Weakly Supervised Segmentation from Extreme Points

Holger Roth(✉), Ling Zhang, Dong Yang, Fausto Milletari, Ziyue Xu,
Xiaosong Wang, and Daguang Xu

NVIDIA, Bethesda, USA
{hroth,lingz,dongy,fmilletari,ziyuex,xiaosongw,daguangx}@nvidia.com

Abstract. Annotation of medical images has been a major bottleneck for the development of accurate and robust machine learning models. Annotation is costly and time-consuming and typically requires expert knowledge, especially in the medical domain. Here, we propose to use minimal user interaction in the form of extreme point clicks in order to train a segmentation model that can, in turn, be used to speed up the annotation of medical images. We use extreme points in each dimension of a 3D medical image to constrain an initial segmentation based on the random walker algorithm. This segmentation is then used as a weak supervisory signal to train a fully convolutional network that can segment the organ of interest based on the provided user clicks. We show that the network's predictions can be refined through several iterations of training and prediction using the same weakly annotated data. Ultimately, our method has the potential to speed up the generation process of new training datasets for the development of new machine learning and deep learning-based models for, but not exclusively, medical image analysis.

1 Introduction

The growing number of medical images taken in routine clinical practice increases the demand for machine learning (ML) methods to improve image analysis workflows. However, a major bottleneck for the development of novel ML-based models to integrate and increase the productivity of clinical workflows is the annotation of datasets that are useful to train such models. At the same time, volumetric analysis has shown several advantages over 2D measurements for clinical applications [1], which further increases the amount of data (a typical CT scan contains hundreds of slices) needing to be annotated in order to train accurate 3D models. However, the majority of annotation tools available today for medical imaging are constrained to annotation in multiplanar reformatted views. The annotator needs to either brush paint or draw boundaries around organs of interest, often on a slice-by-slice basis. Classical techniques like 3D region growing or interpolation tools can speed up the annotation process by starting from seed points or allowing the user to skip certain slices. However, their usability is often limited to certain types of structures and might not work well in general.

L. Zhou et al. (Eds.): LABELS 2019/HAL-MICCAI 2019/CuRIOUS 2019, LNCS 11851, pp. 42–50, 2019.
https://doi.org/10.1007/978-3-030-33642-4_5

Here, we propose to use minimal user interaction in form of extreme point clicks, together with iterative training and refinement. Starting from user-defined extreme points in each dimension of a 3D medical image, an initial segmentation is produced based on the random walker algorithm. This segmentation is then used as a weak supervisory signal to train a fully convolutional network that can segment the organ of interest based on the provided user clicks. We show that the network's predictions can be iteratively refined by using several iterations of training and prediction using the same weakly annotated data.

Related Work: Fully convolutional networks (FCNs) [2] have established themselves as the state-of-the-art methods for medical image segmentation in recent years [3–5]. However, a major drawback is that they are very data hungry, limiting their application in healthcare where data annotation is very expensive. In order to reduce the cost of labeling, semi-automated/interactive and weakly supervised methods have been proposed in the literature.

Building on recent advances in deep learning (DL), several methods have been proposed to integrate it with interactive segmentation schemes. DL has been used in [6] for the DeepIGeoS algorithms, which leverages geodesic distance transforms and user scribbles to allow interactive segmentation. Such a method does not exhibit robust performance when seeking segmentation for unseen object classes. An alternative method [7] uses image-specific fine-tuning and leveraging both bounding boxes and scribble-based interaction. In [8], the authors utilize point clicks that are modeled as Gaussian kernels in a separate input channel to a segmentation FCN in order to model user interactions via seed-point placing. Finally [9] proposes to use user-provided scribbles with random walks [10] and FCN predictions to achieve semi-automated segmentation of cardiac CT images. This method differs from our proposed method in that we only expect the user to provide extreme points rather than scribbles as initial input to the random walker algorithm and uses a different approach when iteratively refining the segmentations.

One of the first approaches using bounding box based weakly supervised training of deep neural networks in medical imaging was by [11]. They used a patch-based classification CNN to segment brain and lung regions using an initial *GrabCut* segmentation. After several rounds of predictions using CNN plus Dense CRF post-processing, the network's segmentation performance could be improved. Weakly-supervised or self-learning in medical image analysis can also make use of measurements readily available in the hospital picture archiving and communication system (PACS) such as measurements acquired during evaluation of the RECIST criteria [12]. However, these measurements are typically constraint to 2D and might miss adequate constraints for more complex three-dimensional shapes. In [13], unsupervised segmentation results are used to train a deep segmentation network on cystic lung regions, again in a slice-by-slice fashion. This approach might work well for certain organs, like the lungs, where an unsupervised technique can have good enough initial performance due to the good image contrast. However, completely unsupervised techniques might fail to generalize to organs where the boundary information is not as clear.

More recently, [14] introduced inequality constraints based on target-region size and image tags in the loss function of a CNN in order to train the network for weakly supervised segmentation.

2 Method

In this work, we approach initial interactive segmentation using user-provided clicks on the extreme points of the organ of interest. The overall proposed algorithm for weakly supervised segmentation from extreme points can be divided into the following steps which are detailed below:

1. Extreme point selection
2. Initial segmentation from scribbles via random walker algorithm
3. Segmentation via deep fully convolutional network
4. Regularization using random walker algorithm

Steps 2, 3, and 4 will be iterated until convergence. Here, convergence is defined based on the differences between two consecutive rounds of predictions as in [13].

1. Extreme point selection: Defining extreme points on the organ surface will allow the extraction of a bounding box around the organ (plus some padding $p = 20$ mm in all our experiment). Bounding box selection significantly reduces the image content that the 3D FCN has to analyze and simplifies the machine learning problem, as previous work on cascaded approaches has shown [15]. Bounding boxes and extreme points on objects have been widely studied in the computer vision literature [16]. Bounding boxes have a practical disadvantage in that the user often has to select the corners of bounding boxes that lie outside the object of interest. This is especially tricky to do for three-dimensional objects where the user typically has to navigate three multi-planar reformatted views (axial, coronal, sagittal) in order to achieve the task. Recent studies have also shown the time savings using extreme point selection brings for 2D object selection instead of traditional bounding box selection [16,17]. At the same time, extreme points provide additional information to the segmentation model (which can be observed in our experimental section, Table 1. They lie on the object surface and we model them as an additional input channel together with the image intensities. This extra channel includes 3D Gaussians G centered on each point location clicked by the user. This approach is similar to [16] but here we extended it to 3D medical imaging problems.

Figure 1 illustrates our approach. We ask the user to click on six extreme points (here four are shown in axial view) that describe the largest extent of the organ. These points are then used to compute a bounding box B automatically, including some padding p.

2. Initial segmentation from scribbles via random walker algorithm: In order to make use of extreme point clicks as a weak supervision signal, we turn them into a probability map $\hat{Y}$ than can act as a pseudo dense label map for driving a 3D FCN to learn the segmentation task. Based on the initial set

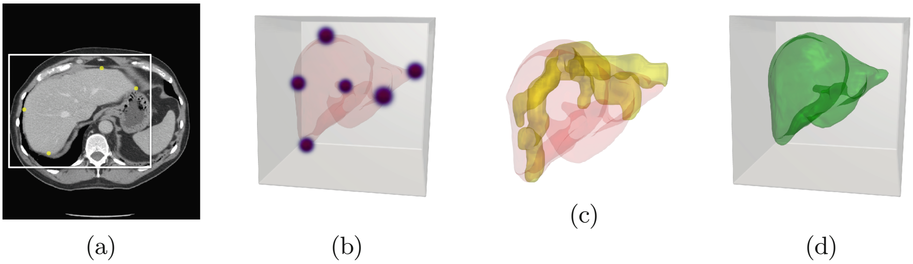

Fig. 1. Our weakly supervised segmentation framework. (a) The user selects extreme points that define the organ of interest (here the liver) in 3D space. (b) Extreme points are modeled as Gaussians in an extra image channel which is fed to a 3D segmentation model. (c) Foreground scribbles are generated automatically to initialize random walker (the ground truth surface is shown in red for reference). (d) Model returns the segmentation results.

of extreme points, we compute a set of foreground and background scribbles that act as the input seeds for the random walker algorithm [10]. We compute Dijkstra's shortest path [18] between each extreme point pair along each image dimension, where we model the distance between neighboring voxels by their gradient magnitude $D = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 + \left(\frac{\partial f}{\partial z}\right)^2}$. Here, the shortest path result can be seen as an approximation of the geodesic distance [6] between the two extreme points in each dimension. Figure 1 shows the foreground scribbles used as input seeds to the random walker algorithm. In order to increase the number of foreground seeds, each path is also dilated with a 3D ball structure element of $r_{\text{foreground}} = 2$. The background seeds are defined as the dilated and inverted version of the input scribbles. While the amount of dilation does depend on the size of the organ of interest, we typically dilate with a ball structure element of radius $r_{\text{background}} = 30$ which achieves good initial seeds for organs like spleen, and liver.

Next, the random walker algorithm [10] is used to generate an initial prediction map $\hat{Y}$ based on the background s_0 and foreground s_1 scribbles described above. The random walker basically solves the diffusion equation between voxels defined as source and sink as defined by the scribbles S. Here, the 3D volume is defined as a graph $G(E, V)$ with edges $e \in E$ and vertices $v \in V$. The edge between two vertices v_i and v_j is denoted as e_{ij} and can be assigned a weight w_{ij} based on the image intensities gradients. Furthermore, the degree of a given vertex is defined by $d_i = \sum w_{ij}$. We solve the diffusion equation in order to get a probability $p(\omega|x) = x_j^{\omega}$ for each vertex v_i to belong to the foreground class ω_1. Here, L is the Laplacian of the weighted image graph G with each element of the matrix defined as:

$$L_{ij} = \begin{cases} d_i, & \text{if } i = j, \\ -w_{ij}, & \text{if } i \text{ and } j \text{ are adjacent voxels}, \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The weights between adjacent voxels are defined as $w_{ij} = e^{-\beta|z_j - z_i|^2}$ to make diffusion between similar voxel intensities z_i and z_j easier. While β is a tunable hyperparameter that controls the amount of diffusion, we keep it fixed at $\beta = 130$ in all our experiments.

3. Segmentation via deep fully convolutional network: Next, given all pairs of images X and pseudo labels $\hat{Y}$, we can train a fully convolutional neural network to segment the given foreground class, with $P(X) = f(X)$. Our network architecture of choice follows the encoder-decoder network proposed in [19], utilizing an-isotropic $(3 \times 3 \times 1)$ kernels in the encoder path in order to make use of pretrained weights from 2D computer vision tasks. As in [19], we initialize from *ImageNet* pretrained weights using a ResNet-18 encoder branch. While the initial weights are learned from 2D, all convolutions are still applied in a full 3D fashion throughout the network, allowing it to efficiently learn 3D features from the image. The Dice loss [4] has been established as the objective function of choice for medical image segmentation tasks. Its properties allow automatic scaling to unbalanced labeling problems. At the same time, it also naturally adapts to the comparing probability maps without any modifications to the original formulation:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} \hat{y}_i^2} \tag{2}$$

Here, y_i is the predicted probability from our network f and $\hat{y_i}$ is the weak label probability from our pseudo label map $\hat{Y}$ at voxel i.

4. Regularization using random walker algorithm: We could stop our learning after the segmentation network f above is trained on the pseudo labels $\hat{Y}$. However, we notice that an additional regularization step by an additional random walker segmentation as described above can be very beneficial to the convergence of our weakly-supervised segmentation approach. This finding is similar in spirit to [11], where a DenseCRF is utilized after each round of CNN training in order to introduce regularization to the segmentation output. In order to increase the amount of regularization the random walker can bring to the network's predictions, we add an area of uncertainty by eroding the foreground prediction $P(X) >= 0.5$ and eroding the background $P(X) < 0.5$ both with a ball structure element of radius $r_{\text{randomwalker}} = 4$ in all our experiments. This allows the random walker to produce new predictions around the boundary of the foreground object that differ from the previous 3D FCN predictions and in turn, help the next iteration to learn new features from the same set of training images, and not to get stuck in a local optimum. In fact, we notice that without this step, our weakly supervised segmentation framework becomes unstable and does not easily converge to a satisfying performance.

3 Experiments and Results

Datasets: We utilize the training datasets (as they include ground truth annotations) from public challenges, specifically, from the *Medical Segmentation*

Decathlon[1] and the *Challenge on Endocardial Three-dimensional Ultrasound Segmentation*[2]. All numbers are reported on 1 mm isotropic images that were generated from the original images using linear interpolation for both CT and MRI images. For ultrasound images, we keep their original resolution as they are close to isotropic. We employ random splits for training and validation for all datasets, resulting 32/9 cases for spleen (CT), 104/27 cases for liver (CT), 26/6 cases for prostate (MRI), and 24/6 cases for left ventricle (LV) in ultrasound (US).

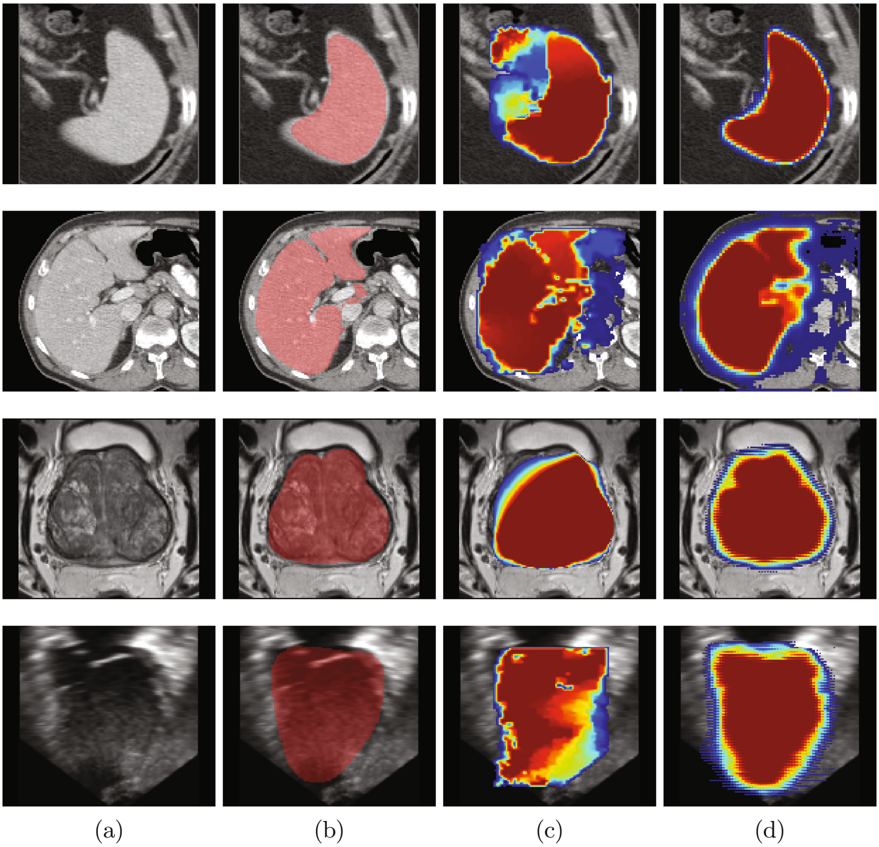

Fig. 2. Our results. We show (a) the image, (b) overlaid (full) ground truth (used for evaluation only), (c) initial random walker prediction, and (d) our final segmentation result produced by the weakly supervised FCN. We show qualitative results for top to bottom: spleen (CT), liver (CT), prostate (MRI), and left ventricle (US) segmentation.

Experiments: In all cases, we iterate our algorithm until convergence on the validation data. We compare both training with and without employing random walker (RW) regularization after each round of 3D FCN training. Furthermore,

[1] http://medicaldecathlon.com.
[2] https://www.creatis.insa-lyon.fr/Challenge/CETUS/.

we quantify the benefit of modelling the extreme points as an extra input channel to the network by running the framework with RW regularization but without the extreme points channel. The results are summarized in Table 1 for all segmentation tasks. It can be observed that the biggest improvements happen in the first round FCN learning after initial random walker segmentation. While random walker regularization does not always improve the average Dice score, it does help to introduce enough "novelty" into our learning framework in order to drive the overall Dice score up in later iterations as shown in Fig. 3. Visual examples of the improvement between from initial random walker to the final FCN prediction is shown in Fig. 2.

Implementation: The training and evaluation of the deep neural networks used in the proposed framework were implemented based on the *NVIDIA Clara Train SDK*[3] using NVIDIA Tesla V100 GPUs with 16 GB memory.

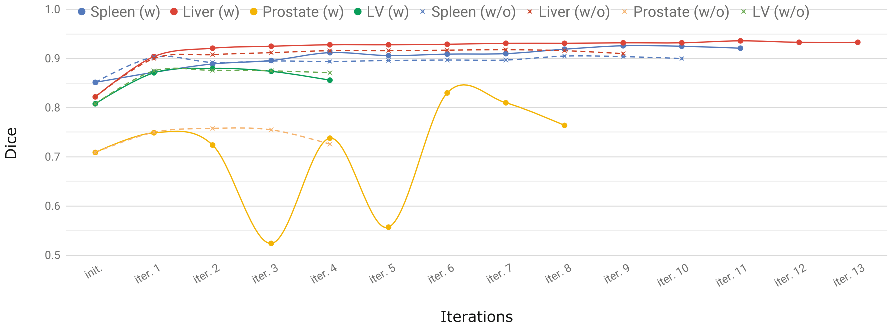

Fig. 3. Weakly supervised training from scribble based initialization. Each segmentation task is shown with (w) and without (w/o) random walker regularization after each round of FCN training.

Table 1. Summary of our weakly supervised segmentation results. This table compares the random walker initialization with weakly supervised training from extreme points with (w) and without (w/o) random walker (RW) regularization, and with RW regularization but without the extra extreme points channel as input to the network (w RW; no extr.). For reference, the performance on the same task under fully supervised training is shown.

Dice	Spleen (CT)	Liver (CT)	Prostate (MRI)	LV (US)
Rnd. walk. init.	0.852	0.822	0.709	0.808
Weak. sup. (w/o RW)	0.905	0.918	0.758	0.876
Weak. sup. (w RW; no extr.)	0.924	0.935	0.779	0.860
Weak. sup. (w RW)	**0.926**	**0.936**	**0.830**	**0.880**
Fully supervised	*0.963*	*0.958*	*0.923*	*0.903*

[3] https://devblogs.nvidia.com/annotate-adapt-model-medical-imaging-clara-train-sdk.

4 Discussion and Conclusions

We presented a method for weakly supervised 3D segmentation from extreme points. Asking the user to select the organ of interest using simple point clicks on the organ's surface in each spatial dimension can reduce the amount of labeling cost drastically. At the same time, the point clicks can describe the region of interest and simplify the machine learning task in 3D. Furthermore, the extreme points can be utilized to generate an initial weak pseudo label based on the extreme points utilizing the random walker algorithm. We found our initial label to be relatively robust to three diverse medical image segmentation tasks involving three different image modalities (CT, MRI, and ultrasound). Occasionally, the random walker can lack robustness for organs showing very diverse interior textures, like some advanced cancer patients in the prostate dataset. Here, a boundary search algorithm could potentially provide a better initial segmentation. Still, our FCN training in is able to markedly improve upon the initial segmentation. Previous work mainly utilized bounding box annotations for weakly supervised learning, e.g. [11]. However, we consider selecting extreme points on the organ's surface to be more natural then selecting corners of a bounding box outside the organ of interest and more efficient than adding scribbles inside and around the organ [6,9]. This is consistent to findings in the computer vision literature [17]. In the future, the region of interest and extreme point selection could be replaced by an automatic proposal network in order to further reduce the manual burden of medical image annotation.

References

1. Devaraj, A., van Ginneken, B., Nair, A., Baldwin, D.: Use of volumetry for lung nodule management: theory and practice. Radiology **284**(3), 630–644 (2017)
2. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, 3431–3440 (2015)
3. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
4. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision, pp. 565–571. IEEE (2016)
5. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
6. Wang, G., et al.: DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 1559–1572 (2018)
7. Wang, G., et al.: Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Trans. Med. Imaging **37**(7), 1562–1573 (2018)

8. Sakinis, T., et al.: Interactive segmentation of medical images through fully convolutional neural networks. arXiv preprint arXiv:1903.08205 (2019)
9. Can, Y.B., Chaitanya, K., Mustafa, B., Koch, L.M., Konukoglu, E., Baumgartner, C.F.: Learning to segment medical images with scribble-supervision alone. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 236–244. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_27
10. Grady, L.: Random walks for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 1768–1783(11) (2006)
11. Rajchl, M., et al.: DeepCut: object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans. Med. Imaging **36**(2), 674–683 (2017)
12. Cai, J., et al.: Accurate weakly supervised deep lesion segmentation on CT scans: Self-paced 3D mask generation from RECIST. arXiv preprint arXiv:1801.08614 (2018)
13. Zhang, L., Gopalakrishnan, V., Lu, L., Summers, R.M., Moss, J., Yao, J.: Self-learning to detect and segment cysts in lung ct images without manual annotation. In: ISBI, pp. 1100–1103. IEEE (2018)
14. Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ayed, I.B.: Constrained-CNN losses for weakly supervised segmentation. Med. Image Anal. **54**, 88–99 (2019)
15. Roth, H.R., et al.: Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Med. Image Anal. **45**, 94–107 (2018)
16. Maninis, K.K., Caelles, S., Pont-Tuset, J., Van Gool, L.: Deep extreme cut: From extreme points to object segmentation. In: CVPR, pp. 616–625 (2018)
17. Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V.: Extreme clicking for efficient object annotation. In: ICCV, pp. 4930–4939 (2017)
18. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische mathematik **1**(1), 269–271 (1959)
19. Liu, S., et al.: 3D anisotropic hybrid network: transferring convolutional features from 2D images to 3D anisotropic volumes. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 851–858. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_94