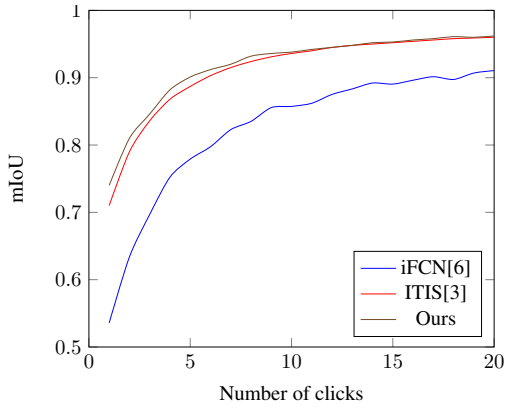
Content-Aware Multi-Level Guidance for Interactive Instance Segmentation

Soumajit Majumder
Institute of Computer Science II
University of Bonn, Germany
majumder@cs.uni-bonn.de

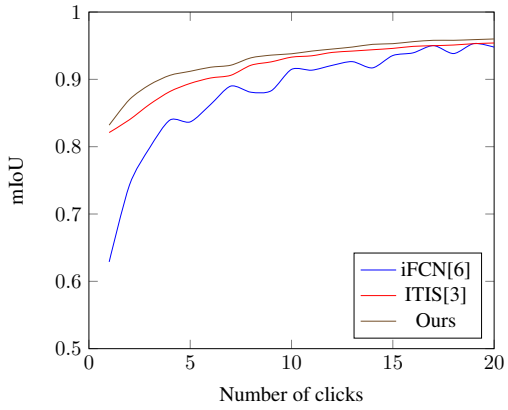
Angela Yao
School of Computing
National University of Singapore
yaoa@comp.nus.edu.sg

1. Experimental Validation

We report the mIoU *vs.* the number of clicks (see Fig. 1) against other methods reported in the literature [6, 3]. In the initial stages of interaction on PASCAL VOC 2012 *val* set, our network outperforms the current state-of-the-art ITIS (as can be seen from Fig. 1).



(a) Pascal VOC 2012 [1]



(b) GrabCut [5]

Figure 1: **mIoU *vs* number of clicks** on the (a) Pascal VOC 2012 *val* set [1] and (b) GrabCut dataset [5].

2. Qualitative Results

Zero Clicks We show via some qualitative examples, the benefits of having the guidance dropout. In several instances, our network is able to produce high quality masks without any user guidance (as shown in Fig. 2).

Multiple Clicks In Fig. 3, we show some examples where undesired objects and background was removed with only a few clicks resulting in a suitable object mask.

Failure Cases We show some examples from PASCAL VOC 2012 *val* set, where our network is unable to generate object masks with $\geq 85\%$ mIoU and exhausts the 20 click budget (see Fig. 4). These failure cases are representative of the problems faced by CNNs while segmenting objects from images such as, small objects [4], occlusion [2], motion blur and objects with very fine structures. In general, we observed that our network had difficulty in handling three object classes from PASCAL VOC 2012 - *chair, bicycle* and *potted plant*. This stems from the inability of CNNs to produce very fine segmentations, most likely due to the loss of resolution from downsampling in the encoder.

References

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

[2] Ke Li and Jitendra Malik. Amodal instance segmentation. In *European Conference on Computer Vision*, pages 677–693. Springer, 2016.

[3] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. Iteratively trained interactive segmentation. In *BMVC*, 2018.

[4] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.

[5] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[6] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *CVPR*, 2016.

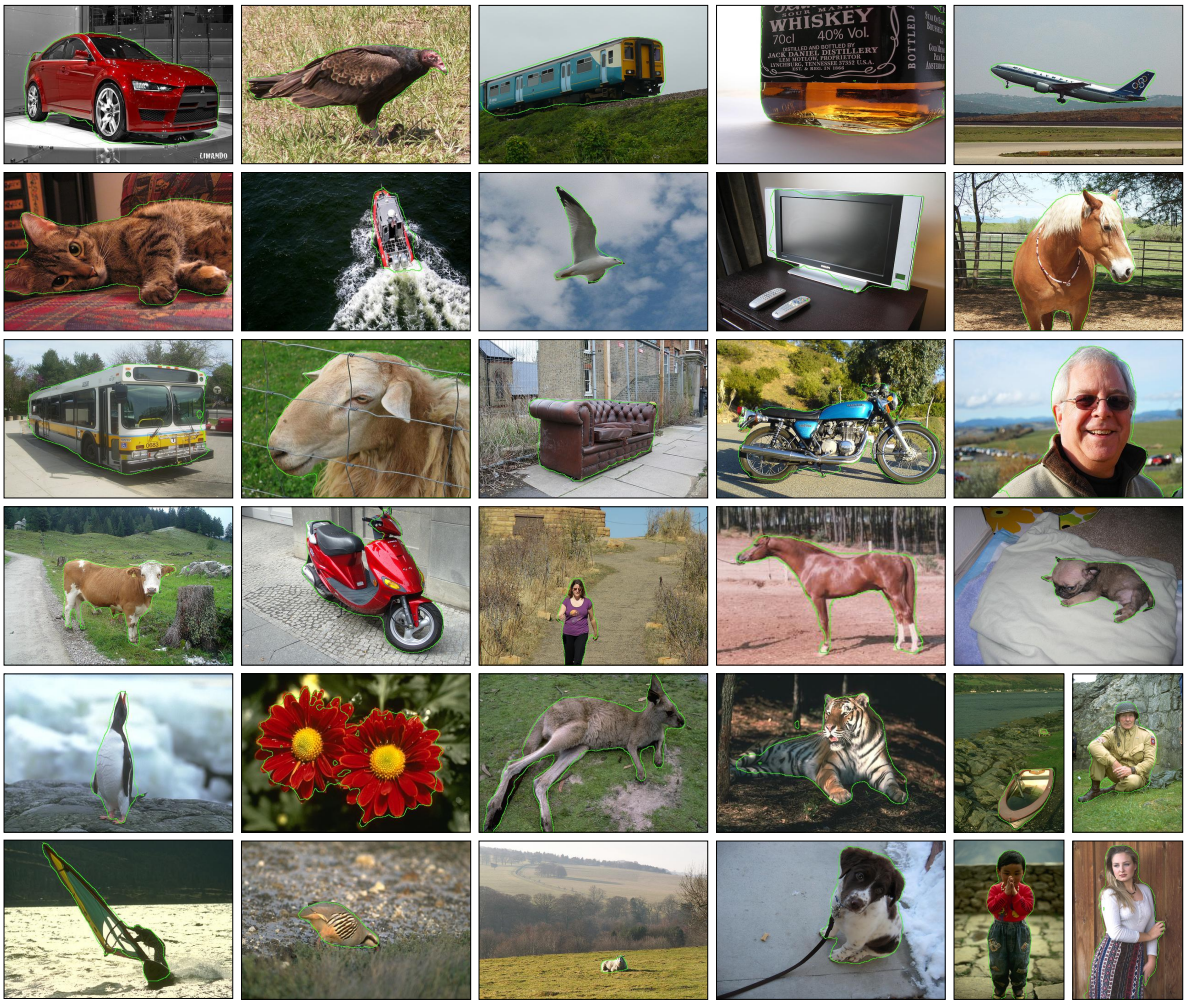


Figure 2: **Zero Clicks.** Examples of high-quality object masks generated without any user guidance. Generated object boundaries are shown in green (Best viewed in color).

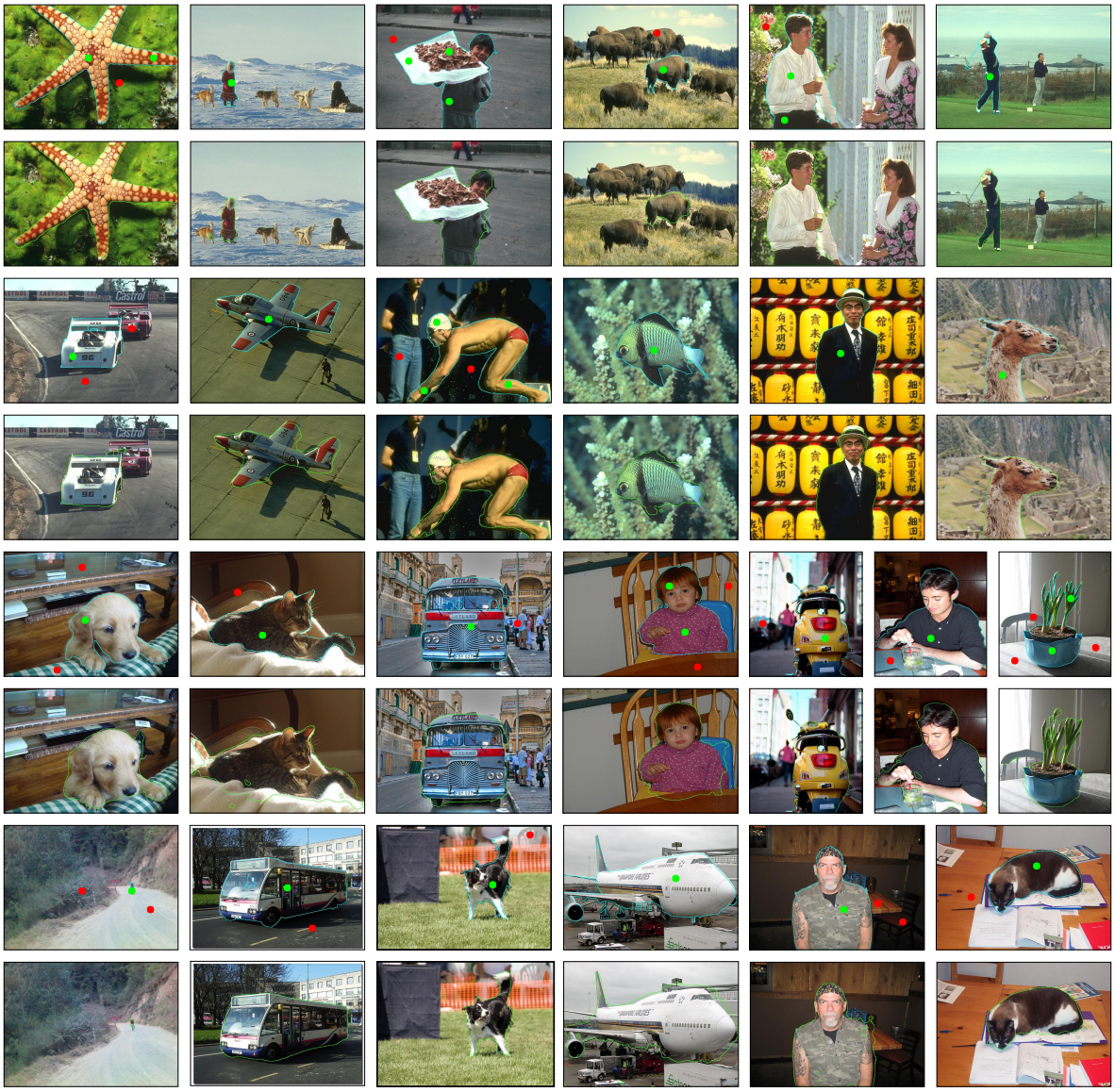


Figure 3: **Multi Clicks.** With a few clicks, background and undesired objects were removed from the final prediction mask. Green dots indicate positive click, red dots indicate negative click. Ground truth object boundaries are shown in cyan and predicted object boundaries are shown in green (Best viewed in color).

Figure 4: **Failure Cases.** Examples of failure cases. Ground truth object boundaries are shown in cyan. Generated object boundaries from the predicted mask are shown in green. In the *dog* example, the network has difficulty distinguishing the fur from the background. For the *car* example, it is either too small (1st row, 2nd column) or too occluded (2nd row, 1st column). For the *bicycle, chair* and *potted plant* example, the error in prediction is due to the inability of the network in handling very fine structures. (Best viewed in color).