Seminar on Interactive Segmentation

Haichao Zhang

Paper List

◆CVPR2019: Interactive Image Segmentation via Backpropagating Refinement Scheme

◆CVPR2020: f-BRS: Rethinking Backpropagating Refinement for Interactive Segmentation

◆CVPR2020: Iteratively-Refined Interactive 3D Medical Image Segmentation with Multi-Agent Reinforcement Learning

◆CVPR2020: Memory Aggregation Networks for Efficient Interactive Video Object Segmentation

Interactive Image Segmentation via Backpropagating Refinement Scheme

Won-Dong Jang
Harvard University
Cambridge, MA
wdjang@g.harvard.edu

Chang-Su Kim
Korea University
Republic of Korea
changsukim@korea.ac.kr

Introduction

Problem:

- The user-annotated locations can be mislabeled in the initial result.

Key Idea:

- Development of a new CNN architecture for interactive image segmentation.
- Proposed backpropagating refinement strategy.
- Generalization of BRS, which can make existing CNNs user-interactive without extra training.
- SOTA performance on the GrabCut [42],Berkeley [34], DAVIS [37], and SBD [12] datasets

Pipeline

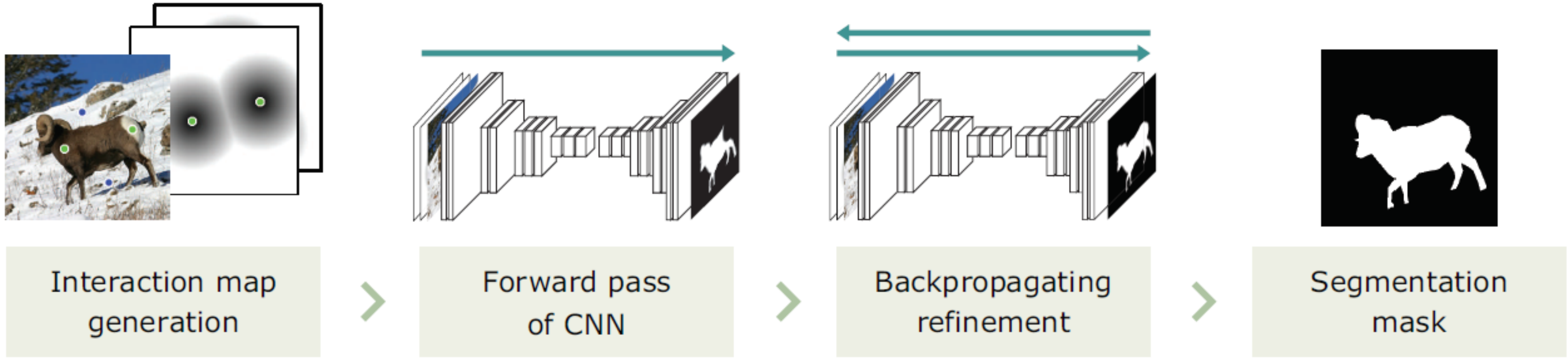


Figure 1. Overview of the proposed algorithm: we perform this segmentation process again when a user provides a new annotation.

1. Feed the input image and the interaction maps into a CNN, get initial probability map.
2. Force the clicked locations to have the user-specified labels by employing the proposed BRS.
3. Repeat 1,2 for servral rounds.

Proposed Architecture

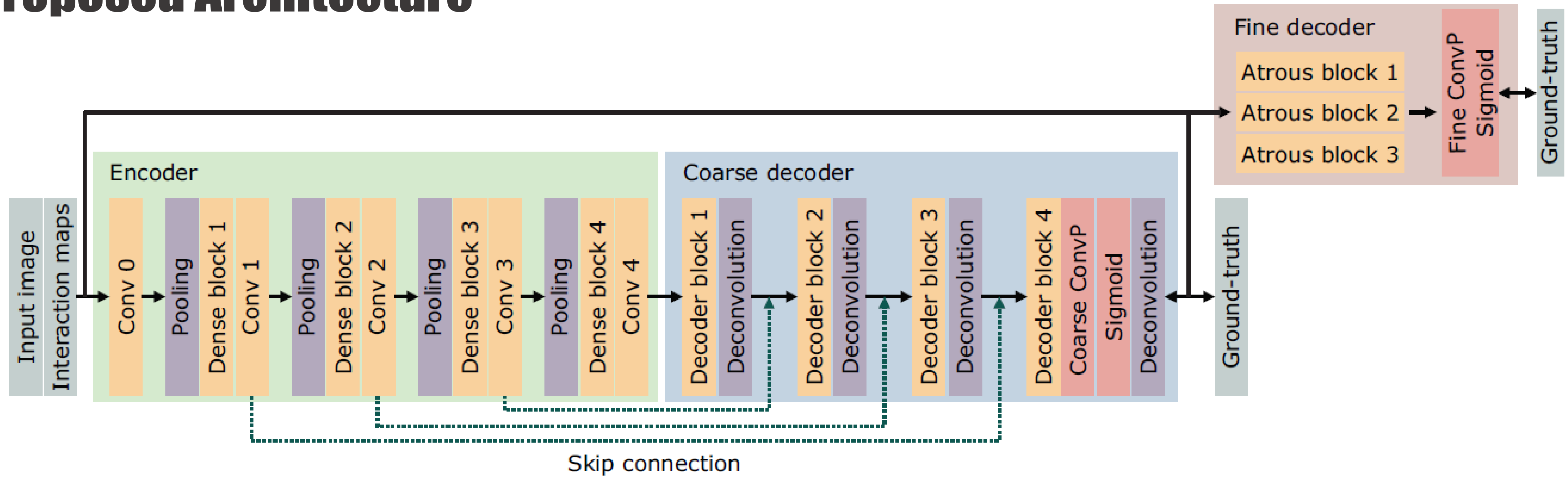
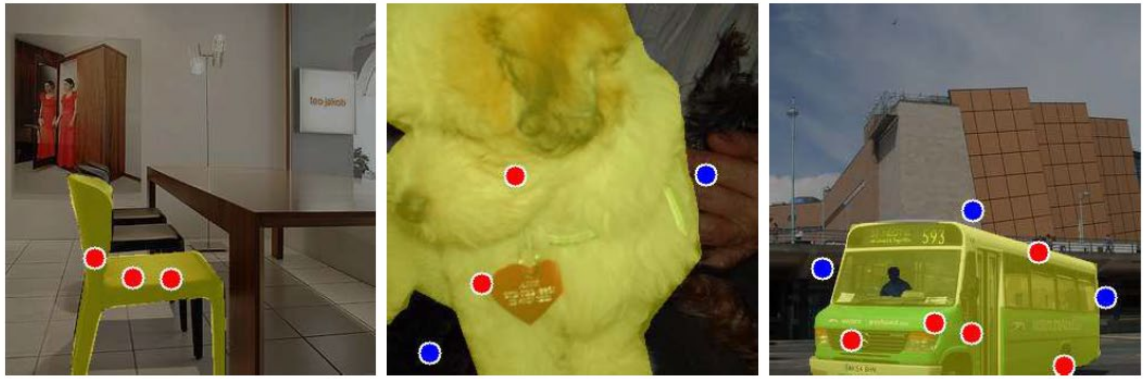


Figure 2. Architecture of the proposed network for interactive image segmentation.

1. the coarse decoder predict a rough segment of a target object
2. the fine decoder improves its detail using low-level features.

User-Annotations Generation



(a) 3 FG / 0 BG (b) 2 FG / 2 BG (c) 5 FG / 3 BG

Figure 3. Examples of generated user-annotations for training. The foreground and background annotations are depicted in red and blue circles, respectively. Also, the ground-truth object masks are highlighted in yellow.

- Use k-medoids algorithm to generate the User-annotations, in which number of FG and BG are random. Set pixels in ground-truth object mask as foreground candidates, meanwhile, set background candidates to be 5-40 pixels away from the boundaries of the object.
- applying the k-medoids algorithm on each set of candidates

Backpropagating Refinement Scheme

$$\mathcal{E}(z^0) = \mathcal{E}_\mathrm{C}(z^0) + \lambda \mathcal{E}_\mathrm{I}(z^0)$$

$$\hat{z}^0 = \arg\min_{z^0} \mathcal{E}(z^0).$$

$$\mathcal{E}_\mathrm{C}(z^0) = \sum_{\mathbf{u} \in \mathcal{U}} \left(l(\mathbf{u}) - y^R(\mathbf{u}) \right)^2$$

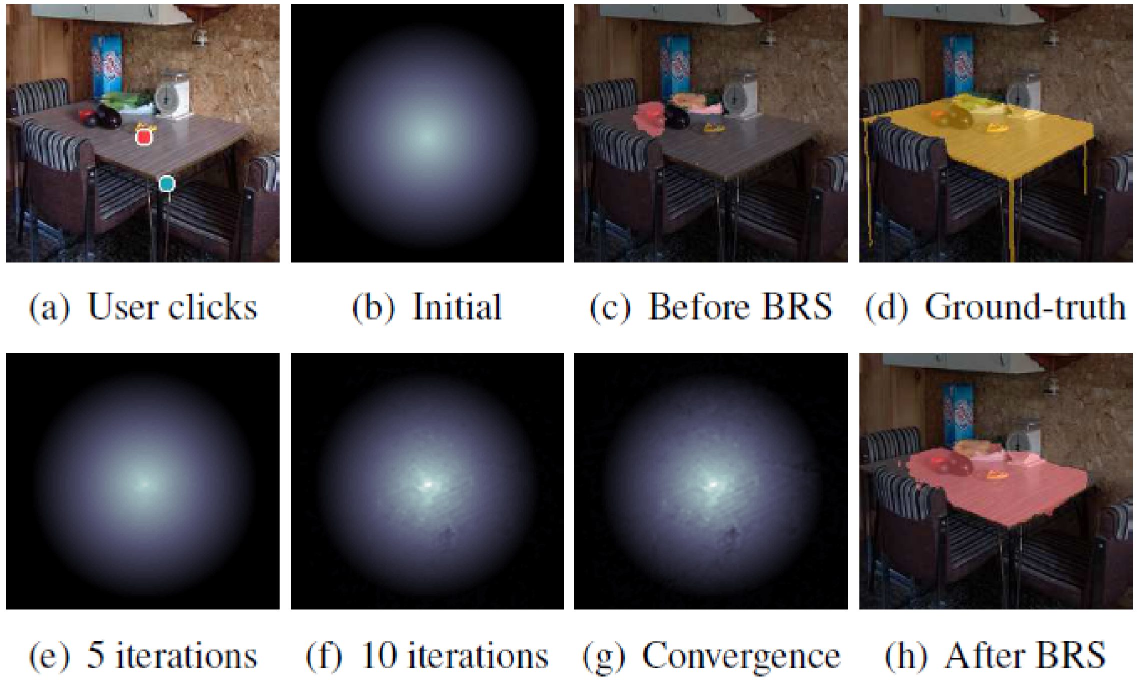
$$\mathcal{E}_\mathrm{I}(z^0) = \sum_{\mathbf{x} \in \mathcal{N}} \left(z^0(\mathbf{x}) - z^0_\mathrm{i}(\mathbf{x}) \right)^2$$

$$\frac{\partial \mathcal{E}_\mathrm{I}}{\partial z^0} = 2 \times \sum_{\mathbf{x} \in \mathcal{N}} \left(z^0(\mathbf{x}) - z^0_\mathrm{i}(\mathbf{x}) \right),$$

$$\frac{\partial \mathcal{E}}{\partial z^0} = \frac{\partial \mathcal{E}_\mathrm{C}}{\partial z^0} + \lambda \frac{\partial \mathcal{E}_\mathrm{I}}{\partial z^0}.$$

□ GOAL: assign correct labels to user-annotated locations by optimizing interaction maps.

Backpropagating Refinement Scheme

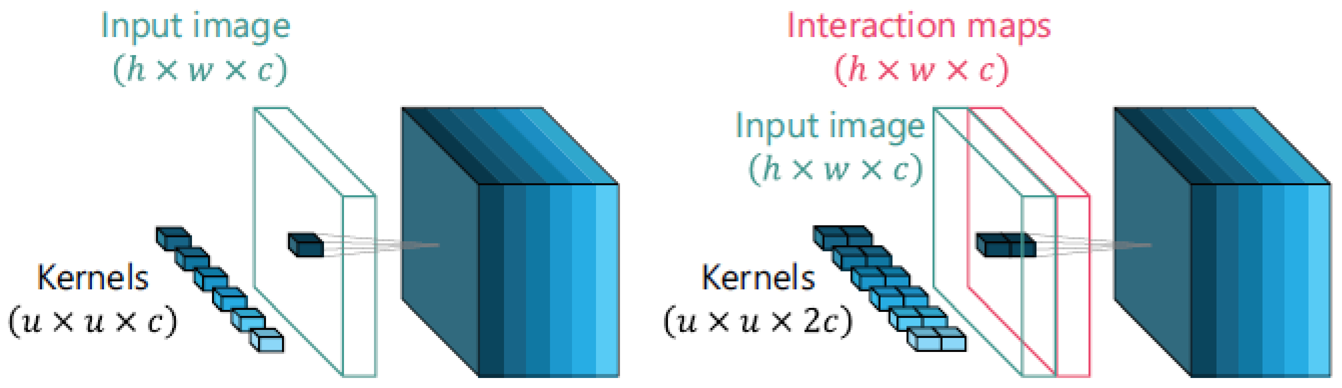

(a) User clicks (b) Initial (c) Before BRS (d) Ground-truth

(e) 5 iterations (f) 10 iterations (g) Convergence (h) After BRS

Figure 5. Foreground and background user-annotations are presented in red and blue dots in (a), respectively. An initial FG interaction map in (b) is updated in (e), (f), and (g). Segmentation results before and after BRS are in (c) and (h). The BG interaction map is not shown due to limited space.

□ GOAL: assign correct labels to user-annotated locations by optimizing interaction maps.

Generalization



(a) Baseline architecture (b) Interactive architecture

Figure 6. Reconfiguration of a network architecture in the first convolution layer. The baseline architecture in (a) is transformed to the interactive one in (b) by the training-free conversion scheme.

BRS can transform existing CNNs into user-interactive ones without extra training.

Results

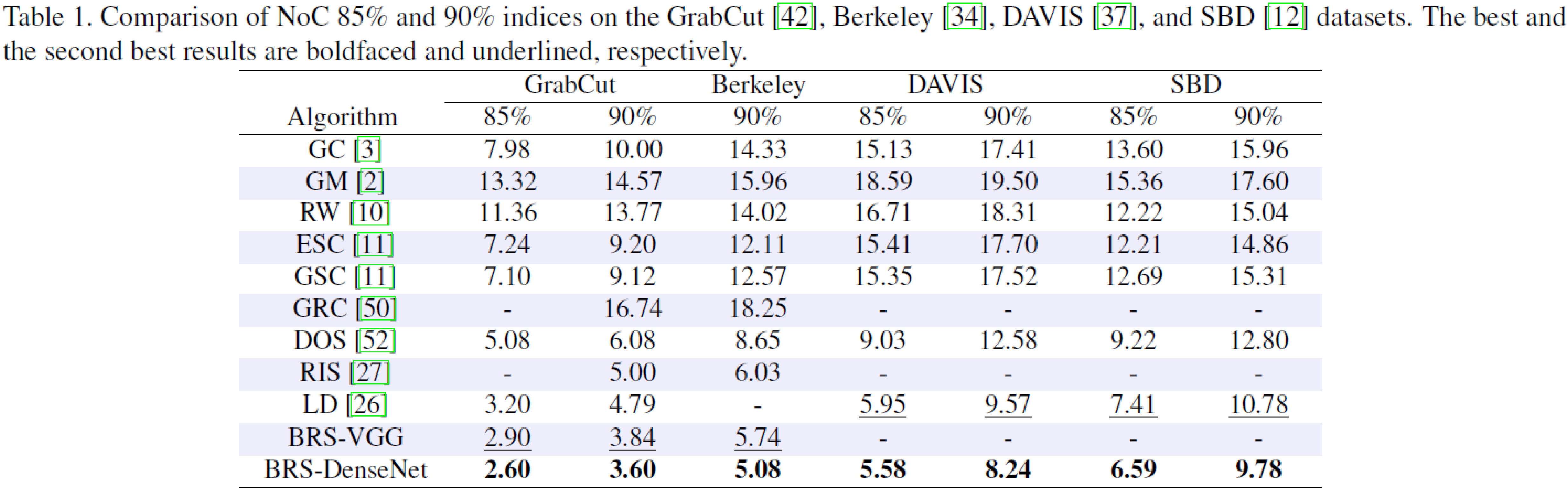


Table 1. Comparison of NoC 85% and 90% indices on the GrabCut [42], Berkeley [34], DAVIS [37], and SBD [12] datasets. The best and the second best results are boldfaced and underlined, respectively.

	GrabCut		Berkeley	DAVIS		SBD	
Algorithm	85%	90%	90%	85%	90%	85%	90%
GC [3]	7.98	10.00	14.33	15.13	17.41	13.60	15.96
GM [2]	13.32	14.57	15.96	18.59	19.50	15.36	17.60
RW [10]	11.36	13.77	14.02	16.71	18.31	12.22	15.04
ESC [11]	7.24	9.20	12.11	15.41	17.70	12.21	14.86
GSC [11]	7.10	9.12	12.57	15.35	17.52	12.69	15.31
GRC [50]	-	16.74	18.25	-	-	-	-
DOS [52]	5.08	6.08	8.65	9.03	12.58	9.22	12.80
RIS [27]	-	5.00	6.03	-	-	-	-
LD [26]	3.20	4.79	-	5.95	9.57	7.41	10.78
BRS-VGG	2.90	3.84	5.74	-	-	-	-
BRS-DenseNet	**2.60**	**3.60**	**5.08**	**5.58**	**8.24**	**6.59**	**9.78**

Results

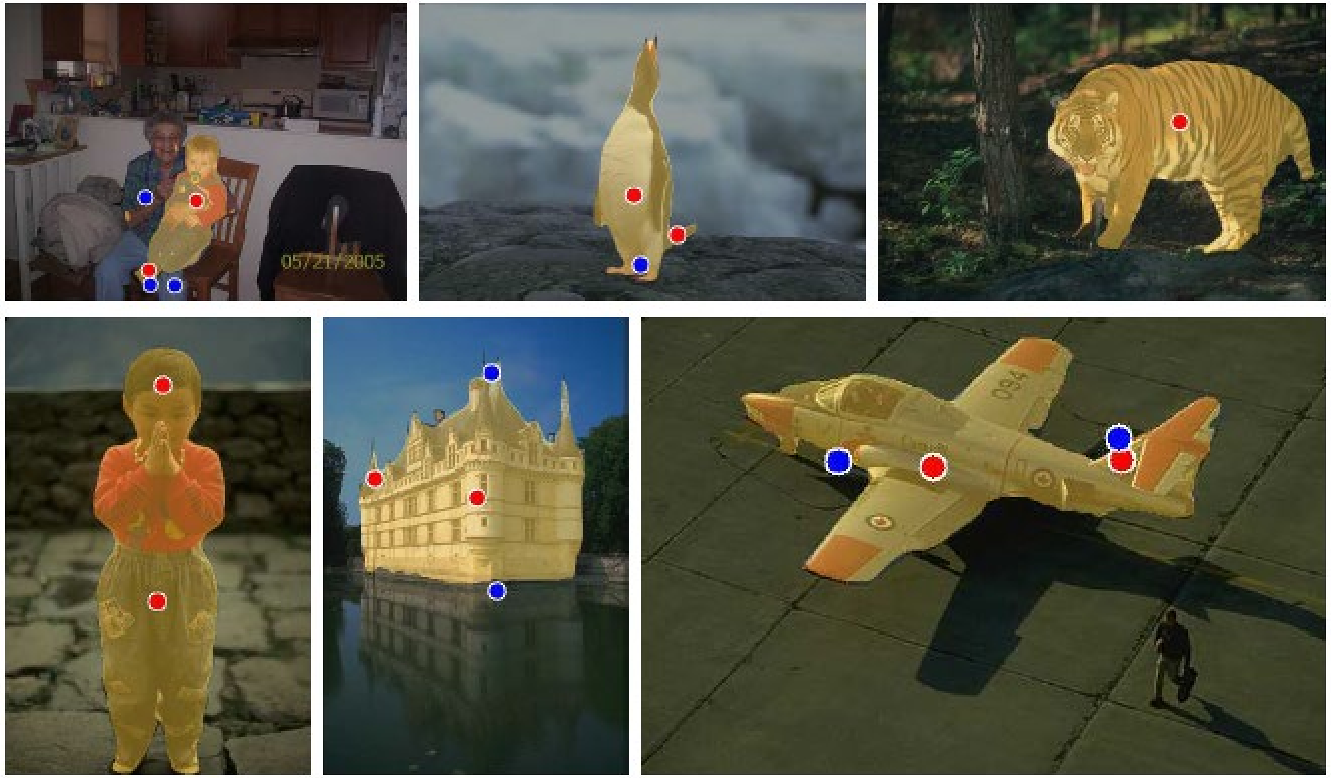


Figure 8. Segmentation results of the proposed algorithm. The segmented object masks are highlighted in yellow masks. Foreground and background user-annotations are depicted in red and blue dots, respectively.

f-BRS: Rethinking Backpropagating Refinement for Interactive Segmentation

Konstantin Sofiiuk Ilia Petrov Olga Barinova Anton Konushin
{k.sofiiuk, ilia.petrov, o.barinova, a.konushin}@samsung.com
Samsung AI Center – Moscow

Introduction

Problem of BRS:

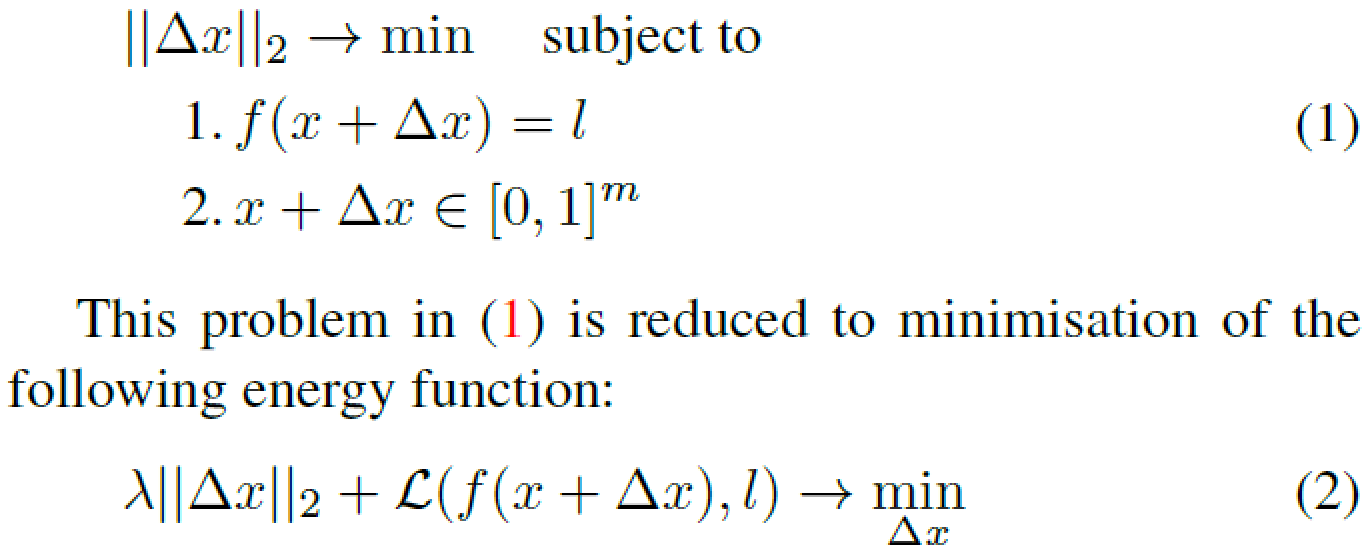
- □ BRS shows significantly better performance for the hard cases. While, BRS requires running forward and backward pass through a deep network several times that leads to significantly increased computational budget per click.

- □ The effect of BRS is based on the fact that small perturbations of the inputs for a deep network can cause massive changes in the network output.

Key Idea of f-BRS:

- □ running forward and backward passes only through a small part of the network.
- □ introduce a set of auxiliary parameters for optimization that are invariant to the position in the image, because the receptive field last convolutions layers relative to the output is too small.
- □ Perform similar effect as the original BRS.

Adversarial examples generation



$$||\Delta x||_2 \to \min \quad \text{subject to}$$
$$1.\ f(x+\Delta x) = l \tag{1}$$
$$2.\ x + \Delta x \in [0,1]^m$$

This problem in (1) is reduced to minimisation of the following energy function:

$$\lambda||\Delta x||_2 + \mathcal{L}(f(x+\Delta x), l) \to \min_{\Delta x} \tag{2}$$

Formulate an optimization problem for generating adversarial examples for an image classification task.

Orignial BRS

□ BRS find minimal edits to the distance maps that result in an object mask consistent with user-provided annotation.

□ Corrective energy function enforces consistency of the resulting mask with userprovided annotation, and inertial energy prevents excessive perturbations in the network inputs.

$$\lambda||\Delta x||_2 + \sum_{i=1}^{n} \left(f(x+\Delta x)_{u_i,v_i} - l_i\right)^2 \to \min_{\Delta x},$$

□ denote the output of a network f for an image x in position (u, v) as $f(x)_{u,v}$ and the set of all user-provided clicks as$\{(\mathrm{u}_i\ , v_i\ , l_i\ )\}_{i=1}^{n}= 1.$

f-BRS

$$\lambda||\Delta p||_2 + \sum_{i=1}^{n} \left(\hat{f}(x, p + \Delta p)_{u_i, v_i} - l_i\right)^2 \to \min_{\Delta p} .$$

□ BRS find minimal edits to the distance maps that result in an object mask consistent with user-provided annotation.

□ Corrective energy function enforces consistency of the resulting mask with userprovided annotation, and inertial energy prevents excessive perturbations in the network inputs.

□ denote the output of a network f for an image x in position (u, v) as $f(x)_{u,v}$ and the set of all user-provided clicks as$\{(\mathrm{u}_i\ , v_i\ , l_i\ )\}_{i=1}^{n}= 1.$

f-BRS

$$\lambda||\Delta p||_2 + \sum_{i=1}^{n} \left(\hat{f}(x, p + \Delta p)_{u_i, v_i} - l_i\right)^2 \to \min_{\Delta p} .$$

$$\hat{f}(x, s, b) = g\big(s \cdot F(x) + b\big),$$

□ Problem：

a)does not have a localized effect on the outputs, b) does not require a backward pass through the whole network for optimization

□ One of the options for such reparameterization may be channel-wise scaling and bias for the activations of the last layers in the network. Scale and bias are invariant to the position in the image, thus changes in this parameters would affect the results globally. Compared to optimization with respect to activations, optimization with respect to scale and bias cannot result in degenerate solutions

Architecture

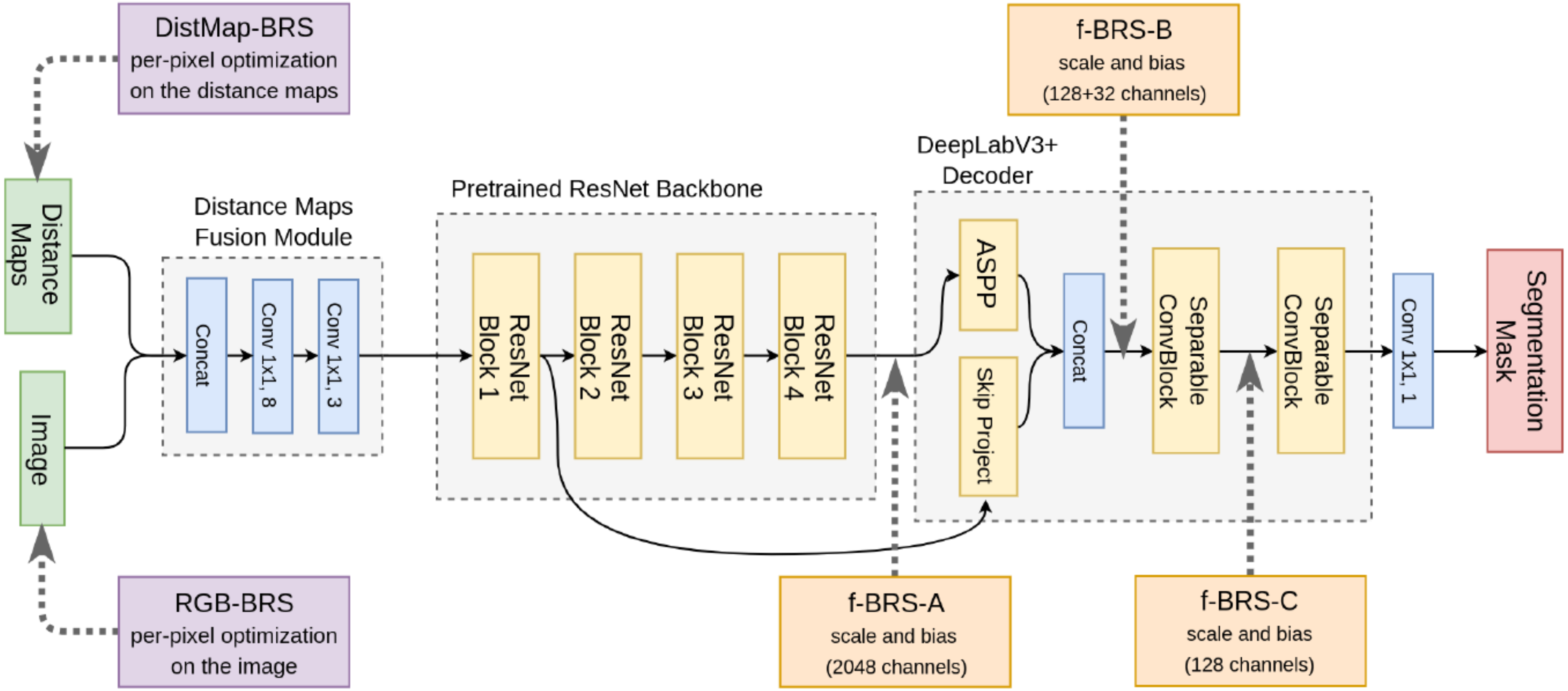


Figure 2. Illustration of the proposed method described in Section 3. f-BRS-A optimizes scale and bias for the features after pre-trained backbone, f-BRS-B optimizes scale and bias for the features after ASPP, f-BRS-C optimizes scale and bias for the features after the first separable convblock. The number of channels is provided for ResNet-50 backbone.

Architecture

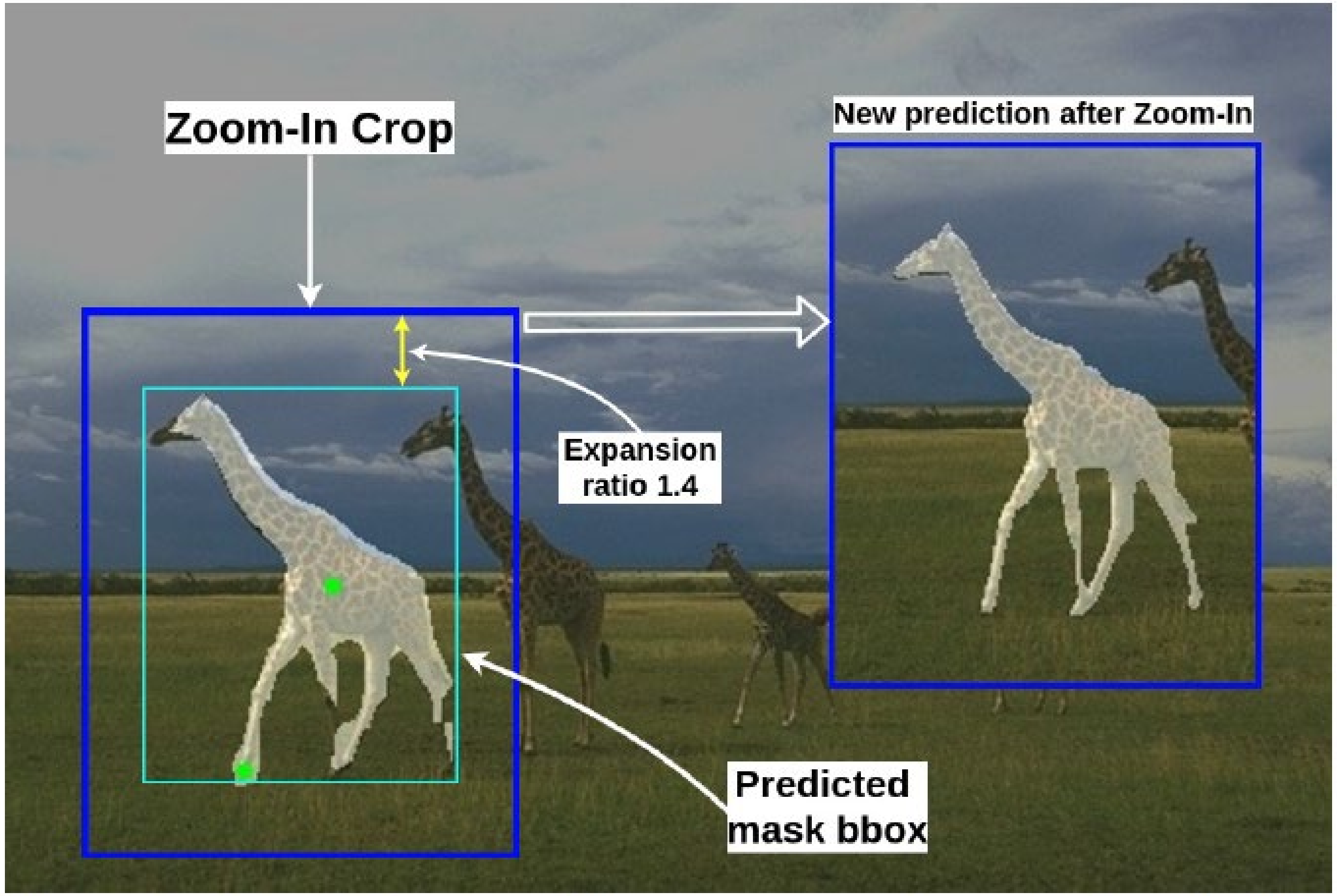


Figure 3. Example of applying zoom-in technique described in Section 4. See how cropping an image allows recovering fine details in the segmentation mask.

Noticed that the first 1-3 clicks are enough for the network to achieve around 80% IoU with ground truth mask in most cases. It allows us to obtain a rough crop around the region of interest.

Results

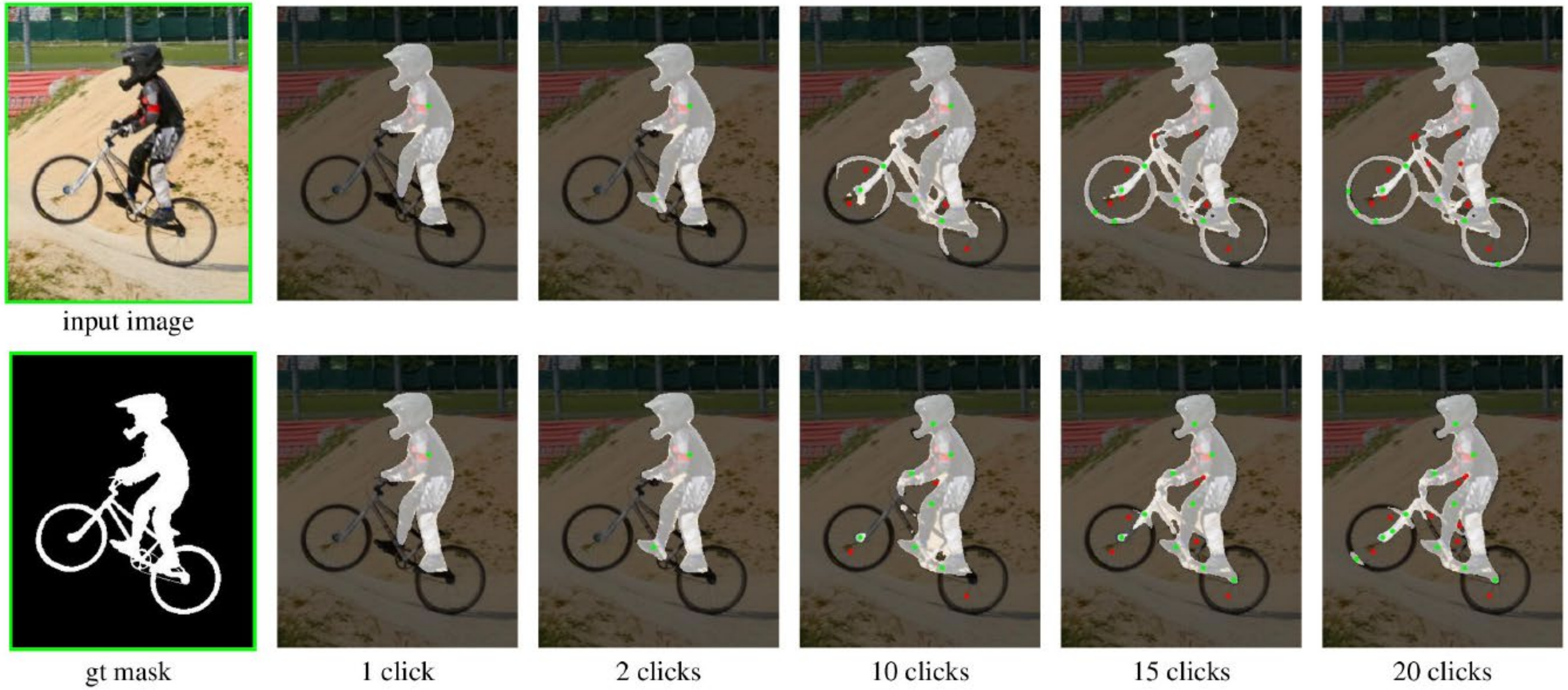


Figure 1. Results of interactive segmentation on an image from DAVIS dataset. First row: using proposed f-BRS-B (Section 3), second row: without BRS. Green dots denote positive clicks, red dots denote negative clicks.

Method		GrabCut		Berkeley	SBD		DAVIS	
		NoC@85	NoC@90	NoC@90	NoC@85	NoC@90	NoC@85	NoC@90
Graph cut [4]		7.98	10.00	14.22	13.6	15.96	15.13	17.41
Geodesic matting [11]		13.32	14.57	15.96	15.36	17.60	18.59	19.50
Random walker [10]		11.36	13.77	14.02	12.22	15.04	16.71	18.31
Euclidean star convexity [11]		7.24	9.20	12.11	12.21	14.86	15.41	17.70
Geodesic star convexity [11]		7.10	9.12	12.57	12.69	15.31	15.35	17.52
Growcut [30]		–	16.74	18.25	–	–	–	–
DOS w/o GC [31]		8.02	12.59	–	14.30	16.79	12.52	17.11
DOS with GC [31]		5.08	6.08	–	9.22	12.80	9.03	12.58
Latent diversity [18]		3.20	4.79	–	7.41	10.78	**5.05**	9.57
RIS-Net [19]		–	5.00	–	6.03	–	–	–
CM guidance [21]		–	3.58	5.60	–	–	–	–
BRS [15]		2.60	3.60	5.08	6.59	9.78	5.58	8.24
Ours w/o BRS	ResNet-34	2.52	3.20	5.31	5.51	8.58	5.47	8.51
	ResNet-50	2.64	3.32	5.18	5.10	8.01	5.39	8.18
	ResNet-101	2.50	3.18	6.25	5.28	8.13	5.12	8.01
Ours f-BRS-B	ResNet-34	**2.00**	**2.46**	4.65	5.25	8.30	5.39	8.21
	ResNet-50	2.50	2.98	**4.34**	5.06	8.08	5.39	7.81
	ResNet-101	2.30	2.72	4.57	**4.81**	**7.73**	**5.04**	**7.41**

Table 3. Evaluation results of GrabCut, Berkeley, SBD and DAVIS datasets. The best and the second best results are written in bold and underlined respectively.