DeepScribble: Interactive Pathology Image Segmentation Using Deep Neural Networks with Scribbles

Sungduk Cho [1] *Hyungjoon Jang* [2] *Jing Wei Tan* [1] *Won-Ki Jeong* [1]

[1] Korea University, College of Informatics,
Department of Computer Science and Engineering, Seoul, Korea
[2] Ulsan National Institute of Science and Technology,
School of Electrical and Computer Engineering, Ulsan, Korea

ABSTRACT

Tumor segmentation is a challenging but crucial task in digital pathology for accurate diagnosis. Recent studies on deep neural networks have shown promising results in various image segmentation problems. However, unlike several medical imaging modalities (such as computerized tomography, CT and magnetic resonance imaging, MRI), the boundary between the normal and the tumor area in pathology images is usually fuzzy and ambiguous, making it difficult to adapt conventional image segmentation methods to these images. In this paper, we propose an interactive segmentation method that corrects the segmented boundaries from deep neural networks using user interaction. The proposed method functions in two stages; the first network initially generates the best prediction of the tumor boundary; the second network then refines the segmentation iteratively by exploiting the user scribble annotations on-the-fly. Our approach leverages the feature learning aspects of deep neural networks in the correction step to reduce user effort. We demonstrate the efficacy of the proposed method on real pathology images.

Index Terms— Interactive segmentation, pathology, convolutional neural networks

1. INTRODUCTION

Tumor region segmentation of digital histopathology whole slide images (WSI) is a challenging task in biomedical image processing and yet has significance in its accomplishment. Automated WSI segmentation could assist pathologists by readily providing the target lesions for further diagnosis. However, it is difficult to generate precise boundaries that classify the tissue into normal and malignant regions in a particular histopathology image. Such images show high similarities in color intensity and texture due to their biological nature as shown in Figure 1. For the highest accuracy and reliability, the segmentation should be performed by users with domain knowledge. However, even the experts take a significant amount of time in the process, as it is difficult to distinguish these regions by the naked eye. Interactive segmentation aims to integrate user annotations with the automated segmentation method to achieve segmentation of best quality with least user interaction.

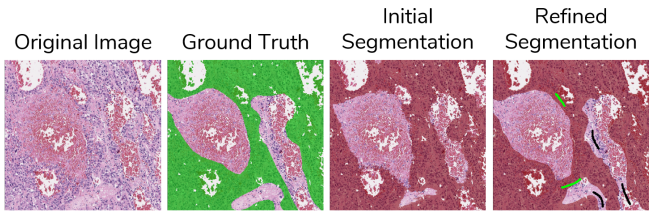

Fig. 1. Exemplary images of our method. The original image and the ground truth mask indicating the tumor region, the initial segmentation output, and the refined segmentation output with user scribbles on top are shown.

Prior to deep learning, conventional interactive segmentation algorithms [1, 2, 3, 4] exploited the characteristics of the image. For example, watershed transform [1] computes the gradients to detect the topographic features of the image. Grady [3] proposed an optimization-based approach in which a random walker starts from user-assigned labels to reach unlabeled pixels. However, conventional methods have not been effective in segmenting medical images compared to natural images owing to their innate features, such as biological traits.

With outstanding performance and recent applications in various fields, deep learning methods [5, 6, 7, 8] are actively used in the field of image segmentation. Long [5] proposed a fully convolutional network (FCN) method to perform pixel-wise predictions for semantic segmentation. U-Net [6] is a variant of a convolutional encoding–decoding network with skip connections and has shown exceptional performance in the segmentation of medical images. Recent works have integrated user interaction and deep learning to enhance the accuracy of deep neural networks. Xu [9] introduced the first deep learning-based interactive segmentation method. They transformed user clicks into two separate distance maps that were then passed into the network along with the RGB channels of the image. Jang [10] proposed a back-propagation refinement scheme, where a network updates the distance maps to match the user click intentions. [11] extended this idea by running

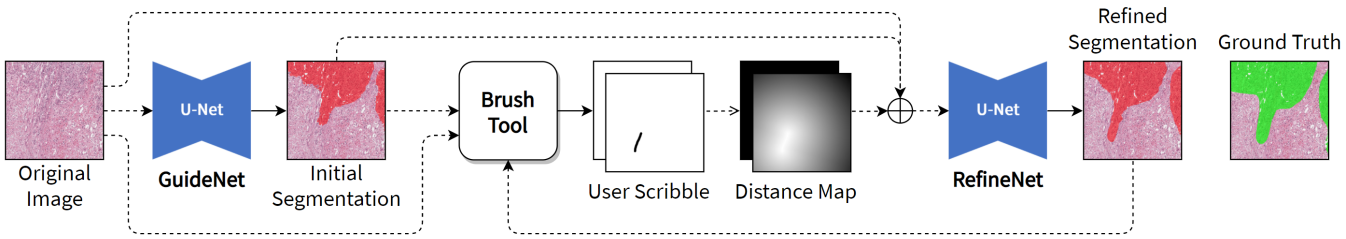

Fig. 2. Overview of proposed network structure. Input to GuideNet is the original image. Generated initial segmentation is then read by the user who uses the interactive Brush Tool to annotate seed points. The foreground and background distance fields converted from user scribbles, output from GuideNet, and the original image are the inputs to RefineNet. The user can iteratively refine the result until target accuracy is reached.

forward and backward passes for a small part of the network, leading to significant speed-up of the process.

In this study, we propose a network to perform segmentation in an interactive manner with given user corrections. The user scribbles are transformed into distance maps and fed into the network as a guide along with the initial prediction. The refinement process is performed iteratively until the target accuracy is met.

Our method takes advantage of the initial prediction output from the general segmentation model; the user then begins annotating near the mislabeled segmentation boundaries, thereby reducing the user burden of starting from scratch. Compared to state-of-the-art methods, we achieve comparable accuracy while achieving fast interaction time and demanding less user interaction. We compared our method with other interactive segmentation methods to show the efficacy of our pathology imaging method in terms of quality and speed.

2. METHOD

2.1. Data Description

We used the hepatocellular histopathology whole slide image (WSI) train dataset from the PAIP2019 challenge [12] (part of the MICCAI 2019 Grand Challenge for Pathology). The dataset is divided into two parts to train GuideNet, the initial segmentation model, and RefineNet, the refinement model. The WSI is scaled to the level of 5x magnification, resulting in approximately 100 to 200 million pixels per image. Patches of size 256×256 and 384×384 are extracted from the WSI. For RefineNet, the patches whose tumor area occupies less than 20% or more than 80% of the entire area is discarded to focus on the refinement process of segmentation near tumor boundaries leaving 5766 patches.

2.2. Proposed System

We used two deep neural networks, GuideNet and RefineNet, for automatic segmentation and refining, along with a brush tool to receive user interaction as shown in Figure 2.

2.2.1. GuideNet

The first network is a general segmentation model in which the input is a three-channel RGB image and the output is a pixel-wise probability map. The model is trained using the patches extracted from WSIs. The last layer outputs a probability map indicating the positive area of the tumor. This probability map is thresholded with a certain value ($\theta = 0.5$) to form a binary segmentation mask. The mask is considered as the tumor segmentation where the binary value 1 is positive for the tumor and 0 is negative for the tumor.

2.2.2. Scribble Input using a Brush Tool

The initial segmentation output is overlaid on the original image to be examined by the user expert. The user will draw scribbles to the mislabeled area. For foreground and background annotations, two single-channel matrices of the same size as the cropped patch are created. Given the binary prediction mask overlaid on the corresponding original image patch, the user looks for a mislabeled area in the segmentation. Once the user decides to correct the segmentation boundary, the scribble is drawn on the mislabeled region with the opposing seed type. The resulting user annotation will be the foreground scribble on the background region of the mislabeled area and vice versa. Next, matrices with scribbles are transformed to distance maps by computing their Euclidean distances. Distance maps are normalized to fit the input range, where greater magnitude indicates proximity to the seed points.

2.2.3. Automated Generation of Correction Scribbles for Training

To train the refinement network, it is required to have correction scribbles for each mislabeled region in the prediction. Therefore, we generated scribbles automatically that would resemble a human user. Given binary prediction, we computed difference with the ground truth label. Then, for each false positive and false negative regions, distance maps were generated. The distance map is thresholded to reduce its size

and skeletonized to generate scribble-like annotations. The aim is to duplicate the annotation behavior where the user generally draw on the center of the mislabeled region.

2.2.4. *RefineNet*

The generated distance maps from the brush tool, the initial probability map from GuideNet, and the three-channel RGB image are stacked into a six-channel input and fed into the second network. The model outputs a refined single-channel probability map as corrected segmentation. For further correction, the user can annotate additional scribbles that will be accumulated with previous scribbles. Only the distance maps are updated for every revision process. Once the target accuracy is met, the refinement process is completed.

2.3. Implementation Details

GuideNet and RefineNet are implemented separately using U-Net [6] with Efficient-Net [13] backbone. The first network, GuideNet, is trained with original WSI patches of shape (256, 256, 3). This network is deployed on a train dataset designated for the second network. The second network, RefineNet, is trained with the images composed of 1) the original WSI patch, 2) the segmentation output of GuideNet, and 3) the distance map generated from each user scribble data for foreground (i.e. tumor area) and background (i.e. non-tumor area) brush strokes, concatenated as an input of shape (384, 384, 6). We provided the ground truth masks as labels and employ binary cross-entropy loss.

We implemented an interactive toolkit in which the user can view the segmentation overlaid image. The user can annotate scribbles on the area to be corrected, using this brush tool. Segmentation is updated instantly after every user interaction.

3. RESULTS

3.1. Experiment

We compared our method with state-of-the-art deep learning-based interactive methods and conventional algorithms. We tested on 20 patches where corresponding initial prediction had a low mean intersection over union (IoU) score (under 75% and above 50%) to evaluate the refinement scheme. We computed the IoU score to measure the quantitative accuracy. Similar to the number of clicks (NoC) metric used in previous works [10, 11], we evaluated the number of interactions (NoI) metric, as our method is based on scribbles, a continuous line, instead of a single click or a point. We set the target IoU as 85% and 90% in a total of 20 interactions. For a fair comparison, we limited the size of scribbles to be less than 0.05% of the total pixels in one patch (< 80 pixels). Because f-BRS [11] and BRS [10] start with no initial prediction and the segmentation, result of the first click is considered as the initial prediction. In the case of conventional algorithms [1, 3] that require both foreground and background seeds for segmentation, we began the comparison after the second click. We recorded the ratio of the images that reached the IoU over 85% after 20 interactions. Finally, we measured the time taken for the response of each interaction as seconds per interaction (SPI) and for total interactions as time in seconds.

We used the available f-BRS code and trained our dataset with the same amount of input data and the same size. For the evaluation, we used the f-BRS-C version for f-BRS and the DistMap-BRS version as BRS. Our method and the conventional algorithms were tested on Windows 10 PC with RTX 2080Ti GPU, whereas f-BRS and BRS were tested on a Linux Server with equivalent RTX 2080Ti GPU.

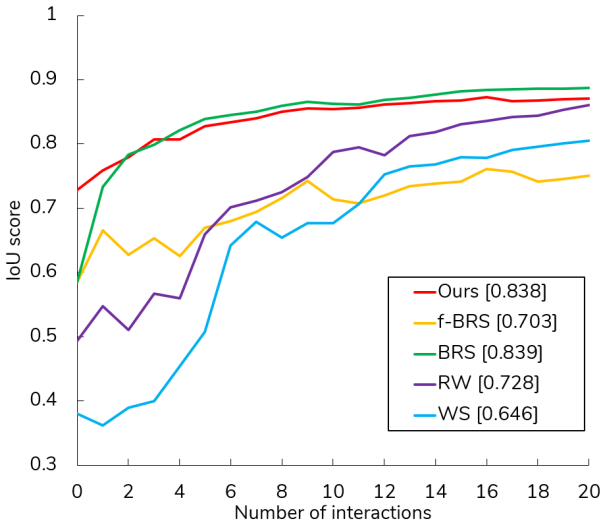

Fig. 3. Comparison of the IoU scores for each interaction; there are a total of 20 interactions. The legend shows average IoU score for each methods.

Table 1. Comparison of the NoI at 85% and at 90%, the ratio of images that reached an IoU of 85% after 20 interactions, total interaction time, seconds per interaction of various methods. The figures set in bold represent the best and the underlined figures represent the second best.

Method	≤ 20 @85	NoI @85	NoI @90	Time, s	SPI
WS [1]	0.5	8.1	11.1	**33**	0.7
RW [3]	<u>0.7</u>	7.1	6.5	47	1.4
BRS [10]	0.5	**4.5**	**4.7**	104	1.7
f-BRS [11]	0.4	<u>5.5</u>	6.6	39	**0.5**
Ours	**0.8**	6.4	<u>5.9</u>	<u>36</u>	<u>0.6</u>

3.2. Comparison

As shown in Figure 3, our method shows comparable accuracy to other methods. The methods in comparison con-

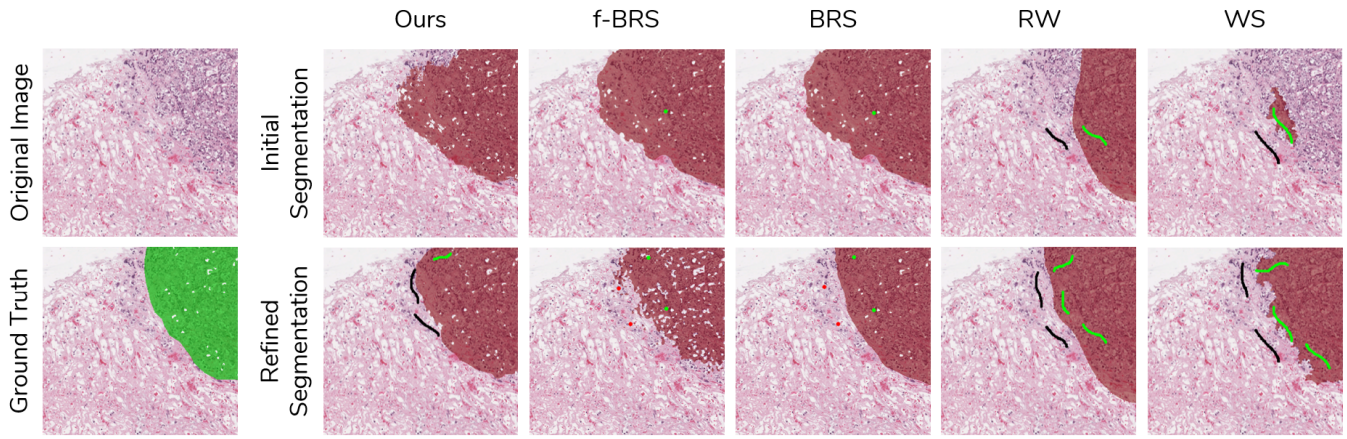

Fig. 4. Segmentation result of each method with equal number of user interactions; one interaction added to f-BRS and BRS, two interactions added to random walk and watershed as an initial prediction. Green and black represent foreground and background user input types, respectively.

verged steadily except for f-BRS. f-BRS aimed to reduce the response time by updating the portion of the network, which updates the global feature per click. As the segmentation encompasses the malignant cells, slight deviation on such points results in a global decrease in accuracy. The f-BRS and watershed (WS) tend to be strong at segmenting holes, whereas BRS and random walk (RW) tend to segment the area as a whole. Our method suggests a middle ground of both methods in a flexible manner. It is shown that BRS achieves the target IoU with fewer interactions. However, the success rate to reach the target IoU is lower, and it takes around three times the interaction time. This shows that our method matches the interaction latency of f-BRS and the segmentation accuracy of BRS, the best of both methods.

3.3. Discussion

Figure 4 visualizes the segmentation output of the compared methods. It takes a significant amount of time for the pathologist to annotate WSI. Therefore, the segmentation output focuses more on encompassing the tumor area as a whole than on pixel-wise separation to reduce the load. As shown in Figure 4, f-BRS has an advantage in distinguishing at the finer (pixel) level. Conversely, BRS has a tendency to produce more robust segmentation. Our method is shown to take the advantage of each method to a certain degree. Finally, starting with the output of general image segmentation affects the success rate of reaching the target IoU.

3.4. Ablation Study

We conducted an ablation study on alternative settings. We varied the input to our model to compare the effectiveness of each input. Each method corresponds to additive setting with the original image (I), the GuideNet prediction (P), the scribble without postprocessing (S), and the distance map of the scribble (D) as a concatenated input. The evaluation data were generated by the whole test set of 1242 patches with the corresponding scribbles automatically generated; the mean IoU score is computed for each setting. As shown in Table 2, our combined input achieves the highest IoU score. Moreover, it is worth noting that the initial prediction alone cannot improve the accuracy significantly, showing the need of user input for better refinement.

Table 2. RefineNet input ablation study

Method	Average IoU
GuideNet Prediction	0.862
RN: I	0.860
RN: I + P	0.870
RN: I + P + S	0.928
RN: I + P + D **(Ours)**	0.933

4. CONCLUSION

Our proposed method integrates user interaction and deep neural networks to efficiently segment pathology images. The process is performed iteratively with user-given annotation as scribbles, along with prior guidance for boundary adjustments. The comparison with other methods through quantitative and qualitative analysis showed comparable accuracy while reducing the user interaction time and demanding less user interaction. In the future, we aim to explore more sophisticated encoding of user input to incorporate into deep network and improve the segmentation quality.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by PAIP2019.

Ethical approval was required as confirmed by the license attached with the open access data.

6. ACKNOWLEDGEMENT

This work is partially supported by the Korea Health Industry Development Institute (HI18C0316), the National Research Foundation of Korea (NRF-2019M3E5D2A01063819), and the Institute for Information communications Technology Planning Evaluation (IITP-2021-0-01819).

7. REFERENCES

[1] Serge Beucher and Fernand Meyer, "The morphological approach to segmentation: the watershed transformation," *Mathematical morphology in image processing*, vol. 34, pp. 433–481, 1993.

[2] Caselles, Ron Kimmel, and Guillermo Sapiro, "Geodesic Active Contours," *International journal of computer vision*, vol. 22, no. 1, pp. 61–79, 1997.

[3] Leo Grady, "Random Walks for Image Segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 11, pp. 1768–1783, 2006.

[4] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," *ACM transactions on graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[5] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for Biomedical Image Segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[7] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[8] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11. Springer, 2018.

[9] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang, "Deep interactive object selection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 373–381.

[10] Won-Dong Jang and Chang-Su Kim, "Interactive Image Segmentation via Backpropagating Refinement Scheme," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5297–5306.

[11] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin, "f-BRS: Rethinking Backpropagating Refinement for Interactive Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8623–8632.

[12] Yoo Jung Kim, Hyungjoon Jang, Kyoungbun Lee, Seongkeun Park, Sung-Gyu Min, Choyeon Hong, Jeong Hwan Park, Kanggeun Lee, Jisoo Kim, Wonjae Hong, et al., "PAIP 2019: Liver Cancer Segmentation Challenge," *Medical Image Analysis*, p. 101854, 2020.

[13] Mingxing Tan and Quoc V Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *arXiv preprint arXiv:1905.11946*, 2019.