2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506185

# LEARNING OR MODELLING? AN ANALYSIS OF SINGLE IMAGE SEGMENTATION BASED ON SCRIBBLE INFORMATION

*Hannah Dröge and Michael Moeller*

University of Siegen
Hölderlinstraße 3, 57076 Siegen, Germany

## ABSTRACT

Single image segmentation based on scribbles is an important technique in several applications, e.g. for image editing software. In this paper, we investigate the scope of single image segmentation solely given the image and scribble information using both convolutional neural networks as well as classical model-based methods, and present three main findings: 1) Despite the success of deep learning in the semantic analysis of images, networks fail to outperform model-based approaches in the case of learning on a single image only. Even using a pretrained network for transfer learning does not yield faithful segmentations. 2) The best way to utilize an annotated data set is by exploiting a model-based approach that combines semantic features of a pretrained network with the RGB information, and 3) allowing the networks prediction to change spatially and additionally enforce this variation to be smooth via a gradient-based regularization term on the loss (*double backpropagation*) is the most successful strategy for pure single image learning-based segmentation.

***Index Terms***— Image Segmentation, Scribbles, Energy Minimization Methods, Machine Learning

## 1. INTRODUCTION

Image segmentation refers to the problem of dividing an image into meaningful non-overlapping regions and is a crucial component in many image processing applications. Famous classical techniques have formulated image segmentation in terms of an energy minimization problem, e.g., in form of graph cuts [1, 2] or variational methods [3]. The *unaries* (or *data terms*) have been modeled in various forms including optimizing for suitable thresholds [3], estimating it via spectral methods [4, 5], or constructing it explicitly via additional annotations such as bounding boxes [6] or user scribbles [7]. Notably, the latter work demonstrated that a faithful segmentation is possible based on a few user scribbles by constructing *spatially varying* color histograms for each label, and estimating the likelihood of a pixel having a certain label by computing probabilities of the pixels' RGB values being a part of the corresponding histogram.

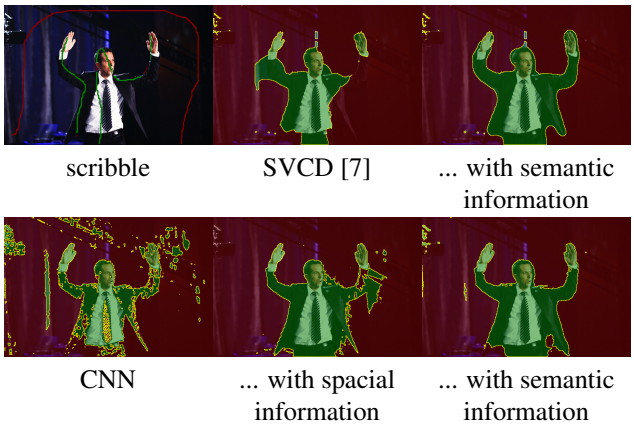


**Fig. 1**. Single image segmentation based on scribbles that rely on image color and on optional spatial or semantic information: The upper row shows the segmentation by SVCD, the second row shows the segmentation by a CNN.

With the rise of deep learning within the past decade, researchers have focused on image segmentation networks with great success [8, 9, 10, 11]. Rather than estimating object properties on separate images, these networks are usually trained on thousands of examples and are therefore able to *learn* common shapes of objects in their training data.

Yet, the limitation to only predict those semantic labels (and objects) present in the training database limits the applicability of such models, particularly because the sole definition of image segmentation as a division into "meaningful" non-overlapping regions makes image segmentation an ill-posed task by definition: What is a meaningful region? The answer, of course, is highly subjective and depends on the specific intended application.

This is the reason why we believe that image segmentation on only one image based on scribbles, i.e., the prediction of regions in a single image that have previously been marked by a few strokes, so-called *scribbles*, by a user (see Fig. 1), remains highly relevant, e.g. for image editing software.

Unfortunately, hardly any works in the area of deep learning focus on single image segmentation from scribbles. This poses the fundamental question if model- or learning-based

approaches represent the state-of-the-art in this field, along with the quest for network architectures and regularization schemes that are well-suited for single image segmentation.

In this work we study these questions in two different scenarios: 1) The case where the current scribbled image represents the sole source of information, and 2) the case in which prior information from a segmentation benchmark may be utilized in the form of transfer learning or accessing features of segmentation networks.

For the first scenario we demonstrate that the additional inclusion of spatial information in a neural network improves the segmentation compared to color-only images. We also present how color and spatial information can be optimally weighted against each other for segmentation using double backpropagation. Yet, the model-based method remains superior to neural networks. In the second scenario, we propose a hybrid technique that combine the (model-based) spatially varying color histogram from [7] with learning-based soft semantic features from [12] and yields results that outperform the segmentation by neural networks and the standalone model-based approach. Here we focus on segmenting one image with scribbles without using any semantic information about the objects in the scene.

## 2. RELATED WORK

Classical approaches such as Graph Cuts [1, 2], or the edge-based segmentation with Snakes [13] phrase the image segmentation problem as the minimization of a suitable cost function including a unary term that drives the segmentation and a smoothness term that yields sufficiently regular regions. Particularly influential in that respect also is the model of Mumford and Shah [14], which forms the basis of the successful two region segmentation method of Chan and Vese [3], and has been extended to multiple regions in various works, see e.g. [15, 16]. A typical form of such approaches is to determine a one-hot representation $\hat{u} \in \mathbb{R}^{n_y \times n_x \times L}$ of an $L$-region segmentation for an image $f \in \mathbb{R}^{n_y \times n_x \times n_c}$ via

$$\hat{u} \in \underset{u_{i,j,l} \in \{0,1\}, \sum_l u_{i,j,l}=1}{\arg\min} \langle u, c(f) \rangle \; + \; \alpha R(u), \qquad (1)$$

where $R$ denotes a suitable regularization that penalizes irregularities, e.g. the (weighted) total variation [17], and $c(f) \in \mathbb{R}^{n_y \times n_x \times L}$ are the unary costs with the entry $(c(f))_{i,j,l}$ having some sort of inverse relation to the estimated likelihood of pixel $(i, j)$ belonging to class $l$.

Due to the difficulty of generating meaningful unaries $c(f)$ without additional information, several works have considered interactive segmentation methods, in which the user provides clues about the object to be segmented, e.g., in form of bounding boxes [6], or scribbles [18]. For instance, in the work [7] (later extended to textural [19] informations and depth segmentation [18]) unaries are estimated by considering a spatially varying color histogram for each object from the user scribbles, and subsequently determining the likelihood of each pixel of belonging to a certain class.

With the rise of deep learning methods in the past decade, end-to-end image segmentation approaches are now largely representing the state-of-the-art, and we refer to [10, 8, 9, 11] for some examples of fully supervised image segmentation networks. Yet, such networks require large training data sets, which are expensive to generate, and the resulting networks limited to exactly those classes they have been trained on.

In weakly supervised methods not every pixel is annotated, but weaker sources of information such as image labels [20], scribbles [21] or bounding boxes [22], are used. Still, the aforementioned approaches require large training data sets and do not generalize to previously unseen categories.

Reducing the amount of supervision even further, several researchers have investigated learning-based clustering methods. Such techniques can also be applied to image segmentation, e.g. in [12], without even knowing the number of classes a-priori. Related to this, the intention of zero-shot segmentation is to segment non-annotated objects that have not previously been seen by a neural network, as in [23, 24], which, however, do not utilize with scribbles.

## 3. SINGLE IMAGE SEGMENTATION

Spatially varying color distributions [7] for $c(f)$ in (1), and using an edge-weighted (or even nonlocal) total variation regularization marked the state of the art in 2014. Since then, deep learning has proven to dominate any segmentation application for which at least a moderate number of training examples is available. Thus, we believe it is high time to ask if modern network architectures are able to outperform model-based approaches *even on a single image*.

We first consider an image segmentation problem for which all methods rely solely on the given scribbled image (such that, on a pixel level, we still have have several hundred training examples). Subsequently, we study how to incorporate prior information in the form of a different data set with ground information. As transfer learning appears to fail (detailed in Sec. 3.3), we propose to instead use *soft semantic features* of Aksoy *et al*. [12], that have been successfully used in soft semantic segmentation without scribble information.

### 3.1. Model- and Learning-based Segmentation Methods

In both settings (with and without prior information), we compare the following approaches: **Spatially varying color distributions (SVCD)** [7] are used to model smoothly changing histograms in space to approach the scribble-based segmentation problem via solving a convex relaxation of (1) followed by a thresholding. It does not involve any learning. To integrate additional semantic information, we concatenate the RGB values with the soft semantic features prior to applying the method.

To mimic the behavior of color histogram-based approaches, we train **pixel-wise networks (PWNs)** $\mathcal{G}(x, \theta)$ with learnable parameters $\theta$ that get the vector $x \in \mathbb{R}^{n_c}$ of RGB values at a single pixel as an input and are suggesting a class label solely based on color. To additionally incorporate the idea of spatially varying color distributions in a learning-based setup, we also train PWNs with a 5-dimensional input vector consisting of the RGB values as well as the xy-coordinates of the image (normalized to a range of $[0, 1]$), resembling an interesting similarity to *implicit* neural representations as e.g. investigated in [25]. Similarly, semantic features are just concatenated with other inputs. In terms of the network architecture, an extensive empirical search resulted in surprisingly small and shallow structure, consisting of two layers with 16 neurons per layer and Leaky-ReLU activations.

As PWNs might have too little spatial context to make faithful predictions, we additionally consider **convolutional neural networks (CNNs)** with larger receptive fields: Using larger convolution kernels and increasing the depth of the networks allows us to provide the network with more and more non-local information. Again, we evaluate networks that use the plain RGB input image as an input and concatenated it in the channel dimension with its xy-coordinates and/or the semantic features. In an ablation study detailed in section 3.4 we again found a rather shallow network of *depth 2*, *width 16*, and a *kernel size of 3* to be most successful.

In our experiments we found that the inclusion of spatial information often dominates the results of the CNNs, such that objects close to the scribble are incorrectly segmented. For this reason, we introduce a regularization term similar to the idea of *double backpropagation* from [26] by computing

$$\min_{\theta} L + 10^{-7}\beta||\nabla_{x_s} L||_1,\ L = CE(G(x_c, x_s, \theta), sc). \quad (2)$$

Here the network loss $L$ is computed by the cross entropy ($CE$) between our scribbles and the output of the network $G$, using the color information $x_c$ and the spatial information $x_s$. We regularize the $l^1$ norm of the gradient of our loss function, a form of *total variation regularization*, in the spatial direction to attenuate the influence of $x_s$ on the our final result.

Fig. 2 visualizes the effect of increasing the regularization on the segmentation, whereby without regularization, structures close to the object are segmented as those, and with a too strong regularization, the color information predominates. We refer to this regularized approach as **CNN+reg.**.

As one of the most famous architectures for semantic image segmentation, we finally consider the **U-Net** architecture from [27]. It is a convolutional architecture with a receptive field that spans large portions of the image while still being able to preserve fine details. While this architecture would clearly be superior in a fully supervised setting with sufficient training data, our investigations aim at an understanding how the strong overparameterization of such an approach in comparison to the small number of labeled (=scribbled) pixels in a single image affects its accuracy. We took the U-Net architecture from [27] as a basis and conducted a study on the number of downscaling steps of the network architecture. We tested U-Net architectures without any downscaling steps up to four steps and observed a decrease in segmentation accuracy for U-Net networks with more downscaling steps.

$\beta = 0$ $\beta = 0.05$ $\beta = 0.1$

**Fig. 2**. Segmentation with color and weighted spacial information via double backpropagation as network input.

### 3.2. Numerical Evaluation

To evaluate the above approaches, we use scribbled images from [21] of the Pascal VOC2012 data set [28] for which ground truth segmentation are available. We tune the hyperparameters of each method on a fixed set of 200 images of this dataset. Table 1 shows the best results we were able to attain for each class of methods, with the left part of the table depicting the single image segmentation results without additional information (using color (**c**) and spatial (**sp**) information only), and the right part additionally allowing the use of semantic (**se**) features from [12].

| | **c** | **c & sp** | **c & se** | **c & sp & se** |
|---|---|---|---|---|
| SVCD | - | 0.775 | - | 0.845 |
| PWN | 0.607 | 0.721 | 0.689 | 0.731 |
| CNN | 0.673 | 0.715 | 0.772 | 0.739 |
| CNN + reg. | - | 0.731 | 0.774 | - |
| U-Net | 0.654 | 0.642 | 0.658 | 0.648 |

**Table 1**. Summary of the best mean IoU each of the pure single image segmentation methods could attain for different types of network input: **c**olor image, **sp**acial pixel position and **se**mantic information. As we can see, the additional inclusion of spatial information improves segmentation, which in turn is outperformed by the inclusion of semantic information in the neural network. SVCD still outperforms the learning based methods.

As we can see, the model-based SVCD approach outperformed all learning-based approaches in terms of pixel accuracy as well as mean intersection over union (IoU). While other fields, e.g., in image reconstruction with pioneering work on deep image priors in [29], indicated that the architecture of common convolutional networks has a regularizing effect that is well suited for natural images and thus allows a self-supervised training on a single image, we cannot confirm that similar effects appear in image segmentation.

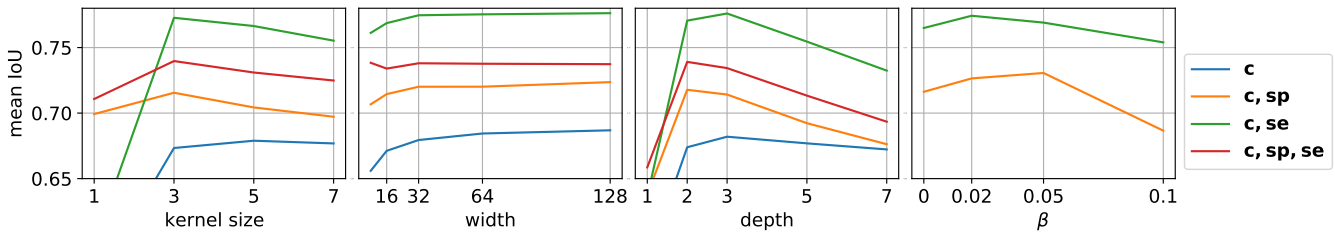


**Fig. 3**. Study of the impact of the convolutional neural network from left to right: kernel size, width , and depth on the segmentation accuracy for different types of network input: **c**olor image, **sp**acial pixel position and **se**mantic information. The most right image shows the regularization of the spacial impact on the segmentation as defined in (2) (orange plot), as well as the corresponding penalty w.r.t. variations of the color channels (green plot).

Interestingly, the inclusion of the spatial coordinates as inputs to the neural network helped to improve all learning-based approaches except the U-Net. Moreover, including our proposed regularization to avoid an overfitting to the spatial information only, gave the best result among the learning-based approaches. Based on the rather small CNNs that our ablation study in Sec. 3.4 found to be optimal, and the surprisingly bad performance of a U-net architecture which is not even influenced by the inclusion of semantic information, we conclude that overfitting remains a significant problem in single image segmentation with neural networks.

Except for U-net, the combination of semantic and color information increases the performance of all methods significantly. In particular, the combination of the model-based creation of spatially varying color histograms with semantic soft features achieves excellent results. Interestingly, the additional inclusion of spatial coordinates on top of semantic features does not appear to be beneficial for CNNs anymore. The slight improvement of CNN+reg. over CNN was obtained similar to (2), but using the gradient with respect to the color input instead of the spatial coordinates. As the proposed approach of using semantic soft features is only possible if a second annotated dataset is available, *transfer learning* is a natural baseline for such approaches.

### 3.3. Transfer Learning

A common method in semantic segmentation is the fine-tuning of a neural network pretrained on a given fully supervised data set. Thus, we finetune the architectures ENet [30] (pretrained on the CityScape [31] dataset by [32]) and DeepLabv3 [33] (pretrained on the CityScape dataset [31] by [34]), on single scribbled images. Despite varying the amount of parameters to freeze and train, the best mean IoUs were found to be 0.52 for ENet and 0.59 for Deeplabv3. Surprisingly, these values are not even close to the results seen in Table 1 even without semantic features. We conclude that - at least for the significantly different datasets of CityScape and VOC2012 - it is not straightforward to utilize transfer learning for single image segmentation with scribbles.

### 3.4. Ablation study for CNNs

To study the impact of the architecture, we train simple CNNs with alternating convolution and ReLu layers of varying width and depth along with a cross-entropy loss on the scribbled pixels only. By expanding the kernel size from $1 \times 1$ convolutions (which is equivalent to our PWNs), the segmentation networks start to include information from neighboring pixels in the predicted segmentation. Fig. 3 shows how the mean IoU depends on the neural networks widths, depths, kernel size, and parameter $\beta$ of our proposed regularization for different inputs using color (c), semantic (se), and spatial (sp) information. Here the non-variable values in the graphs are fixed to *width = 16*, *depth = 2*, *kernel size = 3*, and $\beta = 0$. Given the above parameters and all the input variations shown in Fig. 3 we could measure a variance of the mean IoU of $\pm 0.002$ in our experiments. As we can see, rather shallow networks of only 2 layers are more successful than deep ones, while the width has little effect as long as the network consists of at least 16 channels. Finally, the kernel size was found to be optimal for $3 \times 3$ convolutions, and moderate values of the regularization parameter $\beta$ do allow to increase the mean IoU by over $0.01$.

## 4. CONCLUSIONS

We have shown that image segmentation based on user-drawn scribbles is a challenging problem where model-based approaches still perform better than machine learning. Instead of transfer learning approaches, including soft semantic features as additional input channels to an energy minimization approach using spatially-varying histograms showed the most promising performance. While our modifications to include the spatial coordinates as inputs and simultaneously regularizing their input were able to improve the mean IoU of learning-based approaches, significantly, it remains an interesting challenge for future research to develop networks, regularizations, and training schemes that can outperform model-based approaches even on single image segmentation without prior information.

## 5. REFERENCES


[1] D. M. Greig, B. T. Porteous, and A. H. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 51, no. 2, 1989.

[2] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *ICCV*. IEEE, 2001, vol. 1.

[3] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, 2001.

[4] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems*, vol. 14, 2001.

[5] T. Cour, F. Benezit, and J. Shi, "Spectral segmentation with multiscale graph decomposition," in *CVPR*. IEEE, 2005, vol. 2.

[6] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut" interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, no. 3, 2004.

[7] C. Nieuwenhuis and D. Cremers, "Spatially varying color distributions for interactive multilabel segmentation," *TPAMI*, vol. 35, no. 5, 2012.

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *TPAMI*, vol. 39, no. 12, 2017.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[12] Y. Aksoy, T.-H. Oh, S. Paris, M. Pollefeys, and W. Matusik, "Semantic soft segmentation," *ACM Transactions on Graphics*, vol. 37, no. 4, 2018.

[13] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, 1988.

[14] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Communications on Pure and Applied Mathematics*, vol. 42, no. 5, 1989.

[15] A. Chambolle, D. Cremers, and T. Pock, "A convex approach to minimal partitions," *SIAM Journal on Imaging Sciences*, vol. 5, no. 4, 2012.

[16] E. Bae, J. Yuan, and X.-C. Tai, "Global minimization for continuous multiphase partitioning problems using a dual approach," *International Journal of Computer Vision*, vol. 92, no. 1, 2011.

[17] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, 1992.

[18] J. Diebold, N. Demmel, C. Hazırbaş, M. Moeller, and D. Cremers, "Interactive multi-label segmentation of rgb-d images," in *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer, 2015.

[19] C. Nieuwenhuis, S. Hawe, M. Kleinsteuber, and D. Cremers, "Co-sparse textural similarity for interactive segmentation," in *ECCV*. Springer, 2014.

[20] S. Hong, J. Oh, H. Lee, and B. Han, "Learning transferrable knowledge for semantic segmentation with deep convolutional neural network," in *CVPR*, 2016.

[21] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *CVPR*, 2016.

[22] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *ICCV*, 2015.

[23] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," in *Advances in Neural Information Processing Systems*, 2019.

[24] Z. Gu, S. Zhou, L. Niu, Z. Zhao, and L. Zhang, "Context-aware feature generation for zero-shot semantic segmentation," in *ACM International Conference on Multimedia*, 2020.

[25] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[26] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, 1998.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer Assisted Intervention*. Springer, 2015.

[28] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[29] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *CVPR*, 2018.

[30] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[31] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.

[32] D. Silva, "Pytorch-enet," https://github.com/davidtvs/PyTorch-ENet, 2020.

[33] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[34] D. Dao and Google Inc., "Tensorflow model garden," https://github.com/tensorflow/models, 2021.