Interactive Learning for Semantic Segmentation in Earth Observation

Gaston Lenczner[1,2,*], Adrien Chan-Hon-Tong[1], Nicola Luminari[2], Bertrand Le Saux[3], and Guy Le Besnerais[1]

[1] ONERA/DTIS, Université Paris-Saclay, F-91123 Palaiseau, France
name.surname@onera.fr
[2] Delair, FR-31400 Toulouse, France
name.surname@delair.aero
[3] ESA/ESRIN, Φ-Lab, I-00044 Frascati (RM), Italy
bls@ieee.com

Abstract. Dense pixel-wise classification maps output by deep neural networks are of extreme importance for scene understanding. However, these maps are often partially inaccurate due to a variety of possible factors. Therefore, we propose to interactively refine them within a framework named DISCA (Deep Image Segmentation with Continual Adaptation). It consists of continually adapting a neural network to a target image using an interactive learning process with sparse user annotations as ground-truth. We show through experiments on three datasets using synthesized annotations the benefits of the approach, reaching an IoU improvement up to 4.7% for ten sampled clicks. Finally, we exhibit that our approach can be particularly rewarding when it is faced to additional issues such as domain adaptation.

Keywords: Semantic Segmentation, Interactive Learning, Continual Adaptation

1 Introduction

As massive amounts of data are produced everyday coming from multiple sources such as drones or satellites [20], Earth Observation (EO) data plays a central role in the way we understand our planet. Semantic segmentation, the task of classifying an image pixel-wise, is of primary importance in EO for various purposes such as land-cover classification and is now efficiently addressed by deep neural networks. However, there is no theoretical guarantee of the performances of these networks and they indeed may fail in practice. Besides, they often get worse when they are faced to additional constraints such as limited training database [15], flawed labels [8] or domain shifts [14] (different weather conditions, different geographical positions, different seasons, ...). In many situations, datasets deal with several, if not all, of these issues. Therefore, mistake-free neural networks are still an utopia in such scenarios.

*Corresponding author gaston.lenczner@delair.aero

Depending on the applications, these errors can be controlled or tolerable and data can thus be processed fully automatically at large scale. However, they can also be unacceptable and data processing then needs to be controlled by a human operator who can certify the results in a semi-automatic way [11]. In practice, the user can work hand-in-hand with machine learning models which have been previously trained on large amounts of data. Indeed, as shown by DIOS [21] or DISIR [12], a user can guide a neural network to perform segmentation tasks with clicked annotations given as inputs to the algorithm. Since these approaches do not modify the weights of the networks, their guidance is spatially localized around the annotations. Therefore, we propose in the present paper Deep Image Segmentation with Continual Adaptation (DISCA), a semantic segmentation framework to interactively retrain a neural network to enhance its performances on EO images at an image level using the annotations as a sparse reference. Interestingly, remote sensing seems to be a more relevant setting than natural images for this kind of interactive learning. Indeed, there can be a large variability within a natural image which makes it hard to exploit information provided by an annotation. Inversely, within a remote sensing image, variability tends to be low while the image tends to be large, allowing to take full advantage of the information provided by the user.

Our present contribution is the following. We introduce DISCA, assert its efficiency on three remote sensing datasets and also show that it can be particularly relevant in a domain shift context.

2 State of the art

Weakly supervised learning aims at making algorithms learn on data with flawed or partial labels. In semantic segmentation, these labels can then take many forms such as single points [2,4], bounding-boxes [7] or scribbles [13]. Weak supervision has begun to receive a growing attention in the remote sensing community as labels can be costly to acquire at large scale. [3] proposed an iterative training process based on weak supervision for the task of change detection. Our work is closely related to weak supervised learning as the user annotation maps can be seen as sparse ground-truth maps. Differently than previous works, we only fine-tune in a weakly-supervised fashion the networks otherwise standardly trained.

Continual learning defines the ability of a learning algorithm to continuously learn from a stream of data [16]. It has been approached in remote sensing for semantic segmentation at large scale with new labels appearing over time [19]. Continual learning often implies learning through multiple tasks. This applies to our work if we consider standard semantic segmentation as the first task and then learning from the target image and the annotations as the second one. We also face a challenge inherent to reinforcement learning [9] where an agent (our network) interacts with a dynamic environment (the user annotations). Indeed, we must pay particular attention to the risk of classifier degradation.

Interactive segmentation intends to interactively segment an image into foreground and background pixels with user annotations. It was initially addressed using graph-cut based methods [17] and now mostly by deep neural networks which take as inputs a concatenation of the RGB image and user annotations [21]. In [10], the authors use the annotations as sparse ground truth maps to interactively adapt the neural network to a specific object. Multi-class interactive segmentation broadens interactive segmentation to correct multi-class segmentation maps. [1] proposed a neural network which takes as input a concatenation of the image and the extreme points of each instance in the scene and then lets a user correct the proposed multi-class segmentation using scribbles. We do not assume such extreme point map availability as it is costly to acquire in a remote sensing image with many potential objects. In our previous work [12], we extended [21] to interactively refine remote sensing segmentation maps. We now draw inspiration from [10] to the same purpose.

3 Methodology

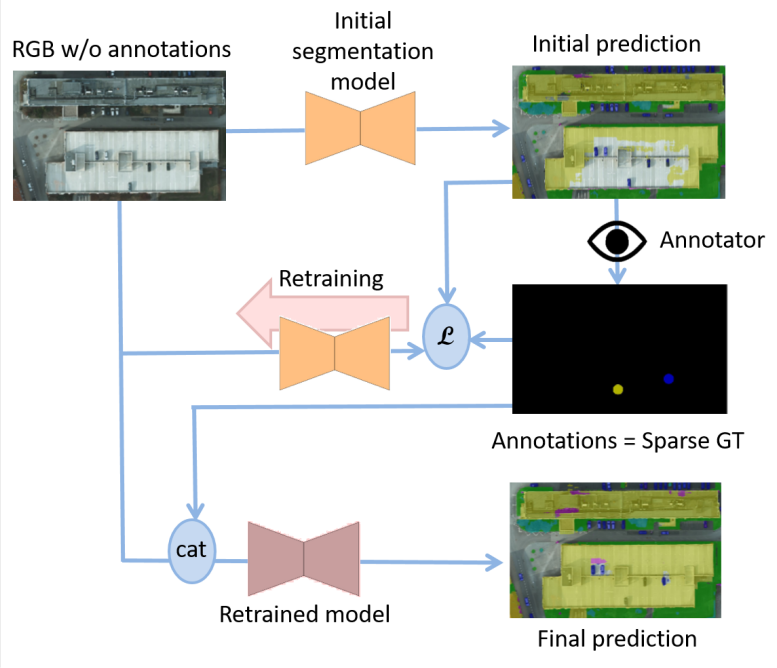

Fig. 1. Visual abstract of DISCA method. Up: initial prediction, middle: retraining using annotations as a sparse ground truth, bottom: final prediction using the retrained model in a DISIR [12] mode.

We rely on DISIR[§] [12] since it is, to the best of our knowledge, the only existing open-source work addressing multi-class interactive segmentation using deep learning in remote sensing. Hence, our network takes as input a concatenation of the RGB image and of the user annotations. Initially under the form of clicked points, these annotations are encoded using distance transforms into a N-dimensional tensor where N is the cardinal of the label space. Sampled randomly from the ground-truth during an initial classic training phase, they are then used by the neural network as guidance to enhance its initial predictions. Since only the network's inputs are modified and not its parameters, the information provided by the annotations does not improve the predictions in the entire image. We propose to bypass this locality constraint by retraining the network with a few back-propagation cycles per annotation.

To fully benefit from the annotations at the image scale, as shown in Figure 1, we use them as a sparse ground-truth to interactively retrain the network using a cross entropy loss on these annotated pixels. We note $\mathbf{f}$ to represent the neural network parameterized by θ and $\mathbf{x}$ its inputs. As only a few pixels are annotated among the millions ones that usually compose a remote sensing image, the ground-truth maps are extremely sparse. In order to deal with this problem and avoid over-fitting, we follow [10] by using the initial prediction $\mathbf{p_0} = f(\mathbf{x}, \theta_0)$ for regularization. This regularization consists of adding a term of cross-entropy loss using the original prediction as ground-truth in order to prevent the model from making a prediction too different from the initial one. Differently from the source paper, we use a regularization function based on a L^1 loss instead of a more permissive cross entropy loss and we do not add an additional regularization over the network parameters. Therefore, our loss during the interactive learning process is defined as follows:

$$\mathcal{L}(\mathbf{x}, \mathbf{c}, \mathbf{p_0}; \theta) = \frac{\mathbf{1}_{[\mathbf{c}=-\mathbf{1}]}}{\|\mathbf{1}_{[\mathbf{c}=-\mathbf{1}]}\|_1} \left\{ -\sum_{i=1}^{N} \mathbf{c}_i \log\left(\mathbf{f}_i(\mathbf{x}; \theta)\right) \right\} + \|\mathbf{f}(\mathbf{x}; \theta) - \mathbf{p_0}\|_1 \qquad (1)$$

where $\mathbf{1}$ represents the indicator function and $\mathbf{c}$ the sparse annotated pixels. In details, $\mathbf{c}$ takes its values in {-1, 0, 1}. For the pixels annotated as belonging to class i, $\mathbf{c}_i = 1$ and $\mathbf{c}_j = 0$ for all $j \neq i$. For the unannotated pixels, $\mathbf{c}_i = -1$ for all i in $\{1, \ldots, N\}$.

The DISIR mechanism is only used in the last inference. Indeed, the fine-tuning DISCA mechanism works on the image only. In other words, during the interactive training process, the annotations are not concatenated with the RGB image at the input of the network. This prevents the network from over-fitting on them. Fortunately, this does not make the network forget how to use the annotations for guidance.

[§] https://github.com/delair-ai/DISIR

Table 1. Mean IoU obtained before and after the interactive processes

	DISIR	DISCA
Before	70.6	
After	71.3	**72.2**

(a) ISPRS

	DISIR	DISCA
Before	85.4	
After	86.4	**86.5**

(b) INRIA

	DISIR	DISCA
Before	85.9	
After	89.5	**90.6**

(c) AIRS

4 Experiments

4.1 Experimental setup

In this section, we study through experiments the scope of our approach and compare it to [12]. We experiment on three semantic segmentation remote sensing datasets: the INRIA Aerial Image Labelling dataset [14] composed of two classes (*buildings* and *not buildings*) and covering more than 800 km^2 in different cities at a 30cm resolution, the Aerial Imagery for Roof Segmentation (AIRS) dataset [6] composed of the same two classes and covering 457 km^2 in New-Zealand at a 7.5cm resolution and the ISPRS Potsdam dataset [18] composed of 6 classes (*impervious surface, buildings, low vegetation, tree, car* and *clutter*) covering 3 km^2 on Potsdam at a 5cm resolution. The initial training sets are divided into a smaller training set and a validation set with a ratio 80%-20%. This allows to synthesise annotations to evaluate the framework. The images are tiled into patches of size 512×512 with an overlap of size 128 to be processed.

We use a neural network based on the LinkNet [5] architecture - but the approach is independent from the neural network backbone. To evaluate the refinement performances on the validation sets, we sample 10 annotations from the ground-truth maps in the largest wrongly predicted areas, adapt the networks in an image-wise fashion and measure the Intersection over Union (IoU) evolution. During the interactive learning phase, we optimize the weights using 10 stochastic gradient descent passes with a learning rate of $2e^{-7}$ and minimize the loss defined in Eq. 1.

4.2 Single datasets experiments

As shown in Table 1, DISCA successfully enhances the initial segmentation maps to reach a higher IoU. Indeed, we observe an average improvement of 2.5% IoU with ten annotation samples. Besides, it also beats DISIR performances on the three datasets.

DISCA efficiently allows the user to make corrections at the image scale: on Figure 2, single annotations enable DISCA to provide a corrected segmentation of the scenes while they are not enough for DISIR to deliver a similar result.

4.3 Domain adaptation experiment

Fig. 2. Results from the ISPRS dataset with single annotations. Best viewed in color.

Aiming to push our approach one step further, we evaluated DISCA on ISPRS with a network trained on AIRS to simulate a domain adaptation scenario. The ISPRS labels were simplified to *building* and *not building* and the images were down-sampled using bi-linear interpolation to match the AIRS resolution. For comparison, we also trained a network to detect buildings only on the ISPRS training set and used it as a control experiment. As shown on Figure 3, even though the performances do not reach the ones from the control experiment, the network within DISCA framework still drastically benefits from the 10 annotations to improve the initially flawed segmentation maps. Indeed, there is a 20% average IoU improvement with DISCA

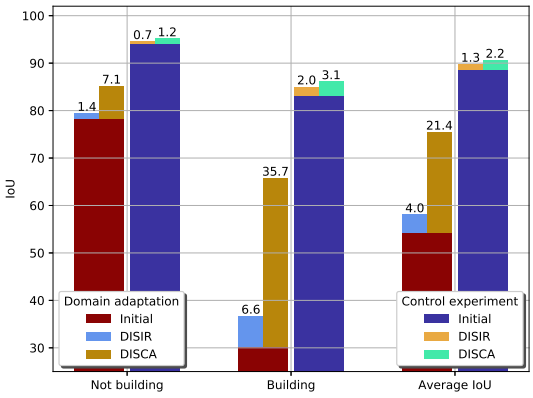

Fig. 3. Domain adaptation (AIRS→ISPRS)

while it is only around 5% with DISIR. This metric improvement is also visually confirmed on the results displayed on Figure 4 where the segmentation maps are globally well enhanced using DISCA with only ten annotation samples. In particular, the network is then able to adapt to buildings with peculiar roofs or of uncommon size with respect to the AIRS dataset. These outcomes suggest that DISCA is specifically adapted for global enhancements when the neural network is uncertain about the initial prediction.

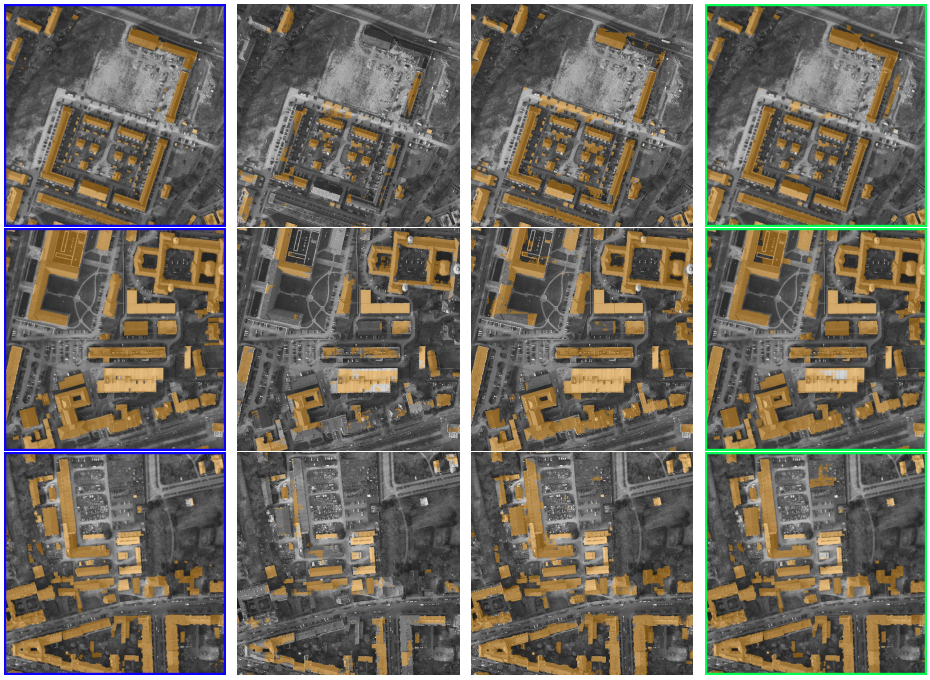

Fig. 4. Building segmentation from the ISPRS validation dataset with a network pretrained on AIRS. The control experiment corresponds to initial predictions made with a network pretrained on ISPRS. Best viewed in color.

5 Conclusion

We have proposed in this paper an interactive learning strategy to incrementally refine segmentation maps and applied it to EO data. It makes use of user annotations as ground-truth to continuously adapt a neural network to a target image. Finally, we have showed on three datasets the benefits of our approach

and that it can be especially adapted to enhance mitigated initial results when dealing with domain adaptation. In the future, we intend to extend our approach to scarce training data and to explore reinforcement policies in order to leverage information provided by the user even better.

References

1. Agustsson, E., et al.: Interactive full image segmentation by considering all regions jointly. In: CVPR. pp. 11622–11631. IEEE (2019)
2. Bearman, A., et al.: What's the Point: Semantic segmentation with point supervision. In: ECCV. pp. 549–565. Springer (2016)
3. Caye Daudt, R., et al.: Guided anisotropic diffusion and iterative learning for weakly supervised change detection. In: CVPR Workshop. IEEE (2019)
4. Chan-Hon-Tong, A., Audebert, N.: Object detection in remote sensing images with center only. In: IGARSS. pp. 7054–7057. IEEE (2018)
5. Chaurasia, A., Culurciello, E.: LinkNet: Exploiting encoder representations for efficient semantic segmentation. In: VCIP. IEEE (2017)
6. Chen, Q., et al.: Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. In: arXiv preprint arXiv:1807.09532 (2018)
7. Dai, J., et al.: BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: ICCV. pp. 1635–1643. IEEE (2015)
8. Heller, N., et al.: Imperfect segmentation labels: How much do they matter? In: MICCAI Workshop, pp. 112–120. Springer (2018)
9. Kaelbling, L.P., et al.: Reinforcement learning: A survey. In: JAIR. vol. 4, pp. 237–285 (1996)
10. Kontogianni, T., et al.: Continuous adaptation for interactive object segmentation by learning from corrections. In: arXiv preprint arXiv:1911.12709 (2019)
11. Le Saux, B., Sanfourche, M.: Rapid semantic mapping: Learn environment classifiers on the fly. In: IROS. pp. 3725–3730. IEEE (2013)
12. Lenczner, G., et al.: DISIR: Deep image segmentation with interactive refinement. In: ISPRS Annals. vol. V-2-2020, pp. 877–884. Copernicus GmbH (2020)
13. Lin, D., et al.: ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation. In: CVPR. pp. 3159–3167. IEEE (2016)
14. Maggiori, E., et al.: Can semantic labeling methods generalize to any city? the INRIA aerial image labeling benchmark. In: IGARSS. pp. 3226–3229. IEEE (2017)
15. Milan, A., et al.: Semantic segmentation from limited training data. In: ICRA. pp. 1908–1915. IEEE (2018)
16. Parisi, G.I., et al.: Continual lifelong learning with neural networks: A review. In: Neural Networks. Elsevier (2019)
17. Rother, C., Kolmogorov, V., Blake, A.: GrabCut: Interactive foreground extraction using iterated graph cuts. In: TOG. vol. 23, no. 3, pp. 309–314. ACM (2004)
18. Rottensteiner, F., et al.: The ISPRS benchmark on urban object classification and 3D building reconstruction. In: ISPRS Annals. vol. 1, no. 1, pp. 293–298. Copernicus GmbH (2012)
19. Tasar, O., et al.: Incremental learning for semantic segmentation of large-scale remote sensing data. In: JSTARS. vol. 12, no.9, pp. 3524–3537. IEEE (2019)
20. Torres, R., et al.: GMES Sentinel-1 mission. In: Remote Sensing of Environment. vol. 120, pp. 9–24. Elsevier (2012)
21. Xu, N., et al.: Deep interactive object selection. In: CVPR. pp. 373–381. IEEE (2016)