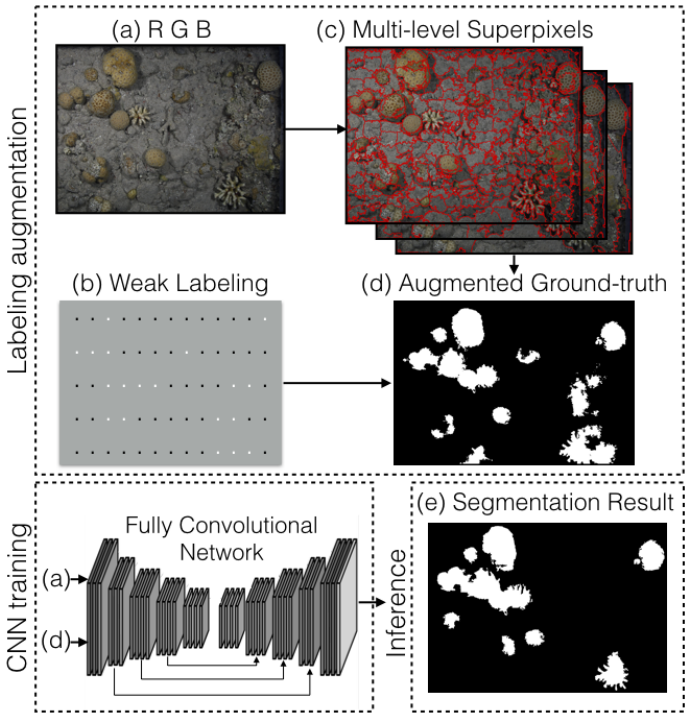


Figure 3.1: Labeling augmentation and segmentation pipeline proposed. Given a RGB image (a), the multi-level superpixels (c) are computed and then combined with the original weak/sparse labeling (b) to get the augmented ground-truth (d). Then, a fully convolutional neural network is trained with this, to get the final segmentation model used for inference (e).

3.2 CNN architectures

Per-patch vs Per-Pixel Classification. The problem of having weak labeling can be addressed using two different classification strategies based of convolutional neural networks:

- The first strategy is training a **classification model** on patches cropped around the labeled pixels. This strategy will have more samples to train (one sample per patch, i.e., n times more, being n the number of crops per image) [BTK+16].
- The second strategy is **augmenting the labeling** in order to get a dense (but approximated) labeling (i.e., all the pixels are labeled, but may have noise) and to train a dense model for semantic segmentation with the augmented labeling.

The first approach has only real labels (actually provided by the human expert) whereas the second one has more labels although these labels are noisy, the augmented labels are not going to be perfect. The second approach will lead to a more detailed result (per pixel information). Other approaches such as training a dense model learning only from sparse ground-truth have also been evaluated, but they work worse than

augmenting the ground-truth, as can be seen in the preliminary results from [ACM+17] (Appendix A shows this approach which is part of our work).

Both approaches consist of classification problems which can be formulated as a minimization of the error between the expected result and the predicted classification $min(error(\hat{y}, y))$. In the neural network, this means that both of them use the cross entropy loss function 3.1 as a general rule.

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} y^{(i)} \ln \hat{y}^{(i)} + \left(1 - y^{(i)}\right) \ln \left(1 - \hat{y}^{(i)}\right) \tag{3.1}$$

N is the number of labeled pixels, $y^{(i)}$ is the label, $\hat{y}^{(i)}$ is the CNN predicted output.

Different architectures for semantic segmentation. Concerning the dense segmentation strategy, we consider architectures from two different types: the FCN [LSD15], which maps the encoder to the expected segmentation in only one step, and the symmetric encoder-decoders [BKC17], which may have shortcut connections.

We have built the three architectures (patch-classification, FCN, symmetric encoder-decoder) on top of the same base model, DenseNet [HLWvdM17]. In particular, the patch-classification architectures uses DenseNet-169 with k=24, the FCN architecture uses the classification configuration (DenseNet-169) combined to an upsampling layer and the symmetric encoder-decoder uses the Tiramisu architecture [JDV+17].

Section 4.2 shows the experiments comparing these architectures, trained from scratch. We also show additional finetuning experiments with the purpose of studying the potential benefits of using pre-trained models as initialization of our training.

3.3 Labeling augmentation with multi-level superpixels

This section describes the proposed strategy for sparse label augmentation. It is based on superpixel segmentation and is able to adapt to any amount/density of labeled pixels.

Superpixel segmentation techniques. These techniques cluster the image pixels creating groups of similar connected pixels (known as superpixels). Our labeling augmentation is based on existing superpixel segmentation techniques. There are many superpixel segmentation techniques using different strategies, as detailed below, but our system is independent of them. The key idea of our augmentation strategy is simple, and follows these steps:

- First, we segment the image in superpixels (see Fig. 3.2).
- Then, we expand the labeled pixels. This expansion propagates the value of the labeled pixels to the rest according to the superpixel segmentation, i.e., all pixels in each superpixel get the label that appears the most within that superpixel.

In the section 4.3 we show some experiments comparing the effectiveness of different superpixel segmentation techniques to augment the sparse ground-truth, in order to pick the one which fits better our purpose. We experiment on SEEDS [VdBBR+12], CRS [CMM13], ERS [LTRC11], SLIC [ASS+10] and PB [ZHMB11]. Figure 3.3 shows some examples.

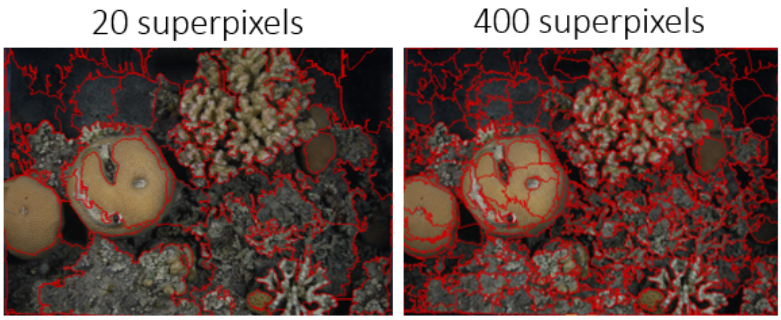


Figure 3.2: Superpixels obtained varying the number of superpixels (clusters) to get, both obtained with SEEDs superpixel technique.

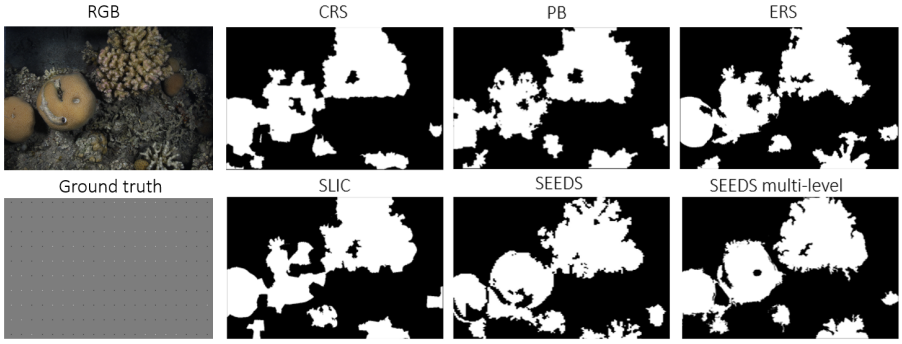


Figure 3.3: Comparison of the augmentation of the sparse labeling using different superpixels segmentation techniques including our contribution, using a multi-level superpixel segmentation. The top-left image is the original image, the rest are binary coral/no-coral segmentation obtained with different techniques to augment the available ground truth (bottom-left). Note it is just a very sparse grid of annotated points.

Multi-level Superpixel Segmentation. The performance of the labeling augmentation depends on the number of labeled pixels and the number of superpixels. The more labeled pixels and superpixels we have, the better performance on the augmentation.

Nevertheless, the standard superpixel segmentation techniques have some flaws and drawbacks:

- The number of superpixels is usually fixed.
- Some superpixels may not have labeled pixels inside, thus, the augmented ground-truth will have unknown regions.

A higher number of superpixels gives better results and it fits better actual shapes. Nevertheless, it increases the number of superpixels with no labeled pixels. Our multi-level technique solves this issue.

The multi-level superpixel segmentation proposed (see Algorithm 1) consists of applying several times the superpixel segmentation techniques to the image, decreasing the number of superpixels to generate in each iteration. In the first iteration, the

number of superpixels should be a very high number, leaving a lot of unlabeled pixels in the augmented labeling. The following iterations continue increasing the size of the superpixels until they manage to fill all the unlabeled pixels.

Algorithm 1: Multi-level Superpixels

function MLsuperpixels $(SparseGT, image)$;
Input : The sparse ground-truth or weak labeling ($SparseGT$) and the RGB image ($image$)
Output: The augmented ground-truth ($augmentedGT$) using multi-level superpixels
$nSuperpixels = getHighNumber()$;
$augmentedGT = blankImage()$;
while $augmentedGT.hasUnlabeledPixels()$ **do**
    $sp = getSuperpixels(image, nSuperpixels)$;
    $aug = getAugmentedLabels(SparseGT, sp)$;
    $augmentedGT = mask(augmentedGT, aug)$;
    $nSuperpixels = decreased(nSuperpixels)$;
end
return $augmentedGT$;

Chapter 4

Experiments

This chapter describes the different experiments performed to evaluate the proposed approach for semantic segmentation.

4.1 Evaluation

First of all, this section explains the datasets considered and the evaluation metrics used.

Datasets. The main dataset used in the following experiments is the Eilat Fluorescence Corals dataset [BTK+16].

This dataset has only 142 training images and 70 for validation. There are 200 labeled pixels per image, assigning to each of them a label from 4 coral and 6 non-coral classes[1].

The other datasets we use in the experiments allow us to demonstrate the generalization of the proposed methods. Besides, they have dense annotations, which allow us to evaluate more accurately the results (Eilat dataset has not). We use three additional datasets:

- Coral: Another coral dataset which has more classes and images.
- RIT: An aerial imagery dataset with multi-espectral information[2].
- Pascal VOC 2012 [EVGW+10] for instance segmentation (Berkeley augmentation)[3].

Ground-truth. Concerning the **Eilat dataset**, we use three different ground-truths to evaluate the results of the segmentation:

- Original-GT: The original sparse ground-truth available with the dataset.
- Augmented-GT: The augmented ground-truth, obtained by our approach.

[1] http://datadryad.org/resource/doi:10.5061/dryad.t4362
[2] https://github.com/rmkemker/RIT-18
[3] http://www.eecs.berkeley.edu/

- Manual annotations: A few manual annotated images, performed by a marine biologist collaborator, for coral vs non-coral segmentation.

The original sparse annotations are the least representative and reliable of this three ground-truth with very few annotations but, on the other hand, they are all true. The augmented ground-truth is an approximated labeling with some noise (94% accuracy against the provided dense manual annotations). The dense manual annotations are the best to compare, but, they are very few images.

The **other datasets** are evaluated with their corresponding available dense labeling. The sparse ground-truth is generated specifying the number of labeled pixels to have. Generating the simulated sparse ground-truth automatically has at least one important drawback to take into account. As no human is involve on the labeling, the labels are picked randomly or using a grid structure. Thus, some small shapes may not be selected leading to its disappearance in the simulated sparse ground-truth.
For these experiments, the simulated ground-truth of the different ground-truth is generated with the 0.1% of the real ground-truth, i.e., for an image of $500x500$ resolution, the simulated sparse ground-truth will have 250 labeled pixels.

Metrics. We use the standard metrics for classification and semantic segmentation: *Accuracy*(mean accuracy per pixel), *mean accuracy per class*, the *mean IoU per class* and the *DICE metric* which is very similar to the IoU.

The IoU stands for Intersection over Union. It is also known as the Jaccard index. Concerning semantic segmentation, this metric is aimed to measure how much of the real/correct segmented object is actually being predicted.

$$Jaccard(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

The X is the real segmentation (the labeling) and the Y is the predicted segmentation.

The DICE metric, also known as Sørensen–Dice coefficient, is similar to the IoU. Both of them imply the concept of the recall metric on their formulas.

$$DICE(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}$$

4.2 Comparison of CNN architectures

These experiments evaluate and compare the three approaches considered, explained in Sec. 3.2. This experiment consists of training different models with our augmented labels as input, to know which CNN architecture and approach gives better results when trained with the sparse labeling.

Configuration of the experiments. The ground-truth augmentation for these experiments is based on SEEDS superpixels [ACM+17] (see Appendix A). The training set up for the three approaches has been the same: 500 epochs (all of them converge), initial learning rate of 0.0005 with an exponential learning rate decay of 0.99. These models are trained from scratch, using the training/test split of the data described in the Section 4.1.

Table 4.1: **CNN model comparison** for binary classification (coral vs non-coral). Input: RGB images. Ground-truth: For FCN and Tiramisu, the Augmented-GT based on [ACM+17]. For patch classification, each patch has only one label pixel which is the label to train.

Model	Avg. accuracy per pixel	Avg. accuracy per class	Dice	IoU
	Evaluation: **Manual annotation** based **dense scores**.			
Patch classif.	74.40	54.36	54.61	43.66
FCN	92.19	81.78	83.50	73.51
Symmetric enc-dec.	**94.02**	**85.10**	**87.87**	**79.02**
	Evaluation: **Augmented-GT** based **dense scores**.			
Patch classif.	89.11	76.11	75.92	64.58
FCN	90.33	70.83	74.27	63.34
Symmetric enc-dec.	92.32	82.07	82.77	73.01
	Evaluation: **Original-GT** based **sparse scores**			
Patch classif.	93.75	90.00	91.78	85.06
FCN	81.01	71.87	73.69	60.27
Symmetric enc-dec.	89.18	86.78	86.44	76.79

Results training from scratch. A summary of these results is shown in Table 4.1. The segmentation approaches yield to better results than the patch-classification method [BTK+16] according to the dense scores, as expected. This implies that the augmentation does its job and works fine. 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 learning effectively due to the huge increase in the number of training data (labeled pixels). The Tiramisu architecture gets the best results, even better than previous work on this dataset and problem [ACM+17]. The Tiramisu architecture is actually pretty suitable for this dataset due to the shortcuts connections which allows to learn better even when training with a very few number of images. This architecture also is a more complex architecture than the FCN. Although the classification approach gives good results on the sparse scores, one thing to point out is, that all the pixels in its ground-truth patches, are treated as if they were of the same class when they are not. Thus, the CNN learns to classify the center of the patch. Training with patches which are not centered on a labeled pixels, leads to a drop of about 10% in all the metrics concerning the sparse scores.

Results with Fine-tuning. In this experiment we only use the Tiramisu architecture. We carried out finetuning experiments both on the binary and multiclass classification with the purpose of answering three questions about the potential benefits of finetuning: *Does the training converge earlier? Can it learn with less amount of data? Does it yield to better results?*.
We use two different datasets to learn the base model, on top of which we will run the finetuning. Moorea[4] dataset, which is similar to the Eilat, and the Camvid dataset[5] which is from a different domain. Concerning both datasets, the finetuning converge earlier than a common inizialitation. Pretraining on the Moorea, a dataset from the same domain, allows to train with less amount of images getting the same results. Concerning the last question, neither of the two datasets allow to get better results through finetuning.

[4]https://www.bco-dmo.org/dataset/676105
[5]http://mi.eng.cam.ac.uk/research/projects/VideoRec/

4.3 Labeling augmentation with multi-level superpixels

The results of this section show the advantages of using the multi-level superpixels approach for labeling augmentation with respect to the basic superpixels augmentation from Alonso et al [ACM+17] as well as a comparison between different superpixels segmentation methods applied to the labeling augmentation.

Table 4.2: **Augmented ground-truth using superpixels** using the augmented ground-truth extracted with the superpixes using **RGB** or **fluorescence**. Evaluated with Coral vs non-coral Eilat dataset.

Augmentations	Avg. accuracy per pixel	Avg. accuracy per class	Dice	IoU
Using fluorescence	Evaluation: **Manual annotation**, **dense scores**			
SEEDS 1-level	93.38	86.86	86.78	77.86
SEEDS multi-level	**94.20**	**87.50**	**88.18**	**79.88**
SLIC multi-level	93.86	85.37	87.10	78.37
Using RGB	Evaluation: **Manual annotation**, **dense scores**			
SEEDS 1-level	92.21	80.20	82.48	72.90
SEEDS multi-level	93.23	84.91	86.03	75.37
SLIC multi-level	92.76	83.60	84.93	75.37

Multi-level vs 1-level. The Eilat dataset has multimodal information. It has RGB and fluorescence images. The first experiment shows this method can work on different types of images.
The table 4.2 shows the augmentation labeling results of the baseline (basic superpixel augmentation) and the two best superpixels augmentations results using multi-level superpixels augmentation. The rest of the superpixel methods we experimented on (PB, ERS, CRS) got around 1% less in all the metrics with respect to SLIC. This methods can work on different multimodal information using the best augmentation to train the CNN. Nevertheless, we focus our efforts and experiments on the RGB information to be able to generalize it to other domains and datasets. The multi-level superpixels augmentation outperforms the baseline by 1-3% in all the metrics. SEEDS superpixels works better for the labeling augmentation due to it fits better to the shapes.

Suerpixel post-processing. The next step is to train the Tiramisu CNN with this augmented ground-truth. The results of training the CNN achieves betters scores (on manual annotation scores) than the augmented ground-truth because the CNN learns even when there is some noise in the labeling. Superpixels can be also used to enhance the quality of the CNN output. Thus, the final result will fit better to the shape of the image to be segmented. We applied SEEDS superpixels to the output to improve it qualitatively and quantitatively (see Fig. 4.1). This idea can be also applied to the manual annotations. Superpixels can fit better the shapes than manual annotations made by human, which they are not perfect. Thus, we use here another metric based on the manual annotations after applying them the SEEDS post-process. Tables 4.3 and 4.4 shows the results of training the Tiramisu CNN with our augmented ground-truths.

Conclusions. The multi-level superpixel augmentation allows the CNN to learn dense predictions getting significant scores on the evaluation. The superpixel post-process,

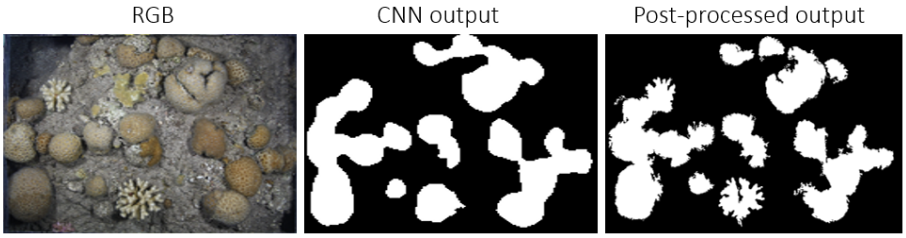


Figure 4.1: A comparison between the result of our pipeline using the multi-level Superpixels for labeling augmentation and the same output applying SEEDS superpixels to enhance it.

Table 4.3: **Segmentation results comparison for binary classification.** Evaluation of the usage of the SEEDS post-process to enhance the results. Tiramisu trained using the augmented ground-truth extracted with the multi-level Superpixes. Input: RGB. Augmentation using RGB. *pp: Post-process with SEEDS superpixels (see Fig. 4.1)

CNN outputs	Avg. accuracy per pixel	Avg. accuracy per class	Dice	IoU
	Evaluation: **Manual annotation**, **dense scores**.			
Our pipeline	94.53	85.51	88.21	79.48
Our pipeline pp*	94.68	84.50	86.59	77.86
	Evaluation: **Manual annotation pp***, **dense scores**.			
Our pipeline	96.53	91.45	91.97	85.72
Our pipeline pp*	97.71	93.72	94.65	90.42

Table 4.4: **Segmentation results comparison for multiclass classification.** Evaluation of the usage of the SEEDS post-process to enhance the results. Tiramisu trained using the augmented ground-truth extracted with the multi-level Superpixes. Input: RGB. Augmentation using RGB. *Post-process with SEEDS superpixels (see Fig. 4.1)

CNN outputs	Avg. accuracy per pixel	Avg. accuracy per class	Dice	IoU
	Evaluation: **Augmented-GT** based **dense scores**.			
Our pipeline	90.96	51.28	50.39	39.44
Our pipeline pp*	91.68	52.76	52.95	42.22
	Evaluation: **Augmented-GT pp*** based **dense scores**.			
Our pipeline	91.32	51.91	50.71	39.68
Our pipeline pp*	92.20	53.67	53.90	43.27

yields to better scores both on the non post-processed annotations and the post-processed ones. This means it actually benefits the segmentation output.

One thing yet to solve is, to compare the results when training on the augmented ground-truth versus training on the real annotations. For answering this question, as this dataset has no dense manual annotations for all the images, in the next section, we use other datasets.

4.4 Generalization to other datasets

This section is aimed to evaluate the proposed method and the conclusions of the previous experiments on different datasets (see Sec. 4.1).

Description. The problem of having good dense annotation data does not happen only in a specific domain. Thus, although we focus our experiments on biological datasets (coral reef's datasets), the solution we propose to this widespread problem can be applied to other domains and datasets.

To facilitate the evaluation, the datasets are densly labeled so that we can properly evaluate the results. This way, we can also simulate the sparse ground-truth (explain in previous Section 4.1) to apply the labeling augmentation proposed based on multi-level superpixel.

Labeling comparison: Augmented vs real. The first experiments consist of comparing the real dense labeling and the augmentation of the simulated sparse labeling through our proposed method. This is done with the three different datasets (see Sec. 4.1). A summary of these results of the is shown in Table 4.5. Fig. 4.2 shows visual examples of these experiments on the different datasets. For more results, please go to the C.2.

The labeling augmentation strategy we propose relies on the labeling. This means, that this method needs the sparse ground-truth to have at least one labeled pixel per object. As we have simulated the sparse ground-truth, if the dataset has images with small object or details and the sparse ground-truth does not cover it, it will not be in the augmented ground-truth. This can be seen in the IoU or DICE metric. The RIT dataset (see Fig. 4.2) is the dataset with more small details and the Table 4.5 shows how the IoU drops when this happens. The proposed method benefits more from covering all the labeled instances in the sparse ground-truth than having more labeled pixels.

Table 4.5: Comparison between the real dense manual annotations and the augmentation from the simulated sparse ground-truth.

Different dataset	Avg. accuracy per pixel	Avg. accuracy per class	Dice	IoU
	Evaluation: **Manual annotation** based **dense scores**.			
Coral	89.08	85.04	82.20	73.63
RIT	97.44	60.64	61.64	54.28
VOC 2012	93.49	82.72	84.90	87.19

Table 4.6: Comparison between the results of training the CNN with real dense manual annotations and the augmentation from the simulated sparse ground-truth.

Dataset and labeling used	Avg. accuracy per pixel	Avg. accuracy per class	Dice	IoU
	Evaluation: **Manual annotation** based **dense scores**.			
Coral (with real labeling)	78.11	19.20	17.69	13.79
Coral (with augmented labeling)	74.03	19.57	18.02	14.54
RIT (with real labeling)	94.23	20.36	20.03	19.16
RIT (with augmented labeling)	89.30	19.65	19.29	17.85

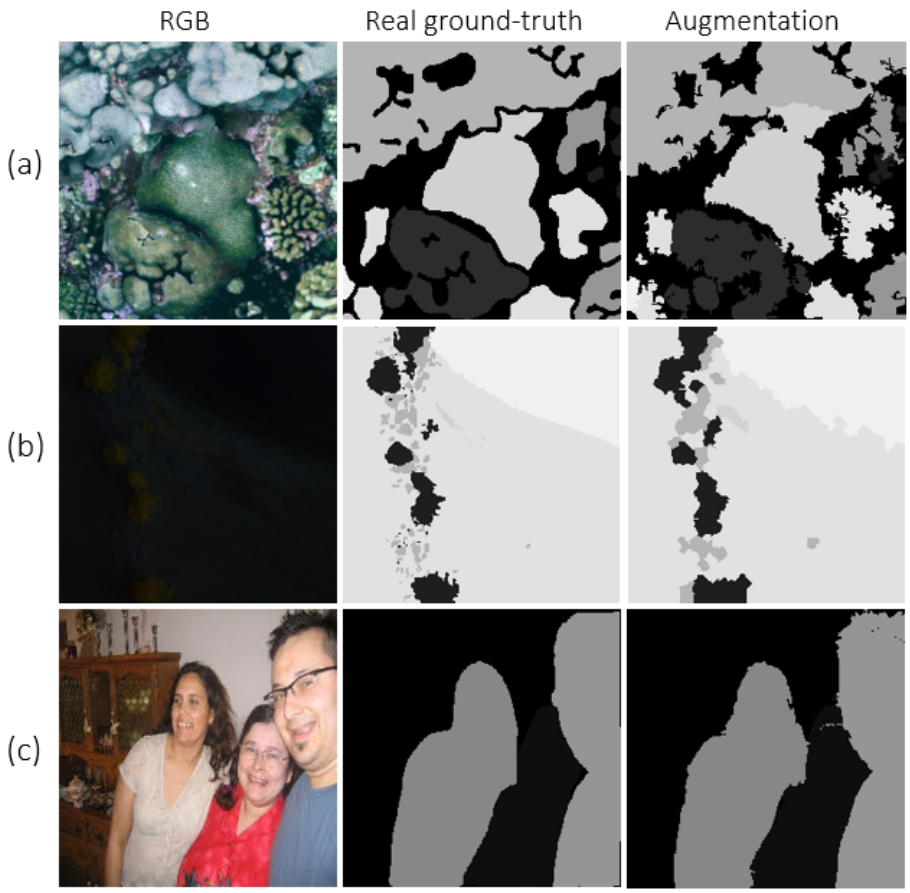


Figure 4.2: Comparison examples between the real dense manual annotations and the augmentation from the simulated sparse ground-truth. Image (a) is from the coral dataset, (b) is an aerial image from the RIT multi-espectral dataset and (c) is an image from PASCAL VOC 2012.

Training with different labeling: Augmented vs real. We also compare the results of training on the real dense segmentation ground-truth and training on the augmented ground-truth via multi-level superpixels. Table 4.6 shows the comparison between training the CNN with the dense ground-truth and training it with the proposed augmented ground-truth. Training with the original dense ground-truth yields to better results as expected. Nevertheless, the results of training with the augmented ground-truth are very similar which is actually a very impressive results.

Chapter 5

Conclusions

5.1 Technical conclusions

Concerning the objectives of the project, all of them have been reached. Next we briefly discuss the most important technical conclusions and possible future steps.

Discussion. We have presented a novel approach 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 segmentation through a econder-decoder CNN getting results comparable to those obtained by training with the real dense ground-truth.

The experiments show the labeling augmentations via superpixel work better than other more direct options. We have analyzed the influence of using different superpixels segmentation methods on the augmentation as well as showed the benefits of applying the proposed multi-level approach which is able to cope with a more variety of sparse labeling and images. Our results show the benefits of using the proposed augmentation of sparse image labels on very different kinds of datasets and domains. We also show how superpixels can enhance bot the outputs of the neural network and the human manual annotations.

Future Work. As future steps, we plan to explore other state-of-the-art CNN architectures for semantic segmentation, as well as studying more sophisticated labeling augmentation methods and probably, to extend these types of methods to 3D data.

.

5.2 Personal conclusions

I must say that this Christmas has been the one I have worked the most, but it has been worth it. Thanks to this project I have learned a lot of things, the most significant are:

- **About deep learning and other related topics reading articles**.

- How to write articles.
- The thinking process and valuable research knowledge
- New programming skills for Deep Learning and Computer Vision applications (it is not the same to know about something than to code it).

As previously mentioned, the project still has a lot of parts that can be improved, and I still have a lot of things to learn on these topics, but both technical and personal goals with this project have been achieved by far. For me, this project and everyday in this lab is really amazing. Working on what you love, surrounded by really amazing people, has no price.

Appendix A

IEEE ICCV-Workshop Publication

Part of the work of this Master thesis has been published and presented in the Workshop on Visual Wildlife Monitoring (held with the IEEE International Conference of Computer Vision 2017, and published in the IEEExplorer conference proceedings). This Appendix includes the whole publication.

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.

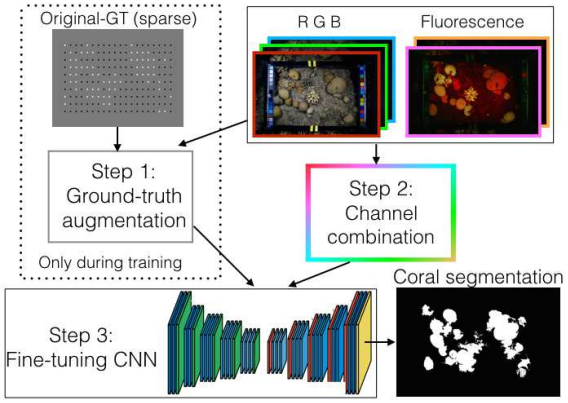


Figure 1. Coral segmentation pipeline based on CNN segmentation model. Step 1: sparse ground truth available is augmented to facilitate training. Step 2: input multimodal data is combined to use the more discriminative channels. Step 3: fine-tuning.

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].

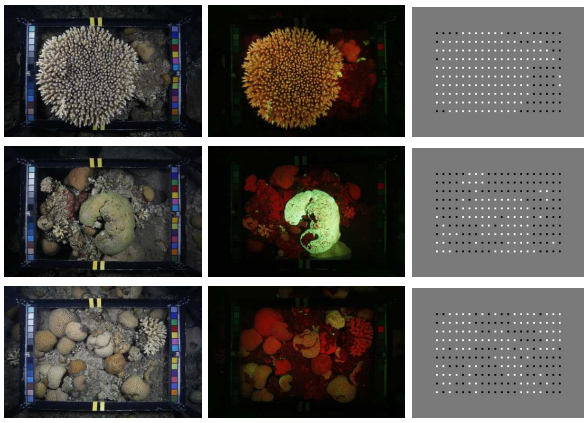
As previously mentioned, our work studies and proposes how to face the challenges to enable latest results on semantic segmentation using CNNs to the segmentation of coral imagery. Prior work has demonstrated how the use of multi-modal data can facilitate this problem, in particular combining RGB images 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 pixels 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.

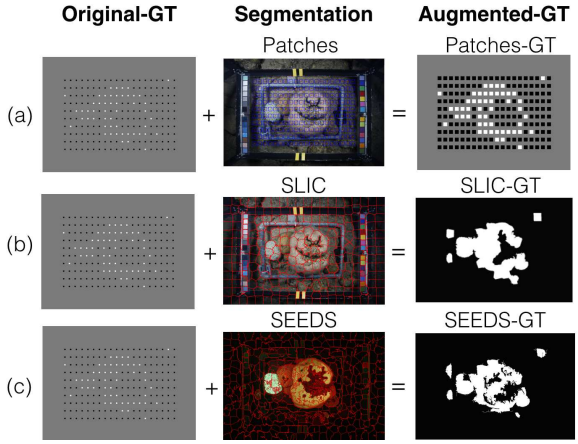


Figure 3. Ground truth augmentation methods that we considered. (a) small patches around original-GT labeled pixels; (b) SLIC and (c) SEEDS superpixels, computed on RGB or fluorescence images, used to expand the original-GT. SLIC and SEEDs can be augmented using either RGB or fluorescence image superpixels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

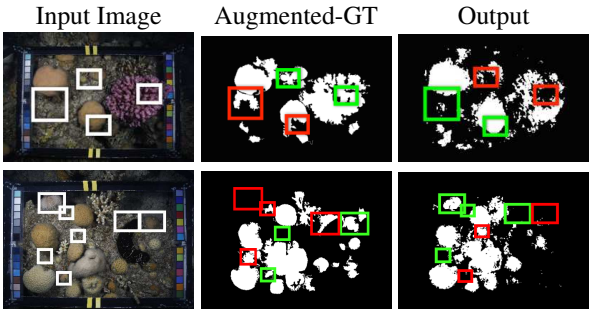


Figure 4. Even though the Augmented-GT (SEEDS-GT in these images) used to train our system is noisy, these examples show that the final segmentation obtained with our trained model detects regions that are missed in the augmented GT computed for those examples. We have manually highlighted incorrect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

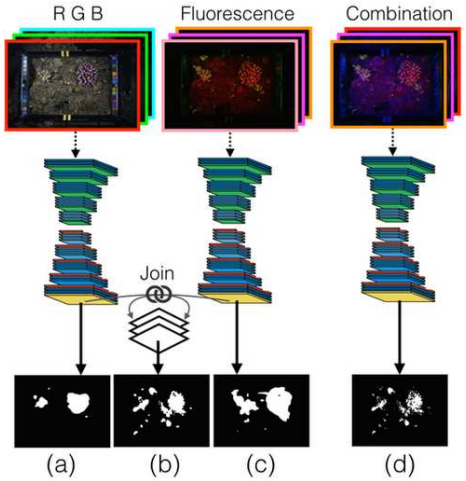


Figure 5. Different strategies to combine available image channels to train an end-to-end segmentation model: (a) fine-tuning with a 3-channel input using RGB data only; (c) using fluorescence data only; (d) combination of both ($Fluor_1 + Fluor_2 + Red$); (b) joining two of the fine-tuned models.

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:

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

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.

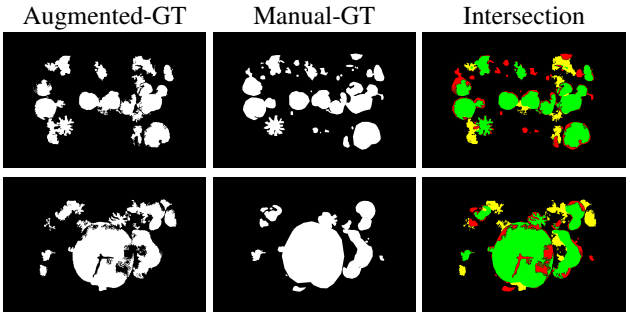


Figure 6. Examples of Augmented-GT and Manual-GT. The intersection of both shows green/red/yellow pixels when labeled as coral in both/only manual-GT/only augmented-GT respectively.

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

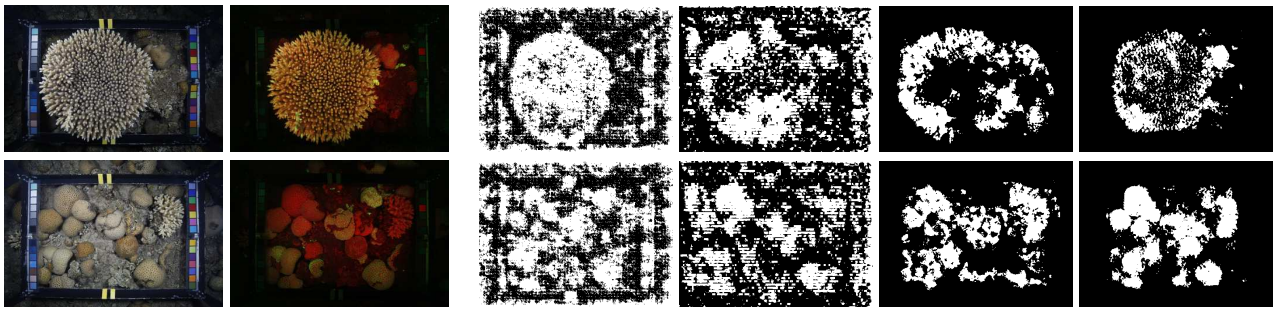
Original-GT based sparse scores. The scores computed based on the original ground truth (Original-GT) are not fully representative, as it will be shown next. Intuitively, 200 pixels labeled out of around a million per image are not a dense ground truth for dense image labeling.

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



(a) (b) (c) (d) (e) (f)

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: *Training*:	Original-GT **(sparse)**	Patches-GT	SLIC-GT **(dense)**	SEEDS-GT **(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).

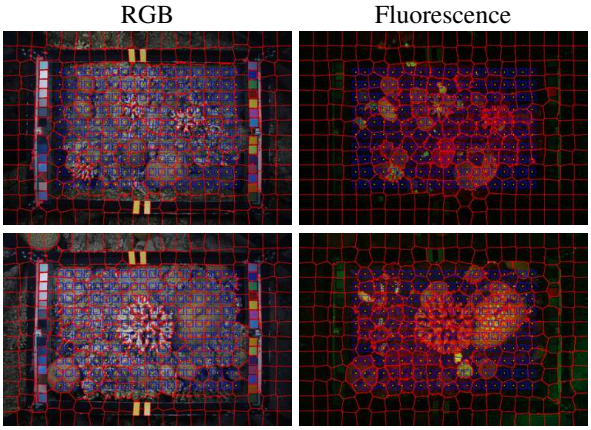


Figure 8. Superpixel segmentation of the image (red boundaries). Segmentation on fluorescence images fits better the coral regions.

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.

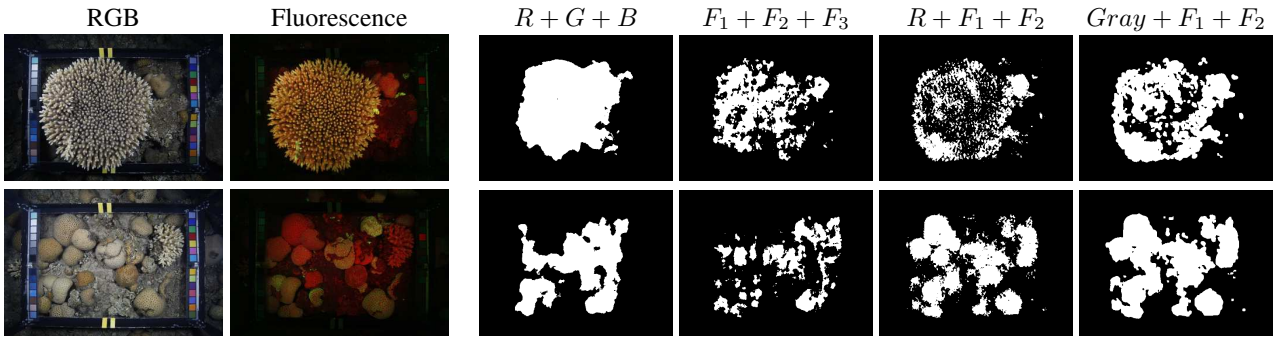


Figure 9. Coral segmentation using the proposed augmented-GT and different input channel combinations. The results of the different combinations are detailed in Table 2. The $Gray + F_1 + F_2$ combination yields the best qualitative results.

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

- 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).

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

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-

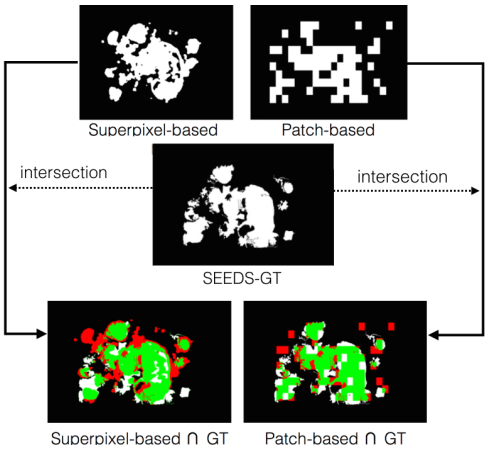


Figure 10. Coral segmentation with patch-based or superpixel-based (ours) approaches compared to augmented ground truth (SEEDS-GT). The patch-based shows the simulated output of results in [4]. Superpixel-based shows results using fine-tuned SegNet using $Gray+F_1+F_2$ input channels. The intersection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

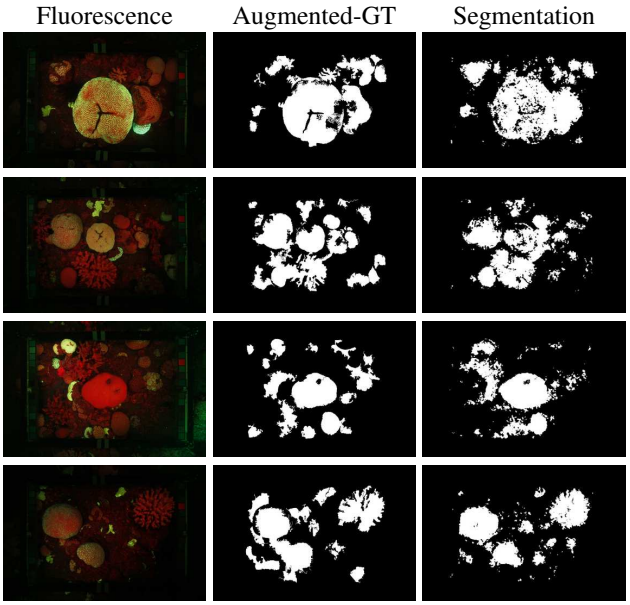


Figure 11. Coral segmentation results obtained from a model trained with the augmented (dense) ground truth and $Gray + F_1 + F_2$ input channels. Each row depicts the fluorescence image, augmented ground truth with SEEDS-GT, and the respective coral segmentation result.

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.

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

Appendix B

Deep Learning and Semantic Segmentation

This appendix contains more detailed background on two important concepts in this work: the *Semantic Segmentation* problem in computer vision, and a few concepts from *Deep Learning*.

Artificial Intelligence is a branch of computer science dealing with the simulation of intelligent behavior in computers, and Machine learning is a branch of Artificial Intelligence which is based on statistics and tries to find patterns in data. Often machine learning is divided in three subcategories depending on the particular problem and the data considered:

- Supervised learning. When the data available has labels of the target classification classes, i.e., the expected label of the data is known. For example, in the case of image classification, the data is the images and the labels are what is in the image (a dog, a cat...).
- Unsupervised learning. When only data is available. When the learning is based on the data structure without having any prior information of its meaning. Clustering is the most common example.
- Reinforcement learning. When there is neither data nor labels, but there is a goal. Here the approach is learning by trial and error.

Deep learning is a machine learning technique now widely used because of the extraordinary results demonstrated lately [LBH15], usually when there is a large amount of data available.

B.1 Deep Learning

Deep Learning can be seen as a concatenation or composition of functions. The *Deep* word mean the number of functions and transformations applied to the input is very high (see Fig. B.1 [1]).

[1] source: https://hackernoon.com/training-an-architectural-classifier-iii-84dd5f3cf51c

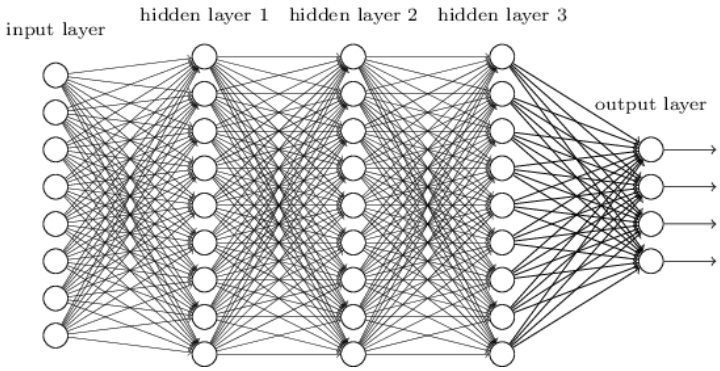


Figure B.1: Simple architecture of a neural network. All neurons are connected with all the neurons of neighbours layers. The common applied function to each relation between two neurons is a weighted the multiplication.

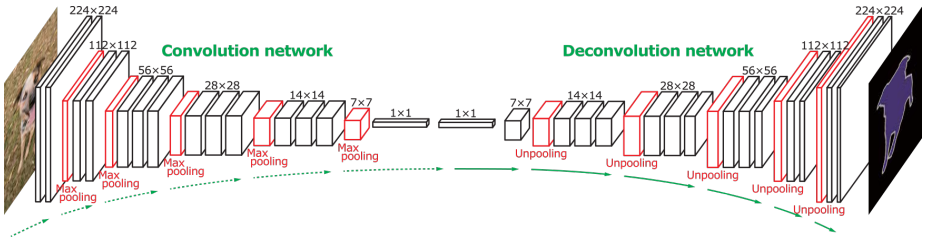


Figure B.2: Encoder-decoder neural network. It has a convolutions part and a deconvolution part. maxpooling layers decrease the size of the output meanwhile the unpooling layers increase it. This is the general architecture for semantic segmentation. (source: `http://cvlab.postech.ac.kr/research/deconvnet/`)

The learning method of Deep Learning is based on back-propagation. As in other supervised learning techniques the learning consists of several steps:

- Given an model, an input and a expected output, compute the loss or error
- Compute the gradients, i.e., see how each learning weight affect to that loss
- Update the learning weights with a small step trying to reduce that loss.
- Repeat the three previous steps with all the dataset till convergence.

The update step is called *learning rate*. In Deep Learning there are a lot of important steps to follow and which matters to properly train a neural net. This are some of the most important steps:

- Find a good codification of the input
- Find a good loss function which really means the optimization you are looking for (between the expected output and the output of the neural network).
- Find a good learning rate
- Select the types of layers and functions, i.e., for images, convolution layers.

The expected output of the network is usually called *ground-truth or labeling*.

A common problem of Machine Learning techniques is the overfitting. Deep Learning copes with this problem using regularization terms like other Machine Learning techniques. But Deep Learning deals with overfitting also using *finetune*. *Finetune* is to train a model with a very large amount of data and then, take that model as a initialization of another training. Depending on the similarity between the pre-trained dataset and the new dataset, you can also freeze the learning weights in order to take advantages of the learned patters on a larger datasets which will generalize more and work better on the majority of the cases.

B.2 Semantic segmentation

This project uses deep learning to solve a semantic segmentation problem. Semantic segmentation (or pixel classification) associates one of the pre-defined class labels to each pixel. The input image is divided into the regions, which correspond to the objects of the scene. So the result of applying to an image semantic segmentation is another image and each object or class of the image has the same values.

Deep Learning approaches try to learn convolutions and deconvolutions. Stacking several layers of learnable convolutions (see Fig. B.2). This functions will learn how to map the RGB image to the semantic segmentation classified image

Appendix C

More detailed results

In this section we show additional results for the experiments from the multi-level superpixel augmentation from Sections 4.3 and 4.4.

C.1 Superpixels

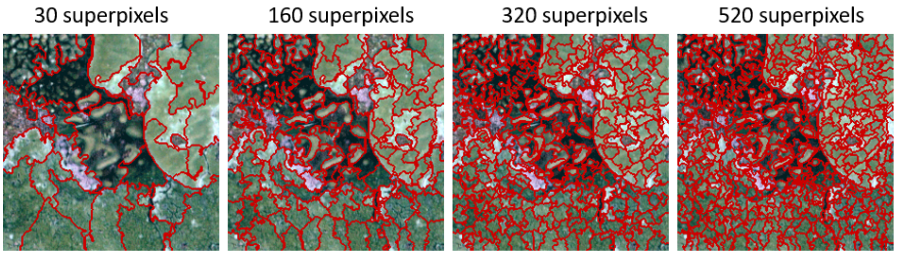


Figure C.1: Example of diferent number of SEEDS superpixels applied to an image.

C.2 Multi-level superpixels augmentation results

Here there are some figures with the augmentation results of the section 4.4 using different datasets.

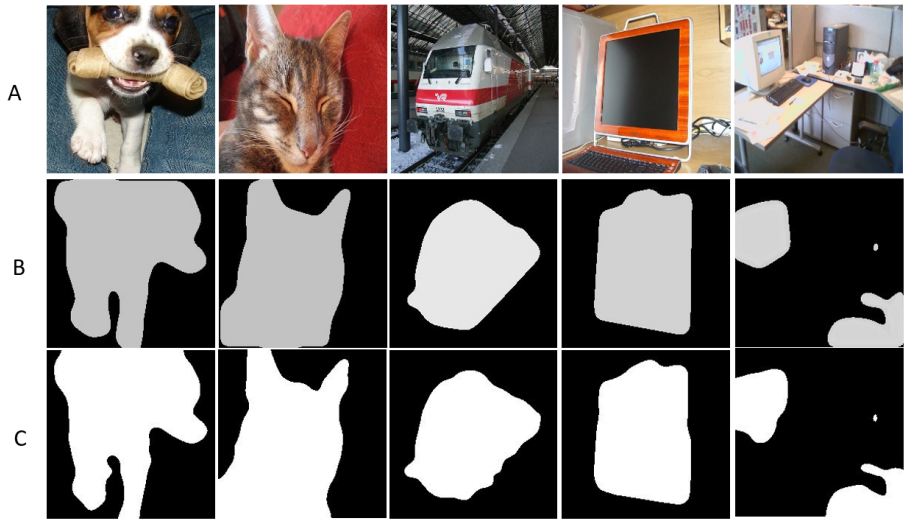


Figure C.2: Augmentations from sparse labeling. The images are from the VOC dataset. Comparison between the iamges (A), the real labeling (B) and the multi-level superpixel augmentation from the sparse labeling (C)

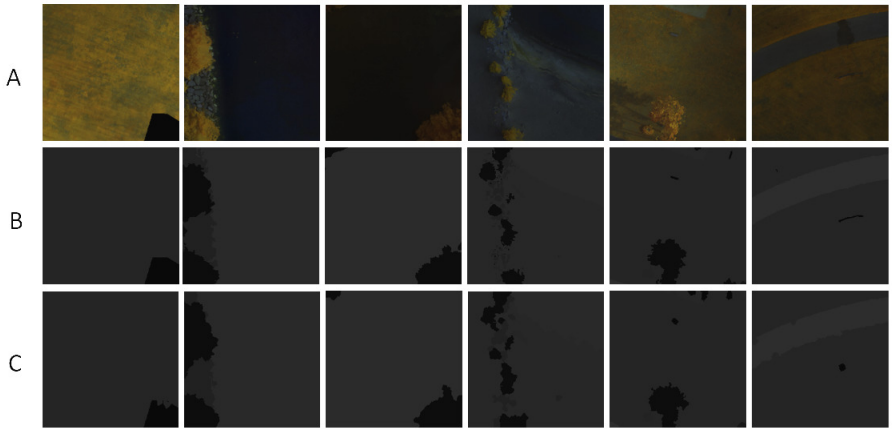


Figure C.3: Augmentations from sparse labeling. The images are from the RIT dataset. Comparison between the iamges (A), the real labeling (B) and the multi-level superpixel augmentation from the sparse labeling (C)

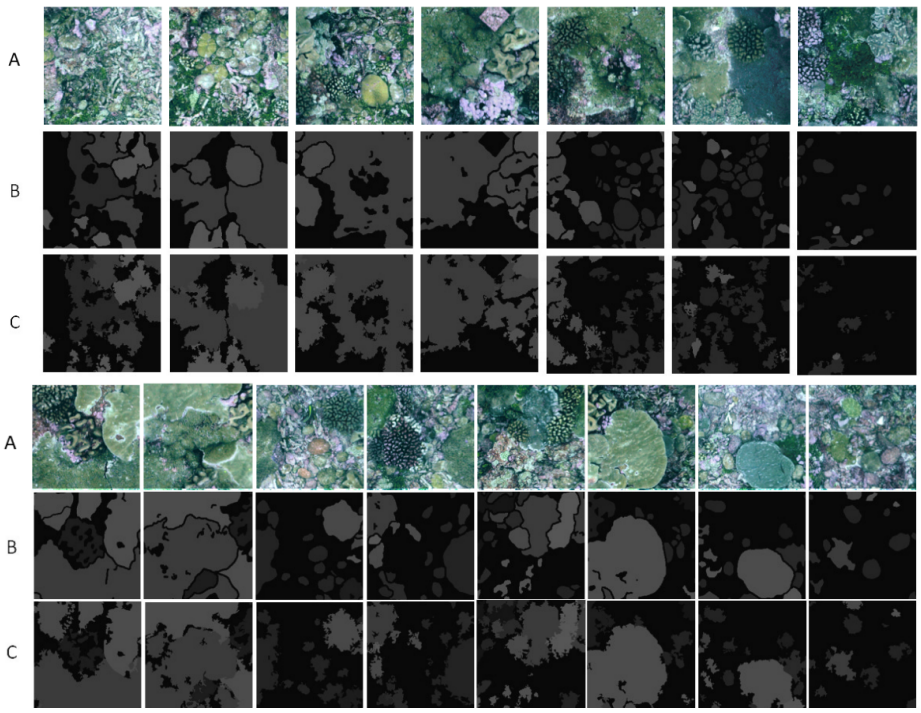


Figure C.4: Augmentations from sparse labeling. The images are from the Coral reefs dataset. Comparison between the images (A), the real labeling (B) and the multi-level superpixel augmentation from the sparse labeling (C)

Bibliography

[ACM+17] Inigo Alonso, Ana Cambra, Adolfo Munoz, Tali Treibitz, and Ana C Murillo. Coral-segmentation: Training dense labeling models with sparse ground truth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2874–2882, 2017.

[ASS+10] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels. Technical report, 2010.

[BDLO] Jean-Nicola Blanchet, Sébastien Déry, Jacques-André Landry, and Kate Osborne. Automated annotation of corals in natural scene images using multiple texture representations. *PeerJ Preprints*, 4:e2026v2.

[BEK+12] Oscar Beijbom, Peter J Edmunds, David I Kline, B Greg Mitchell, and David Kriegman. Automated annotation of coral reef survey images. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1170–1177. IEEE, 2012.

[BKC17] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[BTK+16] Oscar Beijbom, Tali Treibitz, David I Kline, Gal Eyal, Adi Khen, Benjamin Neal, Yossi Loya, B Greg Mitchell, and David Kriegman. Improving automated annotation of benthic survey images using wideband fluorescence. *Scientific reports*, 6, 2016.

[CCCM15] Ping-Yu Chen, Chi-Chung Chen, LanFen Chu, and Bruce McCarl. Evaluating the economic damage of climate change on global coral reefs. *Global Environmental Change*, 30:12–20, 2015.

[Ces00] Herman SJ Cesar. Coral reefs: their functions, threats and economic value. *Collected essays on the economics of coral reefs*, pages 14–39, 2000.

[CMM13] Christian Conrad, Matthias Mertz, and Rudolf Mester. Contour-relaxed superpixels. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 280–293. Springer, 2013.

[DMTC17] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2017.

[EVGW+10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[GGOEO+17] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017.

[GVZ16] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[HDH+17] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. *arXiv preprint arXiv:1711.10370*, 2017.

[HGDG17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *arXiv preprint arXiv:1703.06870*, 2017.

[HLWvdM17] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[JDV+17] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1175–1183. IEEE, 2017.

[KL16] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision (ECCV)*. Springer, 2016.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey 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.

[KVK16] Abhijit Kundu, Vibhav Vineet, and Vladlen Koltun. Feature space optimization for semantic video segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 3168–3175. IEEE, 2016.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[LFX+17] Zhiwu Lu, Zhenyong Fu, Tao Xiang, Peng Han, Liwei Wang, and Xin Gao. Learning from weak and noisy labels for semantic segmentation. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 39(3):486–500, 2017.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[LTRC11] Ming-Yu Liu, Oncel Tuzel, Srikumar Ramalingam, and Rama Chellappa. Entropy rate superpixel segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2097–2104. IEEE, 2011.

[MLD+17] Travis Manderson, Jimmy Li, Natasha Dudek, David Meger, and Gregory Dudek. Robotic coral reef health assessment using automated image analysis. *Journal of Field Robotics*, 34(1):170–187, 2017.

[MYS15] Mohammadreza Mostajabi, Payman Yadollahpour, and Gregory Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[PC15] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1713–1721, 2015.

[PKD15] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1796–1804, 2015.

[PVK17] Maria Papadomanolaki, Maria Vakalopoulou, and Konstantinos Karantzalos. Patch-based deep learning architectures for sparse annotated very high resolution datasets. In *Urban Remote Sensing Event (JURSE), 2017 Joint*, pages 1–4. IEEE, 2017.

[RSM+16] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio 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.

[SCH+17] Florian Shkurti, Wei Di Chang, Peter Henderson, Md. Jahidul Islam, Juan Camilo Gamboa Higuera, Jimmy Li, Travis Manderson, Anqi Xu, Gregory Dudek, and Junaed Sattar. Underwater multi-robot convoying using visual tracking by detection. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4189–4196, Vancouver, Canada, September 2017.

[USS+17] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. *arXiv preprint arXiv:1708.06500*, 2017.

[VC17] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[VdBBR+12] Michael Van den Bergh, Xavier Boix, Gemma Roig, Benjamin de Capitani, and Luc Van Gool. SEEDS: Superpixels extracted via energy-driven sampling. In *European conference on computer vision*, pages 13–26. Springer, 2012.

[XSU15] Jia Xu, Alexander G Schwing, and Raquel Urtasun. Learning to segment under various forms of weak supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3781–3790, 2015.

[YDP12] Suet-Peng Yong, Jeremiah D Deng, and Martin K Purvis. Novelty detection in wildlife scenes through semantic context modelling. *Pattern Recognition*, 45(9):3439–3450, 2012.

[ZHMB11] Yuhang Zhang, Richard Hartley, John Mashford, and Stewart Burn. Superpixels via pseudo-boolean optimization. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1387–1394. IEEE, 2011.

[ZMCL16] Hongyuan Zhu, Fanman Meng, Jianfei Cai, and Shijian 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.