Supplementary Material for InterFormer Real-time Interactive Image Segmentation

You Huang [1], Hao Yang [1], Ke Sun [1], Shengchuan Zhang [1]*, Liujuan Cao[1], Guannan Jiang[2], Rongrong Ji[1]
[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University
[2] Intelligent Manufacturing Department, Contemporary Amperex Technology Co. Limited (CATL)

Implementation Details

We have provided a supplementary video demonstrating the performance of our model, which was finished on a **CPU-only** device. Then, we provide more implementation details.

During training, we employ a sampling strategy whereby a single instance is randomly selected as the ground truth per sample in an image from either the COCO [2] or LVIS [1] dataset, which may belong to either the "thing" or "staff" categories in COCO/LVIS. However, as a consequence of the random resizing and cropping, this strategy may sometimes result in empty annotations. In such cases, we repeat the sampling process until a non-empty annotation is obtained.

Regarding the simulation of clicks, we randomly sample central locations within the erroneous region. Specifically, we define the central locations as points (x, y) that satisfy the condition $d(x, y) \geq \max_{(x',y')} d(x', y')/k$, where $d(x, y)$ is the shortest distance from the point to the boundary of the connected erroneous region that (x, y) is located in, and $k = 1.7$. Notably, our proposed pipeline is efficient, and allows us to simulate clicks with a single inference on the large backbone, and to iteratively perform interactions using the light interactive modules of InterFormer.

I-MSA Stage	Feature Size	Pooling Ratio
1	128	12,16,20,24
2	64	6,8,10,12
3	32	3,4,5,6
4	16	1,2,3,4

Table 1. Configurations of I-MSA's pooling ratios.

To configure the pooling ratios in each stage of the proposed I-MSA, we employ the settings illustrated in Figure 1. It is important to note that this configuration assumes an image size of 512. When dealing with images of varying sizes, we adjust the pooling ratios to ensure that the size of the pooled features remains consistent with that corresponding to an image size of 512. For instance, to maintain low computational complexity for an image of size 1024, we use double pooling ratios.

*Corresponding author

References

[1] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5356–5364. Computer Vision Foundation / IEEE, 2019. 1

[2] 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. *Lecture Notes in Computer Science*, 2014. 1