Weaklier Supervised: Semi-automatic Scribble Generation Applied to Semantic Segmentation

João Pedro Klock Ferreira, João Paulo Lara Pinto, and Cristiano Leite Castro
Graduate Program in Electrical Engineering - Universidade Federal de Minas Gerais
Av. Antônio Carlos 6627, 31270-901, Belo Horizonte - MG, Brazil
Email: {jpklock, joaopaulolara, crislcastro}@ufmg.br

***Abstract*—With many applications regarding semantic segmentation arising, along with the advent of the Deep Semantic Segmentation Networks, the need for large labeled datasets has also largely increased. But labeling thousands of images can be very expensive and time-consuming. Approaches such as weak and semi supervision try do deal with this problem, but the first cannot deal with large datasets and the latter is hard to deal with semantic segmentation. Therefore, in this work we propose a combination of both to create a novel pipeline of weak supervision, with focus in satellite imagery, capable of dealing with large datasets. We propose a pipeline to automatically generate scribbles in images, requiring that the user only label 10% of the images in a given dataset, while a classifier deal with the remaining images. Along with that, we also propose a simple semantic segmentation pipeline, that uses only images with scribbles to train a network. Results show that performance is lower, but similar to a fully supervised pipeline.**

I. Introduction

Semantic segmentation, i.e., automatic pixel-wise classification of images, has been gaining much attention with the advent of self-driving cars and the increasing usage of satellite images for studies related to land cover on Earth's surface. In both cases, the ability to describe the contents of images is of great interest since most applications will require some degree of semantic explainability of what can be seen.

The most accurate and popular way of performing semantic segmentation is using Deep Semantic Segmentation Networks (DSSNs) [1]. Nevertheless, the performance of DSSNs in semantic segmentation tasks is highly dependent on the availability of a fully annotated dataset with segmented masks in which every pixel in the image has its own label. Making these masks can be both very tedious and time-consuming, sometimes even requiring the help of experts to provide the correct annotations, therefore also becoming expensive.

Due to the problems above, many researchers have tried to find ways of generating these ground truth masks more efficiently. In a field called weakly-supervised learning [2], [3], the main idea is to obtain a more straightforward and cheaper way to annotate datasets through weak labels. In the remote sensing literature, such labels may include scribbles [4], [5], binary labels indicating the presence of elements in images [6], bounding boxes [5] and even single pixels [5], [6].

Although the annotation via weak supervision is less time-consuming, in most cases, the interference of an expert is still necessary. This is often a problem for large datasets containing several thousand images. Some studies of semi-supervised image classification have dealt with this problem, including labeling a smaller set of the data, training a model with fewer images, and then classifying the remaining data [3], [7]. However, when using weak labels, this can not be replicated because the labels are pixel level.

To address this problem, automatic weak label generation has been used in the context of image colorization, where automatic scribbles would be generated and manually classified to colorize images [8]. But the labels would still have to be manually allocated, not solving the large dataset problem. In similar works, drawing regions for scribbles are proposed [9], but the problem of manual labeling still remains.

This article presents a solution to generate semantic segmentation masks using only a small set of a large-scale dataset, with focus in satellite imagery. By combining concepts of weak and semi-supervised learning, we propose a pipeline to automatically generate scribbles in images based on regions with similar content. Then, the expert must annotate only a small set of the scribbles, i.e., selecting the scribble label, which we also provide a tool to perform. A classifier will automatically classify (label) the remaining scribbles. We then propose a DSSN architecture and pipeline that, unlike others in the literature, is simple and can be trained using only scribbles, achieving similar results to a network trained with fully supervision. Finally, our main contributions are:

- An automatic scribble generator;
- A semi-supervised labeling pipeline that requires only a small set of images to be labeled;
- A graphical user interface to visualize, draw, edit, remove and annotate scribbles manually;
- A simple weakly-supervised DSSN training pipeline.

II. Methodology

A. Automatic Scribble Generation

For this work, a pipeline for automatic scribble generation is proposed. This idea is based both on works that focus on searching for similar regions in datasets [9], [10] and that generates automatic scribbles [8]. In remote sensing images, there are usually well-defined crops and terrains, especially in rural areas. Given that, it should be possible to group them, based on the similarity of region features (descriptors), in the same way that perhaps a human would unintentionally draw scribbles, given the clarity of the region.

The complete pipeline can be seen in Figure 1. Similar pipelines have been proposed before for grayscale images [8], but every similar step is performed differently in our pipeline.

Starting with an input image in step 1.1, a bilateral filter is applied in step 1.2 to homogenize its contents while preserving edges, further obtaining the variance of the pixels, $ImVar$. The filtered image is only used in this step. In step 1.3 the images are segmented in $1000 * ImVar$ superpixels. Accordingly, images with higher variance (having more artifacts) will have more superpixels than images with homogeneous content.

Next, in step 1.4 we extract color (f_c) and texture (f_t) features from the superpixels, as described in [11]. A graph is mounted in step 1.5 with adjacent superpixels, with the weights, ϕ_{ij}, adapted from [4], but using different features, and calculating their similarities using the features themselves.

Assuming that similar superpixels will have a closer distance in the feature space, we set adjacent superpixels to the same region if their edge weight is higher than 0.9. Groups of superpixels are mounted, and only those with more than 5% of the total are used, avoiding noisy small groups.

To draw the scribbles in step 1.6.1, the shortest path among the superpixels centers is found through a generic solver for the traveling salesman problem, without needing to return to the initial point. Finally, in optional step 1.6.2, after a bicubic interpolation to increase the number of points, a gaussian filter is applied to smooth the scribble, achieving the final result. Approximating the shape of human generated scribbles can be beneficial in cases with curved superpixels, although is most cases only the visual is improved. The process to obtaining the scribbles classes will be described in Section III-A.

One problem observed with this methodology was the difficulty of segmenting small objects in more than one superpixel, such as the "building" class from the chosen dataset, causing them not to reach the minimum amount to originate a scribble. Therefore, we coupled a building detector [12] to the pipeline to improve their representativity, forcing superpixels with buildings to become scribbles, sometimes also adding roads.

B. Scribble-Driven Training

With the scribble set, it is possible to feed the semantic segmentation pipeline designed for this work, as seen in step 3 of Figure 1. Given images with scribbles in step 3.1, each image is segmented in superpixels in step 3.2, using the same pipeline of step 1.2, as described in Section II-A. Then, in step 3.3, the algorithm verifies which superpixels are crossed by each scribble, and assigns the scribble label to their pixels, increasing the number of labeled pixels in the dataset but also adding a small noise. Superpixels not crossed by any annotation lines are assigned to an "undetermined" class.

Step 3.4 consists of calculating and assigning weights for each class present in the dataset to overcome their imbalance during training, along with setting the weight of the "undetermined" class to 0 In order to increase diversity in the training set, in step 3.5 we applied random rotations, flips, brightness changes, zooms, and translations to the images.

The DSSN architecture chosen to perform the semantic segmentation in step 3.6 was DeepLabV3+ [13], which has presented good results for the same dataset [14]. As backbone, ResNet-50 pre-trained on ImageNet was selected for simplicity. Model evaluation is performed with four metrics: pixel accuracy, intersection over union (IoU), mean intersection over union (mIoU), and weighted mean intersection over union (wmIoU). The wmIoU metric, defined as the mean of the IoU weighted by the pixel proportion of each class, was implemented to also consider the class imbalance.

III. Results

A subset of 780 images with a size of 512×512 pixels of the LandCover.ai dataset [14] was used for all experiments, which comprehends areas of different characteristics from Poland. There are five classes: building, woodland, water, road, and background. Those images were selected based on their similarity of saturation, brightness, and resolution since these conditions better represent samples captured in a single flight. Similarly, 276 images were selected to compose the test set. The pixel proportion in which the classes appear in the training set is 0.56% for buildings, 61.17% for woodland, 5.07% for water, 1.89% for roads, and 31.31% for background.

A. Scribble Classifier

When automatically generating the scribbles, they are all assigned a generic class. In step 2 of the pipeline in Figure 1, we propose to annotate the scribbles of a small set of images and classify the remaining ones. With our provided graphical tool, it is possible to draw, edit, remove and change the class of already drawn scribbles. This way, the user can auto-generate, label and fix the worst ones. The graphical user interface is represented in step 2.1 of the pipeline and can be found in https://github.com/jpklock2/Scribble-Editing-Tool.

We explored two levels of human intervention. The first when all scribbles are drawn by an expert, and the second when the scribbles are generated automatically. In both cases, an SVM classifier is trained using the mean features (step 1.4) of every superpixel crossed by a scribble. The scribbles of 10% of the dataset images are used, and the remaining is classified, as illustrated in pipelines' steps 2.2 and 2.3. We then set four experiments to evaluate the process of scribble classification.

- Experiment 1 - 10% of the images with human-made scribbles and 90% classified by SVM.
- Experiment 2 - same as 1 with building detection module.
- Experiment 3 - 10% of the images with automatic generated scribbles and 90% classified by SVM.
- Experiment 4 - same as 3 with building detection module.

Table I shows precision ("Prec.") and recall ("Rec.") achieved by the SVM classifier for each class, with the ground truth provided by an expert. "Back" stands for Background, "Build." for Building, and "Wood." for Woodland.

Generally speaking, the results of experiments 3 and 4 were close to experiments 1 and 2, showing that the proposed method for automatic generation and classification of scribbles can mimic human-made scribbles well.

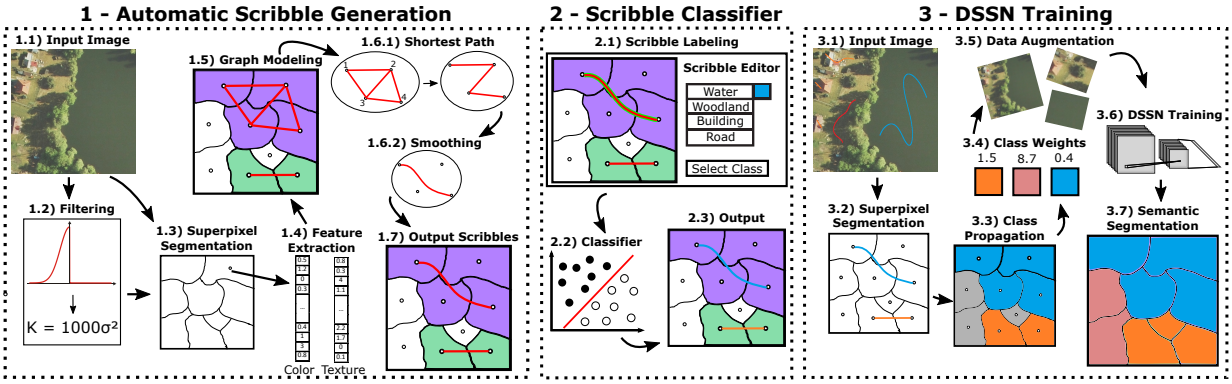

Fig. 1. Complete Pipeline for this work.

TABLE I
SCRIBBLE CLASSIFIER RESULTS.

Exp.	Metric	Back.	Build.	Wood.	Water	Road	Mean
1	Prec.	85.52	85.45	93.83	50.00	87.34	80.43
	Rec.	78.61	90.82	87.49	97.50	66.35	84.15
2	Prec.	87.30	73.29	90.88	50.64	81.01	76.62
	Rec.	73.29	90.48	87.69	96.60	59.92	81.60
3	Prec.	69.89	100.00*	96.62	68.18	65.79	80.10
	Rec.	92.14	100.00*	77.95	85.71	32.05	77.57
4	Prec.	81.26	73.89	92.25	66.90	64.63	75.79
	Rec.	83.94	92.02	88.73	86.61	38.41	77.94

* There were only 3 examples in this class

Analyzing individually by class, the performance was similar for the Background and Woodland; Building and Road performances, in turn, were better for human-made than for automatic-generated scribbles, except for experiment 3, which only had 3 examples and thus is not reliable. This occurred because these classes involve small and narrow objects, making it more challenging to generate their corresponding scribbles automatically. Despite this, we believe this problem could be solved by manually fixing these scribbles, either by removing the bad ones and redrawing or just drawing additional ones.

At last, the building detection module seems to hinder the performance of the scribbles classifier, as can be seen when comparing experiments 1 and 2 or 3 and 4. Some examples of scribbles generated by our methodology can be seen in Figure 2, along with the respective ground truth labels.

B. Semantic Segmentation

This experiment shows the effect of using automatic-generated scribbles in DSSNs training. The resulting training sets of Experiments 1 to 4, presented in Section III-A, were evaluated. As baselines, DSSNs were trained with the ground-truth semantic masks, and semantic masks built from a set of 100% human-annotated scribbles. Segmentation results can be seen in Figure 3, and the test set metrics for Accuracy ("Acc."), IoU, mIoU, and wmIoU are shown in Table II; "HAS" stands for Human Annotated Scribbles and "GT" for Ground Truth.

From the IoU values, it is clear that classifying buildings and roads is the hardest task for this dataset, since there are far fewer scribbles (and pixels) of these classes, and as these

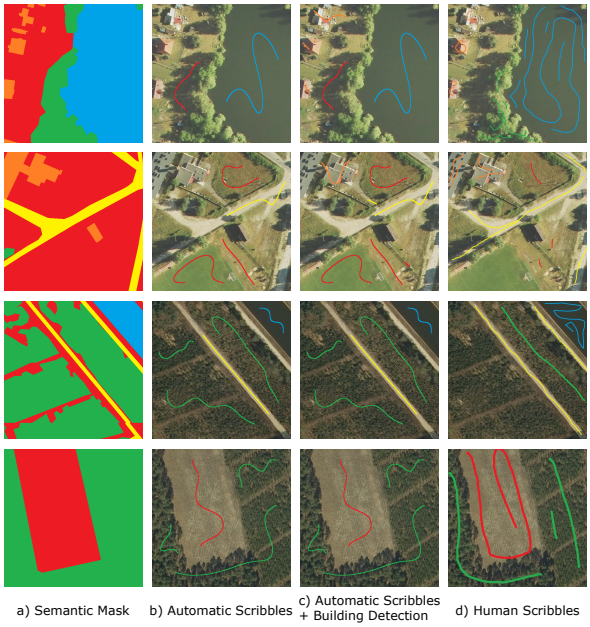

Fig. 2. Different types of scribbles comparison.

TABLE II
SEMANTIC SEGMENTATION METRICS.

Exp.	IoU Back.	Build.	Wood.	Water	Road	mIoU	wmIoU	Acc.
1	0.668	**0.146**	0.825	0.303	**0.098**	0.408	0.747	0.832
2	0.657	0.079	0.826	0.291	0.089	0.388	0.743	0.826
3	**0.679**	0.053	0.827	**0.598**	0.046	**0.441**	**0.760**	**0.857**
4	0.641	0.092	**0.836**	0.406	0.073	0.410	0.750	0.841
HAS	0.635	0.033	0.887	0.792	0.146	0.499	0.792	0.847
GT	0.822	0.256	0.924	0.752	0.423	0.635	0.878	0.929

elements are smaller and narrower, the label propagation step is noisier. Overall, woodland and background were better classified, with consistent IoU values over the four experiments. The results show that the model trained from the scribbles generated by experiment 3 performed best, with its accuracy even surpassing the training with hand-made scribbles.

In experiments 1 and 2, the results got worse with building detection, because when the classifier had an increased number

of automatically generated building scribbles, it probably introduced more error to the dataset, causing the segmentation to decrease performance. As for experiments 3 and 4, the performance for the building and road classes improved substantially, but are still worse than human scribbles.

Some segmentation results can be seen in Figure 3, where the amount of buildings detected increased in the models with building detection, even though the overall results got worse.

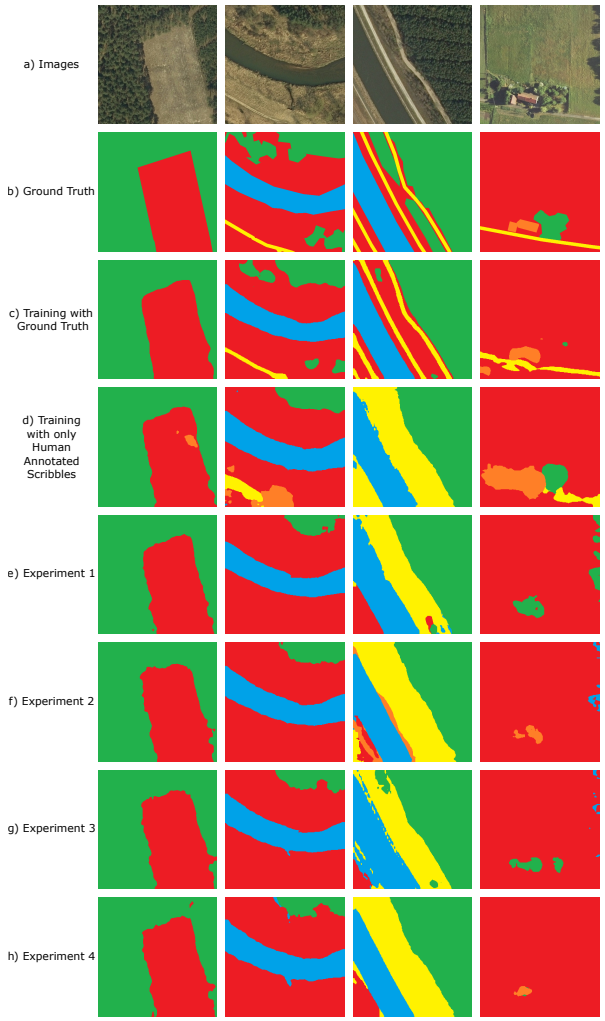

Fig. 3. Network predictions for different conditions. Experiments explanation can be found in Section III-A.

IV. Conclusion

In this work, we present a methodology for automatic scribble generation, a tool to edit and create scribbles, along with a simplified approach for a DSSN supervised by these annotation lines. We show that only 10% of the images with well-distributed classes are enough to label an entire dataset. In addition, those scribbles are used for DSSN training, presenting results promisingly close to those obtained with full supervision. The biggest challenges were the generation and labeling of smaller and more subjective classes, such as buildings and roads scribbles. Some solutions include improving the scribbles using the provided editing tool, performing a study of the method's behaviour when applied in urban area, or analyzing the impact of the superpixel size in these classes.

The presented method was tested using only satellite images. For future works, further tests exploring scene segmentation or autonomous vehicle applications can be performed. Finally, ablation studies can be carried to evaluate the impact of each step of the pipeline in the final results, pointing out where adjustments would be more significant.

Acknowledgment

This work has been supported by CNPq and Capes.

References

[1] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, 2022.

[2] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," *Proceedings of the VLDB Endowment*, vol. 11, no. 3, pp. 269–282, 2017.

[3] Z. H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

[4] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3159–3167.

[5] Y. Hua, D. Marcos, L. Mou, X. X. Zhu, and D. Tuia, "Semantic Segmentation of Remote Sensing Images with Sparse Annotations," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2022.

[6] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell, "Weakly supervised deep learning for segmentation of remote sensing imagery," *Remote Sensing*, vol. 12, no. 2, 2020.

[7] J. Wang, C. H. Ding, S. Chen, C. He, and B. Luo, "Semi-supervised remote sensing image semantic segmentation via consistency regularization and average update of pseudo-label," *Remote Sensing*, vol. 12, no. 21, pp. 1–16, 2020.

[8] X. Ding, Y. Xu, L. Deng, and X. Yang, "Colorization using quaternion algebra with automatic scribble generation," *Lecture Notes in Computer Science*, vol. 7131 LNCS, pp. 103–114, 2012.

[9] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "Interactively co-segmentating topically related images with intelligent scribble guidance," *International Journal of Computer Vision*, vol. 93, no. 3, pp. 273–292, 2011.

[10] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman, "Unsupervised Semantic Segmentation by Distilling Feature Correspondences," in *ICLR 2022*, April 2022.

[11] B. A. Jaimes, J. P. K. Ferreira, and C. L. Castro, "Unsupervised Semantic Segmentation of Aerial Images with Application to UAV Localization," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2022.

[12] B. Sirmacek and C. Unsalan, "A probabilistic framework to detect buildings in aerial and satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 1, pp. 211–221, 2011.

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[14] A. Boguszewski, D. Batorski, N. Ziemba-Jankowska, T. Dziedzic, and A. Zambrzycka, "Landcover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 1102–1110.