Iteratively Trained Interactive Segmentation: Supplementary Material

Sabarinath Mahadevan
mahadevan@vision.rwth-aachen.de
Paul Voigtlaender
voigtlaender@vision.rwth-aachen.de
Bastian Leibe
leibe@vision.rwth-aachen.de

Computer Vision Group
Visual Computing Institute
RWTH Aachen University
Germany

1 Initial Click Sampling

To initialise the click channels, we use the click sampling strategies proposed by [2]. The sampling algorithm works as follows.

Positive clicks. First, the number of positive clicks n_{pos} is sampled from $[1, N_{pos}]$. Then, n_{pos} clicks are randomly sampled from the object pixels, which can be obtained from the ground truth mask. Each of these clicks are sampled such that any two clicks are d_s pixels away from each other and d_m pixels away from the object boundary.

Negative clicks. For sampling negative clicks, we use multiple strategies to encode the user click patterns. Let us define a strategy set $S = \{s_1, s_2, s_3\}$. First, a strategy is randomly sampled from set S and then the sampled strategy is used to generate n_{neg} clicks on the input image. Here, n_{neg} is a number sampled from $[0, N_i]$ where $i \in [1, 2, 3]$ and N_i represents the maximum number of clicks for each strategy. The strategies used here are explained in detail below.

- s_1**:** In the first strategy, n_1 clicks are sampled randomly from the background pixels such that they are within a distance of d_o pixels from the object boundary. The clicks are filtered in the same way as the positive clicks.
- s_2**:** The second strategy is to sample n_2 clicks on each of the negative objects. Here again, the clicks are filtered to honour the same constraints as in the first strategy.
- s_3**:** Here, N_3 clicks are sampled to cover the object boundaries. This helps to train the interactive network faster.

2 Implementation Details

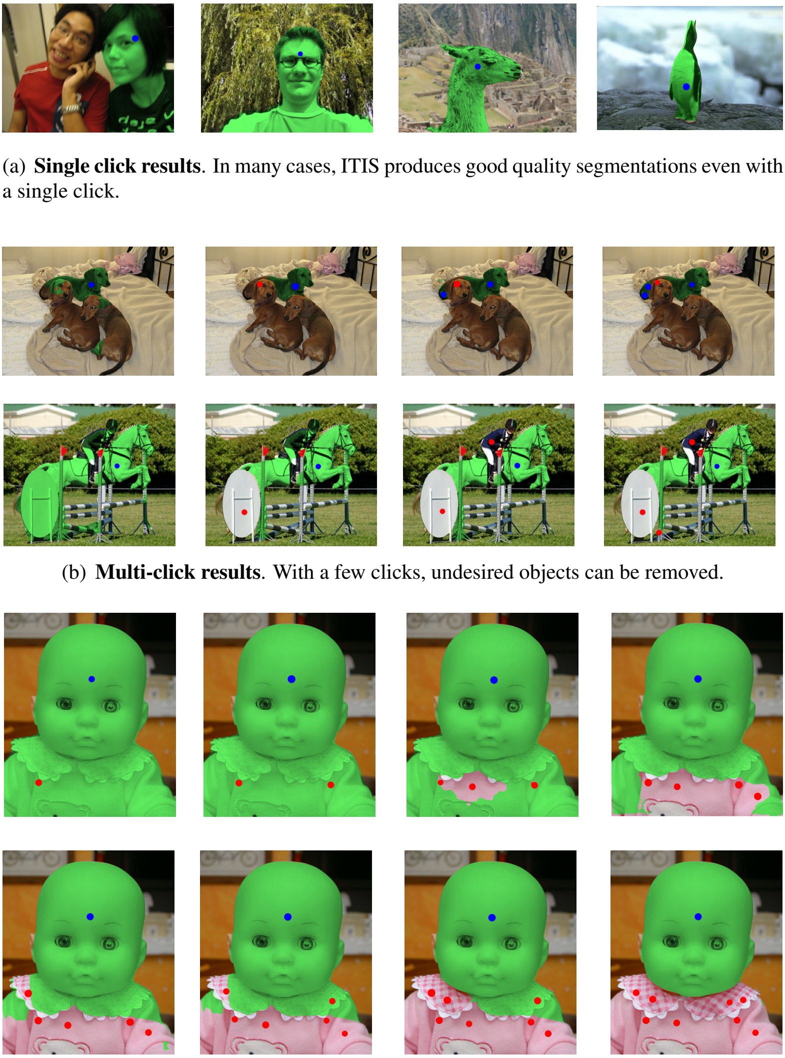
We train with a fixed crop size of 350 × 350 pixels. Input images whose smaller side is less than 350 pixels are bilinearly upscaled such that the smaller side is 350 pixels long. Otherwise, the image is kept at the original resolution. Afterwards, we take a random crop which is constrained to contain at least a part of the object to be segmented. The only form of data augmentations we use are gamma augmentations [1]. We start with a learning rate of

10^{-5} and reduce it to 10^{-6} at epoch 10 and to $3 \cdot 10^{-7}$ at epoch 15. At test time, we use the input image in the original resolution without resizing or cropping.

For the initial click sampling, we set the hyperparameters to $N_{pos} = 5$, $d_m = 5$, $d_s = 40$, $d_o = 40$, $N_1 = 10$, $N_2 = 5$, $N_3 = 10$.

3 Qualitative Results

Figure 1 shows qualitative results of our method.



(a) **Single click results**. In many cases, ITIS produces good quality segmentations even with a single click.

(b) **Multi-click results**. With a few clicks, undesired objects can be removed.

(c) **Failure case**. The initial negative clicks fail to remove the pixels in the body of the doll as the network interprets both the head and the body as a single object. Hence, the network needs more clicks to produce the desired result.

Figure 1: Qualitative results of the proposed iteratively trained interactive segmentation method.

References

[1] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *CVPR*, 2017.

[2] N. Xu, B. L. Price, S. Cohen, J. Yang, and T. S. Huang. Deep interactive object selection. In *CVPR*, 2016.