Interacitve Segmentation

Interactive Image Segmentation via Backpropagating Refinement Scheme

Harvard, Korea Univ. Won-Dong Jang, Chang-Su Kim

2020. 07. 29. Wed.
Taeu

CONTENTS

CVRP 2019, 2020, Interactive-related works

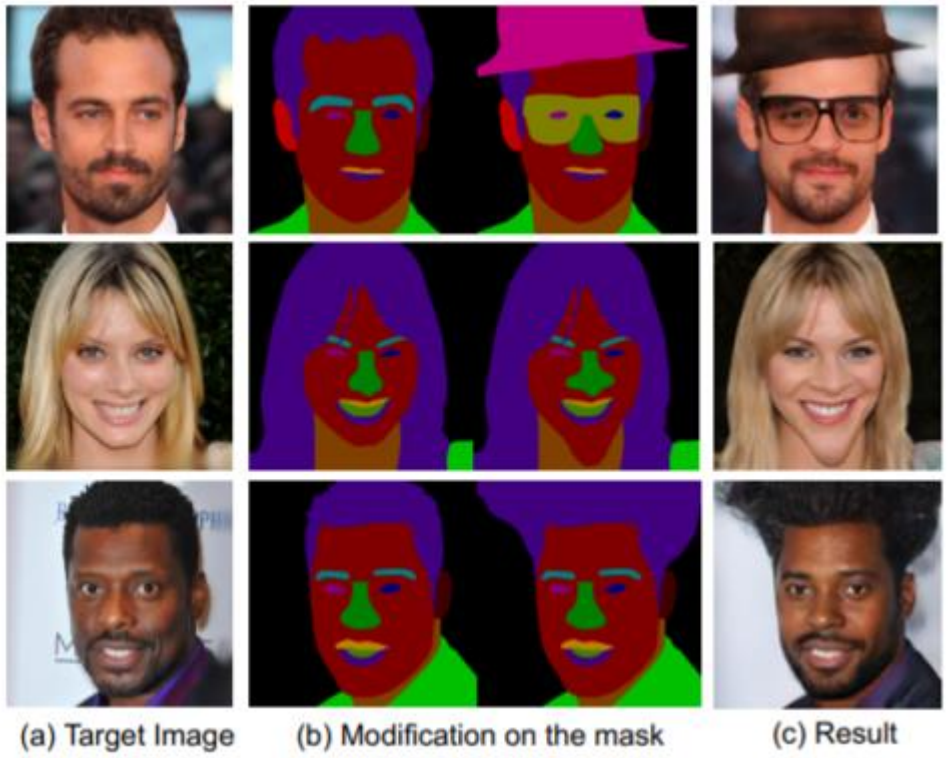
CVPR 2019

No	Title	C.	Institute	Main Idea
1	Fast Interactive Object Annotation With Curve-GCN	33	Toronto Univ, NVIDIA	Graph Conv Network, +
2	Interactive Image Segmentation via Backpropagating Refinement Scheme	17	Havard, Korea Univ	Backpropagation
3	Constrained Generative Adversarial Networks for Interactive Image Generation	3	AirForce.R,USA	Image generation
4	Content-Aware Multi-Level Guidance for Interactive Instance Segmentation	11	Boon,Sigapore Univ	FCN
5	Interactive Full Image Segmentation by Considering All Regions Jointly	13	Google R.	scribble,mask RCNN
6	Large-Scale Interactive Object Segmentation With Human Annotators	25	Google R.	At scale, interaction behavior

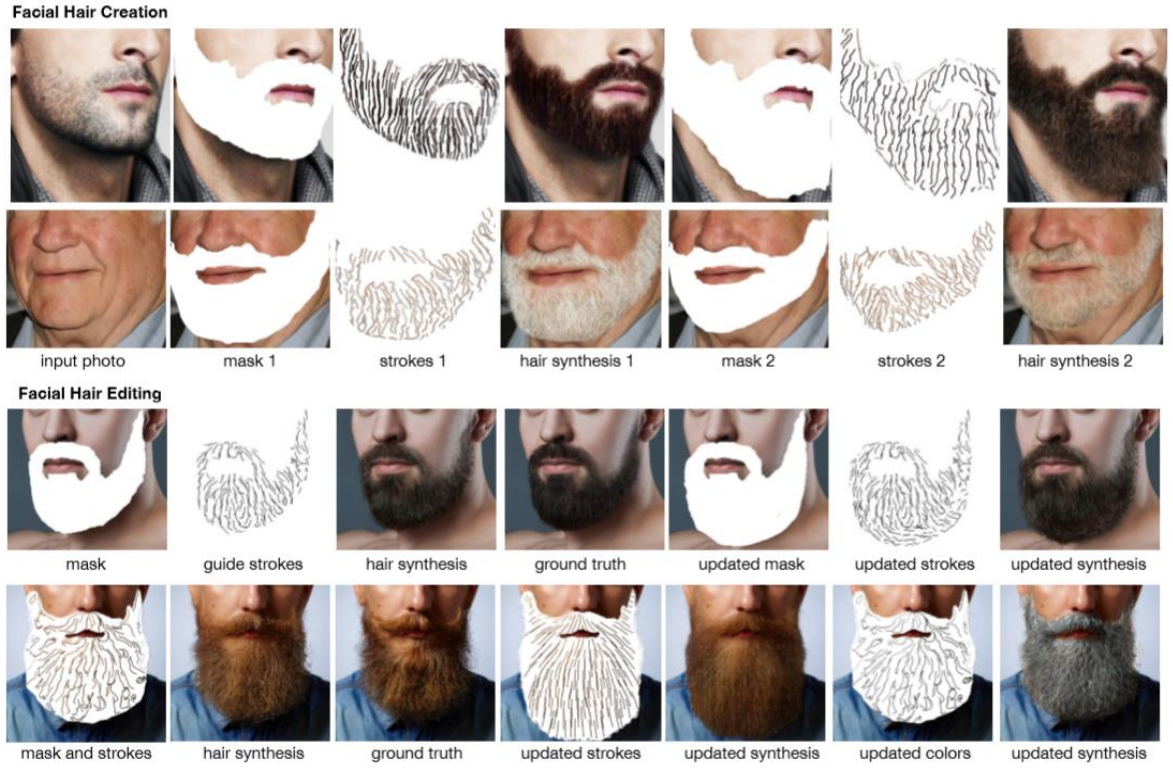
CVPR 2020

No	Title	C.	Institute	Main Idea
1	Interactive Object Segmentation With Inside-Outside Guidance	2	Beijing, Key Lab etc	Inside-Outsize Guidance, Refinement(hint map), FineNet
2	F-BRS: Rethinking Backpropagating Refinement for Interactive Segmentation	2	Samsung AI Center-Moscow	BRS, feature, auxiliary variables
3	Interactive Multi-Label CNN Learning With Partial Labels	2	Northeastern University	interactive learning, new loss function
4	Multi-Scale Interactive Network for Salient Object Detection	3	Dalian University, China	Saliency detection
5	Interactive Two-Stream Decoder for Accurate and Fast Saliency Detection	0	Sun Yat-sen University, China etc	Saliency detection
6	MaskGAN: Towards Diverse and Interactive Facial Image Manipulation	51	SenseTime R. Hong Kong Univ	mask map(semantic mask), interactive conditional GANs
7	Interactive Image Segmentation With First Click Attention	2	Nankai University	first click + concat encoder, fine encoder
8	STINet: Spatio-Temporal-Interactive Network for Pedestrian Detection and Trajecto	0	Waymo LLC, Johns Hopkins University	-
9	Cross-Domain Semantic Segmentation via Domain-Invariant Interactive Relation T	0	Southwestern University, Tencent, China etc	-
10	Memory Aggregation Networks for Efficient Interactive Video Object Segmentatio	3	Baidu, Sydney Tech	video object detection
11	Iteratively-Refined Interactive 3D Medical Image Segmentation With Multi-Agent	1	Shanghai Jiao Tong University	-
12	GAN Compression: Efficient Architectures for Interactive Conditional GANs	6	MIT, Adobe Research etc.	Interactive Conditional GANs
13	SAPIEN: A SimulAted Part-Based Interactive Environment	4	UC San Diego, Stanford Univ, Google R., etc	robotic interaction tasks, heuristic AG and RL
14	Intuitive, Interactive Beard and Hair Synthesis With Generative Models	0	Southern California Univ, Adobe Inc. etc.	Interactive Conditional GANs

Interactive Conditional GANs

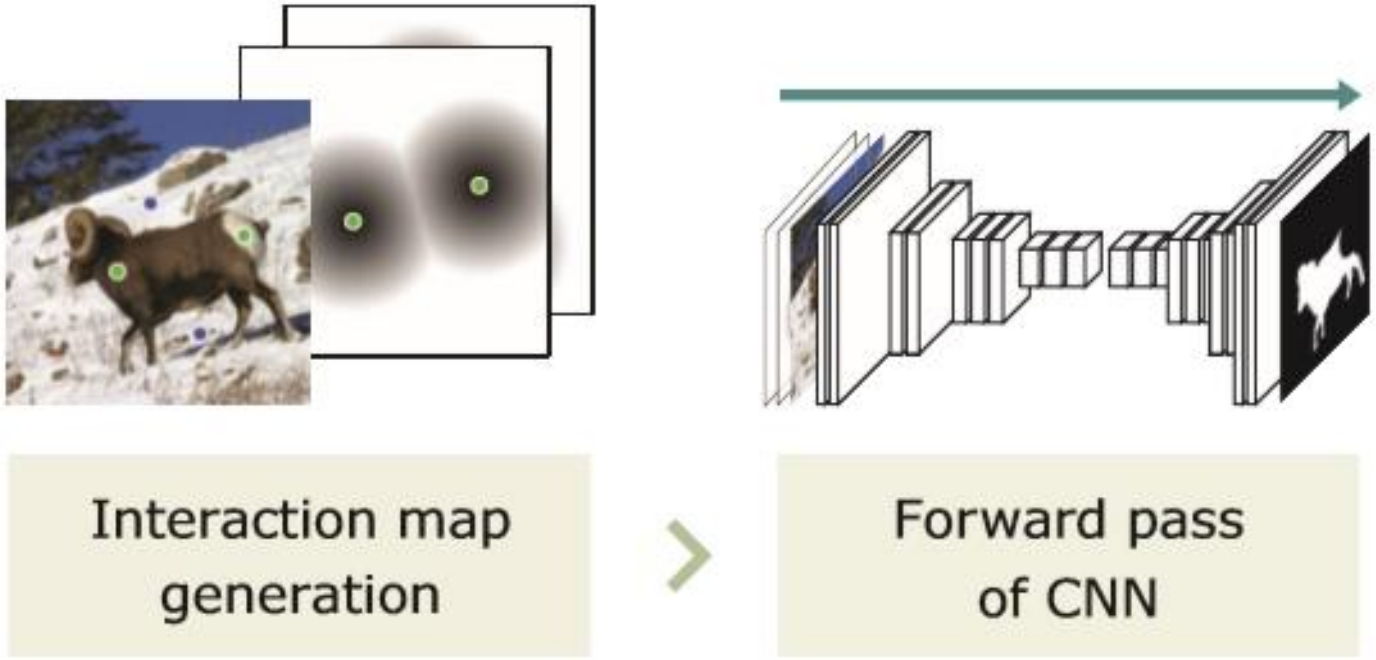


MaskGAN: Towards Diverse and Interactive Facial Image Manipulation



Intuitive, Interactive Beard and Hair Synthesis With Generative Models

Interactive segmentation



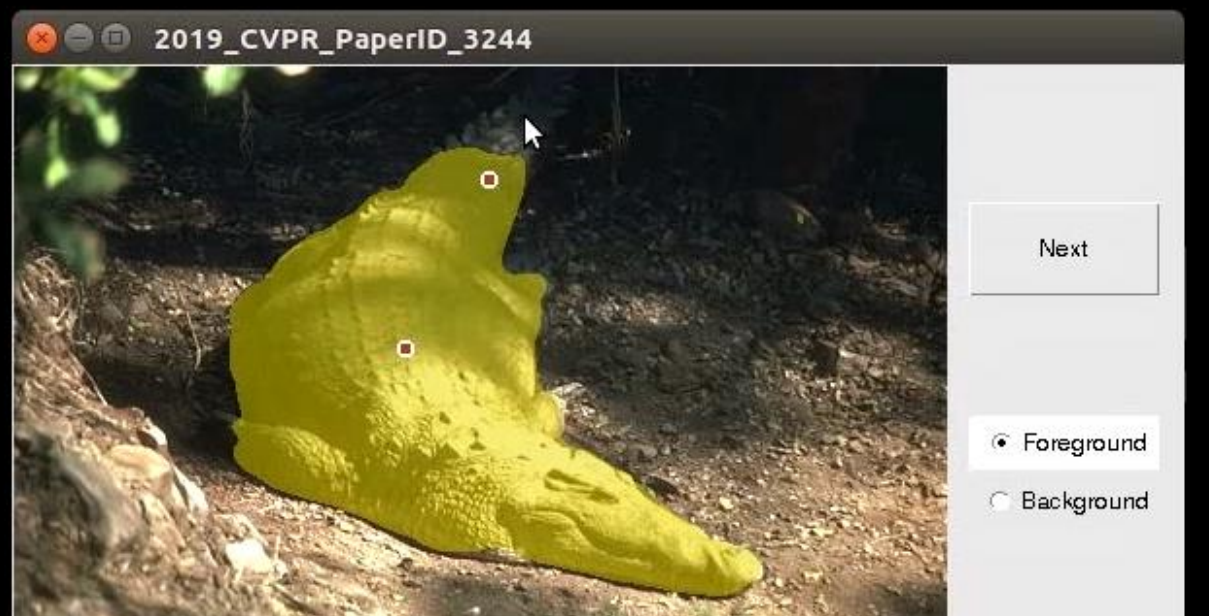
Interactive segmentation

I) Input image with extreme points provided by annotator

II) Machine predictions from extreme points

III) corrective scribbles provided by annotator

IV) Machine predictions from extreme points and corrective scribbles



Abstract

An **interactive image segmentation** algorithm, which accepts user-annotations about a target object and the background, is proposed in this work. We convert user-annotations into interaction maps by measuring distances of each pixel to the annotated locations. Then, we perform the forward pass in a convolutional neural network, which outputs an initial segmentation map.

However, **the user-annotated locations** can be mislabeled in the initial result. Therefore, we develop the backpropagating refinement scheme (BRS), which corrects **the mislabeled pixels**.

Experimental results demonstrate that the proposed algorithm outperforms the conventional algorithms on four challenging datasets. Furthermore, we demonstrate the **generality and applicability of BRS in other computer vision tasks**, by transforming existing convolutional neural networks into user-interactive ones.

Introduction

Interactive image segmentation :
separate **target object or background**

└ annotated in type of click or scribble

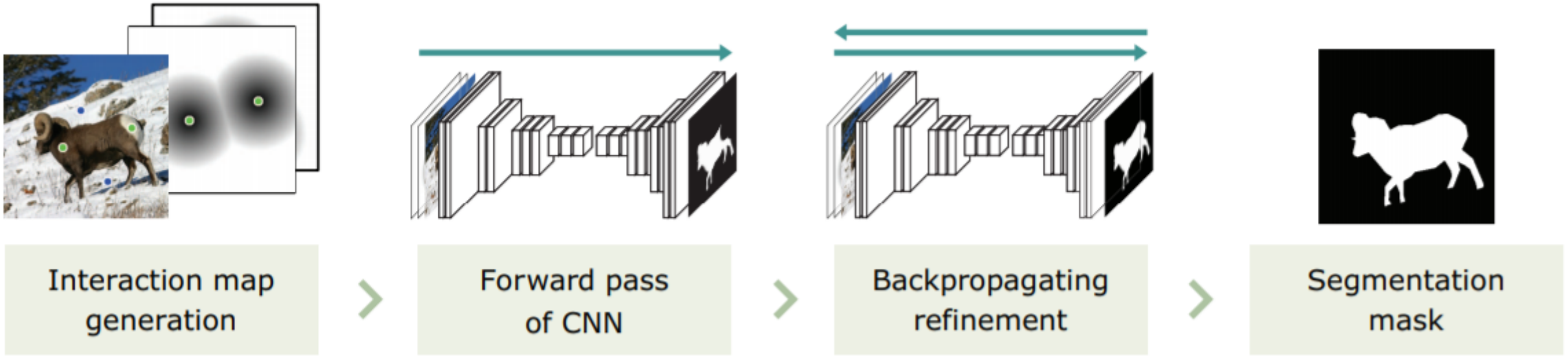
└ to extract an accurate mask of the target using fewer clicks

Semantic segmentation :
(Pretrained Encoder)-(decoder) architecture with skip-connection

backpropagation schemes :
to visualize characteristics of neural networks, for texture synthesis and image style transfer.
They update activation responses backwardly, while freezing parameters, in the networks.

In this paper, interactive image segmentation algorithm, which accepts user scribbles.
In the train phase, use a fully convolutional neural network.
In the test phase, we perform the forward pass in the proposed network using an input image and user-annotations.
With backpropagating refinement scheme (BRS), which constrains user-specified locations to have correct labels
and refines the segmentation result of the forward pass.

Three Main Contributions



Contributions

1. Development of a CNN for interactive image segmentation, which is fully convolutional.
2. Introduction of the backpropagating refinement strategy, which corrects mislabeled locations.
3. Generalization of BRS, which can make existing CNNs user-interactive **without extra training.**

2. Related Work

2.1. Interactive Image Segmentation

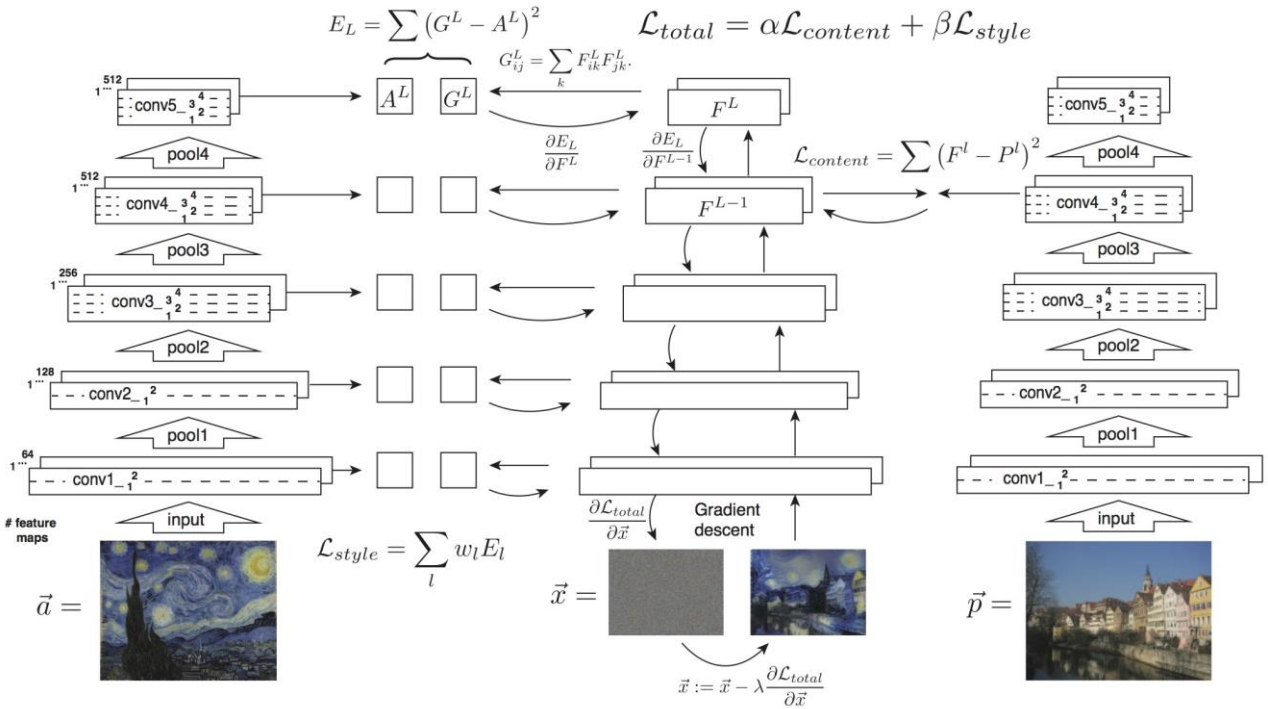
└ [27], [26], [31], [45]

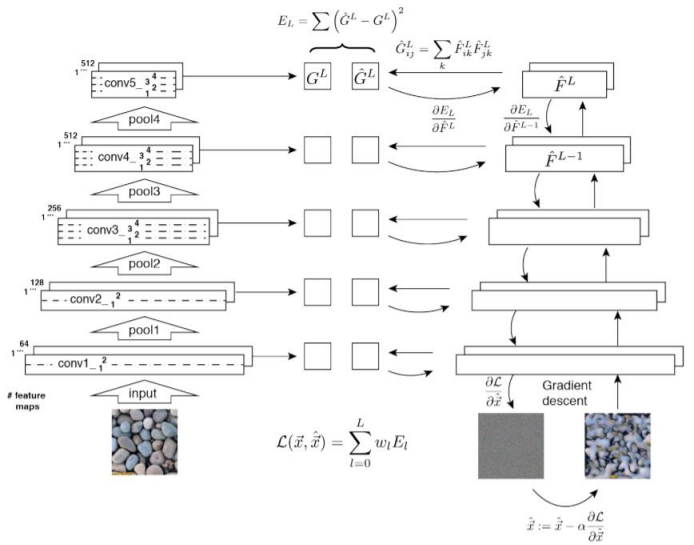
Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S. Huang. Deep interactive object selection. In CVPR, pages 373–381, 2016

2.2. Backpropagation for Activations

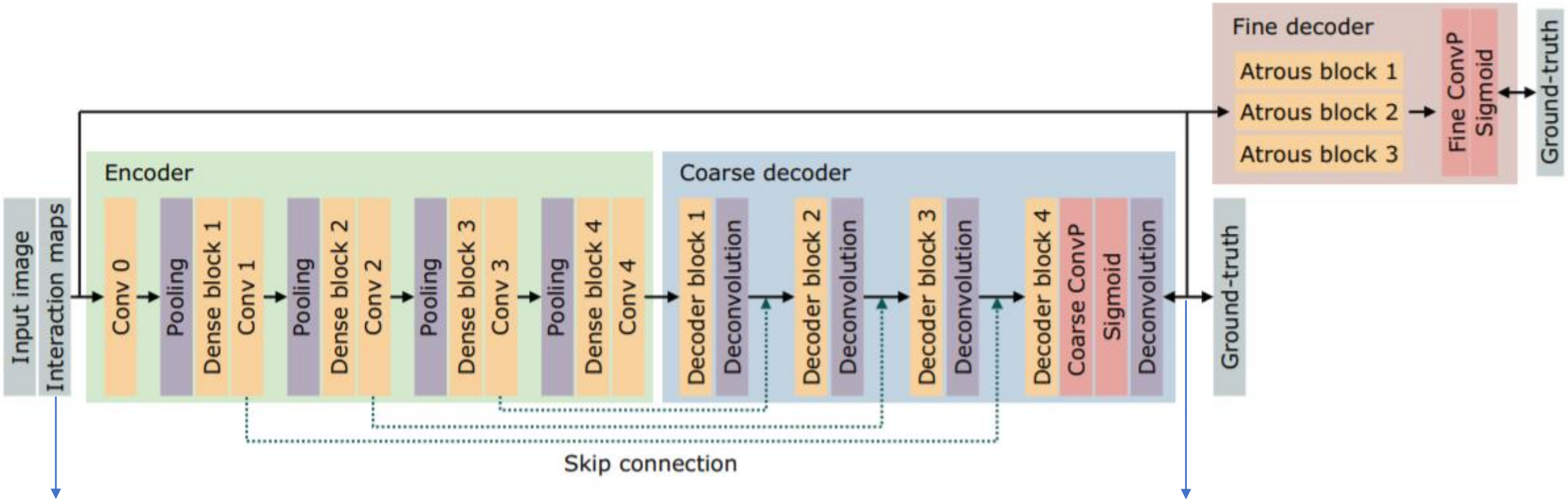
Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In NIPS, pages 262–270, 2015.

Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.





3. Proposed Algorithm



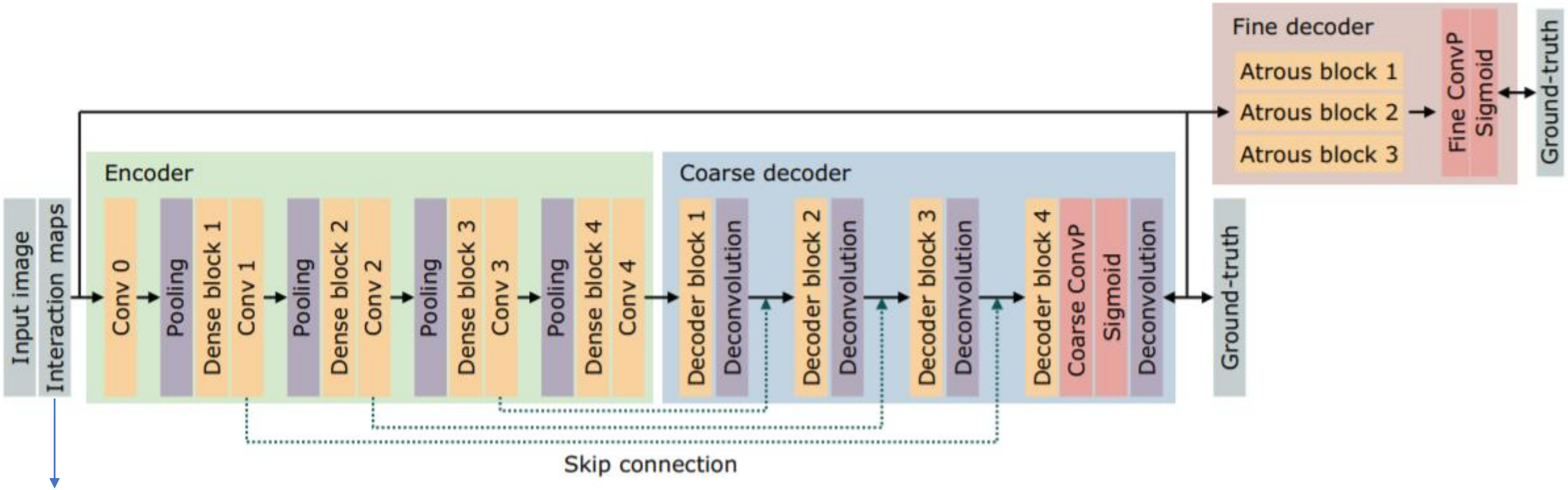
generate foreground and background interaction maps

yields a probability map of a user-specified object

Even though the interaction maps clearly represent the annotated labels in the clicked locations, the probability map may convey wrong information at those clicked locations.

→ we force the clicked locations to have the user-specified labels by employing the proposed BRS

3. Proposed Algorithm



user provides the **first click** on a target object

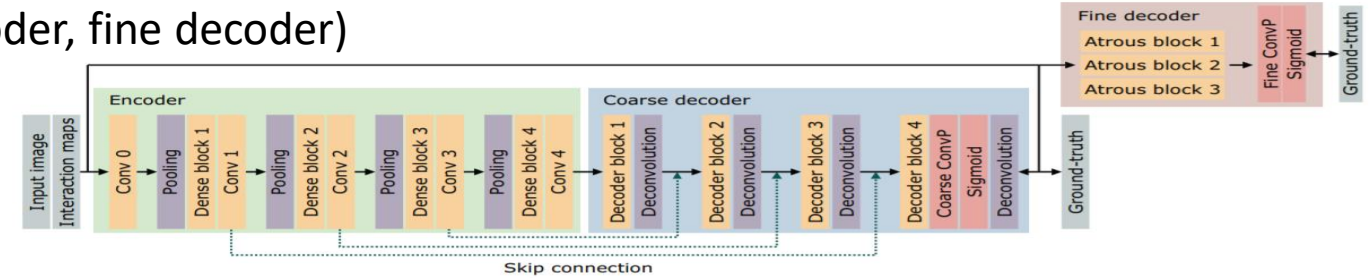
the user may **click a new** location either on the object or the background.

these two steps are conducted recursively until the user stops clicking.

3.1. CNN for Interactive Image Segmentation

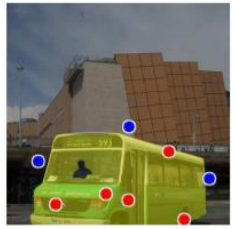
Network architecture

- Encoder(DenseNet – add a squeeze and excitation module at the end of each dense block)
- Decoder(corse decoder, fine decoder)
- Fully convolutional



Training phase

(a) 3 FG / 0 BG



(c) 5 FG / 3 BG

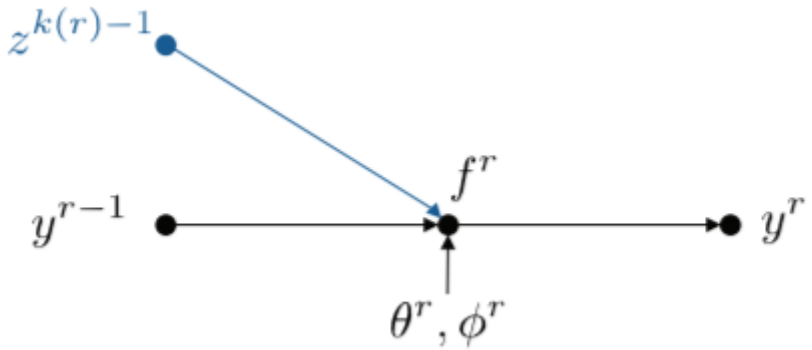
- SBD dataset to train the proposed CNN :
  - randomly crop a 360 × 360 patch to yield pairs of an image patch and its object mask.
  - the center pixel of a cropped patch belongs to foreground in the object mask
  - a simple clustering strategy
    - the numbers of foreground and background clicks are determined randomly within [1, 10] and [0, 10], respectively.
    - we set pixels in a ground-truth object mask as foreground candidates
    - background candidates to be at least 5 pixels and at most 40 pixels away from the boundaries of the ground-truth object
    - By applying the k-medoids algorithm on each set of candidates,
- Training details
  - initialize parameters in the decoders with random values
  - train the network via the stochastic gradient descent
  - train the proposed network for 20 epochs without the fine decoder
  - learning for another 15 epochs with the fine decoder

Inference phase

- Given user clicks, we first update the foreground and background interaction maps by computing the distance of each pixel to the nearest clicks, and then do BRS recursively.

3.2. Backpropagating Refinement Scheme (BRS)

being incapable of guaranteeing that clicked pixels have user-annotated labels.
even clicked pixels may have incorrect labels in the segmentation result.
The proposed BRS performs backpropagation iteratively until all clicked pixels have correct labels.



$$y^r = f^r(y^{r-1}, z^{r-1}, \theta^r, \phi^r).$$

tensors **y^{r−1}** and z^{r−1} are concatenated, and parameters θ^{r} and ϕ^{r} are used to obtain **y^{r}** , which denotes the responses of **the r_th layer** in the network.

y 0 , y R, and z 0 become **an input image, the output of the network**, interaction maps, respectively,
R is the index of the last layer in the fine decoder

3.2. Backpropagating Refinement Scheme (BRS)

Initial interaction maps, which are converted from the user-annotations, may be **imperfect** for making the network yield correct labels in user-annotated locations.

we choose to modify interaction maps, instead of fine-tuning network

The goal of **BRS** is to assign correct labels to user-annotated locations by optimizing **interaction maps z 0** .

By combining a **corrective energy E_C** and an **inertial energy E_I** , the energy function E(z 0) of the interaction maps z 0 is defined as λ : 10−3

$$\mathcal{E}(z^0) = \mathcal{E}_{\mathrm{C}}(z^0) + \lambda \mathcal{E}_{\mathrm{I}}(z^0) \qquad \hat{z}^0 = \arg\min_{z^0} \mathcal{E}(z^0).$$

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

3.2. Backpropagating Refinement Scheme (BRS)

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

we minimize the energy function, by employing **L-BFGS algorithm**, and obtain the optimal interaction map

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

where **U** is the **set of annotated pixels**. Also, l(u) denotes a user-annotated label, which is 1 for foreground and 0 for background, and y R(u) is the output of the proposed network

By employing these backward recursive equations, we obtain the partial derivative, ∂EC/ ∂z0 , of the corrective energy with respect to the interaction maps

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

The inertial energy prevents excessive perturbations of the interaction maps

where **N** is the **set of coordinates in the interaction maps**, **z 0 i** denotes the **initial interaction maps** used in the forward pass

3.2. Backpropagating Refinement Scheme (BRS)

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

The inertial energy yields a high cost when the interaction maps are different from their initial values. We compute the partial derivative of the inertial energy with respect to the interaction maps , which is easily obtainable at the input layer of the network.

We blend the derivatives of the corrective energy and the inertial energy using the parameter λ in (2) as
∂E / ∂z0 = ∂EC / ∂z0 + λ ∂EI / ∂z0

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

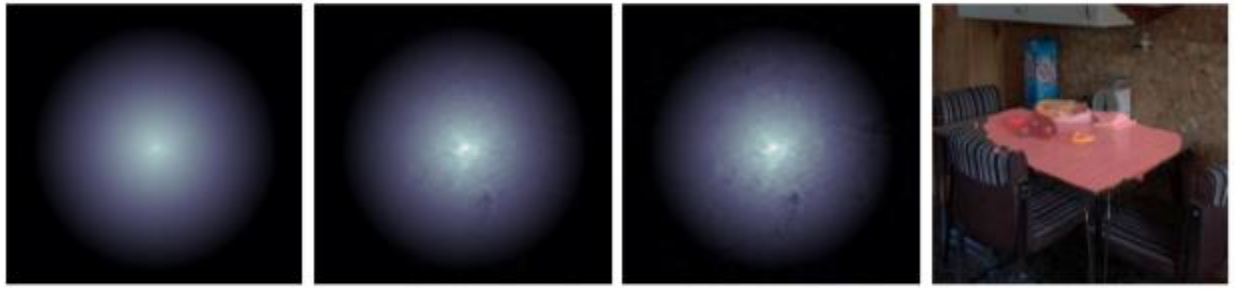
$$\mathcal{E}_{\mathrm{I}}(z^0) = \sum_{\mathbf{x} \in \mathcal{N}} \left(z^0(\mathbf{x}) - z_{\mathrm{i}}^0(\mathbf{x}) \right)^2$$

The inertial energy prevents excessive perturbations of the interaction maps

where N is the set of coordinates in the interaction maps, z 0 i denotes the initial interaction maps used in the forward pass

3.2. Backpropagating Refinement Scheme (BRS)

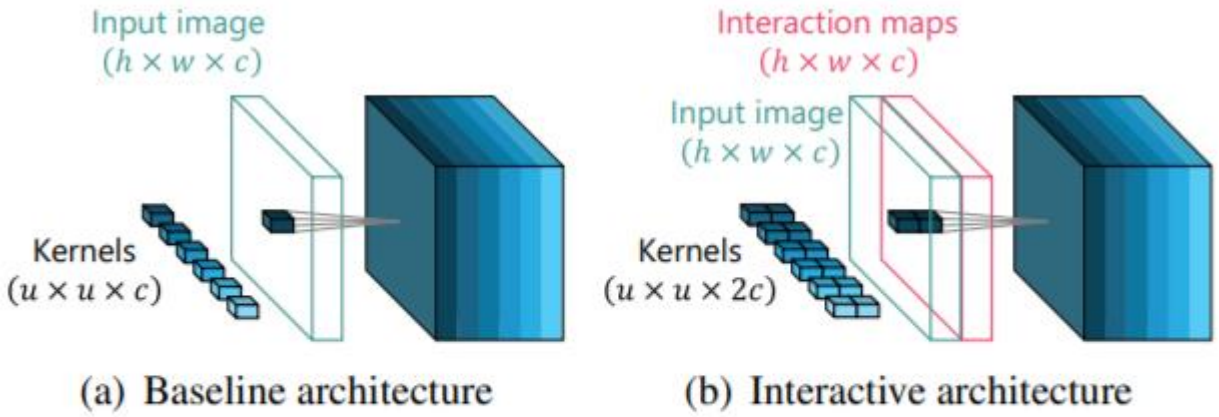
(a) User clicks (b) Initial (c) Before BRS (d) Ground-truth



(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.

3.3. Generalization



(a) Baseline architecture (b) Interactive architecture

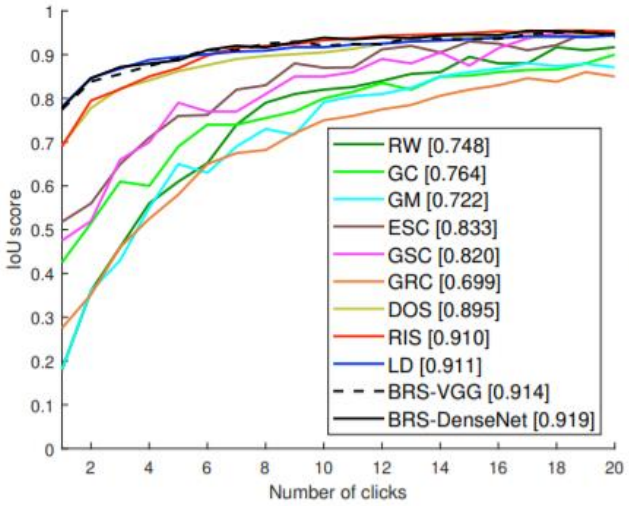
Figure 6. Reconfiguration of a network architecture in the first convolution layer. The baseline architecture in (a) is transformed to the interactive one in (b) by the training-free conversion scheme.

We can employ BRS for general networks that are not trained with interaction map

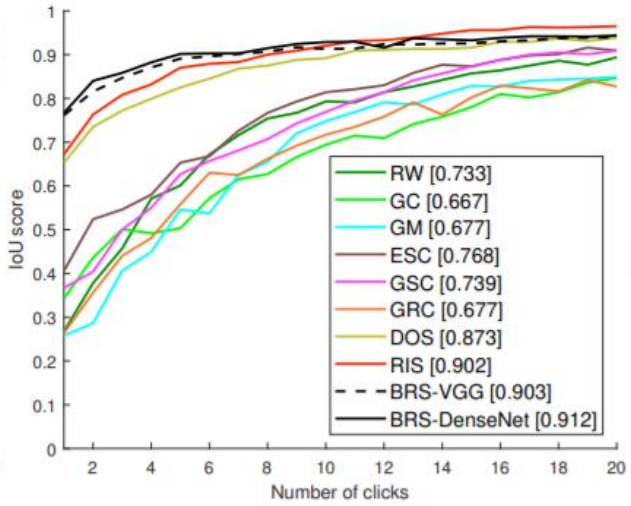
Based on this generality, we show that BRS can transform existing CNNs into user-interactive ones without extra training

The development of interactive algorithms requires time and expertise for training, in terms of composition of training data, network architectures, and hyperparameters. Also, even though interactive algorithms are trained successfully, they often yield inferior results compared to non-interactive algorithms when user interactions are not given.

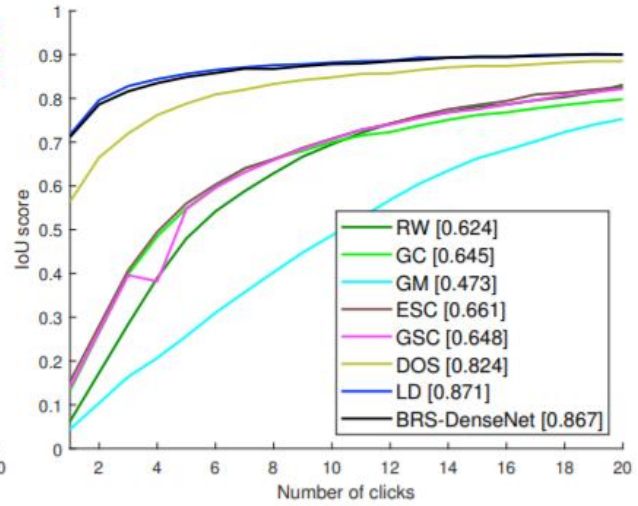
Results



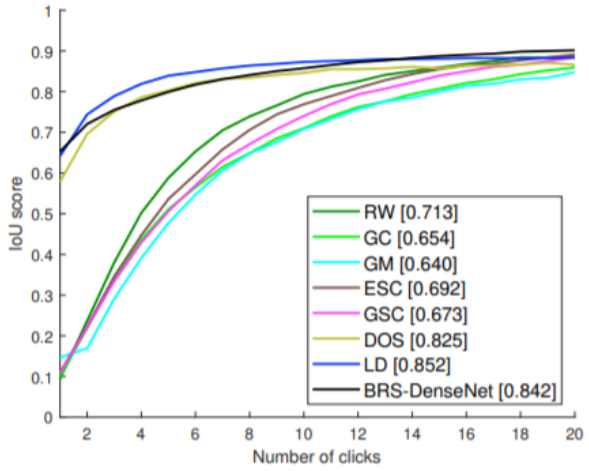
(a) GrabCut



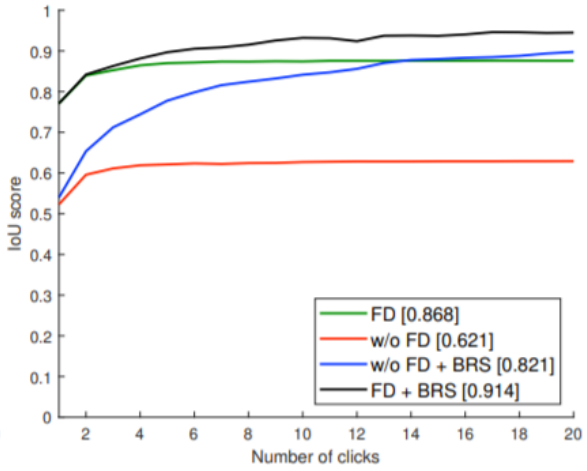
(b) Berkeley



(c) DAVIS



(d) SBD



(e) Ablation study

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**

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.

Table 2. NoC 85% and 90% indices of the proposed algorithm in various settings.

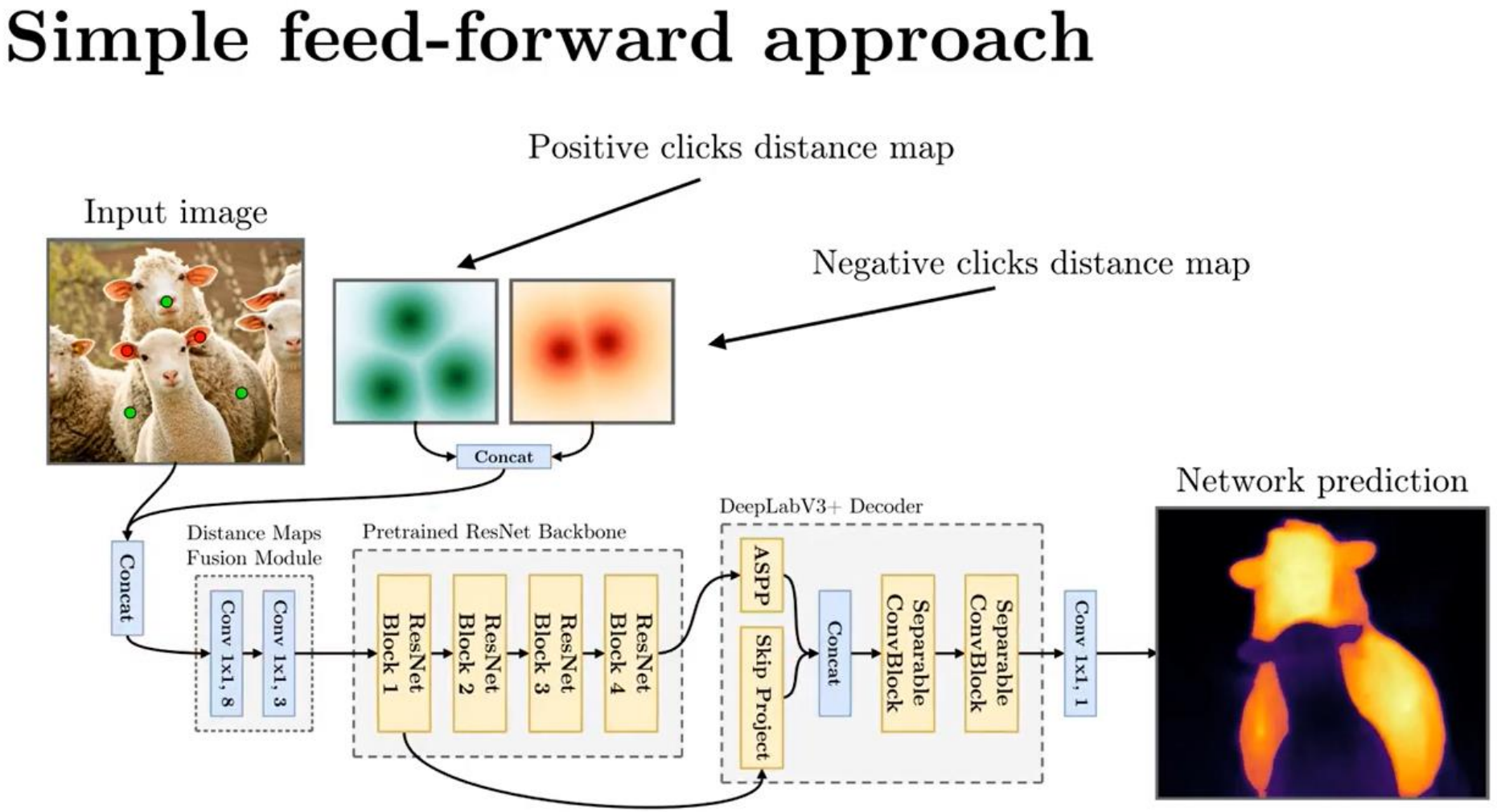
	GrabCut		Berkeley	
Setting	NoC 85%	NoC 90%	NoC 85%	NoC 90%
FD	4.12	6.12	5.33	7.65
w/o FD	14.34	17.4	17.80	19.63
w/o FD + BRS	6.60	10.28	10.09	15.30
FD+BRS	2.60	3.60	3.16	5.08

Q & A

Appendix – f-BRS (CVPR 2020)

f-BRS: Rethinking Backpropagating Refinement for Interactive Segmentation / Samsung AI Center - Moscow

https://arxiv.org/pdf/2001.10331.pdf https://www.youtube.com/watch?v=ArcZ5xtyMCk&feature=youtu.be



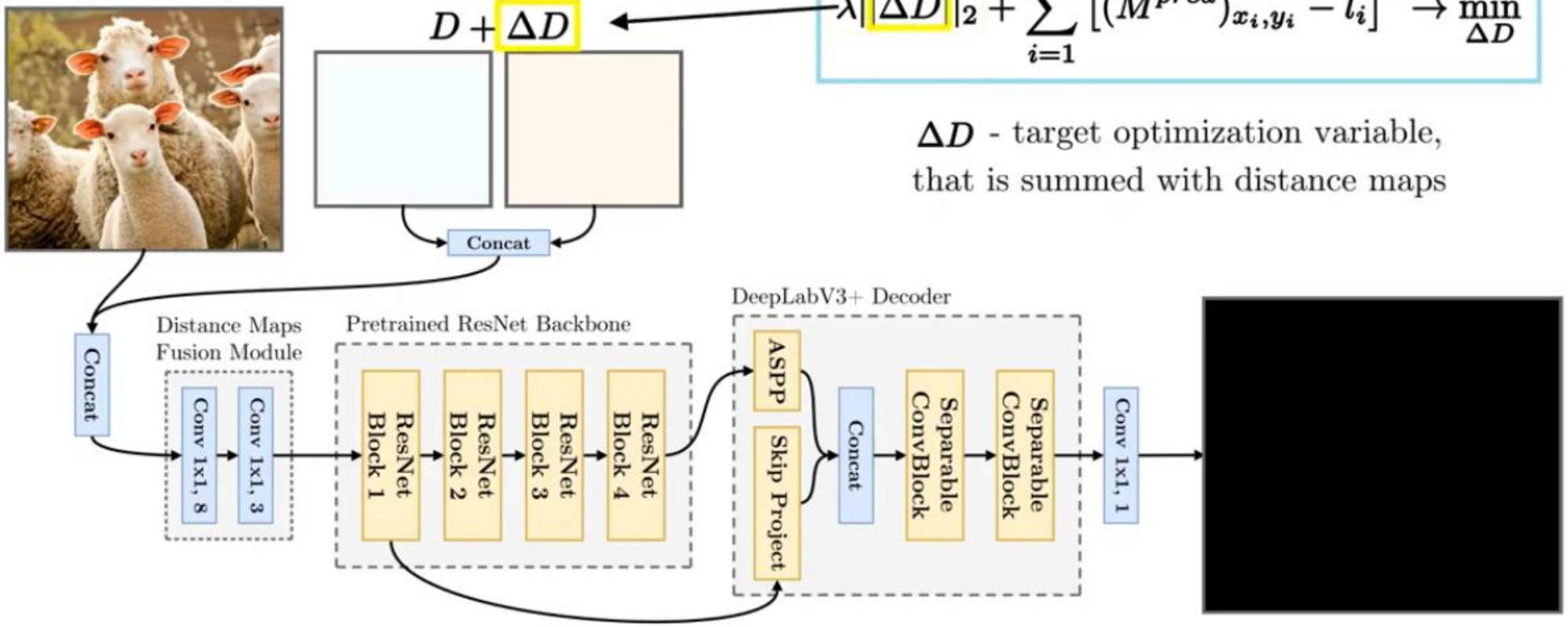
Appendix – f-BRS

Backpropagating Refinement Scheme

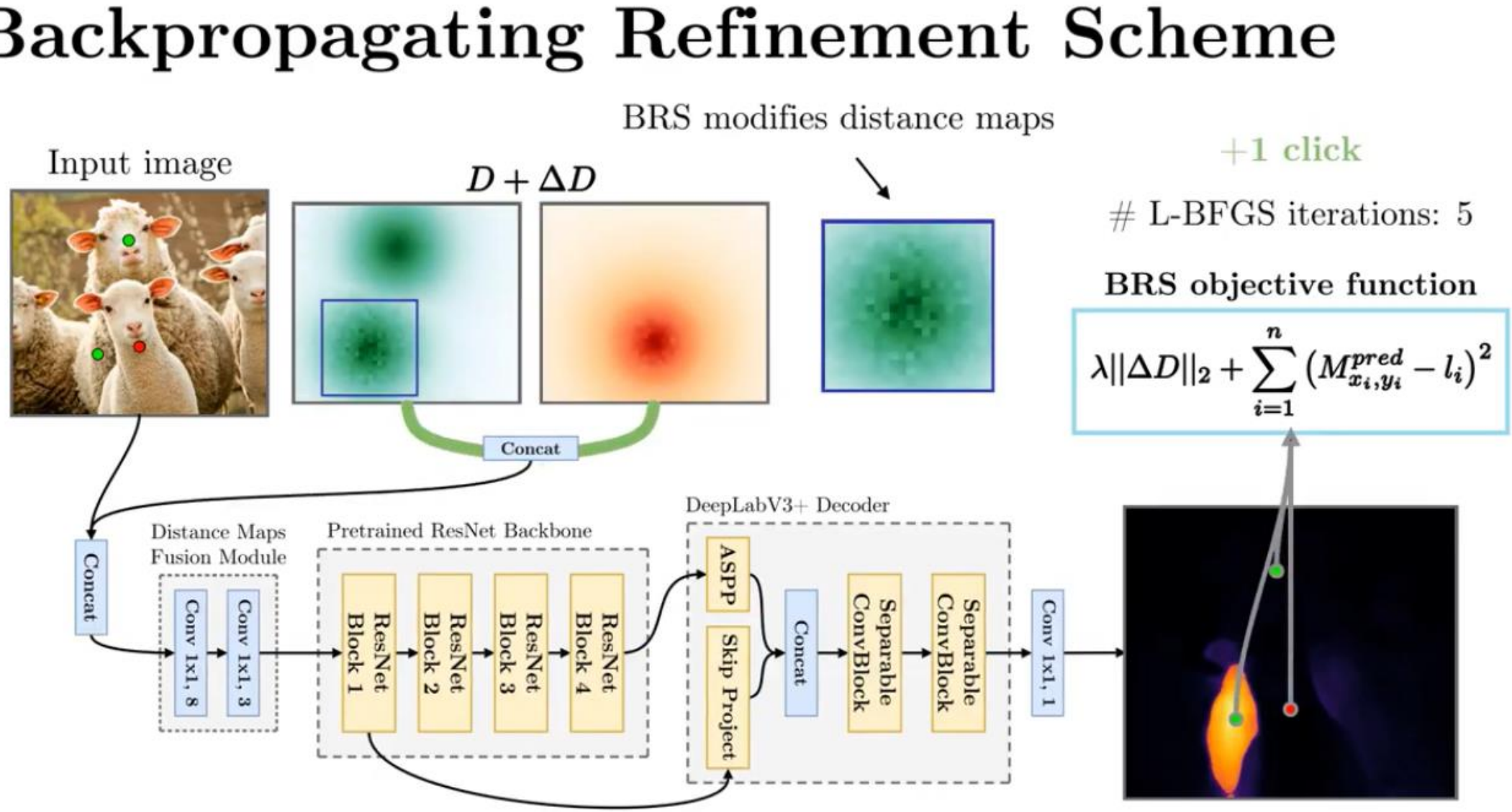
BRS optimization problem:

$$\lambda \left| \left| \Delta D \right| \right|_2 + \sum_{i=1}^{n} \left[(M^{pred})_{x_i, y_i} - l_i \right]^2 \rightarrow \min_{\Delta D}$$

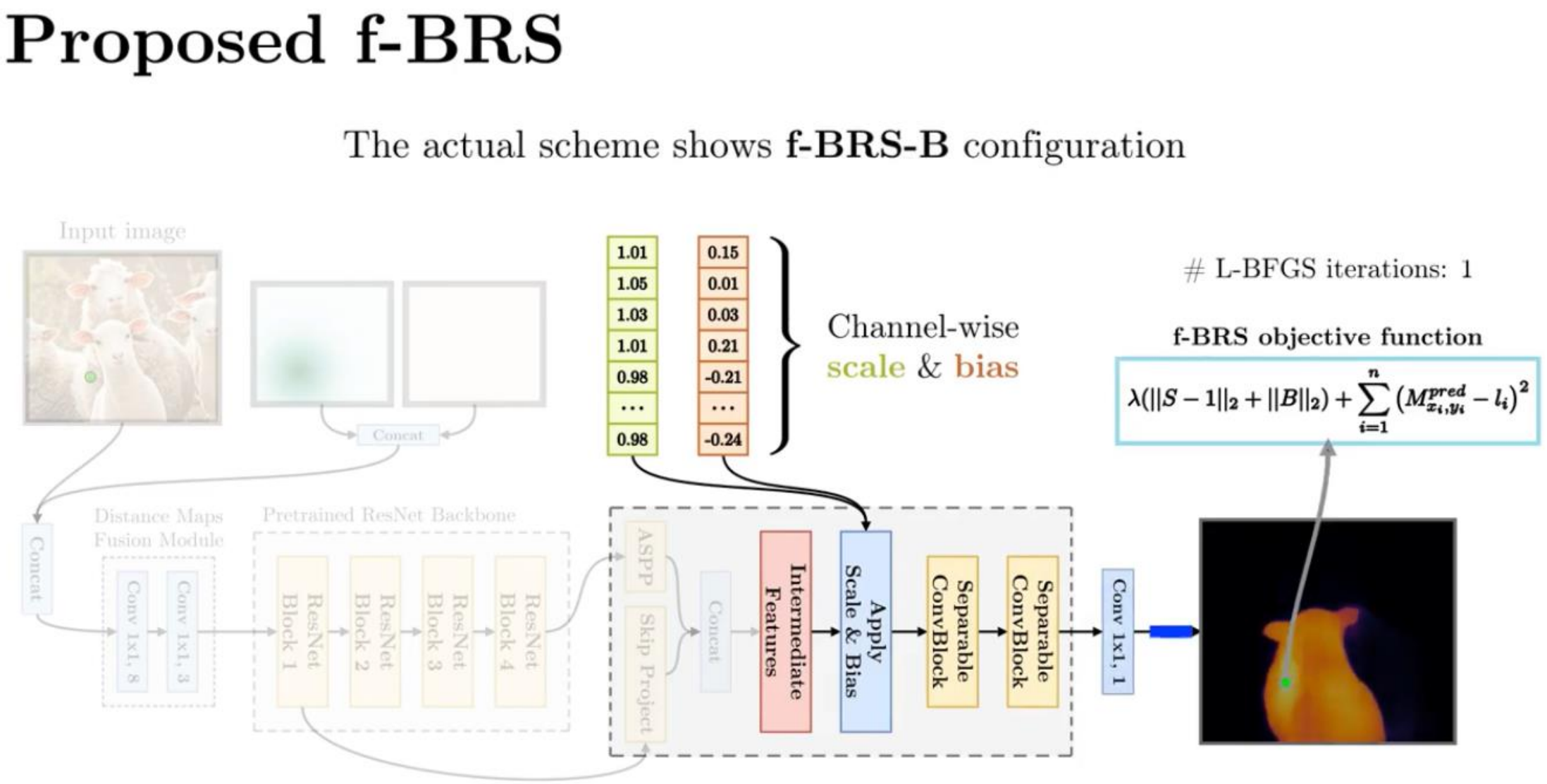
ΔD - target optimization variable, that is summed with distance maps



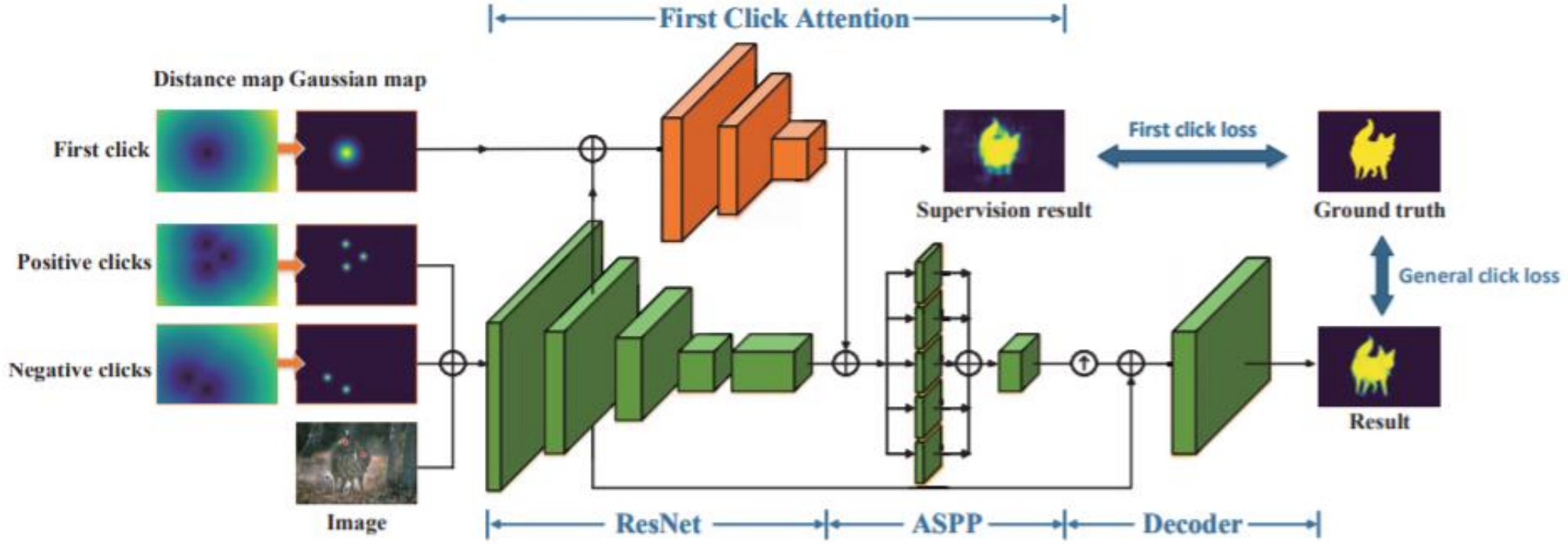
Appendix – f-BRS



Appendix – f-BRS



Appendix – Interactive Image Segmentation with First Click Attention (SOTA)

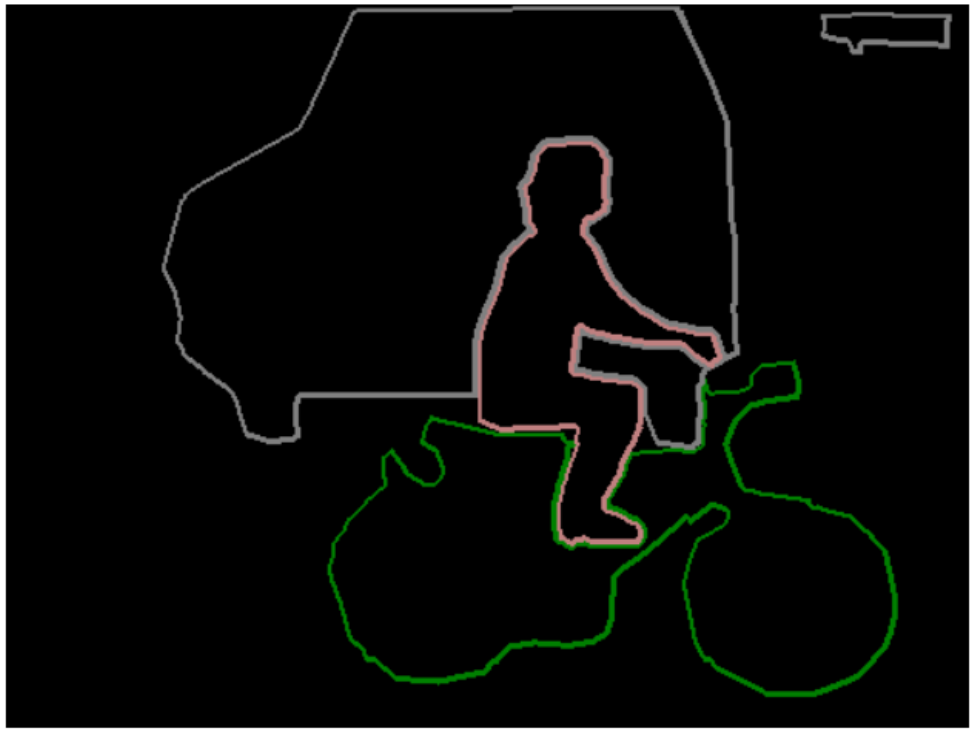


Method	GrabCut @90%	Berkeley @90%	PASCAL VOC @85%	DAVIS @90%	MSCOCO (seen)@85%	MSCOCO (unseen)@85%
BRS [25]$_{CVPR19}$	3.60	5.08	N/A	8.24	N/A	N/A
CMG [37]$_{CVPR19}$	3.58	5.60	3.62	N/A	5.40	6.10
FCA-Net	2.24	4.23	2.98	8.05	4.49	5.54
FCA-Net (SIS)	2.14	4.19	2.96	7.90	4.45	5.33
FCA-Net*	2.16	3.92	2.79	7.64	4.34	5.36
FCA-Net* (SIS)	2.08	3.92	2.69	7.57	4.08	5.01
f-BRS-B	2.46	4.34		7.41		

Appendix - datasets

Semantic Boundaries Dataset and Benchmark (SBD)

http://home.bharathh.info/pubs/codes/SBD/download.html



We created the Semantic Boundaries Dataset(henceforth abbreviated as SBD) and the associated benchmark to evaluate the task of predicting semantic *contours*, as opposed to semantic *segmentations*. While semantic segmentation aims to predict the pixels that lie *inside* the object, we are interested in predicting the pixels that lie on the *boundary* of the object, a task that is arguably harder (or alternatively, an error metric that is arguably more stringent).

Grapcut https://pgram.com/dataset/grabcut/

Berkeley https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/

DAVIS https://davischallenge.org/

Appendix – BRS paper

Paper link https://openaccess.thecvf.com/content_CVPR_2019/papers/Jang_Interactive_Image_Segmentation_via_Backpropagating_Refinement_Scheme_CVPR_2019_paper.pdf

Distance transform, generating hintmap from points

https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.ndimage.morphology.distance_transform_edt.html