






# PseudoClick: Interactive Image Segmentation with Click Imitation (Supplementary Material)

Qin Liu<sup>1,2</sup>, Meng Zheng<sup>2</sup>,  
Benjamin Planche<sup>2</sup>, Srikrishna Karanam<sup>2</sup>, Terrence Chen<sup>2</sup>, Marc  
Niethammer<sup>1</sup>, and Ziyang Wu<sup>2</sup>

<sup>1</sup> University of North Carolina at Chapel Hill, Chapel Hill NC, USA

<sup>2</sup> United Imaging Intelligence, Cambridge MA, USA  
{first.last}@uii-ai.com, qinliu19@cs.unc.edu

## 1 Datasets

This section is supplemented for “Datasets” in the main paper. We evaluate our method on 10 public datasets. The details are as follows:

- **GrabCut** [1]: This dataset contains 50 images with 50 instance masks. It is widely used for the evaluation of interactive segmentation methods.
- **Berkeley** [2]: This dataset contains 96 images with 100 instance masks. It shares a small portion of images with the GrabCut dataset.
- **PASCAL** [3]: We use the validation set for testing, which contains 1,449 images with 3,427 instances. This follows the same training/testing splitting protocols with existing methods.
- **DAVIS** [4]: This dataset contains 50 videos with high quality segmentation masks. Instead of using the entire dataset, we only extract the same 10% of frames that were used in [5] for evaluation.
- **SBD** [6]: The Semantic Boundaries Dataset (SBD) contains 6,671 instance-level masks for 2,820 images. This dataset shares the categories with the PASCAL dataset. We use its training set for training and its validation set for evaluation.
- **COCO** [7]: This dataset contains a total of 1.2M instance masks on 118k training images with 80 object categories. We combine this dataset with the LVIS dataset as a training set.
- **LVIS** [8]: The LVIS dataset shares its images with the COCO dataset but has the highest annotation quality among all the reported datasets on more than a thousand object categories. This dataset is combined with the COCO dataset for training.
- **Cars** [9]: The Cars dataset is only used for qualitative evaluation in this work. Since the Cars dataset only contains image-level labels, we generate the pixel-level masks using publicly available pretrained mask-R-CNN models [10].



Fig. 1: The next human click can be either a positive click (*e.g.*, put in the green circle) or a negative click (*e.g.*, put in the red circle).

- **ssTEM** [11]: This dataset consists of two image stacks. Each stack contains 20 sections from serial section Transmission Electron Microscopy (ssTEM) of the drosophila melanogaster third instar larva ventral nerve cord. We only use the first stack and the mitochondria mask for evaluation.
- **BraTS** [12]: The Brain Tumor Segmentation challenge 2020 (BraTS20) dataset contains 369 training volumes with multi-label annotation masks. We only consider the tumor core; we extract one slice from each volume where the tumor area is the largest. This results in 369 slices with binary masks for evaluation.

## 2 Sharing of Encoding Maps

This section is supplemented for “Limitations” in the main paper.

The human clicks and pseudo clicks can share the same encoding map, i.e., using a 2-channel encoding map instead of two 2-channel encoding maps. We adopted this design in the early state of this project. However, we quickly realized that such an implementation causes accuracy to drop due to possible inaccuracy of the pseudo clicks. Tab. 1 shows the analysis of pseudo clicks generated by a converged **PseudoClick** model. We see that the generated pseudo clicks significantly vary from the simulated human clicks in terms of the click types or location. This variance is caused by the imperfect prediction of the estimated error maps, from which pseudo clicks are extracted. Therefore, naively merging the two types of clicks may confuse the model and lead to sub-optimal performance. To address this issue, we encode the two types of clicks separately, leading to our current design. In this design, inaccuracies of the pseudo clicks are better tolerated during training and are less likely to cause accuracy drops during inference. The comparison results in Tab. 2 show that by separating the pseudo clicks from human clicks the performance improves significantly.

Though we observe in Tab. 1 a huge disagreement between pseudo clicks and human clicks, this won’t be a severe issue in reality. This is because even for a human annotator, there may be many suitable locations to put a click in each interaction, as shown in Fig. 1.

	GrabCut	Berkeley	PASCAL
Total Pseudo Clicks	45	149	5579
Matching Type (Pos/Neg)	20 (44.4%)	70 (47.0%)	2995 (53.7%)
On GT Mask	22 (48.9%)	64 (43.0%)	3124 (56.0%)

Table 1: Analysis of pseudo clicks generated by a converged **PseudoClick** model. The model proposes one pseudo click after each human click till the IoU is greater than 85%. ‘Matching Type’ measures the number of pseudo clicks that have the same type (positive or negative click) with the simulated human clicks. ‘On GT Mask’ measures the number of pseudo clicks that are on the ground truth mask. (*e.g.*, a positive click should be on the background; a negative should be on the foreground).

	GrabCut	Berkeley	PASCAL	SBD
Same Encoding Maps	2.74	3.35	4.01	7.11
Diff. Encoding Maps	1.50	2.08	2.25	5.54

Table 2: Comparison results of two clicks-encoding designs: sharing or separating encoding maps for pseudo clicks and human clicks. Using separate encoding maps significantly improves the performance. The model is trained on the C+L dataset with a HRNet32 backbone. Evaluation measure: NoC@90%.

### 3 More Qualitative Results

This section is supplemented for “Experiments” in the main paper. We show more qualitative results in Fig. 2 and 3. In Fig. 2, we evaluate our model on the PASCAL dataset. In Fig. 3, we evaluate our model on the BraTS and ssTEM datasets.



Fig. 2: Qualitative results on the PASCAL dataset. The segmentation masks for individual instances are overlaid on the image. This is evaluated by a human annotator.

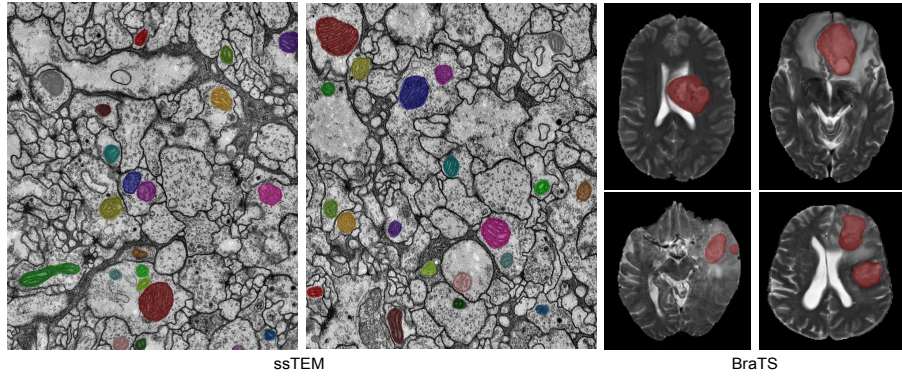


Fig. 3: Qualitative results on the BraTS and ssTEM datasets. Note that the model is trained on the C+L dataset and is evaluated on these datasets without finetuning. This is evaluated by a human annotator.

## References

1. 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.
2. D. Martin, C. Fowlkes, D. Tal, and J. 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*, vol. 2, pp. 416–423, IEEE, 2001.
3. M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VOC) challenge,” *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.
4. F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, “A benchmark dataset and evaluation methodology for video object segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 724–732, 2016.
5. W.-D. Jang and C.-S. Kim, “Interactive image segmentation via backpropagating refinement scheme,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5297–5306, 2019.
6. B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, “Semantic contours from inverse detectors,” in *2011 International Conference on Computer Vision*, pp. 991–998, IEEE, 2011.
7. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in *European conference on computer vision*, pp. 740–755, Springer, 2014.
8. A. Gupta, P. Dollar, and R. Girshick, “Lvis: A dataset for large vocabulary instance segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5356–5364, 2019.
9. J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3d object representations for fine-grained categorization,” in *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, (Sydney, Australia), 2013.
10. W. Abdulla, “Mask r-cnn for object detection and instance segmentation on keras and tensorflow,” 2017.
11. S. Gerhard, J. Funke, J. Martel, A. Cardona, and R. Fetter, “Segmented anisotropic ssTEM dataset of neural tissue,” *figshare*, pp. 0–0, 2013.
12. U. Baid, S. Ghodasara, M. Bilello, S. Mohan, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, *et al.*, “The RSNA-ASNR-MICCAI BRATS 2021 benchmark on brain tumor segmentation and radiogenomic classification,” *arXiv preprint arXiv:2107.02314*, 2021.