SimpleClick: Interactive Image Segmentation with Simple Vision Transformers Supplementary Materials

Qin Liu, Zhenlin Xu, Gedas Bertasius, Marc Niethammer
University of North Carolina at Chapel Hill
https://github.com/uncbiag/SimpleClick

A. Datasets

This section supplements the "Datasets" section in the main paper. Our models are trained either using SBD [8] or the combined COCO [11]+LVIS [7] datasets. Before RITM [16], most of the deep learning-based interactive segmentation models were trained either using the SBD [8] or Pascal VOC [5] datasets. These two datasets only cover 20 categories of general objects such as persons, transportation vehicles, animals, and indoor objects. The authors of RITM constructed the combined COCO+LVIS dataset, which contains 118k training images of 80 diverse object classes, for interactive segmentation. This large and diverse training dataset contributes to the state-of-the-art performance of RITM models. Inspired by RITM and its follow-up works [4,12], we use SBD and COCO+LVIS as our training datasets.

B. Additional Comparison Results

This section supplements Sec. 4.1 "Comparison with Previous Results" in the main paper. Fig. 2 shows convergence results for our models on four datasets: GrabCut [15], Berkeley [13], DAVIS [14], and COCO [11]. Overall, our models perform better than other models on these datasets. However, the results in Fig. 2 are not as compelling as the results on SBD [8] or Pascal VOC [5] (shown in Fig. 3 of the main paper). This is likely due to the limited number of images in these datasets (*e.g.* GrabCut only contains 50 instances, while SBD contains 6671 instances for evaluation).

C. Human Evaluation on Medical Images

This section supplemens Sec. 4.2 "Out-of-Domain Evaluation on Medical Images" in the main paper. In the main paper, we report quantitative results on medical images using an automatic evaluation mode where clicks are automatically simulated. In this section, we perform human evaluations where a human-in-the-loop provides all the clicks. Fig. 1 shows qualitative results on three medical image datasets: ssTEM [6], OAIZIB [1], and BraTS [2]. For simple objects such as cell nuclei in ssTEM, it may take as little as one click for a good segmentation. However, for more challenging objects such as knee cartilage in the OAIZIB dataset or brain tumors in the BraTS dataset, it may take more than ten clicks until a high-quality segmentation is obtained. Considering our models are not finetuned on the label-scarce medical imaging datasets, our observed performance is quite promising. The attached videos demonstrate the evaluation process.

Model	H, W	Patch Size	N	C_0, C_1, C_2
Ours-ViT-B	448, 448	16×16	12	768, 128, 256
Ours-ViT-L	448, 448	16×16	24	1024, 192, 256
Ours-ViT-H	448, 448	14×14	32	1280, 240, 256

Table 1. **Architecture parameters** of `SimpleClick` models. N denotes the number of self-attention blocks. C_0, C_1, and C_2 denote the feature map dimensions at different levels.

D. Implementation Details

D.1. Architectures

Tab. 1 shows the main architecture parameters of our models. By default, our models use an input size of 448×448 during training and evaluation. Our ViT-B and ViT-L models use a patch size of 16×16, while the ViT-H model uses a smaller patch size of 14×14. This leads to a higher resolution representation in terms of the number of patches. Each patch is flattened and projected to an embed dimension of C_0 through the patch embedding layer. The tokens generated by the patch embedding layer are processed by N self-attention blocks, which N is a hyper-parameter inherited from plain ViT models [9]. Inspired by ViTDet [10], we build a simple feature pyramid with the four resolutions $\{\frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}\}$. The $\frac{1}{16}$ resolution uses the last feature map of the ViT backbone. The $\frac{1}{32}$ resolution is built by a 2×2 convolutional layer with a stride of 2. The $\frac{1}{8}$ (or $\frac{1}{4}$) resolution is built by one (or two) 2×2 transposed convolution layer(s) with a stride of 2. We use a 1×1 convolution layer with layer normalization to convert the channels of each feature map

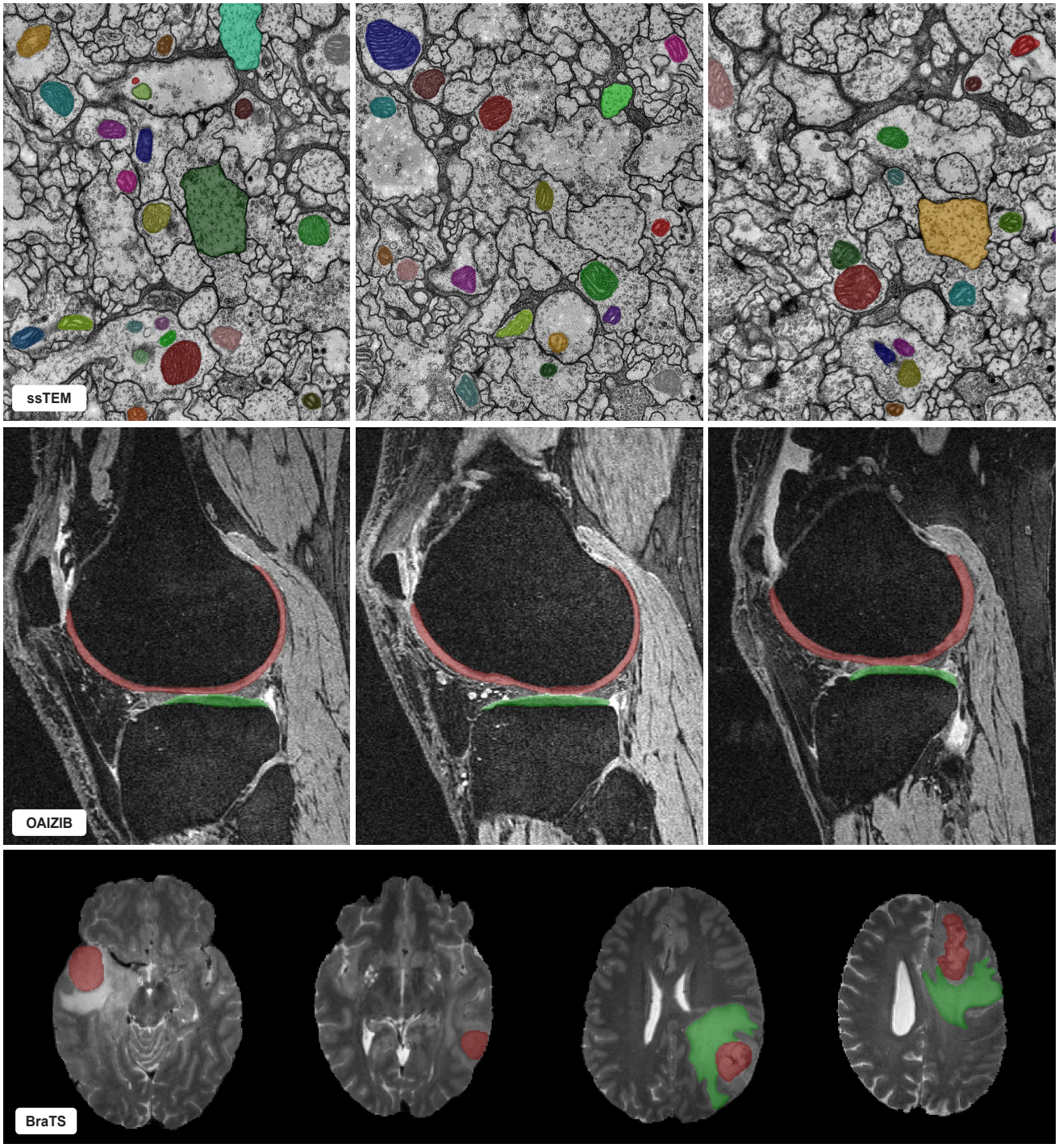

Figure 1. **Qualitative results** of human evaluation on three medical image datasets: ssTEM [6], OAIZIB [1], and BraTS [3]. All the results are obtained by a human-in-the-loop providing the clicks. Though our models are evaluated on medical images without finetuning, they generalize well to all the unseen objects given a few clicks, as shown in the attached videos.

to predefined dimensions. Specifically, feature maps of resolutions $\{\frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}\}$ are converted to channel dimensions of $\{8C_1, 4C_1, 2C_1, C_1\}$, respectively. Each feature map is then converted to the same dimension of C_2 through an MLP layer in the segmentation head, followed by upsampling to the $\frac{1}{4}$ resolution. At this point, the four feature maps have

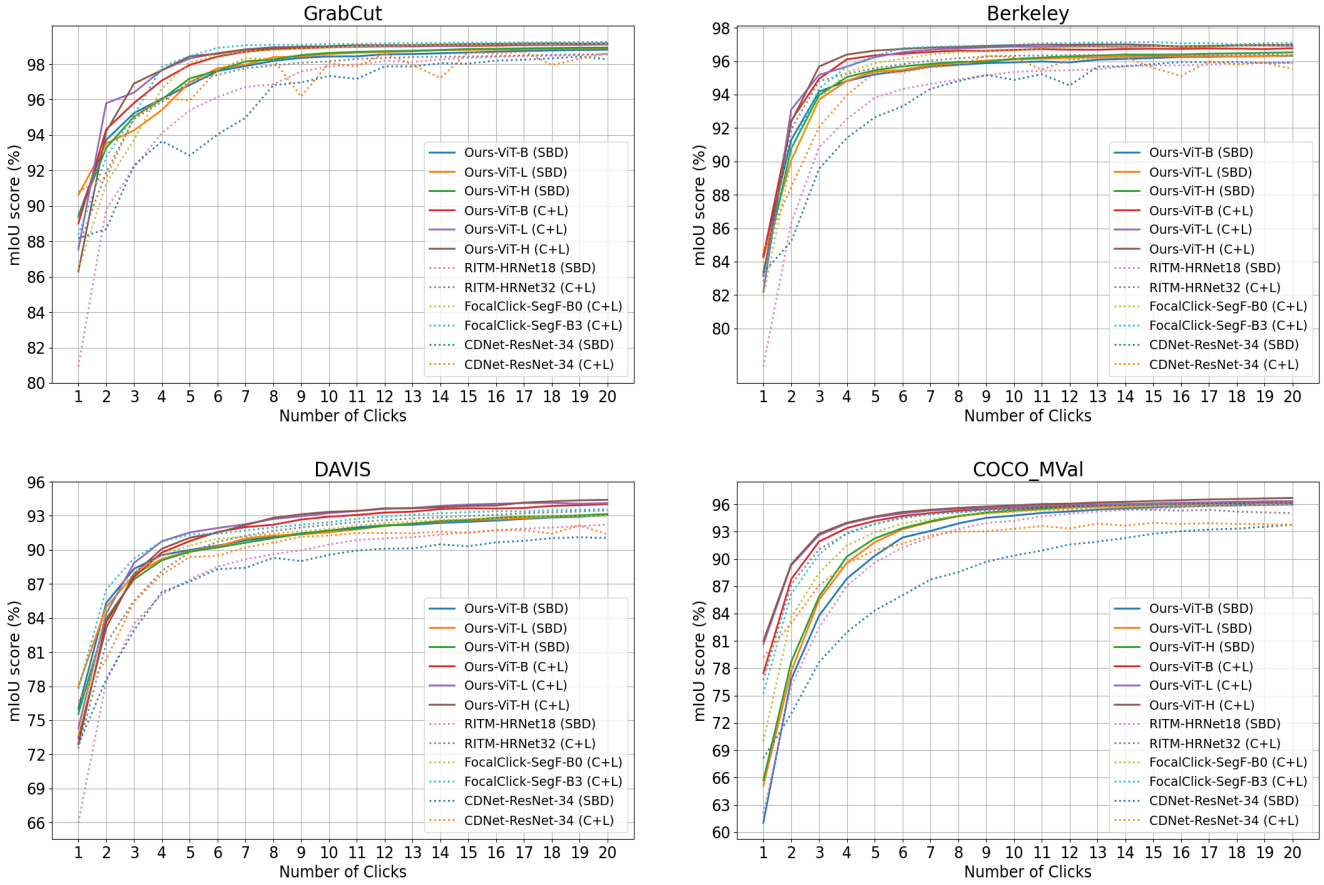

Figure 2. **Convergence analysis** for GrabCut, Berkeley, DAVIS, and COCO. All models are trained on either SBD [8] or COCO [11]+LVIS [7] (C+L). The metric is mean IoU given k clicks (mIoU@k). In general, our models require fewer clicks for a given accuracy level.

the same resolution and the same number of channels. They are concatenated as a single feature map with $4C_2$ channels. Another MLP layer in the segmentation head converts this multi-channel feature map to a one-channel feature map, followed by a sigmoid function to obtain the final binary segmentation. We use C_1 and C_2 as hyper-parameters without tuning.

D.2. Clicks Encoding

We encode clicks, which are represented by the coordinates in an image, as disks with a small radius of 5 pixels. Positive and negative clicks are encoded separately. In our implementation, we also attach the previous segmentation as an additional channel, resulting in a three-channel disk map. Two patch embedding layers, which are of the same structure, process the three-channel disk map and the RGB image separately. The tokens of the two inputs after the patch embedding layers are added element by element, without changing the input dimensions for the self-attention blocks. This design is more efficient than other designs such as concatenation and allows our ViT backbones to be initialized with pretrained ViT weights.

D.3. Finetuning on Higher-Resolution Images

This section supplements Sec. 3.4 "Training and Inference Settings" in the main paper. Our models are pretrained on an image size of 224×224 but are finetuned on an image size of 448×448. We first interpolate the positional encoding to the high resolution. Then, we perform non-overlapping window attention [10] with a few global blocks for cross-window attention. The high-resolution feature map is divided into regular non-overlapping windows. The non-global blocks perform self-attention within each window, while global blocks perform global self-attention. We set the number of global blocks to 2, 6, and 8 for the ViT-B, ViT-L, and ViT-H models, respectively.

E. Statistics for Failure Cases

This section supplements Sec. 5 "Limitations and Remarks" in the main paper. Our method still has much room

Backbone	Training Set	NoC@85	NoC@90	NoF@85	NoF@90
Ours-ViT-B	COCO+LVIS	3.43 ± 4.45	5.62 ± 6.36	267	778
Ours-ViT-L	COCO+LVIS	2.95 ± 4.15	4.89 ± 6.00	223	631
Ours-ViT-H	COCO+LVIS	2.85 ± 4.02	4.70 ± 5.89	206	606

Table 2. **Number of failures (NoF) on the SBD dataset**. We define the interactive segmentation on an image as a failure if NoC@90≥20. The mean ± standard deviation of the NoC metric is also provided for reference.

to improve. As shown in Tab. 2, our method suffers from high variance and a number of failure cases. Note that the standard deviation greater than the mean does not imply negative clicks. It shows to some extend the diversity of the SBD dataset. As a practical annotation tool, Our method needs to be improved in the future to handle challenging cases.

References

[1] Felix Ambellan, Alexander Tack, Moritz Ehlke, and Stefan Zachow. Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the osteoarthritis initiative. *Medical image analysis*, 52:109–118, 2019. 1, 2

[2] Xue Bai and Guillermo Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 1

[3] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 2

[4] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. FocalClick: Towards practical interactive image segmentation. In *CVPR*, pages 1300–1309, 2022. 1

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1

[6] Stephan Gerhard, Jan Funke, Julien Martel, Albert Cardona, and Richard Fetter. Segmented anisotropic sstem dataset of neural tissue. *figshare*, pages 0–0, 2013. 1, 2

[7] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019. 1, 3

[8] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011. 1, 3

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 1

[10] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 1, 3

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1, 3

[12] Qin Liu, Meng Zheng, Benjamin Planche, Srikrishna Karanam, Terrence Chen, Marc Niethammer, and Ziyan Wu. PseudoClick: Interactive image segmentation with click imitation. *arXiv preprint arXiv:2207.05282*, 2022. 1

[13] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001. 1

[14] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016. 1

[15] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004. 1

[16] Konstantin Sofiiuk, Ilia A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. *arXiv preprint arXiv:2102.06583*, 2021. 1