Interactive Image Segmentation with Cross-Modality Vision Transformers Supplementary Material

Kun Li, George Vosselman, Michael Ying Yang
University of Twente, The Netherlands
{k.li, george.vosselman, michael.yang}@utwente.nl

In this supplementary document, we provide detailed explanation on the architecture of the proposed iCMFormer in Sec. A. Additional quantitative results in terms of the mIoU curves and number of failures are provided in Sec. B, together with an ablation study on the number of cross-modality blocks in Sec C. Moreover, we also provide more qualitative results evaluated on the four datasets in Sec. D.

A. Implementation Details

In the main paper, we explain the overall pipeline of the proposed iCMFormer for two different backbones. For better readability and reproducibility, we present the architecture in detail. As the transformer technique is quite popular, we do not expand the multi-head attentions for each block, and only report the dimension as well as the number of corresponding heads. Our iCMFormer for ViT-B and Swin-B backbones are showsn in Tab. 1.

B. Additional Quantitative Results

In the main paper, we report the complete comparison results with respect to the Number of Clicks (NoC). Due to the limited space, here we further provide the evaluation results in terms of mean IoU curves and Number of Failures (NoF) to make the comparison consistent with the employed evaluation protocol.

We report the automatically evaluation results on GrabCut [10] and Berkeley [8] in Fig. 1 for demonstrating the segmentation performance with progressively added clicks. We can see that the proposed methods achieve higher mIoU values within the same number of clicks compared with other models. However, restricted in the sizes of evaluation samples in GrabCut (50) and Berkeley (100), different variants of our methods do not make a huge difference especially when only providing two clicks (already above 90% mIoU).

In addition, we compared the stability of our method with that of others in Tab. 2 using 20 clicks for two thresholds: 85% and 90%. As the previous methods did not report the numbers for GrabCut and Berkeley, we do not add the values in the table (Ours-Swin-B only gets both 0 failure on GrabCut and 0, 1 failure on Berkeley for 85% and 90% IoU, respectively). The models trained in SBD [4] and COCO [6] + LVIS [3] are divided into two parts for fair comparison. We only report the numbers that are provided by the original papers and their released models considering we cannot re-implement their models perfectly. As shown in the table, our method reduces the numbers of failure cases for both thresholds, which show the potential to be a practical annotation tool with robust predictions.

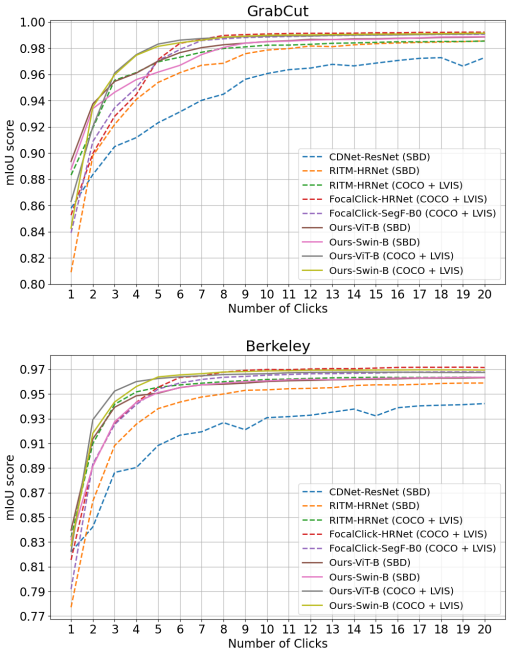

Figure 1: Convergence analysis of mean IoU curves for varying number of clicks compared with other methods on GrabCut [10] and Berkeley [8].

Table 1: The detailed architecture of the iCMFormer with ViT-B and Swin-B backbones. Numbers in square brackets [] mean the input and hidden dimensions, respetively, while the numbers in parentheses () denote the dimension changes in the Conv2d or ConvTranspose2d (only utilized in the Neck) or Linear operations. We set 8 as the numbers of heads for all blocks in ViT-B, and 4, 8, 16, 32 for 4 original stages in Swin-B. The number is set 8 for all cross-attentions for both backbones. We adopt the original position embeddings for both backbones.

Layer Name	Ours-ViT-B		Ours-Swin-B	
Patch-Embed	(3, 768)		(3, 128)	
Shared Group	[768, 2304] + (768, 3072, 768)	x6	[128, 384] + (128, 512, 128)	x2
			[256, 768] + (256, 1024, 256)	x2
Cross-Attention	[768, 2304] + [768, 2304] + (768, 3072, 768)	x3	[512, 1536] + [512, 1536] + (512, 2048, 512)	x4
Combined Group	[768, 2304] + (768, 3072, 768)	x6	[512, 1536] + (512, 2048, 512)	x18
			[1024, 3072] + (1024, 4096, 1024)	x2
Neck	(768, 384, 192, 128) (768, 384, 256) (768, 512) (768, 1536, 1024)		-	
Head	(128, 256) (256, 256) (512, 256) (1024, 256) (256×4, 256, 1)		(128, 256) (256, 256) (512, 256) (1024, 256) (256×4, 256, 1)	

C. Number of Cross-Modality Blocks

We further evaluate the impact of different number of the proposed cross-modality blocks on the performance of our backbones. Simply, we train all the models on SBD [4] and evaluate the results on four datasets with the NoC metric. Tab. 3 shows the corresponding results. As the number of layers increases, the trend of the number of clicks (NoC) shows an initial rise followed by a subsequent decline. Due to the better overall performance, we set 3 and 4 as the default numbers for ViT-B and Swin-B backbones, respectively.

D. More Qualitative Results

We also provide more segmentation results of our iCMFormer on the four datasets. Fig. 2 shows the common cases from GrabCut [10] and Berkeley [8], and Fig. 3 represents common cases from SBD [4] and DAVIS [9]. As shown in Fig. 4, we display some challenging cases where it requires more than the average number of clicks to get the target IoU. We report the segmentation results in the middle stages until reaching 90% IoU. However, there still exist some bad cases due to the limitations of our method, and Fig. 5 shows two examples from DAVIS.

Normally, the qualitative results are collected from the human evaluation while the clicks are based on his/her subjective evaluation (different every time). In other words, the qualitative comparison with other methods could be unfair considering the judgement of the results and potential added clicks could be totally different for different users. To complete the visualization, we only show the compared results within only one positive click in Fig. 6. We randomly pick several examples from the four datasets and put the positive click in the same place for the fair comparison. These figures also verify the superiority of our proposed method.

Table 2: Comparison with previous models in term of number of failures (NoF) that cannot reach the target IoUs after 20 clicks, denoted as ≥20@85 and ≥20@90, respectively. The results are divided into 2 sections on the basis of the training datsets: SBD [4] (represented as †) and COCO [6] + LVIS [3] (represented as ‡). The best results are **bold**.

Method	SBD		DAVIS	
	≥20@85	≥20@90	≥20@85	≥20@90
BRS[5]†	-	-	-	77
f-BRS[11]†	-	1466	-	78
CDNet[1]†	-	-	46	65
FocalClick[2]†	-	-	-	55
FocusCut[7]†	-	-	-	57
FCF[13]†	-	-	-	59
Ours-ViT-B†	**236**	**693**	**30**	**53**
Ours-Swin-B†	242	698	36	**53**
RITM-HRNet-18[12]‡	-	-	52	91
FocalClick-HRNet-18[2]‡	-	-	49	77
FocalClick-SegF-B0-S2[2]‡	-	-	50	86
Ours-ViT-B‡	**225**	695	**20**	49
Ours-Swin-B‡	237	**667**	**20**	**48**

Table 3: Ablation study for the number of proposed cross-modality blocks on GrabCut [10], Berkeley [8], SBD [4] and DAVIS [9] datasets. NoC85 and NoC90 denote the average numbers of clicks to reach a target IoU. All the models are trained on SBD. The best results are **bold** while the second best are underlined.

Method	Layer	Params/M	GrabCut		Berkeley		SBD		DAVIS	
			NoC85	NoC90	NoC85	NoC90	NoC85	NoC90	NoC85	NoC90
Ours-ViT-B	1	105.90	1.46	1.68	1.50	2.56	**3.28**	**5.25**	4.20	5.60
Ours-ViT-B	2	115.36	1.44	1.52	1.46	2.55	3.32	5.31	4.09	5.62
Ours-ViT-B	3	124.81	**1.36**	**1.42**	**1.42**	2.52	3.33	5.31	**4.05**	5.58
Ours-ViT-B	6	153.16	1.52	1.58	1.47	2.54	3.37	5.36	4.17	5.75
Ours-ViT-B	8	172.07	1.54	1.66	1.59	**2.45**	3.32	5.30	4.10	5.54
Ours-Swin-B	1	91.64	1.48	1.56	1.56	2.57	3.31	5.41	4.38	6.07
Ours-Swin-B	2	95.84	1.42	1.62	1.56	2.58	3.28	5.25	**4.18**	5.70
Ours-Swin-B	4	104.25	1.46	**1.50**	**1.52**	**2.32**	**3.21**	**5.16**	4.25	**5.55**
Ours-Swin-B	6	112.66	1.46	1.62	1.55	2.64	3.24	5.29	4.34	5.68
Ours-Swin-B	8	121.06	**1.40**	1.62	1.55	2.50	3.28	5.34	4.25	5.67

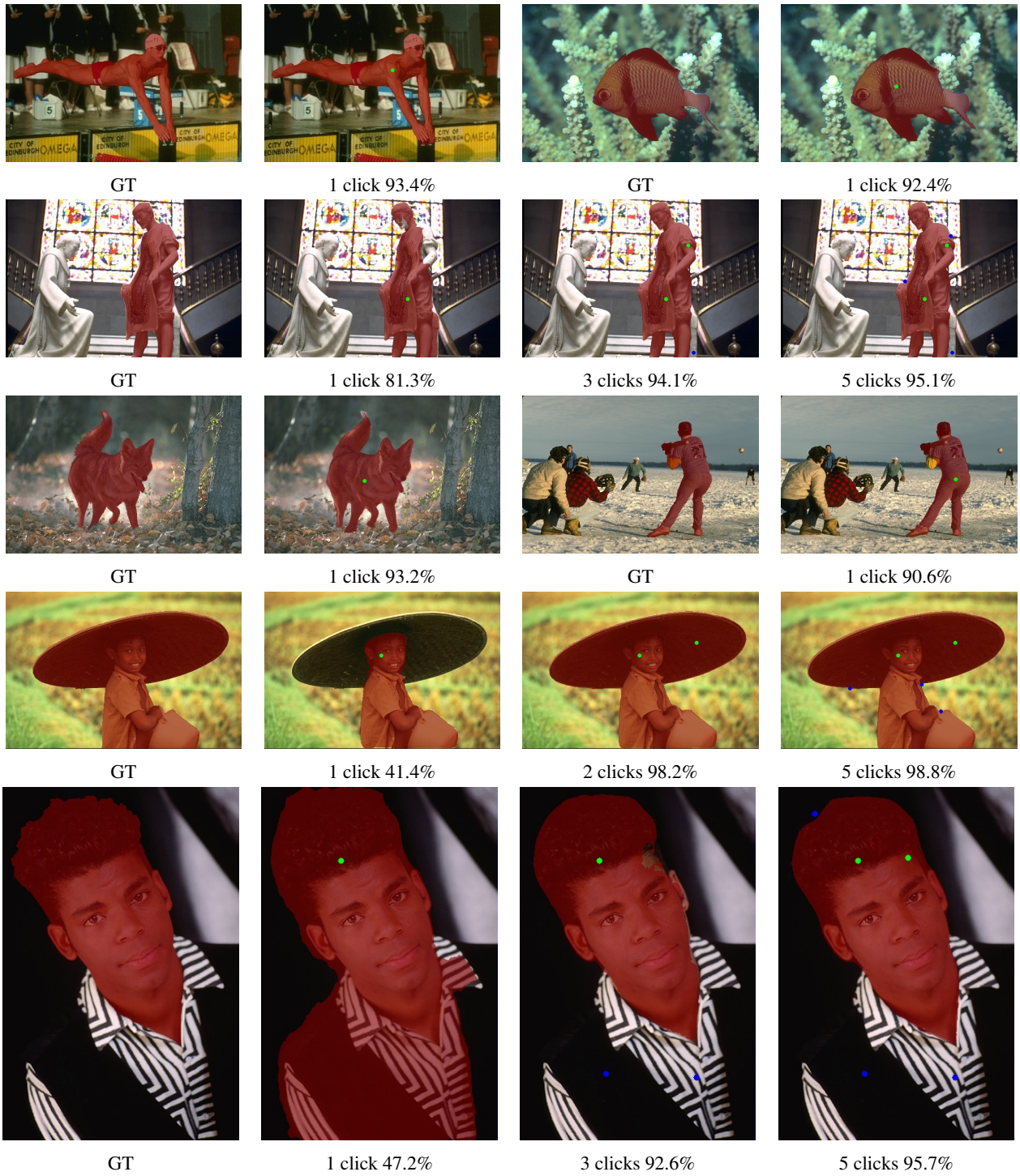

Figure 2: More visualizations of the segmentation results from GrabCut [10] (Row 1-2) and Berkeley [8] (Row 3-5). Green and blue dots denote positive and negative clicks, respectively.

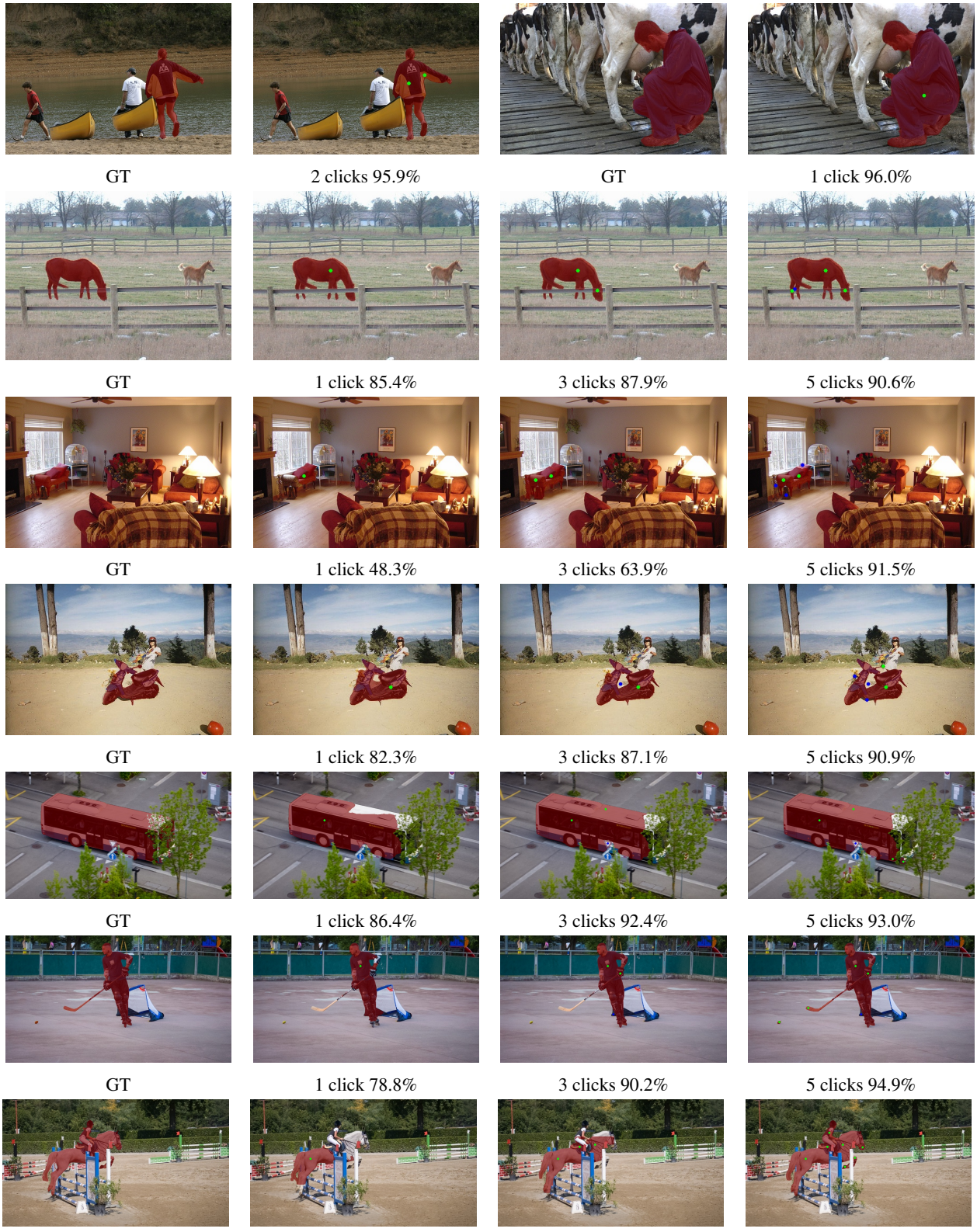

Figure 3: More visualizations of the segmentation results from SBD [4] (Row 1-4) and DAVIS [9] (Row 5-7).

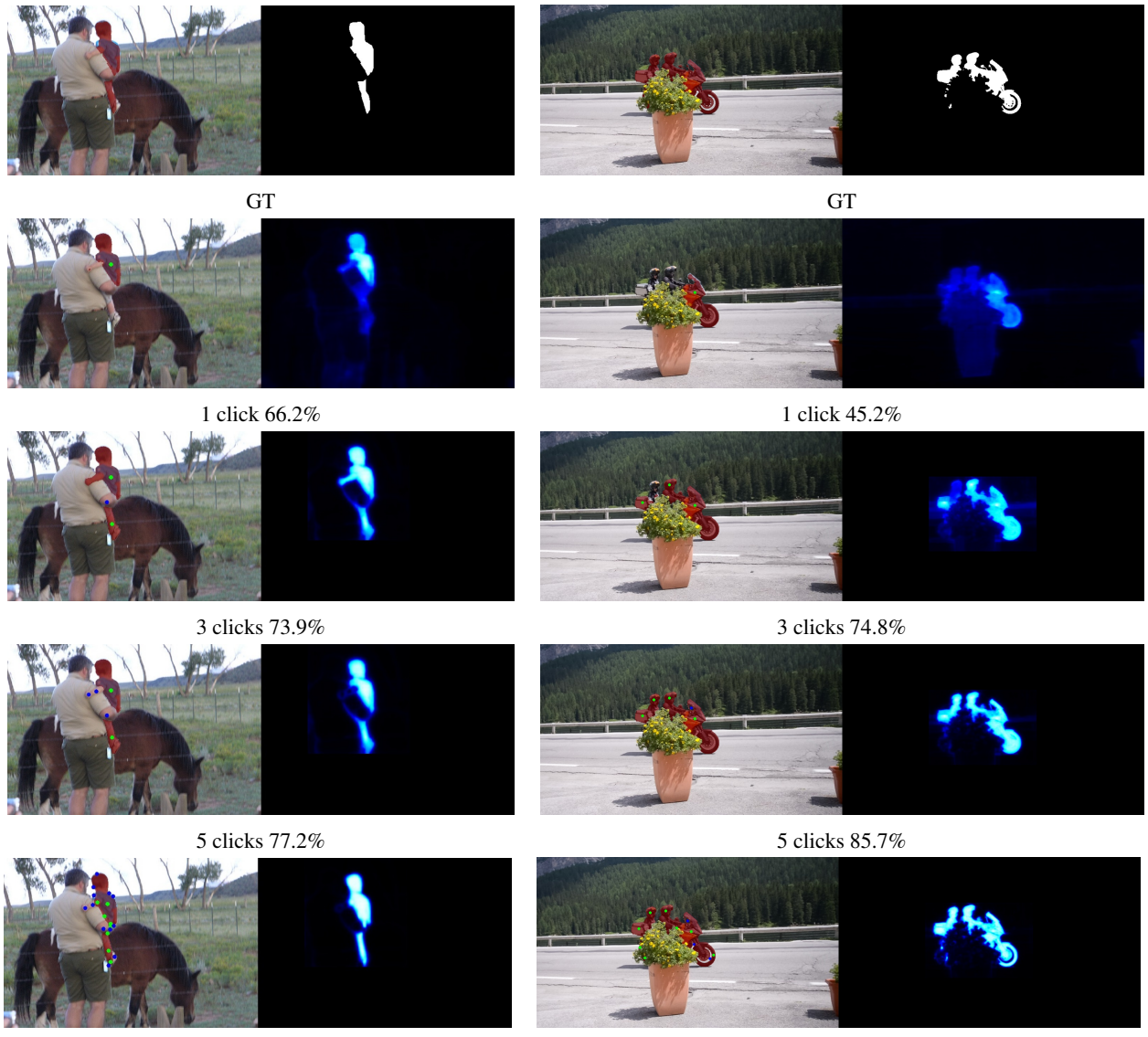

Figure 4: Some of the challenging cases from SBD [4] (left) and DAVIS [9] (right). Green and blue dots denote positive and negative clicks, respectively. The segmentation probability maps are displayed next to the images with overlaid masks.

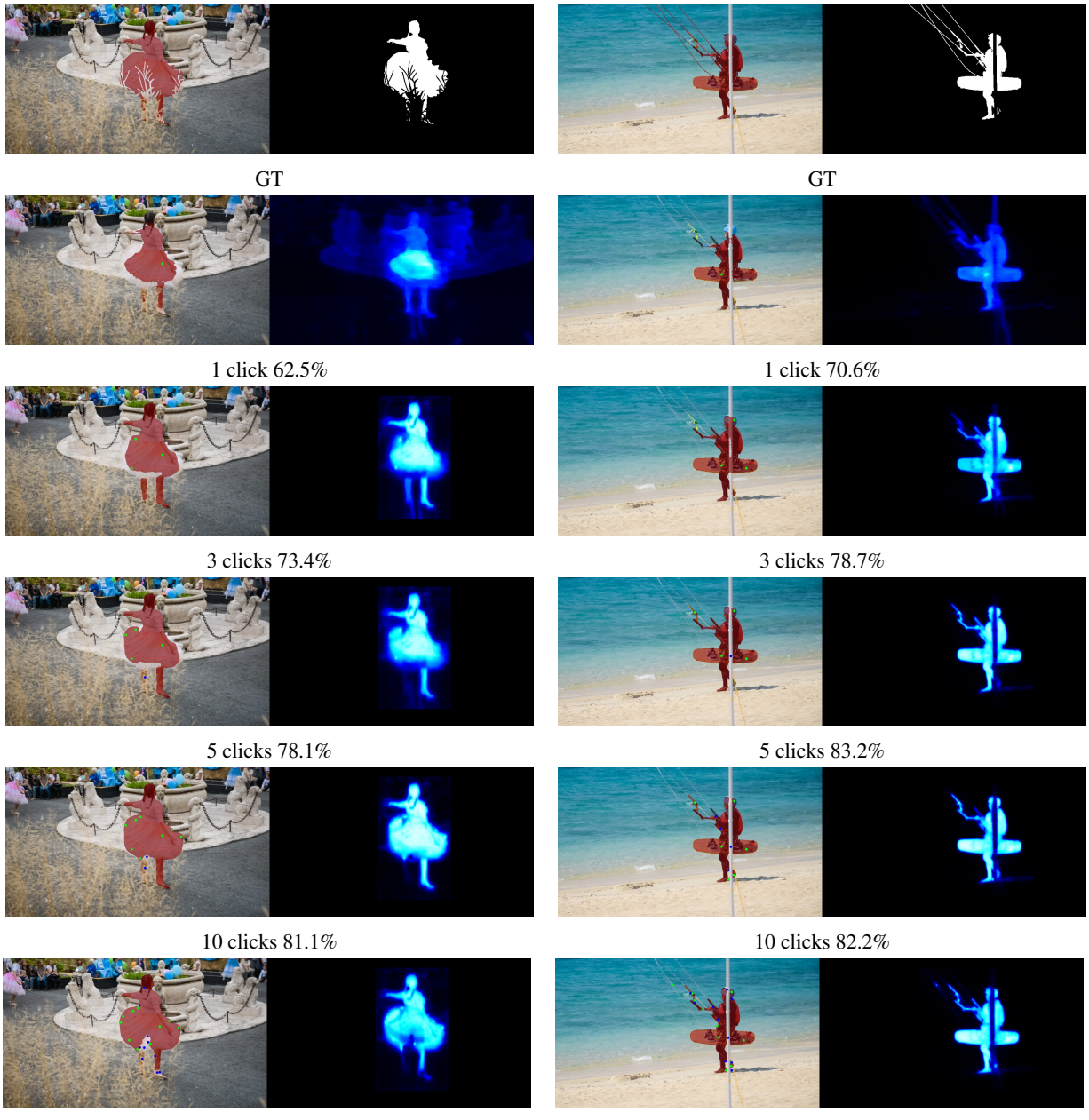

Figure 5: Some of the bad cases from DAVIS [9]. The segmentation probability maps are displayed next to the images with overlaid masks.

Figure 6: More visualizations of the qualitative comparison with other methods within one positive click.

References

[1] Xi Chen, Zhiyan Zhao, Feiwu Yu, Yilei Zhang, and Manni Duan. Conditional diffusion for interactive segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7345–7354, 2021. 3

[2] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: towards practical interactive image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1300–1309, 2022. 3

[3] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 1, 3

[4] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 991–998, 2011. 1, 2, 3, 5, 6

[5] Won-Dong Jang and Chang-Su Kim. Interactive image segmentation via backpropagating refinement scheme. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5306, 2019. 3

[6] 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. 1, 3

[7] Zheng Lin, Zheng-Peng Duan, Zhao Zhang, Chun-Le Guo, and Ming-Ming Cheng. Focuscut: Diving into a focus view in interactive segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2637–2646, 2022. 3

[8] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 416–423, 2001. 1, 2, 3, 4

[9] 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. 2, 3, 5, 6, 7

[10] 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. 1, 2, 3, 4

[11] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8623–8632, 2020. 3

[12] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145, 2022. 3

[13] Qiaoqiao Wei, Hui Zhang, and Jun-Hai Yong. Focused and collaborative feedback integration for interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18643–18652, 2023. 3