Semantic Segmentation from Sparse Labeling using Multi-Level Superpixels

Iñigo Alonso[1] Ana C. Murillo[1]

***Abstract*— Semantic segmentation is a challenging problem that can benefit numerous robotics applications, since it provides information about the content at every image pixel. Solutions to this problem have recently witnessed a boost on performance and results thanks to deep learning approaches. Unfortunately, common deep learning models for semantic segmentation present several challenges which hinder real life applicability in many domains. A significant challenge is the need of pixel level labeling on large amounts of training images to be able to train those models, which implies a very high cost. This work proposes and validates a simple but effective approach to train dense semantic segmentation models from sparsely labeled data. Labeling only a few pixels per image reduces the human interaction required. We find many available datasets, e.g., environment monitoring data, that provide this kind of sparse labeling. Our approach is based on augmenting the sparse annotation to a dense one with the proposed adaptive superpixel segmentation propagation. We show that this label augmentation enables effective learning of state-of-the-art segmentation models, getting similar results to those models trained with dense ground-truth. We demonstrate the applicability of the presented approach to different image modalities in real domains (underwater, aerial and urban scenarios) with publicly available datasets.**

I. Introduction

Solutions for semantic segmentation have witnessed a significant boost in recent years thanks, in big part, to convolutional neural networks [11]. A lot of applications in the robotic field have seen the impact of these improvements, such us autonomous driving [24] or object detection and manipulation [35]. Unfortunately, many applications and domains do not have available the large amount of labeled training data required by successful existing techniques. Semantic segmentation, or dense labeling models, typically need pixel-level annotations in order to be trained [11], [37]. This type of labeling is very time consuming and often needs human experts, therefore it is not always available. For example, we find monitoring data from underwater regions in the CoralNet project [5] which only provide sparse labels provided by marine biology experts. Another example is DeepSat [4], a large dataset of satellite images which only provides image-level labels. These and plenty of similar monitoring projects would benefit from strategies to learn semantic segmentation models from image-level annotations or a few pixel labels. In particular, our work is focused on the challenge of *how to train dense labeling models from sparse labels*, as illustrated in Fig. 1. Solving this challenge enables the application of recent advances of semantic segmentation CNNs to a lot of domains, which can then enhance their processing and extract more detailed information and conclusions from their data.

[1] I. Alonso and A.C. Murillo are with RoPeRt group, at DIIS - I3A, Universidad de Zaragoza, Spain. {inigo, acm}@unizar.es

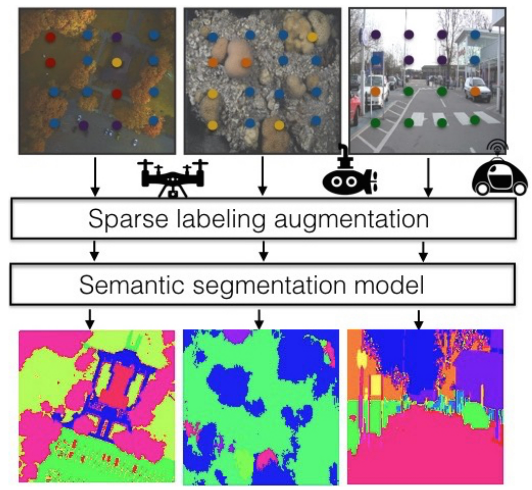

Fig. 1. Robotic platforms have enabled easy collection of plenty of monitoring datasets in different domains. This work demonstrates that augmenting sparse input data labels enables effective training of semantic segmentation models to process them with much lower labeling effort.

Inspired by the success of different data augmentation techniques used to train CNNs, we explore an strategy to enable better training from sparse labeling. The two main contributions presented in this work are:

- A **sparse labeling augmentation** method, based on propagation using superpixels, which enables effective training of fully convolutional neural networks, for pixel-level classification, when only very sparse ground truth labels are available.
- A comparison of different CNN based strategies for **semantic segmentation** from sparse ground-truth labels. We demonstrate their performance in realistic domains (underwater, aerial and urban scenarios) with publicly available datasets.

Our experimental results demonstrate that the proposed augmentation from sparse ground truth labels, despite having a few incorrect pixel labels, provides valuable and effective information to train an end-to-end segmentation model. The proposed approach benefits are twofold: 1) it provides better results than prior work, which was based on individual patch classification or simpler labeling augmentations, training on the same sparsely labeled data; 2) it achieves comparable results to approaches trained on densely labeled images, while having the advantage of less intensive annotation requirements.

II. Related Work

This section discusses related work from the most relevant topics to our work: state-of-the art semantic segmentation and strategies to deal with weak and sparse labeling.

A. Semantic image segmentation

Semantic segmentation has received significant attention in the recent years. As in many other applications, convolutional neural networks (CNN) have achieved the state-of-the-art with approaches such us Mask-RCNN [13], combining the idea of regions of interest (ROI) with per class segmentation and classification of every ROI, or Tirasmisu architecture [17], a fully convolutional extension of DenseNet [16], a state-of-the-art architecture for classification. Among earlier approaches, we find numerous solutions based on superpixel segmentation techniques [29]. This type of semantic segmentation approaches perform a superpixel classification or superpixel based label propagation [26], [31]. This recent survey on image segmentation by Zhu et al. [37] provides a detailed discussion of classical solutions for this problem, while Garcia et al. [11] present a compilation of recent contributions on semantic segmentation focused on deep learning, including new architectures and common datasets. Our approach joins both recent CNN based semantic segmentation models with superpixel segmentation algorithms. As we discuss later, while the CNN based models are the core of the segmentation process, the superpixels are shown to be very effective to augment sparse training labels.

Autonomous systems have facilitated the collection of extremely large amounts of data for monitoring tasks, such as target following in unstructured 3-D environments [28], autonomous surveillance of coral reef ecosystems [25] or wildlife monitoring from aerial systems [14]. To enable automatic processing of their content, semantic segmentation models for the different domains target classes are needed. This requires training new models, either from scratch or fine-tuning existing related models. In either case, a significant bottleneck is the lack of dense labeling to train semantic segmentation models, specially in domains where an expert is needed to label ground truth images. For example, CoralNet [5] is a collaboration project focused on coral reef monitoring, which shares a lot of datasets from underwater regions from all over the world. Unfortunately, they only provide a few pixels labeled per image, clicked manually by an expert. This common situation leads to a key challenge for our work: how to deal with the lack of labeled training data.

B. Lack of training data and labels

Scarce 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.

Data augmentation, i.e., generating additional data by altering the original labeled data, is a very common solution. Many works have followed this strategy, including for example the well know *Alexnet* model [20].

A more recent solution to augment the training data is to generate **synthetic data** [12], [27]. This strategy provides perfect ground-truth labels, at the cost of generating the scenes on the simulation platform. This brings additional challenges due to the difficulty of generating realistic images which cover all the variations from real world.

Models for weakly labeled data. Another common strategy to deal with a lack of accurate training labeled data is to build approaches that can learn from weakly labeled data. Lu et al [23] carried out a survey on different approaches to train semantic segmentation from noisy and weakly labeled data, which discusses these problems in detail and presents many related solutions. We discuss a few examples of the most frequently weak labels used for semantic segmentation: image-level annotations and sparse pixel annotations.

Several recent approaches study how to use **per image labels**, as opposed to per pixel labels, to obtain per pixel image segmentation models. Kolesnikov and Lampert [19] propose a new composite loss function to train semantic segmentation CNN models directly from image-level labels. Durand et al. [9] propose a classification neural network, trained from image-level labels to learn good representations, and then work with its feature maps to get an accurate segmentation result. Other scenario, closer to our case of study, consists of having **sparse labels** available when the model to be trained requires dense per pixel annotations. Several recent works have approached this challenge from different perspectives. Uhrig et al [32] propose a new CNN architecture, Sparsity Invariant CNNs, focused on reconstructing a dense depth map from sparse LIDAR sensor information. They work with sparse convolutions to learn directly from sparse labeling, and show successful results with levels of sparsity between 5% and 70%). Vernaza et al [34] 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. Hu et al. [15] propose to train from partially labeled data, introducing a new partially supervised training paradigm and weight transfer function.

Differently previously discussed approaches, which propose specific architectures, we study an alternative but complementary path. Inspired by the good results of data augmentation at image-level in many deep learning based approaches, we propose to *augment* the labeled data at pixel level by using propagation and study its effects. Preliminary results of training dense segmentation models with augmented sparse labels were shown in [2]. Here we improve those results thanks to a better strategy to augment the sparse labels, which is more robust, regardless of the modality of the input images, than the preliminary results. We present significantly better performance both in binary and multi-class segmentation, and a more exhaustive validation using more recent CNN architectures and additional application domains.

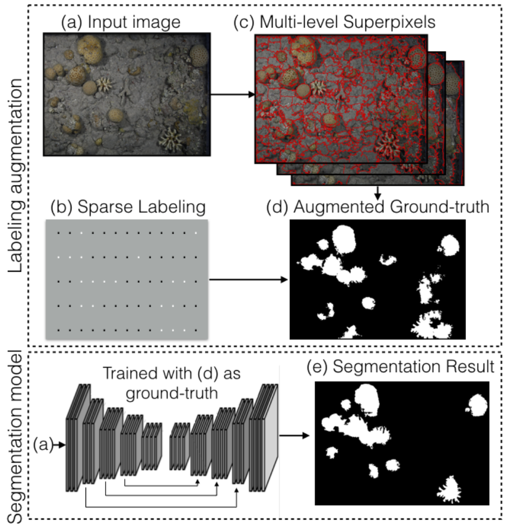

Fig. 2. Sparse labeling augmentation to train dense semantic segmentation models. Given an image (a) with sparse labels (b), the multi-level superpixel propagation proposed (c) obtains an augmented ground-truth (d). This augmented labeling is shown to be very effective to train a fully convolutional neural network for semantic segmentation (e).

III. Proposed Approach

This section describes the pipeline proposed to overcome the challenges discussed in previous sections. Fig. 2 represents the main ingredients from the presented approach: end-to-end CNN model for semantic segmentation trained with our augmented labeling.

A. Semantic Segmentation Formulations

We consider two common formulations of the semantic segmentation from sparse annotations problem: *classification of small image patches*, to combine them to obtain the final image segmentation, and *per pixel classification*, to directly obtain the image semantic segmentation.

1) Per patch classification: Semantic segmentation can be formulated as a patch classification problem. If we have a set of labeled pixels as ground truth, we can train a classification CNN on patches cropped around those labeled pixels and obtain the final image segmentation combining the classification result for each patch. This strategy, which has been successfully applied in existing approaches [6], is trained on n labeled patches, one per labeled pixel in the training images. The training pairs used are of the form:

$$(\mathbf{X}_{(i,j)},\ y_{(i,j)}),$$

where $\mathbf{X}_{(i,j)}$ is a patch of dimensions $d \times d$ centered around each a labeled pixel with coordinates (i, j), and $y_{(i,j)}$ is a scalar representing the label of this pixel.

2) Per pixel classification: More frequently, semantic segmentation is formulated as a pixel classification problem. In this case, an end-to-end CNN architecture is trained from dense input ground truth labels to obtain directly the classification for each pixel, i.e., the final semantic segmentation. We consider the most common fully convolutional architectures for this problem: the FCN architecture [22] and the symmetric encoder-decoder [3]. In both architectures, the network is trained with pairs of images:

$$(\mathbf{X},\ \mathbf{Y}'),$$

where $\mathbf{X}$ is the original input image, a $m \times n \times 3$ array, and $\mathbf{Y}'$ is a $m \times n$ array with a label for each pixel.

Both approaches are a classification problem, whose model is obtained by minimizing the error between predicted and expected value $(\min(|\hat{y} - y|))$ for the corresponding pairs of training input. Both strategies are trained using the common cross entropy loss function described in (1).

$$\mathcal{L} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{c=1}^{M} y_{c,j}\ln(\hat{y}_{c,j}), \tag{1}$$

N is the number of labeled pixels and *M* is the number of classes. $Y^{(i)}$ is a binary indicator (0 or 1) of belonging to a certain class c for pixel j and $\hat{y}^{(i)}$ is the CNN predicted probability of belonging to a certain class c for pixel j. In the per pixel approach each i represents a pixel, while in the per patch approach each i represents a patch, so $N = 1$ since we only have one label per patch.

We have built the three architectures (patch classification and two architectures for end-to-end semantic segmentation) on top of the same base CNN model, DenseNet [16]. In particular, for the patch-classification architecture we used DenseNet-169 with k=24, the FCN architecture uses the same classification architecture (DenseNet-169) combined with an up-sampling layer and the symmetric encoder-decoder uses the FC-Densenet103 architecture [17]. Sec. IV-B discusses the results obtained with our trained models of these three alternative architectures, both trained from scratch and exploring some finetuning options.

B. Labeling augmentation with multi-level superpixels

This section describes the proposed strategy for sparse label augmentation[1]. The goal is not to the propagation itself, but to augment our available training data to boost the training of a CNN for semantic segmentation.

a) Superpixel based label propagation: Our strategy for label augmentation is based on existing superpixel segmentation techniques. These techniques cluster image pixels into groups of similar connected pixels (named superpixels).

The basic *single-level* superpixel based augmentation strategy has two steps. First, the image is segmented into superpixels, as shown in the examples in Fig. 3. Then, the labeled pixels information is propagated following the superpixel segmentation, i.e., all pixels in each superpixel get the label that appears the most within that superpixel. Fig. 4 shows some examples using different superpixel segmentation algorithms: CRS [8], PB [36], ERS [21], SLIC [1] and SEEDS [33]. Section IV-C compares the effectiveness of different existing superpixel segmentation techniques when applied to the proposed label augmentation.

[1] Code at https://github.com/Shathe/ML-Superpixels

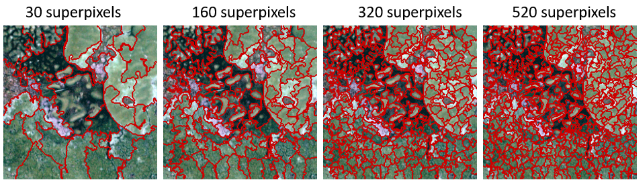

Fig. 3. Superpixel segmentation obtained varying the number of superpixels (clusters) to get using SEEDs technique.

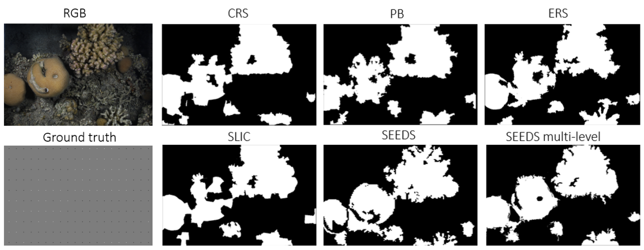

Fig. 4. Sparse ground truth labeling augmentation obtained with different superpixel segmentation techniques. The top-left view is the original image and the bottom left view is the sparse available ground truth. The rest are binary (coral/no-coral) labeling augmentations.

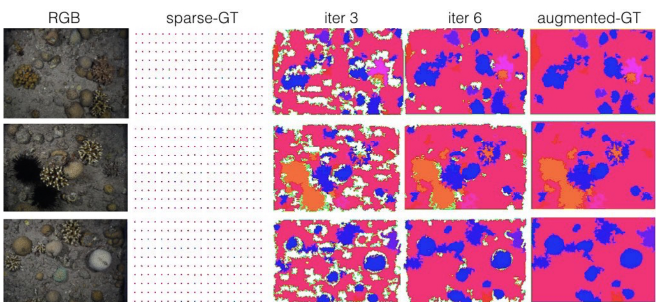

Fig. 5. Multi-level superpixel augmentation algorithm. From left to right: input image, available labels (sparse GT), augmented labels after 3 and 6 iterations, and final augmented labels (augmentedGT) after 10 iterations.

This basic single-level superpixel approach, used in prior work with promising results [2], has some drawbacks:

- The number of used superpixels is fixed.
- Some superpixels may not have labeled pixels inside, thus, they will generate unlabeled regions.

This generates a strong trade-off between proper contour fit and amount of unlabeled regions: higher number of superpixels gives better results and fits better the actual shapes, but it increases the number of superpixels that turn out unlabeled. The multi-level superpixel extension described next solves these issues.

b) Multi-level Superpixel Segmentation: The multi-level superpixel segmentation proposed (see Algorithm 1) consists of applying iteratively the superpixel image segmentation, progressively decreasing the number of superpixels generated in each iteration. In the first iteration the number of superpixels is very high, leaving a lot of unlabeled pixels in the augmented labeling but being able to find small regions. The following iterations continue increasing the superpixel size until they get to fill all the unlabeled pixels (see Fig. 5).

Algorithm 1: Propagation with Multi-level Superpixel Segmentation

```
function MLsuperpixels (SparseGT, img)
  Input  : RGB image (img) and corresponding sparse
           ground-truth labeling (SparseGT)
  Output : Augmented ground-truth
nSuperpixels = getHighNumber();
augmentedGT = blankImage();
while augmentedGT.hasUnlabeledPixels() do
  | sp = getSuperpixels(img, nSuperpixels);
  | aug = getAugmentedLabels(SparseGT, sp);
  | augmentedGT = mask(augmentedGT, aug);
  | nSuperpixels = decreased(nSuperpixels);
end
return augmentedGT;
```

IV. Experiments

Experiments in this section evaluate different aspects of the proposed approach to augment sparse input labels for semantic segmentation model training. They demonstrate the effectiveness of the proposed strategy compared to existing ones with more costly labeling requirements.

A. Evaluation

This section details the datasets and the evaluation metrics used in the following experiments.

1) Datasets: We use four different datasets in our experiments, from four different domains. The first one, Eilat Fluorescence Corals dataset [6], is the main dataset used in our experiments, because it is a real use case of data published with only sparse labels available. The other datasets used allow us to demonstrate the generalization of the proposed methods to different domains. Besides, since they have dense annotations, we can do a direct comparison of training results with augmented labels vs. original dense labels.

- **Eilat** [6] has 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.
- **Camvid** [7] is an autonomous driving dataset with 11 different classes, frequently used to train existing state-of-the-art approaches for urban areas image segmentation models.
- **RIT** [18] is an aerial imagery dataset with multi-espectral data from 18 classes. RIT does not provide test images labeling, so we evaluate its results by separating part of the evaluation set they provide.
- **Pascal VOC 2012** [10] is a well known general purpose dataset with 20 different classes.

2) Ground-truth and annotations considered: We use several types of annotations, or ground truth labels, to evaluate the results of the segmentation models obtained.

The **Eilat dataset** is evaluated with metrics computed with respect to three different reference annotations:

- *Original-GT*: original sparse ground-truth labels available with the dataset.

TABLE I

BINARY (CORAL VS NO-CORAL) CLASSIFICATION WITH DIFFERENT SEMANTIC SEGMENTATION APPROACHES

	Metrics		
Model	PA	MPA	MIoU
Evaluation based on **Manual annotation**			
Patch classification*	74.40	54.36	43.66
FCN	92.19	81.78	73.51
Symmetric encoder-decoder	**94.02**	**85.10**	**79.02**
Evaluation based on **Augmented-GT**			
Patch classification*	89.11	76.11	64.58
FCN	90.33	70.83	63.34
Symmetric encoder-decoder	**92.32**	**82.07**	**73.01**
Evaluation based on **Original-GT**			
Patch classification*	**93.75**	**90.00**	**85.06**
FCN	81.01	71.87	60.27
Symmetric encoder-decoder	89.18	86.78	76.79

* Our implementation of Beijbom et al. [6].

- *Augmented-GT*: augmented ground-truth obtained by our approach.
- *Manual annotations*: a few manual annotated images for binary (coral vs non-coral) segmentation obtained by a marine biologist.

The *Original-GT* is the least representative and reliable of the three ground-truth options, since it has very few annotations per image, but it is necessary to perform direct comparisons with previous results using it. The *Augmented-GT* is an approximated labeling and, as we measure next, contains some noise (94% accuracy against the obtained manual annotations). However it provides a very representative reference labeling [2], as we discuss next. The *Manual annotations* are the most reliable and representative to compare, but we do not have them available for all images.

The **other three datasets** are evaluated with their available dense labeling ground truth (*Manual annotations*). The sparse labels of these datasets are obtained automatically by sampling the dense labeling following a grid (random distributions leads to similar augmentation results). Note that small objects or instances may not get any ground truth label in the simulated sparse ground-truth, which corresponds to only a 0.1% of the dense ground-truth (e.g., from a $500x500$ image, the simulated sparse ground-truth will contain just 250 labeled pixels).

3) Metrics: standard metrics for semantic segmentation are used: **PA** (Pixel Accuracy), **MPA** (Mean Pixel Accuracy) and the **MIoU** (Mean Intersection over Union).

B. Semantic Segmentation Architectures Evaluation

This first experiment is intended to evaluate and compare the results of common CNN based architectures when trained from the sparse available labels.

1) Experiment setup: This experiment is run with the Eilat dataset, and considers the three common architectures for semantic segmentation detailed in Sec. III-A. For the *patch classification* we set the dimension of the patches around each labeled pixel to $d = 50$ pixels. To train the two end-to-end semantic segmentation architectures, the ground-truth augmentation used is the basic version, as in [2] to allow direct comparisons with those results.

2) Results training from scratch: For the three architectures, a model is trained from scratch, using image augmentation (horizontal and vertical flips and crops of the data) and the same training set up (all of them converge): 500 epochs, initial learning rate of 0.001 and exponential learning rate decay of 0.99. Table I shows a summary of the performance of the three models obtained. It includes results using the original sparse labels (Original-GT) for completion, but as previously discussed [2], those only evaluate a few pixels, while the other metrics compute a score considering all pixels and therefore are more significant.

The end-to-end semantic segmentation approaches yield to better results than the patch-classification method. Then, in the rest of the experiments, results are shown only with the best performing architecture, the symmetric encoder-decoder based on FC-Densenet103.

3) Results with Fine-tuning: We have explored the benefits of finetuning existing models on the target data, since it is a common strategy to improve results when there is not a lot of labeled data available for training.

Finetuning experiments were carried out on both 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 obtain a base model, which we finetune later on the Eilat data. The first one, is another dataset from the CoralNet project [5], Moorea[2] dataset, and belongs to a similar domain to Eilat (corals). Therefore, this is expected to work better for finetuning. The second one, Camvid dataset, is from a different domain, urban scenes. For both cases, finetuning for the Eilat data converged 2.5 times earlier to a similar quality model than training from scratch. As expected, pre-training on the similar domain of Moorea data allows finetuning for Eilat with less amount of images, while obtaining the same results. Therefore, finetuning saves training time and labeled data requirements, but does not improve the quality of the obtained models.

C. Labeling augmentation quality evaluation

The experiments in this section show the advantages of using the presented multi-level superpixels based labeling augmentation with respect to other augmentation approaches in different image modalities and domains.

1) Experiment setup: For all the following experiments, the multi-level superpixel based augmentation starts with the number of superpixels set to 1500, and decrease it to the 80% in each iteration. This augmentation costs an average of 15 seconds per image on a 500x500 resolution (1 second per superpixel level). To evaluate the quality of the obtained label augmentation, we compare it with the actual ground truth segmentation of the test images.

[2]https://www.bco-dmo.org/dataset/676105

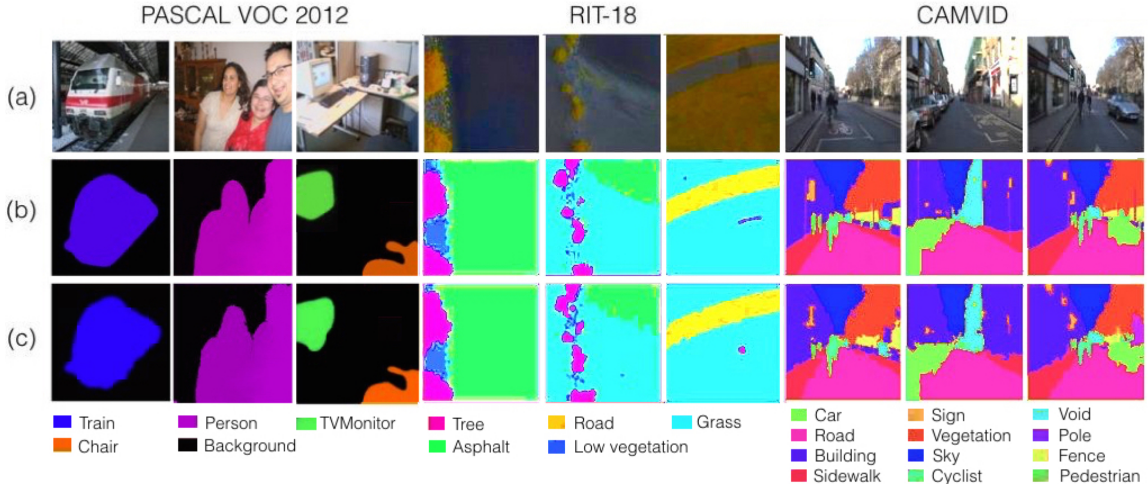

Fig. 6. Examples for labeling augmentation evaluation with different datasets. (a) input images, (b) original dense labeling, (c) augmented labeling recovered from just a 0.1% of the original labeled pixels.

TABLE II
LABELING AUGMENTATION ON RGB AND FLUORESCENCE IMAGES

Augmentation Approach	**Metrics**		
	PA	MPA	MIoU
Using fluorescence	Evaluation based on **Manual annotation**		
SEEDS single-level	93.38	86.86	77.86
SEEDS multi-level	**94.20**	**87.50**	**79.88**
SLIC multi-level	93.86	85.37	78.37
Using RGB	Evaluation based on **Manual annotation**		
SEEDS single-level	92.21	80.20	72.90
SEEDS multi-level	93.23	84.91	75.37
SLIC multi-level	92.76	83.60	75.37

TABLE III
SEEDS MULTI-LEVEL LABELING AUGMENTATION COMPARED TO ORIGINAL DENSE MANUAL ANNOTATIONS ON DIFFERENT DOMAINS

Datasets	**Metrics**		
	PA	MPA	MIoU
	Evaluation based on **Manual annotation**		
Camvid	91.95	76.91	65.05
RIT	97.44	72.31	59.18
VOC 2012	96.87	95.77	93.31

2) Multi-level vs single-level: First, we compare the quality of the labeling augmentation with the multi-level approach with respect to the basic version, single-level, used in earlier results [2]. Table II shows the augmentation labeling results of the baseline (single-level superpixel augmentation) and the two best superpixel segmentation results using our multi-level approach. The rest of the superpixel segmentation methods we experimented on (PB, ERS, CRS) got around 1% less in all the metrics with respect to SLIC. We evaluate the results with the two image modalities, RGB and fluorescence, available on the Eilat dataset.

The multi-level superpixels augmentation outperforms the baseline by 1-3% in all the metrics. SEEDS superpixels work slightly better because they fit better to contours, therefore SEEDs is used for the rest of the experiments.

TABLE IV
LABELING AUGMENTATION USING PASCAL VOC 2012 DATASET

Augmentation from traces	**MIoU**
Vernaza et al. (SPCON)[34]	76.50
Vernaza et al. (RAWKS v1) [34]	75.80
Vernaza et al. (RAWKS v2)[34]	81.20
Augmentation from sparse pixel labels	**MIoU**
Alonso et al. from 0.1% of pixels (300 pixels) [2]	86.36
Multi-level approach from 0.01% of pixels (30 pixels)	74.40
Multi-level approach from **0.1%** of pixels (300 pixels)	**93.31**
Multi-level approach from 1% of pixels (3000 pixels)	**97.25**

3) Label augmentation in different domains data: This experiment evaluates the proposed label propagation method on different domains data: Camvid, RIT and VOC (described in Sec. IV-A.1). We compare the original dense labeling available on each dataset and the results from applying our approach to augment a *simulated* sparse labeling. Fig. 6 shows qualitative results of these experiments, and Table III summarizes the quantitative comparison of the augmented labeling with the original dense labeling, with very good results in the three different domains. The proposed augmentation methods performs a propagation of existing sparse labels, therefore it needs to have at least one labeled pixel per object or instance. The sparse labeling simulation (sampling) can miss samples from very small instances. So the RIT dataset, which is the one with more small details (see Fig. 6), gets the lowest scores.

Table IV compares our approach, *Multi-level*, using different sparsity as input for the propagation, with other recent label augmentation or propagation methods using the PASCAL VOC 2012 dataset. Vernaza et al. [34] uses traces as the input of the augmentation process as well as the learned boundaries (learnt by a neural network) using the RAWKS algorithm (v1) to augment the traces sparse labeling. V2 means the evaluation is done on the 94% of the pixels, where the model is confident enough. Alonso et al. [2], our baseline version, uses the same grid structure of sparse pixels as our

TABLE V

RESULTS TRAINING ON ORIGINAL DENSE MANUAL ANNOTATIONS (REAL) AND OUR SIMULATED GROUND-TRUTH (AUGMENTED)

	Metrics		
	PA	MPA	MIoU
Datasets	Evaluation based on **Manual annotation**		
Camvid (real)	**88.68**	48.81	44.36
Camvid (augmented)	87.70	46.97	42.95
RIT (real)	**94.23**	20.36	19.16
RIT (augmented)	89.30	19.65	17.85

TABLE VI

MULTICLASS SEMANTIC SEGMENTATION RESULTS ON EILAT DATASET

	Metrics		
	PA	MPA	MIoU
Evaluation on dense scores: **Augmented-GT**			
Beijbom et al.[6] v1	—	—	—
Beijbom et al.[6] v2	73.61	25.32	17.89
Alonso et al.[2]	85.88	42.25	31.12
Ours v1	90.96	51.28	39.44
Ours v2	**91.68**	**52.76**	**42.22**
Evaluation on sparse scores: **Original-GT**			
Beijbom et al.[6] v1	87.80	48.50	—
Beijbom et al.[6] v2	**90.20**	53.10	43.66
Alonso et al. [2]	81.23	41.97	28.14
Ours v1	84.96	56.96	42.94
Ours v2	84.54	**59.26**	**44.10**

multi-level superpixels augmentation.

The augmented labeling obtained with our approach is very close to the original ground truth, as shown in the different results in this section. The presented multi-level superpixel augmentation outperforms related methods for labeling augmentation with different levels of label sparsity.

D. Performance of models trained with augmented labeling

This last group of experiments demonstrates that training state-of-the-art CNNs with the proposed augmented labeling gets similar results than with the original dense labeling.

1) Experiment Setup: For these experiments we use the augmented labeling generated from the 0.1% of the dense labeling, simulated as explained in Sec. IV-A.2. We use the Eilat, RIT-18 and Camvid datasets for evaluation. For training we use the hyper-parameters described in Jegou et al. [17]. This configuration applies to the training with the augmented and original dense labeling. We also use image augmentation (horizontal and vertical flips, data crops).

2) Training results: Augmented vs real: We compare the quality of the segmentation obtained from a model trained on the original dense segmentation ground-truth or from a model trained on the augmented ground-truth using multi-level superpixels. Table V shows a summary of the results with Camvid and RIT, which provide ground truth segmentation.

The results obtained training with our augmented ground-truth are comparable to training with original ground-truth. This could be expected since we already validated that the augmented labeling is very close to the original one. The differences between the original and augmented labeling can be seen as noisy labels, and neural networks have been shown to be capable to learn from noisy data [30]. Fig. 7 shows some visual examples of this comparison.

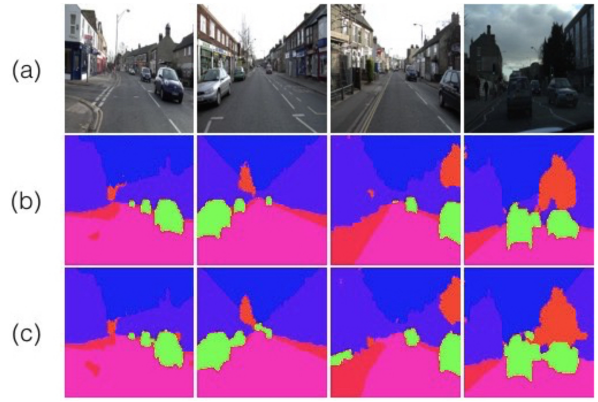

Fig. 7. Semantic segmentation on Camvid. (a) original images, (b) results using a model trained on original dense labeling, (c) results using a model trained with our proposed augmented labeling.

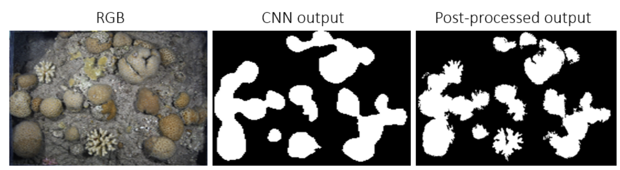

Fig. 8. A comparison between the result of our pipeline using the multi-level Superpixels for labeling augmentation and the same output applying SEEDS superpixels to enhance it.

Eilat dataset does not have original dense labeling for all images, therefore we do a separated evaluation for it. We compare our approach with prior work published by the authors of the dataset for multi-class segmentation (10 classes). Table VI summarizes these results. We compare the original results in [6] (v1) and our implementation of it with a newer base model (v2). This method consists of a patch-classification approach. Note that (v2) does perform equal or better than original (v1), and that (v1) is shown only where original publication included results. Results also include our previous work baseline with the *single-level* label augmentation [2], and two versions of the work presented here (*Ours v1* and *Ours v2*). *Ours v1* consists of training the FC-DenseNet103 with our multi-level superpixels labeling augmentation. *Ours v2* applies an additional SEEDS superpixel post-processing to refine the *Ours v1* segmentation result (see Fig. 8). We show the original-GT scores because some related work only has published results using this. However, note how the proposed method significantly outperforms previous work on the more significant dense scores. These results also point that a final superpixel based smoothing can help to enhance the final result.

V. CONCLUSION

We have presented a novel approach to augment labeled data, at pixel level, to facilitate training semantic segmentation models. As shown, our approach enables the training of state of the art architectures for semantic segmentation in scenarios where there are only sparse labels available. More generally, it also benefits any scenario by lowering the labeling requirements to train new models on new domain datasets which still need to be labeled.

Our experiments analyze different aspects of the proposed approach: different superpixel segmentation techniques, relevant architectures for semantic segmentation, and the influence of different density in the available labels. We have demonstrated that the proposed augmented labeling is effective to train CNN models for segmentation, reaching comparable results to those obtained by training with dense ground-truth labels, much more costly to obtain. The use of grid-based sparse data was motivated because it is actually available in real world use cases. As future steps, we plan to explore different types of sparsity distributions and extend the applicability to other data types like 3D information.

ACKNOWLEDGEMENTS

The authors would like to thank A. Cambra, A. Muñoz and T. Treibitz for their help in earlier versions of this work, and NVIDIA Corporation for the donation of a Titan Xp GPU used in this work. This research has been partially funded by the Spanish Government project DPI2015-69376-R, UZCUD2017-TEC-06 and Aragón regional government (Grupo DGA T45_17R).

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 Trans. on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[2] I. Alonso, A. Cambra, A. Munoz, T. Treibitz, and A. C. Murillo. Coral-segmentation: Training dense labeling models with sparse ground truth. In *IEEE Int. Conf. on Computer Vision Workshops*, pages 2874–2882, 2017.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2017.

[4] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani. Deepsat: a learning framework for satellite imagery. In *Proc. of the 23rd SIGSPATIAL Conf.*, page 37. ACM, 2015.

[5] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman. Automated annotation of coral reef survey images. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[6] 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.

[7] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[8] C.Christian, M.Mertz, and R.Mester. Contour-relaxed superpixels. In *Computer Vision and Pattern Recognition Workshops*, pages 280–293, 2013.

[9] T. Durand, T. Mordan, N. Thome, and M. Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017.

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *Int. journal of computer vision*, 88(2):303–338, 2010.

[11] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017.

[12] A. Gupta, A. Vedaldi, and A. Zisserman. Synthetic data for text localisation in natural images. In *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2016.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *IEEE Int. Conf. on Computer Vision*, pages 2980–2988, 2017.

[14] J. C. Hodgson, S. M. Baylis, R. Mott, A. Herrod, and R. H. Clarke. Precision wildlife monitoring using unmanned aerial vehicles. *Scientific reports*, 6:22574, 2016.

[15] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick. Learning to segment every thing. *arXiv preprint arXiv:1711.10370*, 2017.

[16] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017.

[17] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *CVPRW*, pages 1175–1183. IEEE, 2017.

[18] R. Kemker, C. Salvaggio, and C. Kanan. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018.

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

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.

[21] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. Entropy rate superpixel segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011.

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

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

[24] P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun. Predicting deeper into the future of semantic segmentation. In *Int. Conf. on Computer Vision*, page 10, 2017.

[25] T. Manderson, J. Li, N. Dudek, D. Meger, and G. Dudek. Robotic coral reef health assessment using automated image analysis. *Journal of Field Robotics*, 34(1):170–187, 2017.

[26] B. Mičušík and J. Košecká. Multi-view superpixel stereo in urban environments. *Int. journal of computer vision*, pages 106–119, 2010.

[27] 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 *Conf. on Computer Vision and Pattern Recognition*, 2016.

[28] F. Shkurti, W. Chang, P. Henderson, M. Islam, J. Gamboa Higuera, J. Li, T. Manderson, A. Xu, G. Dudek, and J. Sattar. Underwater multi-robot convoying using visual tracking by detection. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2017.

[29] D. Stutz, A. Hermans, and B. Leibe. Superpixels: an evaluation of the state-of-the-art. *Computer Vision and Image Understanding*, 166:1–27, 2018.

[30] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE Int. Conf. on Computer Vision*, 2017.

[31] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *E. Conf. on Computer Vision*, 2010.

[32] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger. Sparsity invariant CNNs. In *IEEE Int. Conf. on 3D Vision*, pages 11–20, 2017.

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

[34] P. Vernaza and M. Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2017.

[35] J. M. Wong, V. Kee, T. Le, S. Wagner, G.-L. Mariottini, A. Schneider, L. Hamilton, R. Chipalkatty, M. Hebert, D. Johnson, et al. Segicp: Integrated deep semantic segmentation and pose estimation. *Int. Conf. on Intelligent Robots and Systems*, 2017.

[36] Y. Zhang, R. Hartley, J. Mashford, and S. Burn. Superpixels via pseudo-boolean optimization. In *IEEE Int. Conf. on Computer Vision*, pages 1387–1394, 2011.

[37] 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.