Appendix: Cascade Sparse Feature Propagation Network for Interactive Segmentation

Chuyu Zhang*[1,3]
zhangchy2@shanghaitech.edu.cn
Chuanyang Hu*[1]
huchy3@shanghaitech.edu.cn
Hui Ren[1]
renhui@shanghaitech.edu.cn
Yongfei Liu[1]
liuyf3@shanghaitech.edu.cn
Xuming He[1,2]
hexm@shanghaitech.edu.cn

[1] ShanghaiTech University
Shanghai, China
[2] Shanghai Engineering Research Center of Intelligent Vision and Imaging
Shanghai, China
[3] Lingang Laboratory
Shanghai, China

A Datasets

We evaluate our method over a wide range of datasets including GrabCut, Berkeley, DAVIS, COCO and SBD, by following the standard evaluation protocol.

GrabCut[6] is a typical interaction segmentation dataset, which contains 50 images with distinguishable foreground and background.

Berkeley[9] contains 96 images with 100 object masks from its test subset.

DAVIS[5] is originally introduced for video segmentation. Only 345 randomly sampled frames with finely labeled objects are used in our method by following [2].

COCO[3] is a typical semantic segmentation dataset, containing more complex scene and multiscale objects. Following [8], we split the dataset into COCO(seen) and COCO(unseen) according to their object class whether in PASCAL VOC or not. And finally, 10 images are sampled randomly for each category. For simplicity, we denote COCO(seen) and COCO(unseen) as COCO^s and COCO^u

SBD[1] contains 6671 object masks for 2820 images.

B More comparison with SAM on iShape Dataset

To further demonstrate our advantage compared to SAM, we evaluate our model and SAM on the iShape dataset [9], which has many thin structure and requires model to segment details. As in Fig.1, our approach significantly surpass SAM on iShape dataset when number of points exceeds 5, which indicates that out approach is better at segmenting the detailed

(slender and long) part. Additionally, as shown in Fig.2, our approach is better at segmenting the detailed (slender and long) parts.

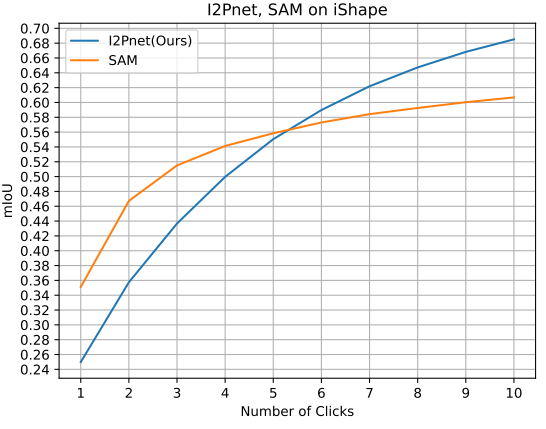

Figure 1: Comparison with SAM on iShape dataset. It is obvious that although SAM has higher mIoU at first, our I2Pnet quickly catches up and surpass SAM as the number of points increases.

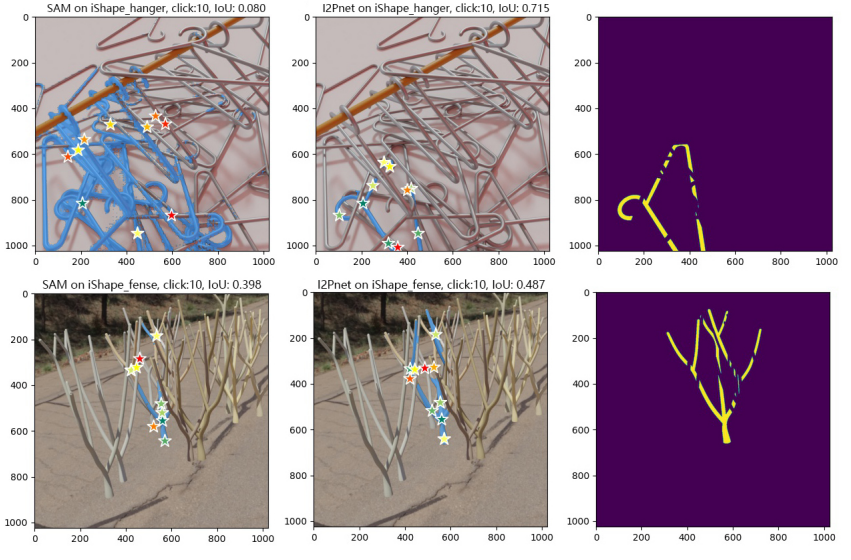

Figure 2: Examples of comparison with SAM on iShape dataset. They are two examples of "hanger" and "branch" classes in iShape dataset. Points with color from dark green to light green are the positive ones added from early to late, while the ones with color from red to yellow are negative points. The rightmost images are the ground truth.

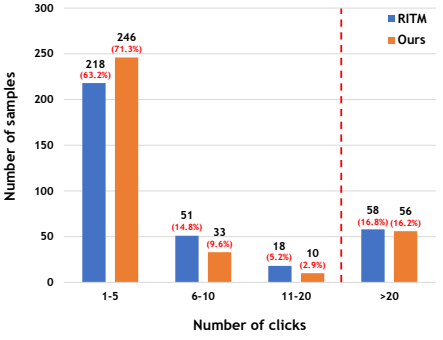

Figure 3: The distribution of the number of clicks on DAVIS dataset. The experiments are on the ResNet-50 backbone. The left of red dotted line is the success samples and the right of red dotted line is the failed samples, which can't reach 90% IoU within 20 clicks. 1-5 means the number of samples that need at least one click and at most five clicks to reach 90% IoU.

Table 1: The performance on DAVIS. Boundary IOU(B-IOU) is evaluated under 20 interactive step.

Method	NoC@90	B-IOU(20 steps)
RITM+CascadePSP	6.82	85.30
RITM	6.68	86.25
FocusCut	6.22	**87.40**
Ours	**5.80**	87.07

C Success samples with different number of clicks

To further analysis our improvement, we report the distribution of success samples with different number of clicks on the challenging fine-grained DAVIS dataset, which contains 345 samples in total. As in Fig.3, our approach can successfully segment 71.3% (246) samples within **5 clicks**, which greatly outperforms the RITM by 8.1%. It indicates that we can better utilize user-provided sparse click information.

D The performance of boundary

To validate our design, we utilize cascadePSP to refine the prediction of RITM in each step. We evaluate NoC@90 and boundary IOU when each image is interacted for 20 steps on the challenging DAVIS dataset. The results in Tab.1 below show that CascadePSP actually hurts the performance. The main reason is that CascadePSP ignores the human input, which is essential for interactive segmentation. Such smoothing may lead to failure cases when the target region includes multiple object classes, such as segmenting people and horses in Fig.1 (main text). More importantly, compared with FocusCut which refine the boundary progressively, we achieve comparable results on boundary IOU and much lower NoC, illustrating our method can capture user intention and segment the boundary simultaneously.

Table 2: **Ablation study of HSGN on ResNet-34 backbone**. Baseline(BS) is our implementation of RITM[7]. The Fuse means directly concatenating the high-resolution feature with the high-level feature and then going through a convolution layer with ReLU.

#	Comparison	DAVIS	SBD
1	BS + SGN	6.62	5.57
2	BS + SGN + Fuse	6.75	5.51
3	BS + SGN + Fuse + SGN	6.52	5.39
4	BS + SGN + HSGN	**6.39**	**5.36**

Table 3: The results on DAVIS under different perturbations.

Method	$a=0$	$a=3$	$a=6$
RITM	6.68	6.89	7.15
Ours	5.75	5.77	6.01

E HSGN analysis

To further investigate the proposed HSGN, we conduct several additional ablations on its design. As shown in Tab.2, when directly fusing the multi-scale features in HSGN without any sparse message propagation, the results in 1st & 2nd row indicate that such design will not bring any performance gain. Furthermore, we apply another SGN to conduct sparse click message propagation on the high-resolution feature map after fusion. The results in 3rd row show that the SGN can propagate annotated click features effectively and benefit the final segmentation results. However, it still underperforms our proposed HSGN due to its inaccurate affinity calculated on low-level features. Overall, our HSGN performs best compared with other variants and achieves 6.39 and 5.36 NoC@90 on DAVIS and SBD respectively.

F Undesired annotated point

We apply random perturbation to the ideal click positions based on a uniform distribution in $[-a,a]^2$. We re-evaluate RITM and our method on DAVIS under $a=\{0,3,6\}$ pixels. The results below indicate that the undesired annotations do cause performance drop. But our method still outperforms the RITM by a large margin and is more robust to the click noises.

References

[1] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011.

[2] Won-Dong Jang and Chang-Su Kim. Interactive image segmentation via backpropagating refinement scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5297–5306, 2019.

[3] 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. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[4] Kevin McGuinness and Noel E O'connor. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2):434–444, 2010.

[5] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016.

[6] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23 (3):309–314, 2004.

[7] Konstantin Sofiiuk, Ilia A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. *arXiv preprint arXiv:2102.06583*, 2021.

[8] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 373–381, 2016.

[9] Lei Yang, Yan Zi Wei, Yisheng He, Wei Sun, Zhenhang Huang, Haibin Huang, and Haoqiang Fan. ishape: A first step towards irregular shape instance segmentation. *arXiv preprint arXiv:2109.15068*, 2021.