한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 29, No. 3, P. 127~135

ZoomISEG: 조직 병리학 전체 슬라이드 영상 분할을 위한 대화형 다중스케일 융합

민성희° 정원기*

고려대학교 컴퓨터학과

{qhrh01, wkjeong}@korea.ac.kr

ZoomISEG: Interactive Multi-Scale Fusion for Histopathology Whole Slide Image Segmentation

Seonghui Min ° Won-Ki Jeong*

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

요 약

조직병리에서 전체 슬라이드 영상의 정확한 분할은 질병 진단과 치료 계획에 매우 중요한 작업이다. 그러나 전체 슬라이드 영상은 크기가 크고 조직의 형태, 염색 및 촬영 조건이 다양하기 때문에 기존의 자동 영상 분할 알고리즘을 항상 적용하는 것은 어렵다. 최근 인간의 전문 지식과 알고리즘을 결합한 대화형 영상 분할 기술의 발전은 전체 슬라이드 영상 분할의 효율 성과 정확성을 개선할 수 있는 가능성을 보여주었다. 그러나 이러한 접근 방식은 동시에 어려운 과제를 제기하기도 했다. 본 논문에서는 다중 해상도 전체 슬라이드 영상을 활용하는 새로운 대화형 분할 방법인 ZoomISEG를 제안한다. 기존의 단일 스케일 방법과의 비교 및 ablation study를 통해 제안된 방법의 효율성과 성능을 입증한다. 실험 결과, 제안된 방법은 사람의 개입을 줄이면서도 최고 해상도 데이터를 사용하는 방식에 필적하는 정확도를 달성함을 확인했다.

Abstract

Accurate segmentation of histopathology whole slide images (WSIs) is a crucial task for disease diagnosis and treatment planning. However, conventional automated segmentation algorithms may not always be applicable to WSI segmentation due to their large size and variations in tissue appearance, staining, and imaging conditions. Recent advances in interactive segmentation, which combines human expertise with algorithms, have shown promise to improve efficiency and accuracy in WSI segmentation but also presented us with challenging issues. In this paper, we propose a novel interactive segmentation method, ZoomISEG, that leverages multi-resolution WSIs. We demonstrate the efficacy and performance of the proposed method via comparison with conventional single-scale methods and an ablation study. The results confirm that the proposed method can reduce human interaction while achieving accuracy comparable to that of the brute-force approach using the highest-resolution data.

키워드: 대화형 영상 분할, 디지털 병리학, 다중 해상도
Keywords: Interactive segmentation, Digital pathology, Multi-resolution

1. Introduction

WSI is commonly used in digital pathology for disease diagnosis and analysis. Pathologists use WSI to analyze and make critical

*Corresponding author: wkjeong@korea.ac.kr

*corresponding author: Won-Ki Jeong/ Department of Computer Science and Engineering, Korea University (wkjeong@korea.ac.kr)

Received : 2023.06.17./ Review completed : 1st 2023.06.30. / Accepted : 2023.07.05.
DOI : 10.15701/kcgs.2023.29.3.127
ISSN : 1975-7883(Print)/2383-529X(Online)

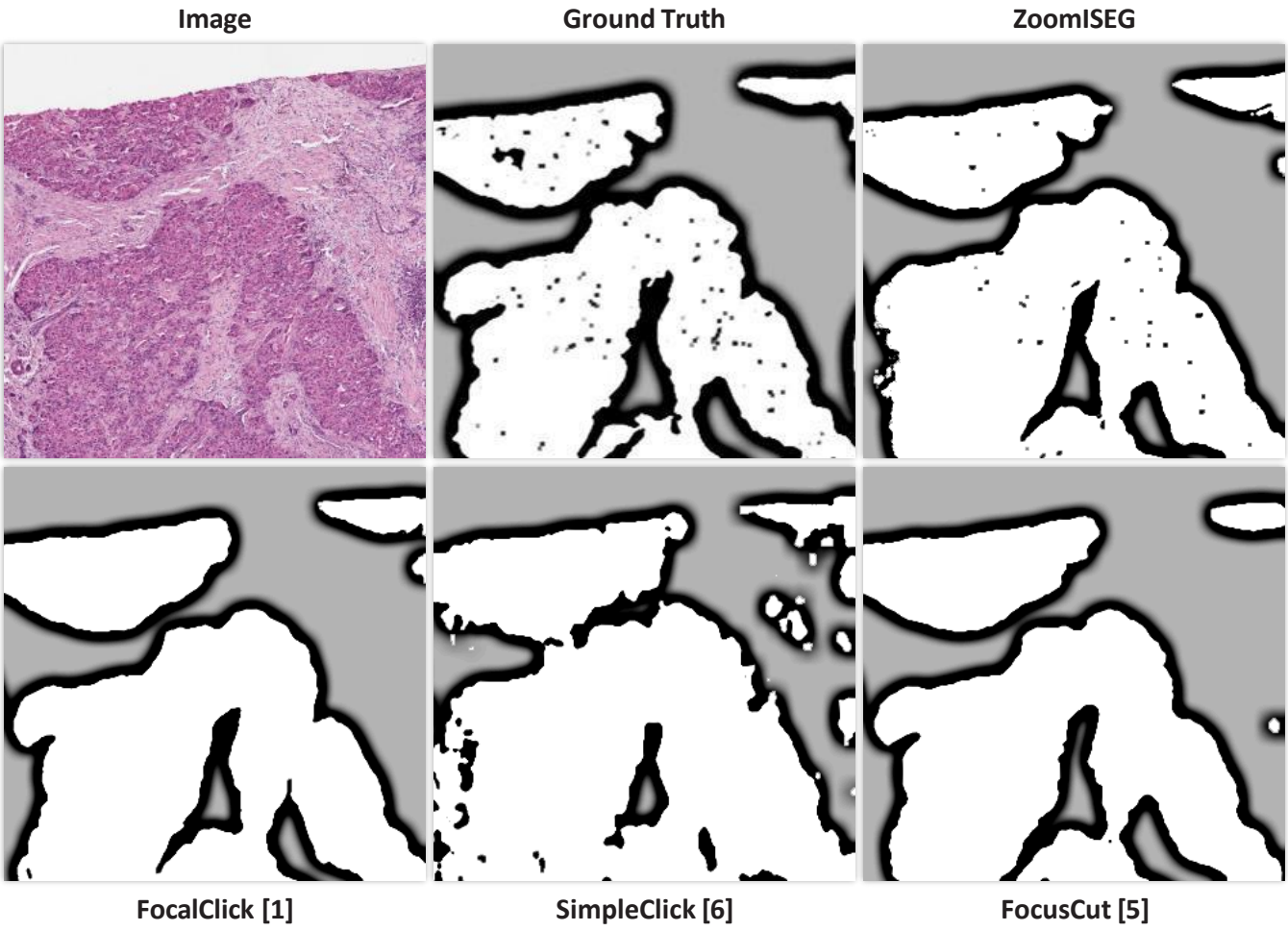

Figure 1: Comparison of segmentation masks generated by each method. With single-resolution WSIs, even the SOTA methods only generate robust masks that encompass tumor regions as a whole. In contrast, `ZoomISEG` satisfies pixel-wise separation of the tumor region required in ground truth by additionally utilizing high-resolution images in areas that require detail.

decisions based on the segmentation of tissue regions of interest (ROIs). Accurate segmentation is essential for the precise diagnosis of diseases such as cancer, as it can affect treatment plans and patient outcomes. However, due to the complexity and variability of tissue samples, automatic segmentation of WSI can be challenging. Conventional image segmentation algorithms [1, 2] mainly rely on edges, which may not work well on WSIs where regions are separated based on texture similarity. Leveraging deep learning segmentation methods [3] is also limited due to the large size of WSI and the limited computational resources.

The recent development of deep-learning-based interactive segmentation in the computer vision field [4, 5, 6] can be a promising solution to address the challenging issues in WSI segmentation, e.g., DeepScribble [7] and CGAM [8]. Interactive segmentation allows the user to interact with the segmentation algorithm by selecting or correcting the segmentation results, which improves the accuracy of automatic segmentation and provides pathologists with a more efficient and intuitive tool for analyzing WSI. However, the adaptation of existing interactive segmentation methods in WSI segmentation is still challenging. The state-of-the-art (SOTA) methods [4, 5, 6] incorporate additional focus views on ROIs or use modern architectures such as vision transformers. However, as shown in Fig. 1, they still may not capture all the necessary details required for accurate segmentation due to the use of *single-resolution* images. This is especially problematic for ROIs with complex and heterogeneous structures in WSIs, such as tumor margins or infiltrating immune cells.

In this paper, to overcome this limitation, we propose a novel interactive segmentation method, `ZoomISEG`, that can effectively utilize multi-resolution WSIs. We enable information transfer from an image of one resolution to an image of another resolution by using the mask-type input of the network. We prevent the issue of user interaction uncertainty in WSIs, where segmentation is difficult due to ambiguous boundaries, by adding a new loss term referred to as click loss to the network training process. We implemented an algorithm that mimics real user behavior to quantita-

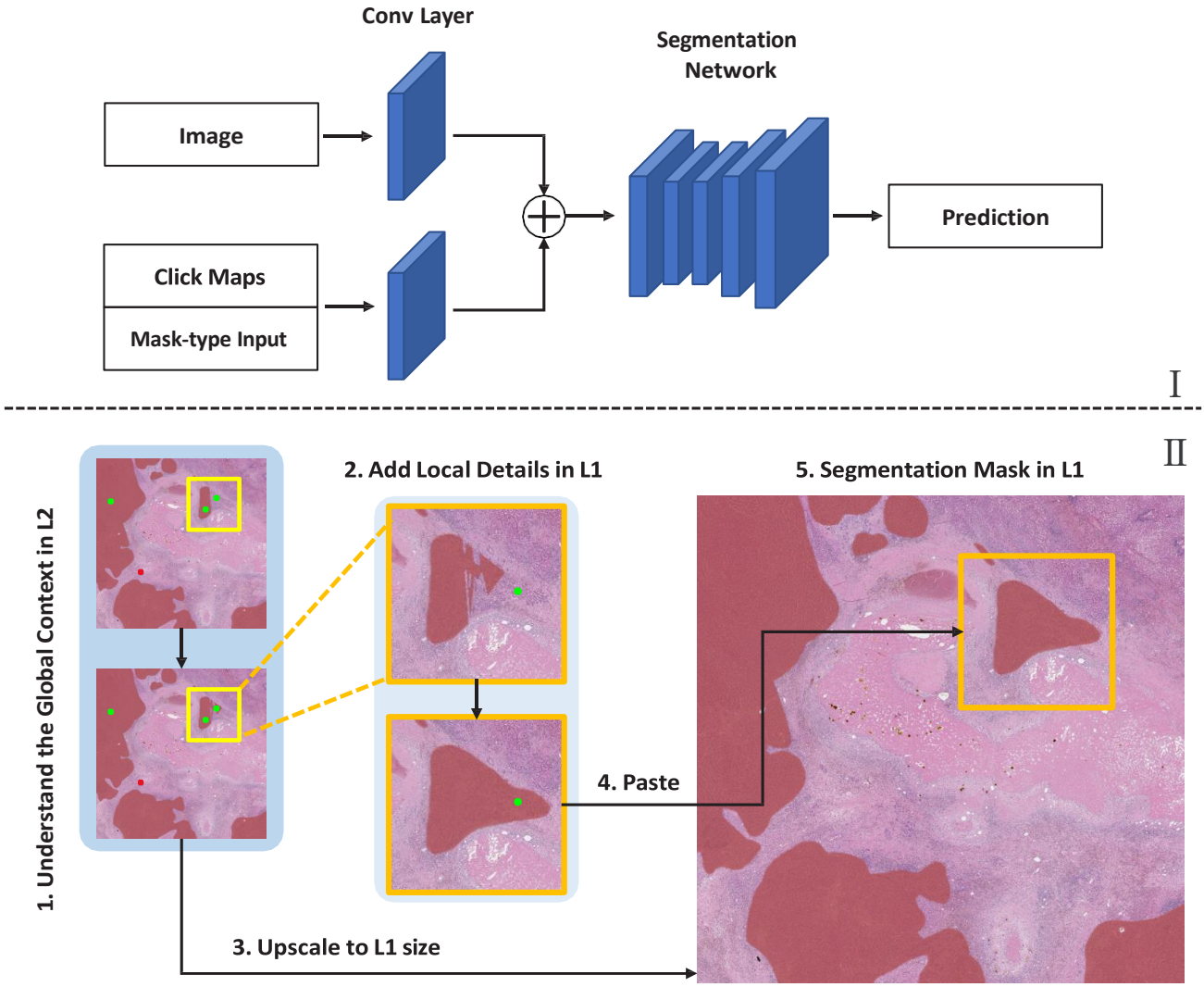

Figure 2: Overview of the proposed ZoomISEG. ZoomISEG deals with two different magnification levels of WSI. L1 refers to a magnification level of 20×, while L2 refers to a magnification level of 5×. I represents the network architecture used for interactive segmentation. II describes the process by which a segmentation mask is generated for a WSI using ZoomISEG. Red and green represent positive and negative clicks, respectively.

tively evaluate the proposed method on a pathology image dataset. The proposed method showed competitive performance by appropriately combining the efficiency of the model using low-resolution images and the high accuracy of the model using high-resolution images.

2. Methods

An overview of the proposed method is shown in Fig. 2. We define the WSIs at a magnification level of 5× as level 2 (L2) and the WSIs at a magnification level of 20× as level 1 (L1). The process of generating a segmentation mask for a WSI through ZoomISEG is as follows: First, a segmentation mask reflecting global context is generated from a low-resolution L2 image with a wide receptive

field. Then, additional adjustments to the mask are made using de- tailed information from a high-resolution L1 image for areas that require fine-tuning. Finally, the L2 mask is upscaled to the size of the L1 image using a cubic interpolation scheme, and partial patch masks generated from L1 are overlaid onto it to complete the pre- diction. This method enables the generation of a comprehensive mask that captures the entire context at a low cost, without sacri- ficing the capture of crucial details in significant areas.

2.1 Network Architecture

In deep-learning-based interactive segmentation, the neural net- work learns to incorporate click-type user interaction into segmen- tation masks. In this work, we modified two conventional segmen- tation neural networks, U-Net [3] and UCTransNet [9] for inter-

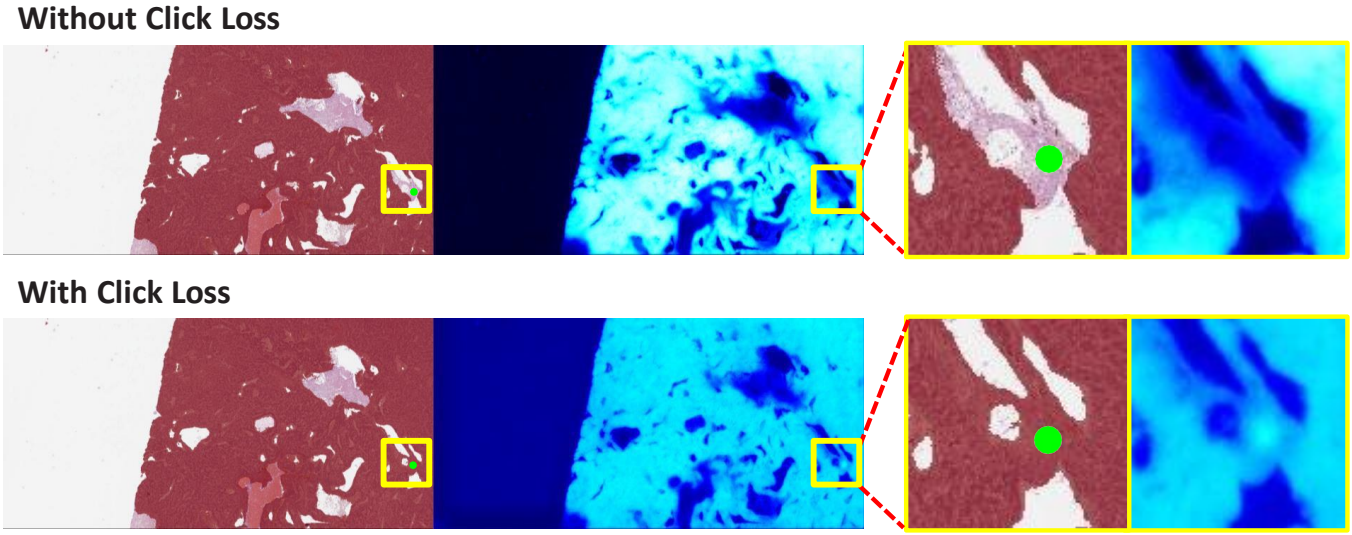

Figure 3: Comparison of predictions generated by models trained with and without click loss. The first column shows an overlay of the input image and segmentation mask, with green indicating the foreground click. The second column shows the output probability map. The model trained with click loss guarantees the class specified by the user for the clicked area.

active segmentation of L2 and L1 WSIs, respectively. The network takes an image along with click maps generated from user clicks as input. Click maps are 2-channel inputs that include positive clicks generated for the foreground and negative clicks generated for the background. The network also takes a mask-type input such as the previous mask generated for the previous click or an externally available external mask.

As shown in Fig. 2, the click maps and mask-type input are concatenated into a 3-channel. This 3-channel input and the input image are each passed through a separate branch to generate a 64-channel feature map, which is then combined with the other feature map via element-wise addition. U-Net and UCTransNet have multiple skip connections of different depths. To prevent the excessive dominance of click maps in generating segmentation masks (limited class changes confined to clicked regions), we made the following modification to the networks: The skip connection located at the shallowest part, which forms a connection before the 64-channel feature map generated from the input image, is added to the feature map generated from the click maps, and the second shallowest skip-connection is removed.

2.2 Click Loss

Adding mask-type input, such as an externally generated mask or the prediction mask generated for the previous click, to the network input is a commonly used method [4, 5] in interactive segmentation of natural image datasets. However, this degrades the segmentation performance for WSI. We address this issue by giving more weight to the impact of clicks through the incorporation of a new loss component, referred to as click loss, during model training.

We define a set of user-provided clicks as $\{(u_i, v_i, l_i)\}_{i=1}^{N}$ where (u, v) and $l \in \{-1, 1\}$ represent the coordinates and label of each click, respectively. Assume f is a function implemented by the network. With an input image X and click maps C, click loss is calculated in the form of squared hinge loss, as follows:

$$L_{click} = E_{i\in[1,N]}[max(l_i - f(X, C)_{u_i,v_i}, 0)]^2. \quad (1)$$

The effect of training with click loss can be observed in Fig. 3.

2.3 Low-to-High Information Transfer

Proper utilization of multi-resolution images enables rapid generation of high-quality, high-resolution segmentation results with less user interaction. To achieve this, it is important to capture the global context through a wide receptive field in low-resolution images that are reduced to small sizes and to leverage rich local features in high-resolution images only for important parts that require detail. Combining information from images of different resolutions can also have a synergistic effect when analyzing an image at a specific resolution. Low-to-high information transfer (L2H) provides how the region is segmented in a global context, including neighbor patch information, which is unknown in the corresponding single high-resolution patch that occupies a small portion of the entire image. The process of L2H is as follows: First, a segmentation mask for the entire WSI is generated in L2, and additional inference is performed in L1 for ROIs that require more detail. At this time, the L2 segmentation mask in the area corresponding to the L1 patch is

Algorithm 1 Zoom Simulation

Require: L2 patch X_2, label l_2

for $i = 1$ **to** NoC_{max} **do**
  Update click maps C_2
  Get prediction P_2=network$_2$(X_2, C_2)
  $d \leftarrow$ the minimum Euclidian distance between clicks
  $center\ coords \leftarrow$ the midpoint between the two clicks that generate d
  $q \leftarrow$ the IoU of P_2 and l_2
  if $d < D_{thr}$ and $Q_{thr} < q < Q_{max}$ **or** $d < D_{min}$ **then**
    Get L1 patch X_{1i} with $center\ coords$
    Get L2 mask M_{2i} by cropping P_2 corresponding to the location of X_{1i}
    for $j = 1$ **to** NoC_{max} **do**
      Update click maps C_{1i}
      Get prediction P_{1i}=network$_1$(X_{1i}, C_{1i}, M_{2i})
    end for
  end if
end for
Get P_{coarse} by scaling P_2 to the size of L1
Get prediction P by pasting P_{1i} to P_{coarse} where $i \in [1, NoC_{max}]$
return P

cropped and entered as a mask-type input for the first click to the L1 model.

3. Experiments

3.1 Settings

3.1.1 Dataset

We utilized the PAIP2019 challenge [10] dataset, which consisted of hepatocellular histopathology whole slide images and tumor region labels. A total of 441 WSIs were scaled to a magnification level of 5× for the L2 model, and a magnification level of 20× for the L1 model. Patches of size 1024 × 1024 were extracted from the WSIs. Among the patches with a magnification level of 5×, those that did not contain tumor regions were discarded, leaving a total of 1749 patches. At a magnification level of 20×, 12,480 patches were selected, where the tumor area accounted for 10% to 90% of the entire area. Patches at each magnification level were split into a 9:1 ratio to train and evaluate the corresponding model. Five WSIs were used to evaluate the entire process through zoom simulation.

3.1.2 Metrics

Two metrics were used to measure the efficiency and performance of the proposed method. Mean intersection over union (mIoU) of the segmentation results was used to demonstrate the accuracy performance of the methods. The total number of clicks (tNoC) was used to show the efficiency of the methods by indicating how many user clicks were required to complete high-quality segmentation masks for the WSIs.

3.1.3 Implementation Details

We implement our models in PyTorch and test with a single NVIDIA RTX A6000 GPU. We set the batch size to 4. We implement modified U-Net and UCTransNet described in Subsection 2.1 for L2 and L1 models. We trained our models using a combination of normalized focal loss proposed in [11] and click loss described in Subsection 2.2, where the scaling constant for click loss was 0.05. We sampled the clicks during training following the procedure of [12]. We use the Adam optimizer with β_1 = 0.9, β_2 = 0.999. We set the learning rate to 5×10^{-5}. We trained the L1 network for 40 epochs and the L2 network for 100 epochs.

3.1.4 Zoom Simulation

We implement automatic zoom simulation that simulates user behavior to quantitatively validate the interactive multi-resolution WSI segmentation method. The need for refinement in L1 is determined by the distribution of the clicks and the quality of the mask generated in L2. The distance between clicks is used to represent the distribution of clicks and predict which part of the image should be zoomed in for segmentation at higher magnification. The process moves on to L1 after ensuring that the quality of the L2 mask is satisfactory enough. This simulates users sequentially analyzing multi-resolution images starting from low resolution. Zoom simulation is conducted according to Algorithm 1. If the quality of the L2 mask is better than Q_{max}, it is judged that additional modifications are unnecessary. If the minimum distance between clicks is less than D_{min}, it is judged that the ROI is too small for analysis in L2, and zoomed in using L1. Otherwise, if the minimum distance between clicks is less than D_{thr} and the quality of the L2 mask is better than Q_{thr}, it is judged that zooming is required. The default D_{min}, D_{thr}, Q_{thr}, Q_{max}, and NoC_{max} is set to 50, 250, 0.85, 0.95, and 20 respectively.

3.2 Results

We compare our method `ZoomISEG` with single-resolution models of different magnification levels. Since the L2 model is designed to handle relatively small and low-resolution images, it tends to infer only a small number of patches when the image is divided into patches of the same size. Therefore, it is possible to generate a segmentation mask with only a small number of clicks, but this approach shows relatively lower accuracy. On the other hand,

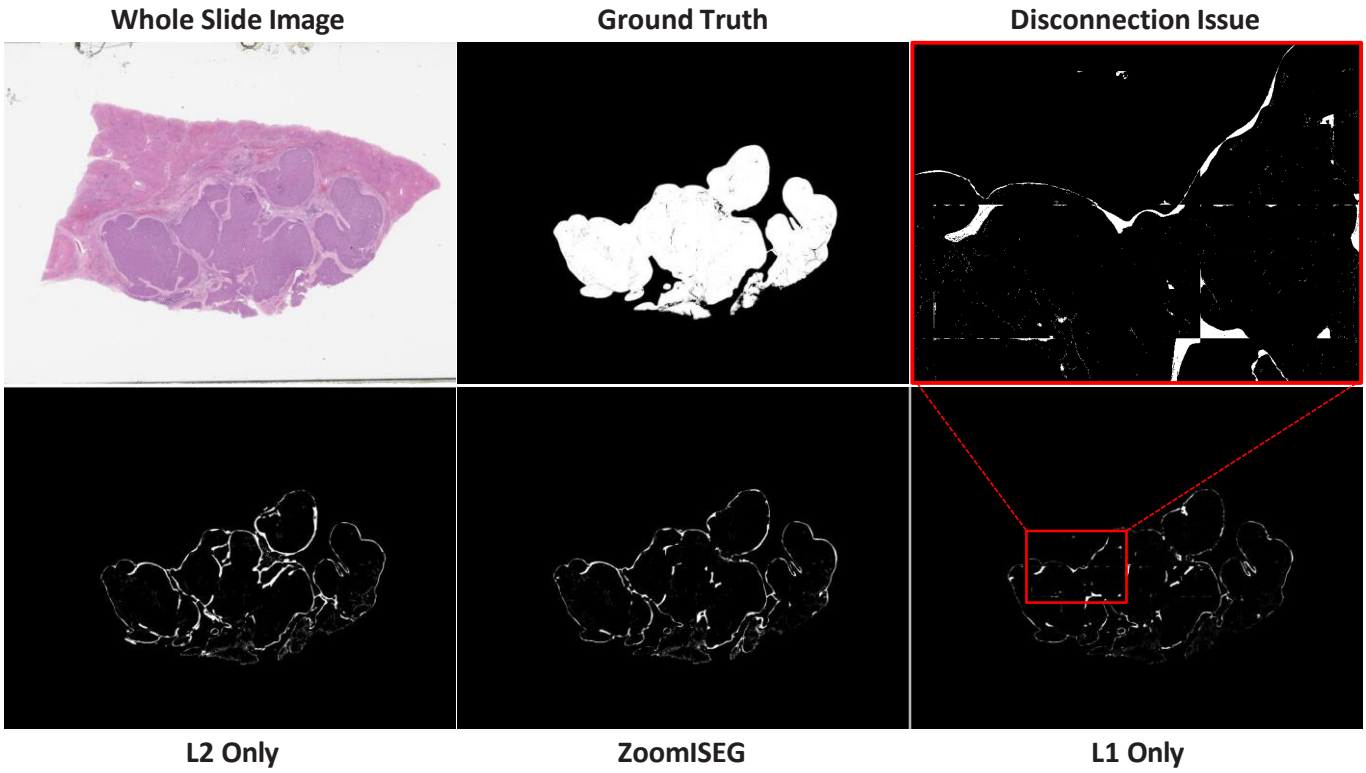

Figure 4: Comparison of ZoomISEG and single-resolution methods. The second row is the error mask for the ground truth of the predictions generated by each method. The red box represents the disconnection issue that occurs in the L1 model.

Table 1: Comparison between ZoomISEG and single-resolution methods, and ablation study on L2H and click loss. The top result is indicated in bold, and the second result is underlined.

	Comparison of Methods			Ablation Study	
Method	L2 only	ZoomISEG (L2+L1)	L1 only	w/o L2H	w/o click loss
mIoU	0.894	<u>0.913</u>	**0.942**	0.909	0.868
tNoC	**138**	<u>385</u>	867	394	1579

the L1 model is able to distinguish the boundaries of tumor regions more precisely as it is designed to handle high-resolution images. However, as the image size increases, the number of patches that need to be processed also increases, and the number of clicks that users need to input also increases. In addition, the process of independently processing a large number of patches and then reassembling them can lead to visual artifacts (disconnection issues) similar to the example shown in Fig. 4. ZoomISEG combines the advantages of each model to increase the accuracy of predictions at a reasonable cost. For regions of interest where details are important, ZoomISEG utilizes high-resolution images to achieve high accuracy, while performing inference on low-resolution images for the rest of the areas to maximize efficiency. As shown in Table 1, compared to the L2 model, the L1 model increased its performance by 5.37%p by adding 739 more clicks, while ZoomISEG achieved a 2.13%p performance increase with only 247 additional clicks.

3.3 Ablation Study

We conducted an ablation study to examine the effectiveness of low-to-high information transfer and click loss. As shown in Table 1, both L2H and click loss contributed to improving the model. In the case of the L1 model, only high-resolution local areas are independently inferred as patches, so even adjacent areas cannot be fully known. However, by receiving global context information from larger areas generated by the L2 model through L2H, it is possible to more effectively create high-quality masks. Click loss helps ensure that the model assigns the designated class to the area

where the user has clicked, without confusion caused by the difficulty of learning ambiguous boundaries of cancerous regions in pathological images during training.

4. Conclusion

In this paper, we introduce `ZoomISEG`, the interactive multi-resolution WSI segmentation method. By utilizing WSIs of different magnification levels, `ZoomISEG` can generate predictions more efficiently than a single high-resolution model and more accurately than a single low-resolution model. The limitation of this study is that currently only uni-directional information transfer from low-to-high is possible. In the future, we plan to develop a bi-directional information propagation scheme. We expect `ZoomISEG` can provide convenience in pathologists' workflow when employed in the analysis or annotation tools they use.

감사의 글

본 연구는 과학기술정보통신부 재원의 정보통신기획평가원의 ICT명품인재양성사업 (IITP-2023-2020-0-01819), 과학기술정보통신부 재원의 한국연구재단의 바이오의료기술개발사업 (NRF-2019M3E5D2A01063819), 교육부 재원의 한국연구재단의 기초과학연구사업 (NRF-2021R1A6A1A13044830), 보건복지부 재원의 한국보건산업진흥원의 한국보건기술연구개발사업 (HI18C0316), 한국과학기술연구원 기본사업 (22E32210, 2E32211), 그리고 고려대학교의 지원을 받아 수행되었음.

References

[1] L. Grady, "Random Walks for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1768–1783, 2006.

[2] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, 2015.

[4] X. Chen, Z. Zhao, Y. Zhang, M. Duan, D. Qi, and H. Zhao, "FocalClick: Towards Practical Interactive Image Segmentation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1300–1309, 2022.

[5] Q. Liu, Z. Xu, G. Bertasius, and M. Niethammer, "SimpleClick: Interactive Image Segmentation with Simple Vision Transformers," *arXiv preprint arXiv:2210.11006*, 2022.

[6] Z. Lin, Z.-P. Duan, Z. Zhang, C.-L. Guo, and M.-M. Cheng, "Focuscut: Diving into a Focus View in Interactive Segmentation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2637–2646, 2022.

[7] S. Cho, H. Jang, J. W. Tan, and W.-K. Jeong, "DeepScribble: Interactive Pathology Image Segmentation Using Deep Neural Networks with Scribbles," *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 761–765, 2021.

[8] S. Min and W.-K. Jeong, "CGAM: Click-Guided Attention Module for Interactive Pathology Image Segmentation via Backpropagating Refinement," *arXiv preprint arXiv:2307.01015*, 2023.

[9] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-Wise Perspective with Transformer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 2441–2449, 2022.

[10] Y. J. Kim, H. Jang, K. Lee, S. Park, S.-G. Min, C. Hong, J. H. Park, K. Lee, J. Kim, W. Hong, *et al.*, "PAIP 2019: Liver cancer segmentation challenge," *Medical Image Analysis*, vol. 67, p. 101854, 2021.

[11] K. Sofiiuk, O. Barinova, and A. Konushin, "AdaptIS: Adaptive Instance Selection Network," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7355–7363, 2019.

[12] K. Sofiiuk, I. A. Petrov, and A. Konushin, "Reviving Iterative Training with Mask Guidance for Interactive Segmentation," *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3141–3145, 2022.

〈저자소개〉

민 성 희

- 2022년 고려대학교 바이오의공학부 학사
- 2022년~현재 고려대학교 컴퓨터학과 석사과정
- 관심분야: 인공지능, 딥러닝, 영상처리, 컴퓨터 비전
- https://orcid.org/0009-0008-6509-3072

정 원 기

- 1999년 고려대학교 수학과 학사
- 2001년 고려대학교 컴퓨터학과 석사
- 2008년 University of Utah, Computer Science 박사
- 2008년~2011년 Harvard University 연구원
- 2011년~2020년 울산과학기술원 전기전자컴퓨터공학부 부교수
- 2020년~현재 고려대학교 컴퓨터학과 정교수
- 관심분야: 가시화, 영상처리, 기계학습, 고성능컴퓨팅
- https://orcid.org/0000-0002-9393-6451