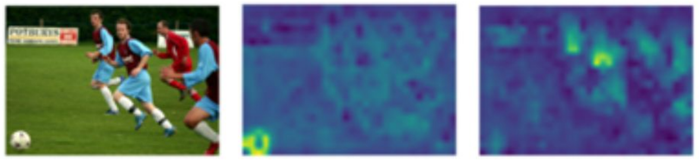
Text and Click inputs for unambiguous open vocabulary instance segmentation

Nikolai Warner[1] , Meera Hahn[2] , Jonathan Huang[2], Irfan Essa[12] , Vighnesh Birodkar[2]
[1]Georgia Institute of Technology [2] Google Inc

Introduction

Deep learning models like SOLOv2 and MaskFormer have advanced image segmentation but are constrained by predefined object categories. We introduce "Text + Click Segmentation," a method that merges a user click with a text prompt to guide the segmentation. Our approach leverages saliency maps from image-text models like MaskCLIP [2], enabling generalization to unseen object classes. We validate its performance on differentiating overlapping semantic categories and demonstrate its effectiveness across datasets like refCOCO, COCO, VOC, and OpenImages.



Input Image “Ball” “Shirt”

We use MaskCLIP to generate rough guesses of where objects are in a scene.

Input Image No Text Baseline “Tie” “Person”

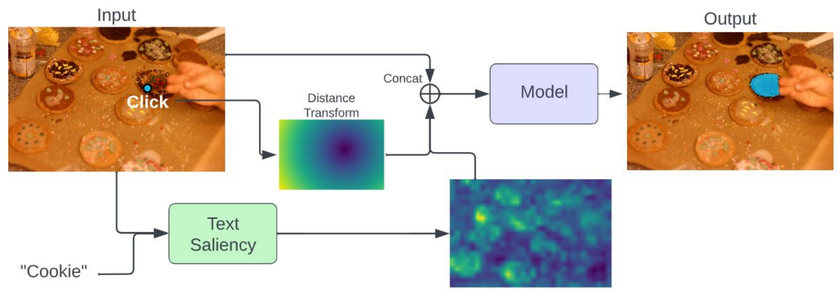
The same click could represent multiple user intents. A text query helps to disambiguate what to segment.

Contributions

- We introduce text-conditioned segmentation using pre-trained CLIP via MaskCLIP, enabling effective segmentation for previously unseen categories.
- Our method outperforms or matches PhraseClick [4], a prior approach that uses text-based cues but is limited to specific classes, while offering broader category generalization.
- We also surpass the Segment Anything (SAM) [1] model, which specializes in click-based segmentation, even when our model is trained on a smaller dataset.

Methods and Experiments

Our approach takes an RGB image, a foreground click, and a text prompt as inputs to produce a class-agnostic segmentation mask.
We train separate models for Pascal VOC, COCO, refCOCO, and OpenImages datasets. Our models are trained in two configurations: zero-shot segmentation and fully-supervised segmentation.



An overview of our full pipeline. A click is combined with a text query to produce a rough guess; which is fed into a class-agnostic segmentor.

Dataset	Text Input	mIoU		
		Overall	Seen	Unseen
refCOCO	✓	66.02 (+3.03)	70.30 (+1.86)	56.35 (+5.68)
refCOCO		62.99	68.44	50.67
VOC	✓	57.76 (+4.52)	59.31 (+3.2)	50.73 (+10.45)
[illegible]		53.24	56.11	40.28
COCO	✓	38.42 (+3.89)	42.06 (+1.72)	33.45 (+6.98)
COCO		34.53	40.34	26.47
OpenImages	✓	57.05 (+4.40)	67.03 (+3.35)	53.92 (+4.74)
OpenImages		52.65	63.68	49.18

We outperform a no-text baseline by up to 10.45 mIoU on unseen class segmentation, when both models are trained on a portion of all classes present in a dataset.

References
1. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. arXiv:2304.02643, 2023
2. Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from CLIP. CoRR, abs/2112.01071, 2021.
3. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision
4. Henghui Ding, Scott Cohen, Brian Price, and Xudong Jiang. Phraseclick: Toward achieving flexible interactive segmentation by phrase and click. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, Computer Vision – ECCV 2020, pages 417–435, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58580-8.

Qualitative Results

Input Image + Click Baseline Ours

Input Image + Click Baseline Ours

Input Image + Click Baseline Ours

Our model is able to better localize novel objects compared to a no-text baseline.

Input Image + Click SAM Ours

Input Image + Click SAM Ours

We perform well against SAM, which struggles to disambiguate overlapping objects.