MultiSeg: Semantically Meaningful, Scale-Diverse Segmentations from Minimal User Input - Supplementary Materials

Jun Hao Liew[1] Scott Cohen[2] Brian Price[2] Long Mai[2] Sim-Heng Ong[1] Jiashi Feng[1]
[1] National University of Singapore [2] Adobe Research
liewjunhao@u.nus.edu {scohen,bprice,malong}@adobe.com {eleongsh,elefjia}@nus.edu.sg

1. Effects of Clicks on the Proposals Generated

In this supplementary material, we first demonstrate the consistency of the proposals generated by our proposed MultiSeg model with the user inputs. As shown in Figure 1, when only a single positive click (green) is given, our MultiSeg produces a variety set of proposals since the target segmentation remains ambiguous (first row of each example). As more clicks are provided, all the proposals respond accordingly to conform with the newly added user inputs (*e.g.* the left pig is removed from all the proposals when a negative click (blue) is added to it, as shown in the second row of the first example). Lastly, when the segmentation target is no longer ambiguous, all the proposals eventually converge to a single solution as depicted in the last row of each example.

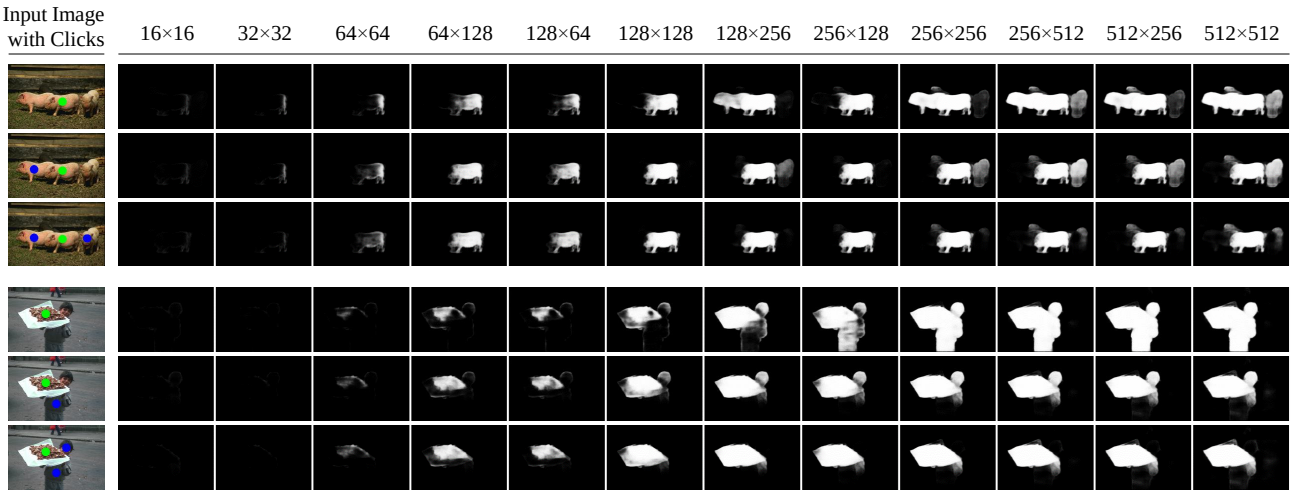

Figure 1: The segmentation results generated by our MultiSeg model respond accordingly in order to be consistent with the user inputs. The values 16×16, 32×32, ..., 512×512 represent the two-dimensional scales of each proposal. Note that all the proposals eventually converge to a single solution (last row) when more clicks are provided.

2. Comparing MultiSeg and DIOS

Next, we present some qualitative comparison between our MultiSeg and DIOS [3] in term of segmenting multiple objects (Figure 2) and object parts (Figure 3). It should be noted that the two models are based on the same DeepLabv3+ [1] backbone architecture as described in the main paper for fair comparison.

Segmenting Multiple Objects: For DIOS, as shown in the top row in Figure 2, despite its excellence in segmenting single object, we observe that the segmentation quality degrades significantly (highlighted with red boxes) when trying to segment a group of objects given more clicks. On the other hand, given only a single positive click, our MultiSeg model can produce diverse, high quality segmentation results as shown in the bottom row.

Segmenting Object Parts: Similarly, as shown in the left column of Figure 3, DIOS requires a significant number of background clicks for deselecting the person's body when it comes to segmenting the smaller parts. On the other hand, since our MultiSeg incorporates a set of scale priors to constrain each proposal to extract the most likely segmentation within a predefined scale, it is therefore capable to generate both a smaller object part (bag, pants) and the larger full object (person) using just a single positive click (right column). We would also like to emphasize that our MultiSeg was not trained with any parts annotations before.

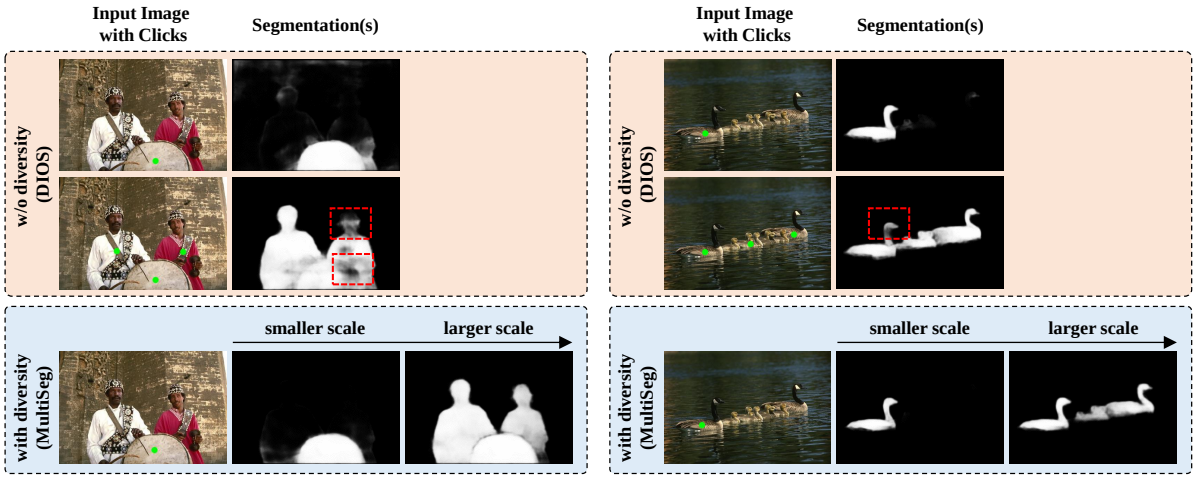

Figure 2: **(Top row)** Despite its excellence in segmenting single object, DIOS struggles with segmenting multiple objects where the segmentation quality of each individual object degrades (red boxes). **(Bottom row)** Our MultiSeg can produce both results with just a single positive click.

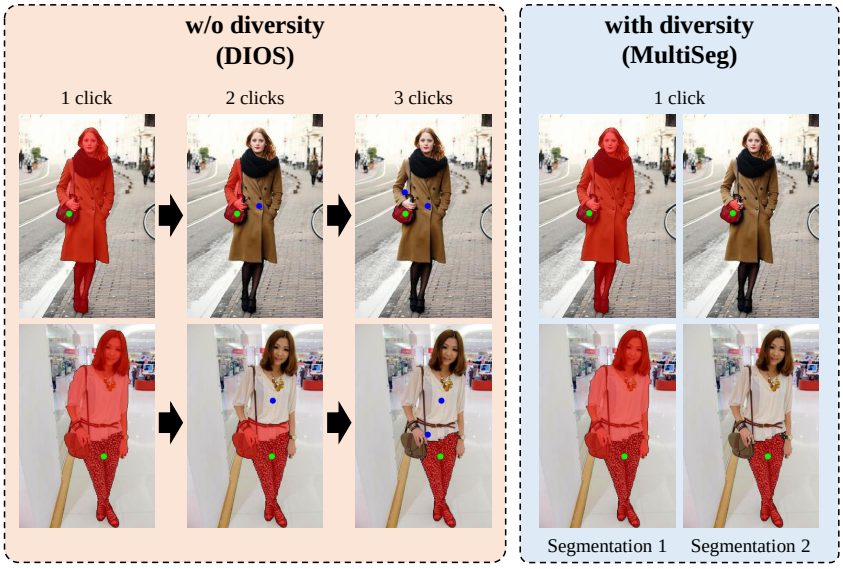

Figure 3: **(Left column)** DIOS typically requires extensive amount of background clicks for deselecting the incorrect regions when the target segmentation is a smaller part. **(Right column)** Our MultiSeg can generate both a smaller part and the larger full object given only a positive click. Note that our MultiSeg was not trained with any parts annotations before.

3. Comparing MultiSeg and [2]

Next, we also qualitatively compare our MultiSeg with the recent work from Li *et al*. [2][1]. The results are shown in Figure 4. Unlike our scale-diversity approach which imposes a set of scale priors into the network for constraining each proposal to respect a predefined scale, the diversity training framework in [2] is unconstrained in that there is no mechanism to encourage different branches to be either meaningful or different from one another.

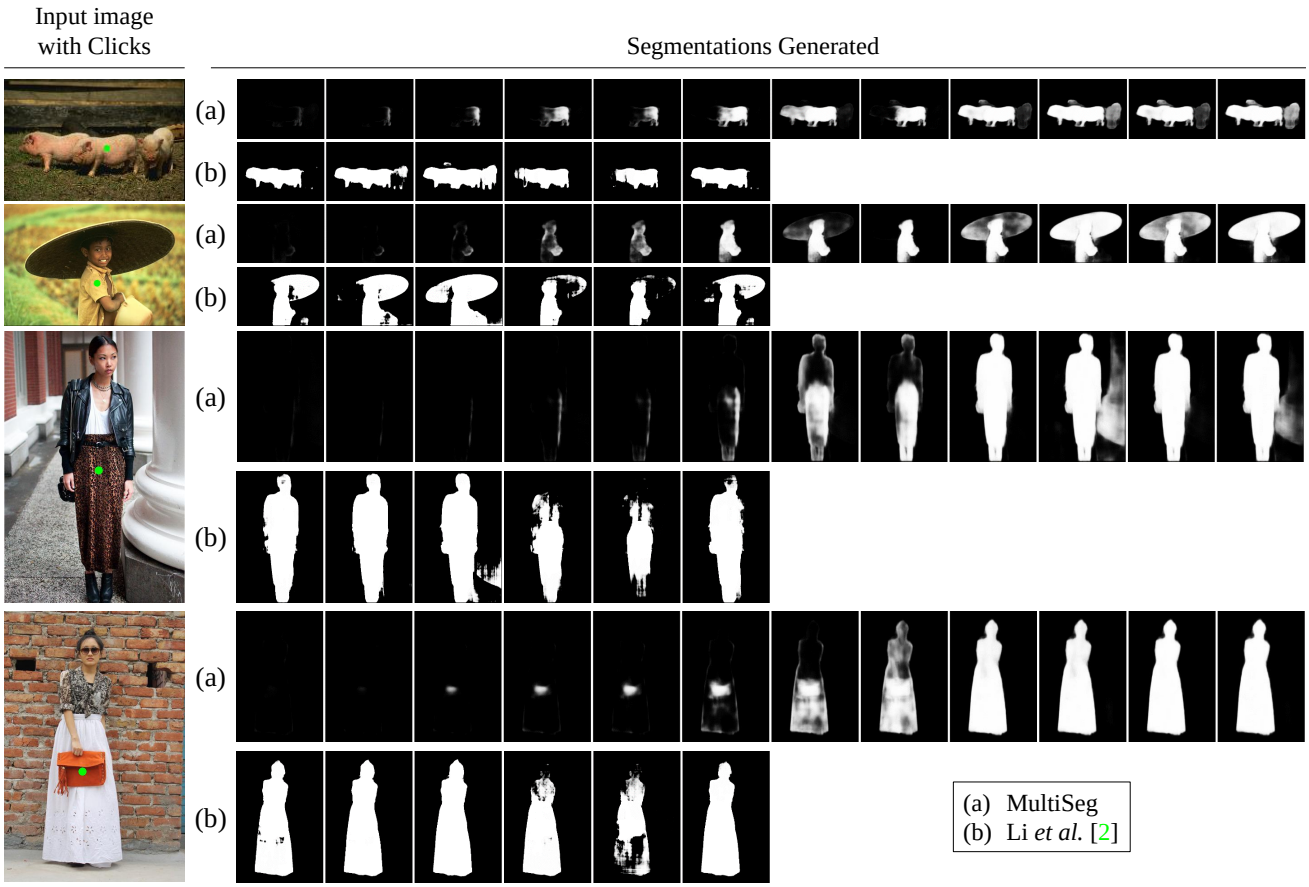

Figure 4: For each example, given a single positive click, the proposals generated by (a) MultiSeg and (b) Li *et al*. [2] are visualized. Note that [2] produces only 6 outputs since they observed that the performance plateaus when the number of solutions increase beyond 6.

4. More Qualitative Results

In addition to the results presented in the main paper, we show more qualitative results of our proposed MultiSeg in Figure 5. Interestingly, **without** being trained with any parts annotations before, our MultiSeg can still segment object parts, such as the car on the champagne bottle, tail light, airplane engine, hand, car wheel, cow nose, hat, motorcycle headlight *etc.* This demonstrates that our MultiSeg generalizes well to unseen object parts and is thus suitable for the task of interactive image segmentation.

In this supplementary material, we also provide some visualization results of our method on the the Fashionista dataset [4] in Figure 6. Note that our MultiSeg can segment skirt, bag, head, boots, lower body *etc.* despite it has not seen such parts-annotated training examples before.

[1] https://github.com/IntelVCL/Intseg

Figure 5: Given one positive (red) and one negative click (blue), the top segmentation results of our proposed MultiSeg after non-maximum suppression (NMS) and graph cut optimization are presented. Interestingly, without being trained with any parts annotations before, our MultiSeg can still segment object parts (*e.g.* tail light, the car on the champagne bottle, airplane engine, arm, car wheel, cow nose, hat, motorcycle headlight *etc.*)

Figure 6: More qualitative results on the Fashionista dataset. Note that our MultiSeg can segment skirt, bag, boots, face, lower body *etc.* despite it has not seen such parts-annotated training samples before.

5. Failure Cases

A possible limitation of our MultiSeg is that it cannot distinguish between two different segmentations if both of them are represented by the same scale. For example, as shown in Figure 7, given only a positive click at the lady, our model cannot segment the lady alone without segmenting the cat (first row). Fortunately, this can be alleviated by adding negative click(s) to deselect the unwanted object (second row).

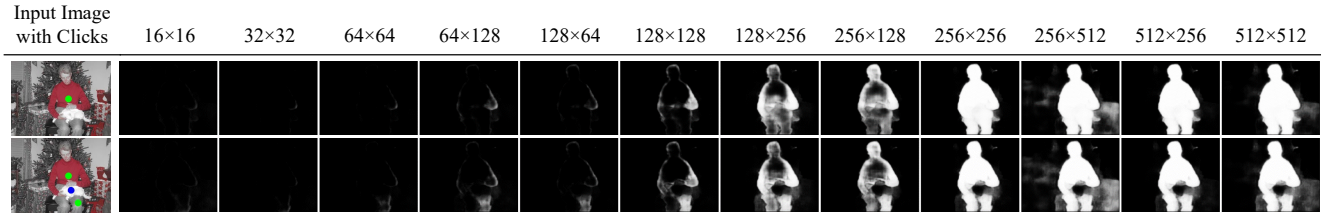

Figure 7: Failure case. Our MultiSeg fails to distinguish two different segmentations if they are represented by the same scale (*e.g.* both the segmentations of (i) the lady alone and (ii) the lady with the cat). However, this problem can be alleviated by introducing additional negative click(s) to deselect the unwanted object.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018.

[2] Zhuwen Li, Qifeng Chen, and Vladlen Koltun. Interactive image segmentation with latent diversity. In *CVPR*, 2018.

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

[4] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. Dataset: https://github.com/lemondan/HumanParsing-Dataset.