CGAM: CLICK-GUIDED ATTENTION MODULE FOR INTERACTIVE PATHOLOGY IMAGE SEGMENTATION VIA BACKPROPAGATING REFINEMENT

Seonghui Min   *Won-Ki Jeong* *

Department of Computer Science and Engineering, Korea University, Seoul, Korea

ABSTRACT

Tumor region segmentation is an essential task for the quantitative analysis of digital pathology. Recently presented deep neural networks have shown state-of-the-art performance in various image-segmentation tasks. However, because of the unclear boundary between the cancerous and normal regions in pathology images, despite using modern methods, it is difficult to produce satisfactory segmentation results in terms of the reliability and accuracy required for medical data. In this study, we propose an interactive segmentation method that allows users to refine the output of deep neural networks through click-type user interactions. The primary method is to formulate interactive segmentation as an optimization problem that leverages both user-provided click constraints and semantic information in a feature map using a click-guided attention module (CGAM). Unlike other existing methods, CGAM avoids excessive changes in segmentation results, which can lead to the overfitting of user clicks. Another advantage of CGAM is that the model size is independent of input image size. Experimental results on pathology image datasets indicated that our method performs better than existing state-of-the-art methods.

Index Terms— Interactive segmentation, digital pathology

1. INTRODUCTION

Segmenting tumor area in whole-slide images (WSI) is an important task in digital pathology, as it serves as a basis for the diagnosis of a target lesion. However, the difference in visual features, including the color and texture of malignant and normal regions, is insignificant in particular histopathological images. Because of this innate biological property, even experts in this domain need considerable time to accurately distinguish these areas with the naked eye. In addition, it is difficult to capture precise boundaries for classifying malignant regions using conventional automated segmentation methods that mostly rely on edges. To this end, interactive segmentation modifies automated segmentation methods to enable user interactions [1]. This allows users to quickly obtain high-quality segmentation results by providing interactions that reflect their intentions.

Commonly used user-interaction types include bounding boxes [2, 3], scribbles [4, 5], and clicks [6, 7, 8, 9]. Among these, we specifically focused on click-based interactive segmentation in which users provide positive/negative clicks to differentiate between foreground and background regions. Prior to deep learning, conventional approaches [3, 4, 5] considered interactive segmentation as an optimization problem. Because semantic information has not been fully exploited with many built-in heuristics, these methods require large amounts of user interactions. Deep learning-based interactive segmentation methods [9, 10] improve the segmentation accuracy of deep neural networks [11, 12] by incorporating user interactions. While showing outstanding performance compared to conventional methods, existing deep learning-based interactive segmentation methods rely heavily on high-level semantic priors and perform poorly for object classes not seen during training.

Recently, backpropagating refinement scheme (BRS) [6] addressed this issue by integrating optimization-based and deep learning-based methods. BRS sets the interaction maps entered into the network as trainable parameters. By backpropagating the loss calculated by prediction and user clicks, BRS fine-tunes the interaction maps in an online manner. Feature backpropagating refinement scheme (f-BRS) [7] is an improvement to the previous method in terms of inference time and computational budget by inserting a set of auxiliary parameters after the intermediate network layer for optimization. Backpropagation only through a subpart of the network improves the efficiency of f-BRS. In a follow-up study, Lin *et al.* [8] proposed generalized backpropagating refinement scheme (G-BRS), advanced layer architectures that enable more delicate refinement. However, the above methods optimize additional modules only through the minimization of loss calculated by limited user interactions, causing unwanted overall changes owing to overfitting.

In this study, we propose a new click-guided attention module (*CGAM*) for BRS-based interactive segmentation. CGAM addresses the overfitting issue of existing BRS-based methods by directly receiving click maps and a feature map from where it is inserted as a module input and utilizing them for optimization. CGAM enforces the desired specification on
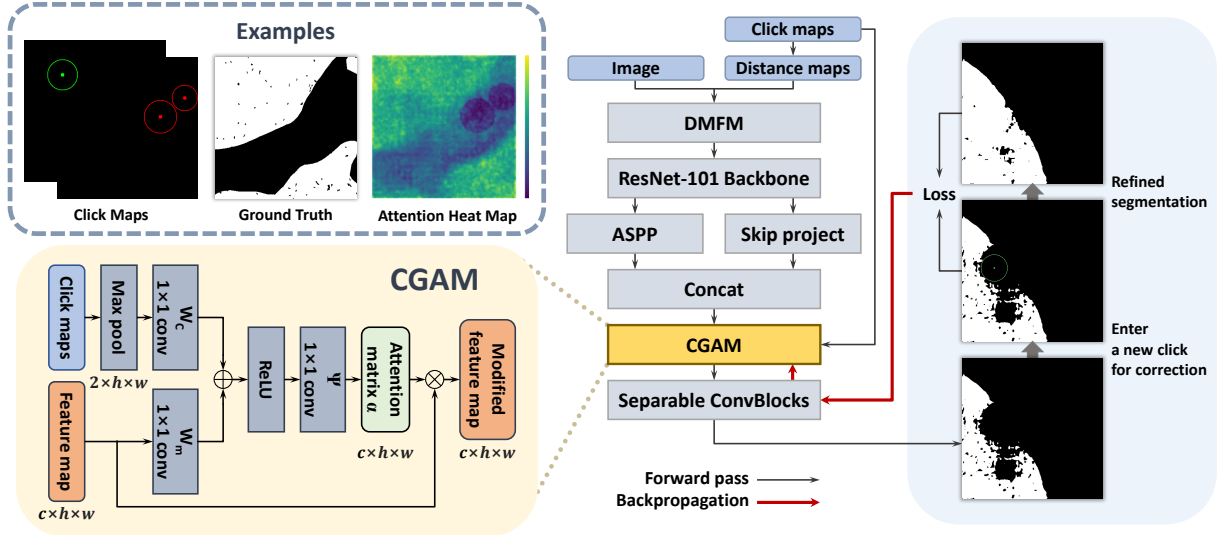
---

*Corresponding author: wkjeong@korea.ac.kr

Fig. 1. Overview of the proposed method, and the structure of CGAM. The user inputs a click where correction is required in the segmentation result. CGAM takes click maps and an intermediate feature map as input. It weights the feature map to highlight the target region of segmentation. By backpropagating the loss computed from user-provided clicks and output prediction, the parameters of CGAM are updated to produce an improved segmentation result. This process is repeated each time the user provides a new click. Red and green represent positive and negative clicks, respectively. The attention radius is visualized with a circle.

the result of a deep learning model by restricting the feature space with self-attention and the additional guidance of click maps. Furthermore, in contrast to G-BRS, the model size of CGAM is independent of the input image size, allowing the method to easily handle large-scale images. We demonstrate the segmentation performance of CGAM over existing deep learning-based interactive segmentation refinement methods on the PAIP2019 challenge dataset [13].

2. METHOD

2.1. Architecture Overview

An overview of the proposed method is shown in Fig. 1. To compare the proposed model, f-BRS, and G-BRS directly, we chose the standard DeepLabV3+ with ResNet-101 containing a distance maps fusion module (DMFM) proposed in [7] as the basis architecture. As in [7, 8], CGAM is inserted after the atrous spatial pyramid pooling (ASPP) layer in the DeepLabV3+ decoder. CGAM modifies the feature map of the corresponding location by receiving the feature map and click maps as inputs. The output logits are generated as the modified feature map passes through the rest of the network. In this study, we define interactive segmentation as an optimization problem and solve it with respect to the parameters of CGAM. Thus, the optimization loss is computed from the output logits and user-provided clicks. By backpropagating the loss, CGAM is updated for better segmentation performance in an online manner.

Assume f is a function implemented by the basis network. With input image I and click maps C, the intermediate feature map of the location where CGAM is inserted is defined as $g(I, C)$. Using h to denote a function that the network head implements, f can be represented as $f(I, C) := h(g(I, C))$. We express the whole process $\hat{f}$ with CGAM parameterized by θ inserted as follows:

$$\hat{f}(I, C; \theta) = h(\theta(g(I, C), \ C)). \tag{1}$$

We define a set of user-provided clicks as $\{(u_i, v_i, l_i, r_i)\}_{i=1}^N$ where (u, v), $l \in \{-1, 1\}$, and r represent the coordinates, label, and radius of each click, respectively. $M \in \{0, 1\}^{H \times W}$ is a binary mask of the newest click and selects the region outside of r. The optimization problem is formulated as a minimization for the following loss similar to that of [8]:

$$\mathcal{L}_t(I, C_t) = \min_{\theta_t} \mathbb{E}_{i \in [1,t]} [max(l_i - \hat{f}(I, C_t; \theta_t)_{u_i, u_v}, 0)]^2$$
$$+ \lambda \|M \odot (\hat{f}(I, C_t; \theta_{t-1}) - \hat{f}(I, C_t; \theta_t))\|_2^2, \tag{2}$$

where $\odot$ is Hadamard product, and $t \in [1, N]$ is an interaction step. The first term enforces the correct output segmentation corresponding to the user-provided clicks, and the second term prevents excessive modification to avoid overfitting. The scaling constant λ regulates the trade-off between the two terms.
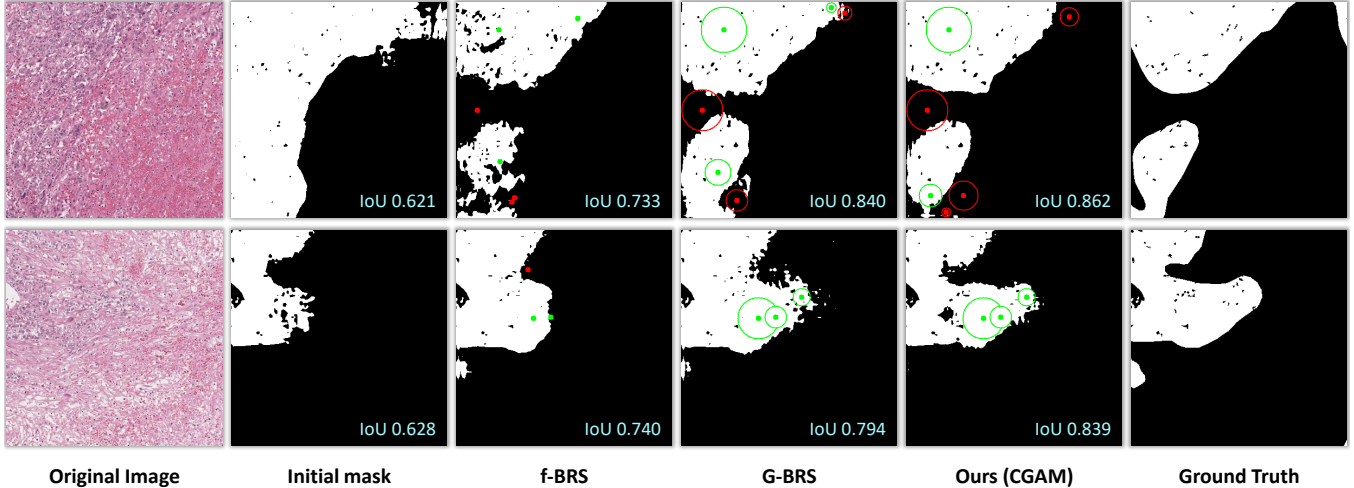
Fig. 2. Comparison of various BRS-based interactive segmentation results. Each method used the same number of clicks.

2.2. Click-Guided Attention Module

CGAM is a self-attention module that specializes in interactive segmentation, inspired by self-attention methods [14, 15]. By receiving assistance from the additional guidance of click maps, CGAM highlights feature responses in regions reflecting user intention. Fig. 1 left illustrates the pipeline of CGAM. By denoting the input feature map as $g(I, C) = m \in \mathbb{R}^{c \times hw}$, attention matrix $\alpha \in \mathbb{R}^{c \times hw}$ is obtained as follows:

$$\alpha = \psi^T(ReLU\left(W_C^T C_d + W_m^T m\right)), \qquad (3)$$

where $C_d \in \mathbb{R}^{2 \times hw}$ represents click maps downsampled to the resolution of m. The linear transformations with weight matrices $W_C \in \mathbb{R}^{2 \times \frac{c}{2}}$, $W_m \in \mathbb{R}^{c \times \frac{c}{2}}$, and $\psi \in \mathbb{R}^{\frac{c}{2} \times c}$ are implemented as a 1×1 convolution. The output of CGAM, a modified feature map $\hat{m} \in \mathbb{R}^{c \times hw}$, is then finally calculated as follows:

$$\hat{m} = m \odot \alpha. \qquad (4)$$

CGAM preserves the initial behavior of the network before it learns through backpropagation by setting the initial value of α to one.

It can be observed from the attention heat map in Fig. 1 that CGAM focuses on important regions by exploiting information in the click map. The attention matrix assigns element-wise weights to the feature map, which enables local refinement. This operation sets the CGAM free from the dependency of the input image size on G-BRS.

3. EXPERIMENT

3.1. Data Description

The whole-slide image (WSI) dataset used in our experiment was from the PAIP2019 challenge [13]. After scaling at $5\times$ magnification, because interactive refinement of segmentation results is mainly required at the boundary of the tumor, patches with tumor areas accounting for 20% to 80% of the total area were considered boundary regions and extracted.

3.2. Implementation Details

We trained our network on the pathology dataset with 5190 patches using the normalized focal loss proposed in [16]. We sampled the clicks during training following the standard procedure of [9]. The maximum number of clicks per image was set as 20, limiting the number of positive and negative clicks to less than 10. We used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and trained the networks for 120 epochs. We set the learning rate as 5×10^{-4} for the first 100 epochs, and 5×10^{-5} for the last 20 epochs.

For the inference time optimization, we also used the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We performed back-propagation for 20 iterations for each click. We set the learning rate as 5×10^{-2} and λ as 1. We conducted an experiment on a Windows 10 PC equipped with an NVIDIA RTX 3090 GPU.

3.3. Evaluation Protocol

For fair comparisons, we used the automatic click generation strategy proposed in [8]: The class of the following click was determined based on whether the dominant prediction error type was false positive or false negative. The click was placed at the point where the corresponding error region had the maximum Euclidean distance from its boundary. The distance was set as the radius of the click. This process was repeated until the target Intersection over Union (IoU) or the maximum number of clicks was reached.
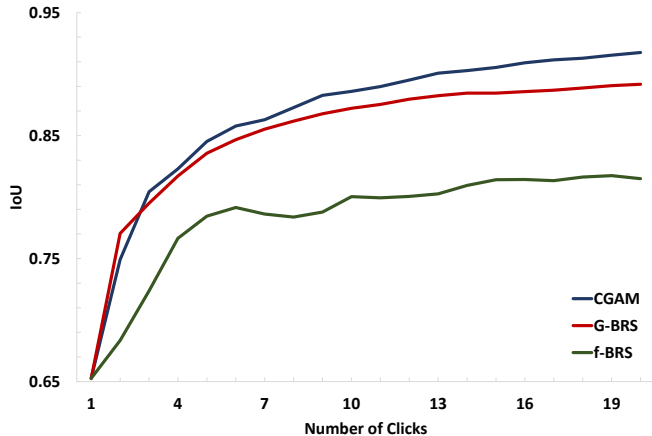
Fig. 3. Comparison of the IoU scores with respect to the number of clicks added by a user.

Table 1. Evaluation results of pathology image dataset. The best results are in bold.

Method	NoC@85	NoC@90	NoF@85	NoF@90	SPC (s)	Time (s)
f-BRS	11.59	15.43	55	89	0.053	107
G-BRS	7.70	12.67	23	58	**0.049**	81
CGAM	**6.38**	**11.32**	**8**	**36**	0.052	**77**

3.4. Evaluation Metrics

We set the target IoU as 85% and 90%. We limited the maximum number of clicks to 20. We reported the mean number of clicks (NoC) required to achieve the target IoU. We reported the number of failures (NoF) indicating the number of cases in which the target IoU was not reached with the maximum number of clicks. We reported the second per click (SPC) to measure the response time for each click. Finally, we reported the total time required to process the entire dataset.

4. RESULTS

4.1. Comparison

We evaluated 131 patches of WSI whose initial predictions had IoU scores between 50% and 70%. We compared CGAM with f-BRS and G-BRS, state-of-the-art BRS-based methods. For G-BRS, we selected the G-BRS-bmconv layer with the best performance reported in [8].

Table 1 presents the average NoC, NoF, SPC, and total time of the three methods for target IoUs of 85% and 90%. CGAM outperformed the other methods in nearly all metrics. The NoC results show that users can obtain satisfactory segmentation masks with less effort using CGAM. From the NoF results, as compared to the other methods, it is observed that CGAM successfully reached the target IoU in most cases. For

Table 2. Ablation study of click map guidance. In each method, CGAM takes proper click maps, random click maps, and zero tensor as input.

Method	NoC@85	NoC@90	NoF@85	NoF@90	SPC (s)	Time (s)
Proper	**7.18**	**11.82**	**11**	**36**	**0.050**	**77**
Random	7.31	12.43	19	53	0.060	95
Zero	7.37	12.23	15	48	0.060	95

speed-related metrics, the SPC of CGAM was slightly slower than that of G-BRS. However, CGAM reduced the total time required to reach the target IoU with fewer clicks.

4.2. Ablation Study

We conducted an ablation study to assess how click maps contribute to the performance of CGAM. We tested the following three scenarios; The first scenario received clicks on the appropriate coordinates and classes, generated by an automatic click-generation strategy described in Subsection 3.3. The second scenario assumed inappropriate (incorrect) clicks by generating clicks on random coordinates and classes. The third scenario assumed that no clicks were provided using a zero tensor with the same shape as that of the click maps. As shown in Table 2, CGAM achieved the best performance in all metrics when proper click maps were provided. Considering that the results of random clicks are worse than those of zero-tensor case, we can confirm that CGAM leverages the information in the click maps for optimization.

4.3. Discussion

CGAM outperformed other methods in terms of both accuracy and time. In particular, the model size of CGAM is fixed and constant regardless of the input size, unlike G-BRS, which expands linearly in proportion to the height and width of the input image; this decreases the number of parameters from 93k to 84k even for small images with 400×400 pixels. These results show the efficiency of CGAM and further demonstrate its potential for extending to multi-scale and large image segmentation tasks (e.g., segmentation of WSI).

5. CONCLUSION

In this study, we proposed CGAM for interactive image segmentation through back-propagating refinement. Exploiting the information in click maps by using it as input, CGAM increases the utility of user-provided clicks in interactive segmentation tasks. Experiments showed the improved performance of CGAM in pathology image segmentation as compared to other state-of-the-art methods. In future work, we plan to extend the current framework to address entire WSI such that it can be flexibly applied in real-world situations.

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by PAIP2019. Ethical approval was not required as confirmed by the license attached with the open access data.

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] Sungduk Cho, Hyungjoon Jang, Jing Wei Tan, and Won-Ki Jeong, "DeepScribble: Interactive Pathology Image Segmentation Using Deep Neural Networks with Scribbles," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 761–765. 1

[2] Jiajun Wu, Yibiao Zhao, Jun-Yan Zhu, Siwei Luo, and Zhuowen Tu, "MILCut: A Sweeping Line Multiple Instance Learning Paradigm for Interactive Image Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 256–263. 1

[3] 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. 1

[4] Leo Grady, "Random Walks for Image Segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 11, pp. 1768–1783, 2006. 1

[5] Varun Gulshan, Carsten Rother, Antonio Criminisi, Andrew Blake, and Andrew Zisserman, "Geodesic Star Convexity for Interactive Image Segmentation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3129–3136. 1

[6] 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. 1

[7] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin, "f-BRS: Rethinking Backpropagating Refinement for Interactive Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8623–8632. 1, 2

[8] Fanqing Lin, Brian Price, and Tony Martinez, "Generalizing Interactive Backpropagating Refinement for Dense Prediction Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 773–782. 1, 2, 3, 4

[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. 1, 3

[10] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool, "Deep Extreme Cut: From Extreme Points to Object Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 616–625. 1

[11] 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. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. 1

[13] 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*, vol. 67, pp. 101854, 2021. 2, 3

[14] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., "Attention U-Net: Learning Where to Look for the Pancreas," *arXiv preprint arXiv:1804.03999*, 2018. 3

[15] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803. 3

[16] Konstantin Sofiiuk, Olga Barinova, and Anton Konushin, "AdaptIS: Adaptive Instance Selection Network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7355–7363. 3