Reviving Iterative Training with Mask Guidance for Interactive Segmentation

Konstantin Sofiiuk, Ilia A. Petrov* and Anton Konushin**

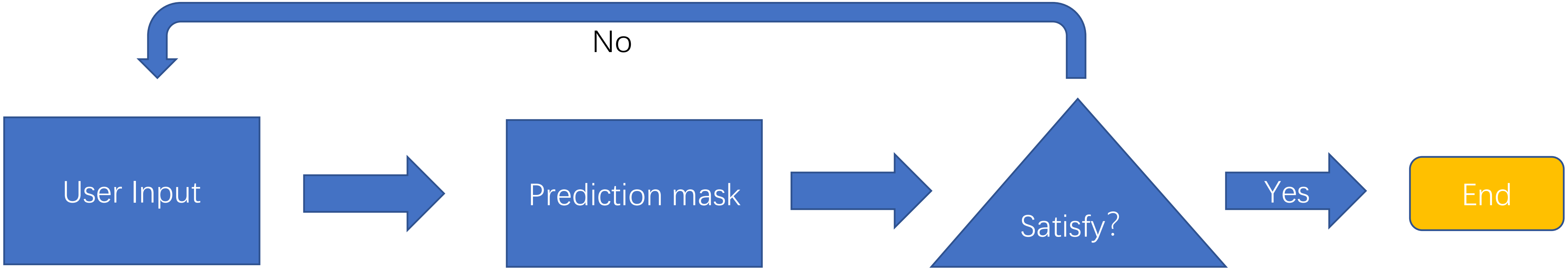
Visual Understanding lab., AI Center Moscow, Samsung Electronics Co., Lesnaya 5C, Moscow, Russia

RITM	**icip2022**	**36**
EdgeFlow	iccv2021	17
CDNet	iccv2021	13
FocalClick	cvpr2022	6
FocusCut	cvpr2022	4
PseudoClick	eccv2022	2

Task

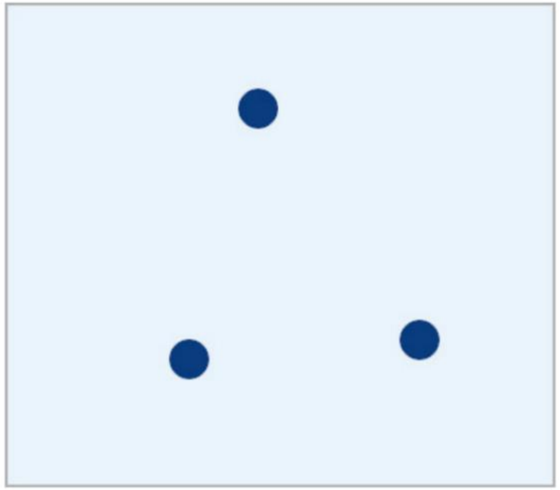
class agnostic segmentation with user's input

Pipeline

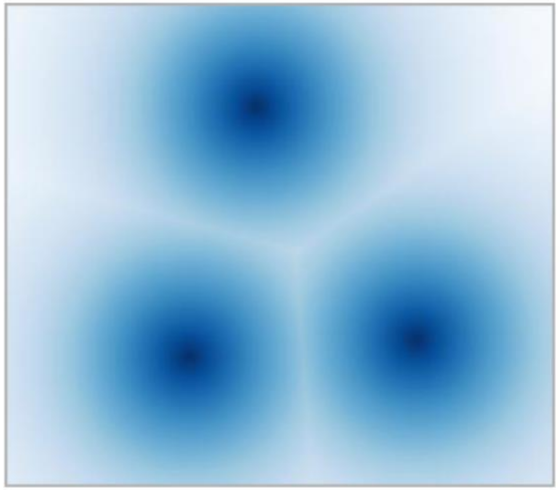


The key difference is in the user input:
its main aspects are the **encoding** and **processing of the encoded input**

encoding



Disks with a small radius

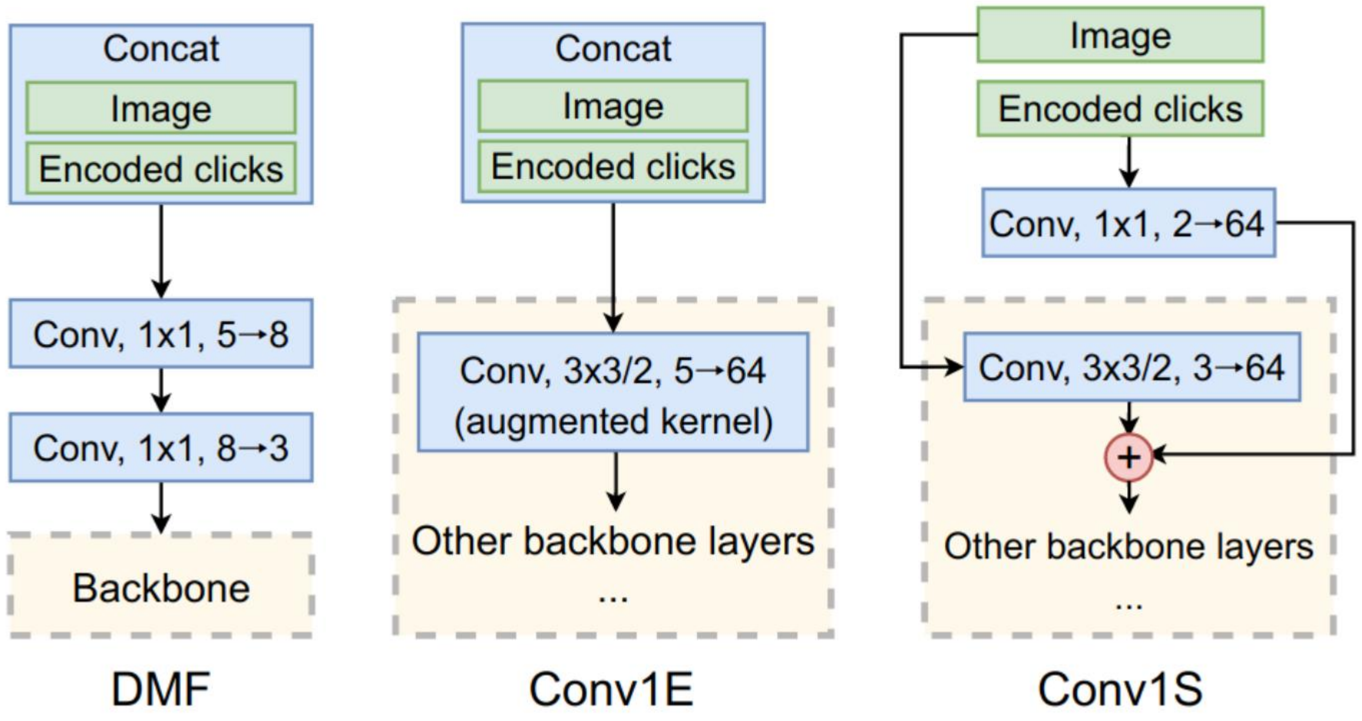


Distance Transform (DT)

The changes in disk encoding caused by adding new points or moving existing ones are always **local** and only **slightly** affect the encoding map.

A distance transform map can **change drastically** when a new point is added, especially if there are only a few points. In turn, such sudden considerable changes might **confuse a network**.

processing of the encoded input

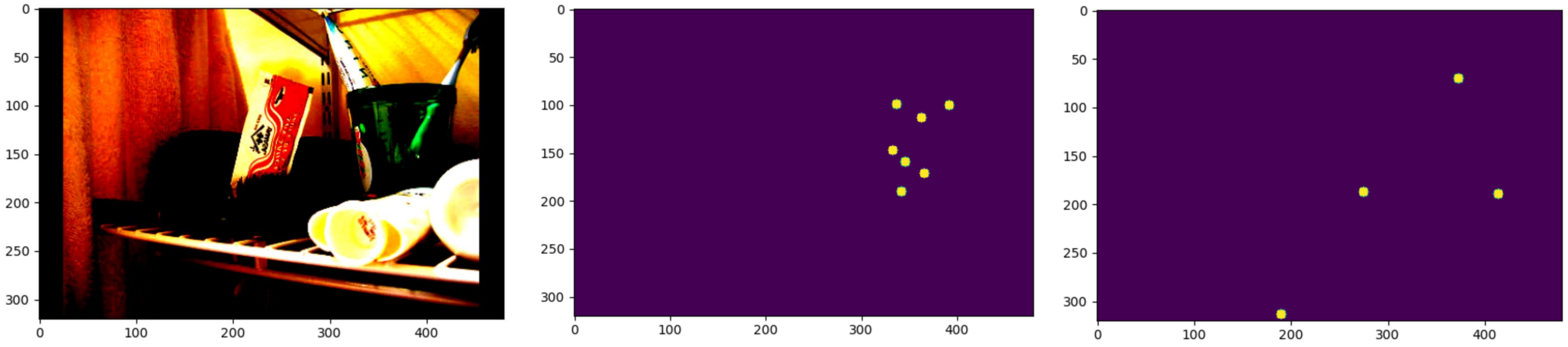


Iterative Sampling Strategy

```
for click_indx in range(num_iters):
        output = self.net(img, points)
        points = get_next_points(output, gt_mask, points)
```

Iterative Sampling Strategy + Random Sampling Strategy

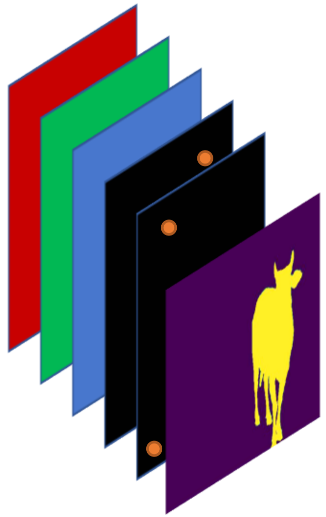
random sampling is used for initialization and then
a few clicks are added using the iterative sampling procedure



Incorporating Masks From Previous Steps

- providing additional prior information that can help improve the quality of prediction

- Our model takes this mask as the third channel together with two channels for positive and negative encoded clicks, respectively.



[R, G, B, foreground click, background click, previous mask]

Normalized Focal Loss

$$FL(i,j) = -(1-p_{i,j})^{\gamma} \log p_{i,j}$$

$$P(\hat{M}) = \sum_{i,j} (1-p_{i,j})^{\gamma}$$

$$NFL(i,j,\hat{M}) = -\frac{1}{P(\hat{M})}(1-p_{i,j})^{\gamma} \log p_{i,j}$$

$P(\hat{M})$ decreases when the accuracy of the prediction increases

The gradient of NFL does not fade over time due to normalization

Evaluation metric

Number of Clicks (NoC)：

the number of clicks required to achieve the predefined IoU

NoC@85

NoC@90

eg：

NoC@85%	NoC@90%
1.54	1.68

clicks limit = 20

Method		GrabCut		Berkeley	SBD		DAVIS		Pascal VOC
		NoC@85	NoC@90	NoC@90	NoC@85	NoC@90	NoC@85	NoC@90	NoC@85
GC [15]		7.98	10.00	14.22	13.60	15.96	15.13	17.41	–
GM [17]		13.32	14.57	15.96	15.36	17.60	18.59	19.50	–
RW [16]		11.36	13.77	14.02	12.22	15.04	16.71	18.31	–
ESC [17]		7.24	9.20	12.11	12.21	14.86	15.41	17.70	–
GSC [17]		7.10	9.12	12.57	12.69	15.31	15.35	17.52	–
DIOS with GC [1]		–	6.04	8.65	–	–	–	–	6.88
Latent diversity [19]		3.20	4.79	–	7.41	10.78	5.05	9.57	–
RIS-Net [20]		–	5.00	6.03	–	–	–	–	5.12
ITIS [14]		–	5.60	–	–	–	–	–	3.80
CAG [36]		–	3.58	5.60	–	–	–	–	3.62
BRS [2]		2.60	3.60	5.08	6.59	9.78	5.58	8.24	–
FCA-Net (SIS) [22]		–	2.08	3.92	–	–	–	7.57	2.69
IA+SA [3]		–	3.07	4.94	–	–	5.16	–	3.18
f-BRS-B [4]		2.50	2.98	4.34	5.06	8.08	5.39	7.81	–
Ours SBD	H18	1.96	2.41	3.95	4.12	6.66	5.08	7.17	2.94
	H18 IT-M	1.76	2.04	3.22	**3.39**	**5.43**	4.94	6.71	<u>2.51</u>
Ours C+L	H18	1.54	1.70	2.48	4.26	6.86	4.79	6.00	2.59
	H18s IT-M	1.54	1.68	2.60	4.04	6.48	4.70	5.98	2.57
	H18 IT-M	**1.42**	**1.54**	<u>2.26</u>	3.80	6.06	<u>4.36</u>	<u>5.74</u>	**2.28**
	H32 IT-M	<u>1.46</u>	<u>1.56</u>	**2.10**	<u>3.59</u>	<u>5.71</u>	**4.11**	**5.34**	2.57

Ablation Studies

Table 1
Ablation studies of the network architecture choices described in Section 3.1. Each cell consists of two results "X/Y", where "X" and "Y" correspond to evaluation without and with f-BRS-B[4], respectively. "DT" stands for the distance transform clicks encoding. All models are trained on SBD.

Backbone	Input Scheme	Clicks Encoding	NoC_{20}@90	
			Berkeley	DAVIS
ResNet-34	DMF [4]	DT	5.50/4.32	8.45/8.34
	Conv1E	DT	4.79/4.43	7.56/7.60
	Conv1S	DT	4.98/4.16	7.41/7.28
	Conv1S	Disk3	4.52/4.04	7.27/7.18
	Conv1S	Disk5	4.09/3.89	6.92/7.22
HRNet-18	DMF [4]	DT	4.93/4.35	8.59/8.00
	Conv1E	DT	4.41/3.95	7.50/7.43
	Conv1S	DT	3.99/3.81	7.16/7.24
	Conv1S	Disk3	3.63/3.47	7.14/7.04
	Conv1S	Disk5	3.52/3.50	6.90/6.97

HRNet-18 and ResNet-34 models with Conv1S show better performance

Disk encoding significantly improves results of both HRNet-18 and ResNet-34

Disk + Conv1S + HRNet

Ablation Studies

Method	$NoC_{20}@90$			
	GrabCut	Berkeley	SBD	DAVIS
BCE	1.82	3.13	7.58	6.31
Soft IoU	2.02	3.03	7.94	6.45
FL	1.80	3.28	7.56	6.40
NFL	**1.70**	**2.48**	**6.72**	**5.90**

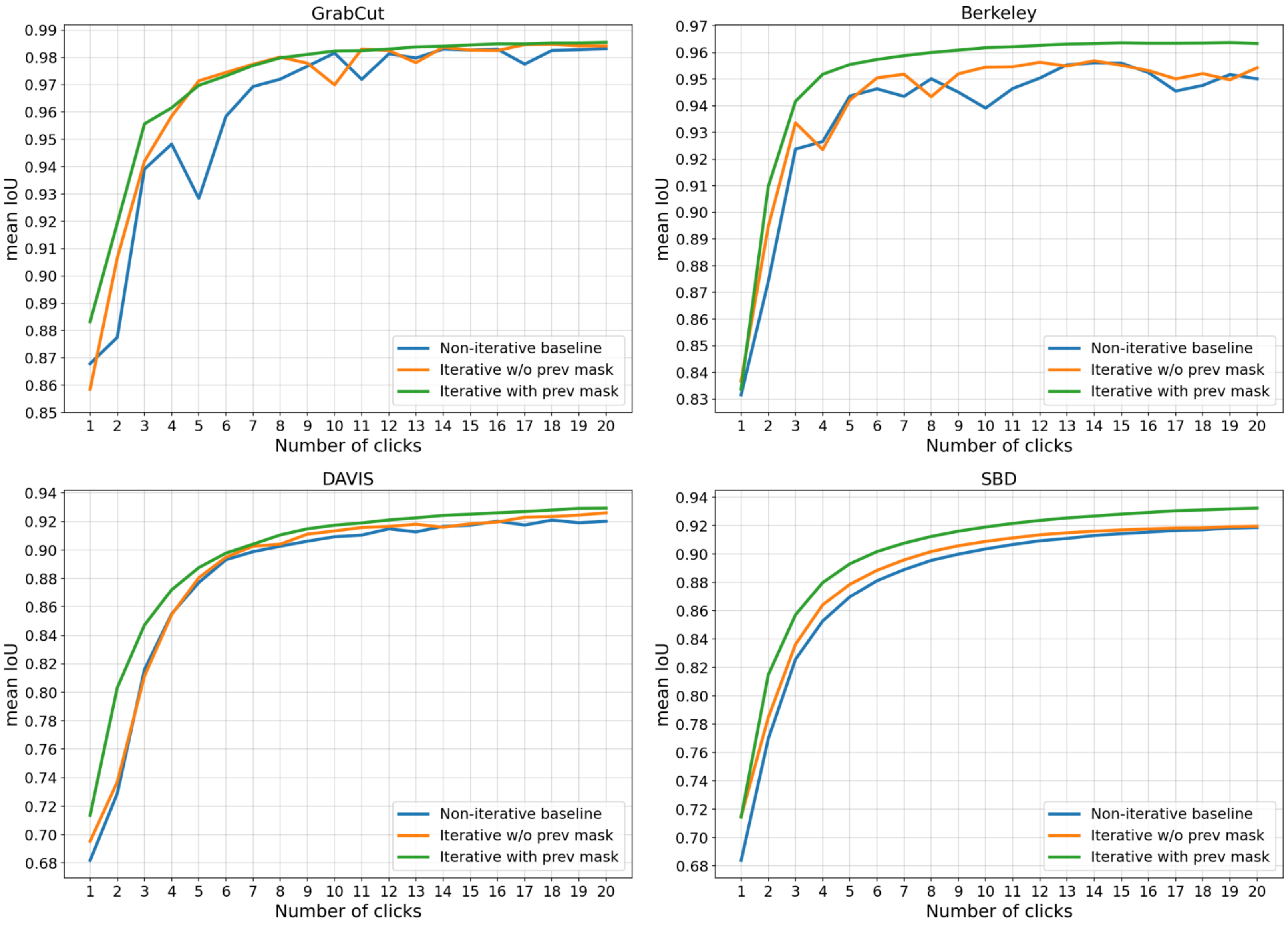
NFL leads to better accuracy and convergence on all 4 datasets

Ablation Studies

N_{iters}	Prev Mask	NoC$_{20}$@90		
		Berkeley	DAVIS	SBD
3	−	2.38	5.92	6.49
3	+	2.26	5.74	6.06
1	+	2.57	5.81	6.15
2	+	2.48	**5.70**	6.10
3	+	**2.26**	5.74	6.06
4	+	2.52	6.03	**6.04**
5	+	2.49	5.98	6.24
6	+	2.55	6.11	6.82

too high N values (> 4) lead to instability during training and to worse results

Ablation Studies



model that takes a mask from a previous step is much more stable and converges to a better IoU

Method	Train Data	GrabCut [34]		Berkeley [31]	SBD [15]		DAVIS [33]	
		NoC 85	NoC 90	NoC 90	NoC 85	NoC 90	NoC 85	NoC 90
Graph cut [3]	/	7.98	10.00	14.22	13.6	15.96	15.13	17.41
Geodesic matting [12]	/	13.32	14.57	15.96	15.36	17.60	18.59	19.50
Random walker [11]	/	11.36	13.77	14.02	12.22	15.04	16.71	18.31
Euclidean star convexity [12]	/	7.24	9.20	12.11	12.21	14.86	15.41	17.70
Geodesic star convexity [12]	/	7.10	9.12	12.57	12.69	15.31	15.35	17.52
DOS w/o GC [44]	SBD [15]	8.02	12.59	–	14.30	16.79	12.52	17.11
DOS with GC [44]	SBD [15]	5.08	6.08	–	9.22	12.80	9.03	12.58
Latent diversity [22]	SBD [15]	3.20	4.79	–	7.41	10.78	5.05	9.57
RIS-Net [23]	SBD [15]	–	5.00	–	6.03	–	–	–
CM guidance [30]	SBD [15]	–	3.58	5.60	–	–	–	–
BRS [18]	SBD [15]	2.60	3.60	5.08	6.59	9.78	5.58	8.24
f-BRS-B-resnet50 [35]	SBD [15]	2.50	2.98	4.34	5.06	8.08	5.39	7.81
CDNet-resnet50 [5]	SBD [15]	2.22	2.64	3.69	4.37	7.87	5.17	6.66
RITM-hrnet18 [36]	SBD [15]	1.76	2.04	3.22	**3.39**	**5.43**	4.94	6.71
Ours-hrnet18s-S2	SBD [15]	1.86	2.06	**3.14**	4.30	6.52	**4.92**	**6.48**
Ours-segformerB0-S2	SBD [15]	**1.66**	**1.90**	3.14	4.34	6.51	5.02	7.06
FCANet (SIS) [27]	SBD [15]+PASCAL [9]	-	2.14	4.19	-	-	-	7.90
99%AccuracyNet [10]	SBD [15]+Synthetic	-	1.80	3.04	3.90	-	-	-
f-BRS-B-hrnet32 [35]	COCO [26]+LVIS [13]	1.54	1.69	2.44	4.37	7.26	5.17	6.50
RITM-hrnet18s [36]	COCO [26]+LVIS [13]	1.54	1.68	2.60	4.04	6.48	4.70	5.98
RITM-hrnet32 [36]	COCO [26]+LVIS [13]	1.46	1.56	2.10	3.59	5.71	4.11	5.34
EdgeFlow-hrnet18 [14]	COCO [26]+LVIS [13]	1.60	1.72	2.40	-	-	4.54	5.77
Ours-hrnet18s-S1	COCO [26]+LVIS [13]	1.64	1.82	2.89	4.74	7.29	4.77	6.56
Ours-hrnet18s-S2	COCO [26]+LVIS [13]	1.48	1.62	2.66	4.43	6.79	3.90	5.25
Ours-hrnet32-S2	COCO [26]+LVIS [13]	1.64	1.80	2.36	4.24	6.51	4.01	5.39
Ours-segformerB0-S1	COCO [26]+LVIS [13]	1.60	1.86	3.29	4.98	7.60	5.13	7.42
Ours-segformerB0-S2	COCO [26]+LVIS [13]	**1.40**	1.66	2.27	4.56	6.86	4.04	5.49
Ours-segformerB3-S2	COCO [26]+LVIS [13]	1.44	**1.50**	**1.92**	**3.53**	**5.59**	**3.61**	**4.90**
Ours-hrnet32-S2	Large Dataset	1.30	1.34	1.85	4.35	6.61	3.19	4.81
Ours-segformerB3-S2	Large Dataset	1.22	1.26	1.48	3.70	5.84	2.92	4.52