Soft Self-labeling and Potts Relaxations for Weakly-Supervised Segmentation
Supplementary Material

Zhongwen Zhang & Yuri Boykov
University of Waterloo, Canada
z889zhan@uwaterloo.ca, yboykov@uwaterloo.ca

A. Optimization Algorithm

In this section, we will focus on the optimization of our loss where we iterate the optimization of y and σ. The network parameters are optimized by standard stochastic gradient descent in all our experiments. Pseudo-labels are also estimated online using a mini-batch. To solve y at given σ, it is a large-scale constrained convex problem. While there are existing general solvers to find global optima, such as projected gradient descent, it is often too slow for practical usage. Instead, we reformulate our problem to avoid the simplex constraints so that we can use standard gradient descent in PyTorch library accelerated by GPU. Specifically, instead of directly optimizing y, we optimize a set of new variables $\{l_i \in \mathbb{R}^K, i \in \Omega\}$ where y_i is computed by $softmax(l_i)$. Now, the simplex constraint on y will be automatically satisfied. Note that the hard constraints on scribble regions still need to be considered because the interaction with unlabeled regions through pairwise terms will influence the optimization process. Inspired by [13], we can reset $softmax(l_i)$ where $i \in S$ back to the ground truth at the beginning of each step of the gradient descent.

However, the original convex problem now becomes non-convex due to the Softmax operation. Thus, initialization is important to help find better local minima or even the global optima. Empirically, we observed that the network output logit can be a fairly good initialization. The quantitative comparison uses a special quadratic formulation where closed-form solution and efficient solver [1, 6] exist. We compute the standard soft Jaccard index for the pseudo-labels between the solutions given by our solver and the global optima. The soft Jaccard index is 99.2% on average over 100 images. In all experiments, the number of gradient descent steps for solving y is 200 and the corresponding learning rate is 0.075. To test the robustness of the number of steps here, we decreased 200 to 100 and the mIoU on the validation set just dropped from 71.05 by 0.72. This indicates that we can significantly accelerate the training without much sacrifice of accuracy. When using 200 steps, the total time for the training will be about 3 times longer than the SGD with dense Potts [11].

B. Additional Experiments

Dataset and evaluation We mainly use the standard PASCAL VOC 2012 dataset [5] and scribble-based annotations for supervision [8]. The dataset contains 21 classes including background. Following the common practice [2, 10, 11], we use the augmented version which has 10,582 training images and 1449 images for validation. We employ the standard mean Intersection-over-Union (mIoU) on validation set as the evaluation metric. We also test our method on two additional datasets. One is Cityscapes [4] which is built for urban scenes and consists of 2975 and 500 fine-labeled images for training and validation. There are 19 out of 30 annotated classes for semantic segmentation. The other one is ADE20k [12] which has 150 fine-grained classes. There are 20210 and 2000, images for training and validation. Instead of scribble-based supervision, we followed [7] to use the block-wise annotation as a form of weak supervision.

Method	Architecture	Cityscapes	ADE20k
Full supervision			
Deeplab [3]	V3+	80.2	44.6
Block-scribble supervision			
DenseCRF loss [11]	V3+	69.3	37.4
GridCRF loss* [9]	V3+	69.5	37.7
TEL [7]	V3+	71.5	39.2
$\mathbf{H}_{\text{CCE}} + \mathbf{P}_{\text{CD}}$	V3+	72.4	39.7

Table 1. Comparison to SOTA methods (without CRF postprocessing) on segmentation with block-scribble supervision. The numbers are mIoU on the validation dataset of cityscapes [4] and ADE20k [12] and use 50% of full annotations for supervision following [7]. The backbone is ResNet101. "$*$": reproduced results. All methods are trained in a single-stage fashion.

References

[1] Multilabel random walker image segmentation using prior models. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 763–770. IEEE, 2005. 1

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 1

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[5] 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*, 88:303–308, 2009. 1

[6] Leo Grady. Random walks for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 28 (11):1768–1783, 2006. 1

[7] Zhiyuan Liang, Tiancai Wang, Xiangyu Zhang, Jian Sun, and Jianbing Shen. Tree energy loss: Towards sparsely annotated semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16907–16916, 2022. 1

[8] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016. 1

[9] Dmitrii Marin and Yuri Boykov. Robust trust region for weakly supervised segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6608–6618, 2021. 1

[10] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1818–1827, 2018. 1

[11] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 507–522, 2018. 1

[12] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1

[13] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. *ProQuest Number: INFORMATION TO ALL USERS*, 2002. 1