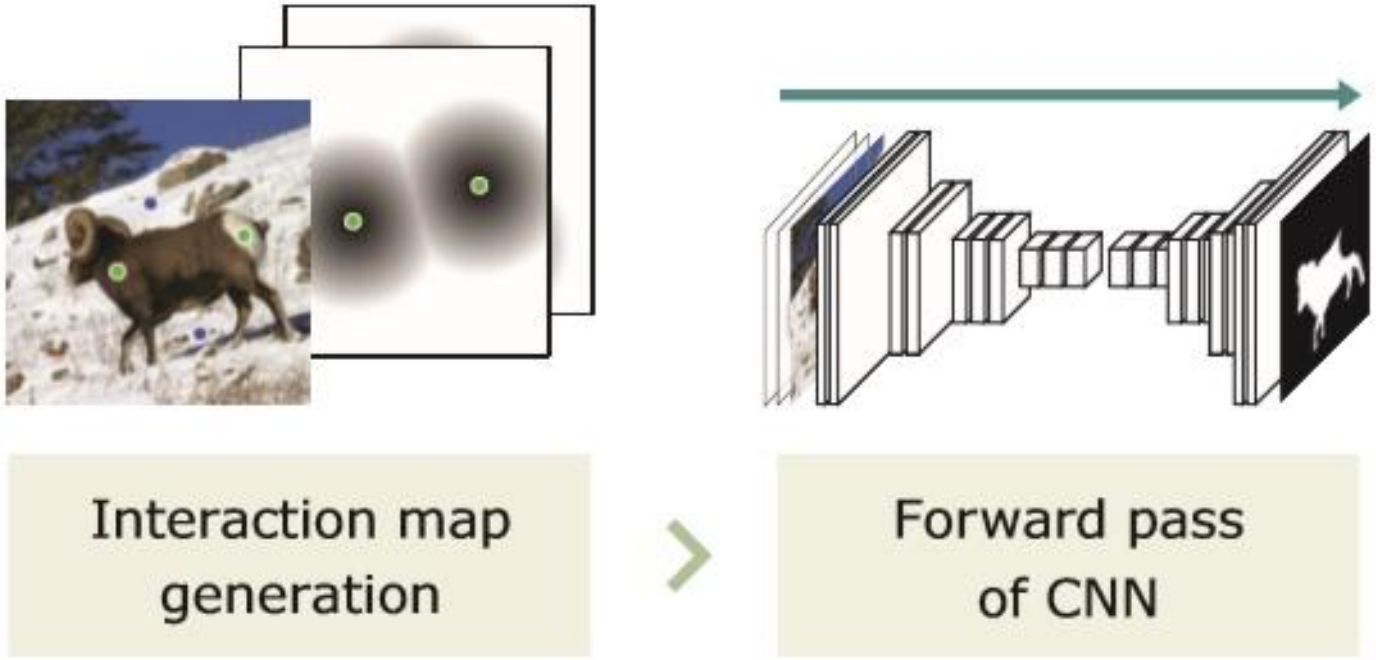
Interactive Image Segmentation with First Click Attention

Nankai University, CVPR 2020

Presenter : TAEU

Interactive Image Segmentation

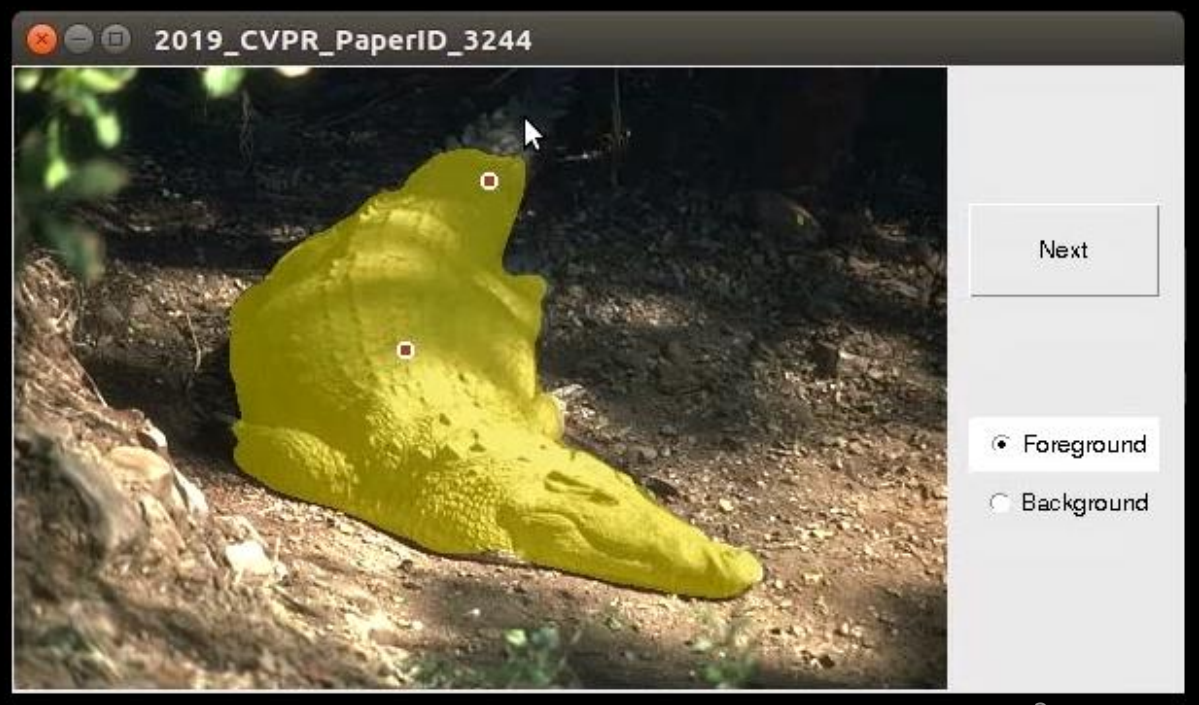


I) Input image with extreme points provided by annotator

II) Machine predictions from extreme points

III) corrective scribbles provided by annotator

IV) Machine predictions from extreme points and corrective scribbles



Related work : Interactive Image segmentation

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	5	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	68	SenseTime R. Hong Kong Univ	mask map(semantic mask), interactive conditional GANs
7	Interactive Image Segmentation With First Click Attention	3	Nankai University	focus on first click
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 Tr	0	Southwestern University, Tencent, China etc	-
10	Memory Aggregation Networks for Efficient Interactive Video Object Segmentation	3	Baidu, Sydney Tech	video object detection
11	Iteratively-Refined Interactive 3D Medical Image Segmentation With Multi-Agent R	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

Abstract

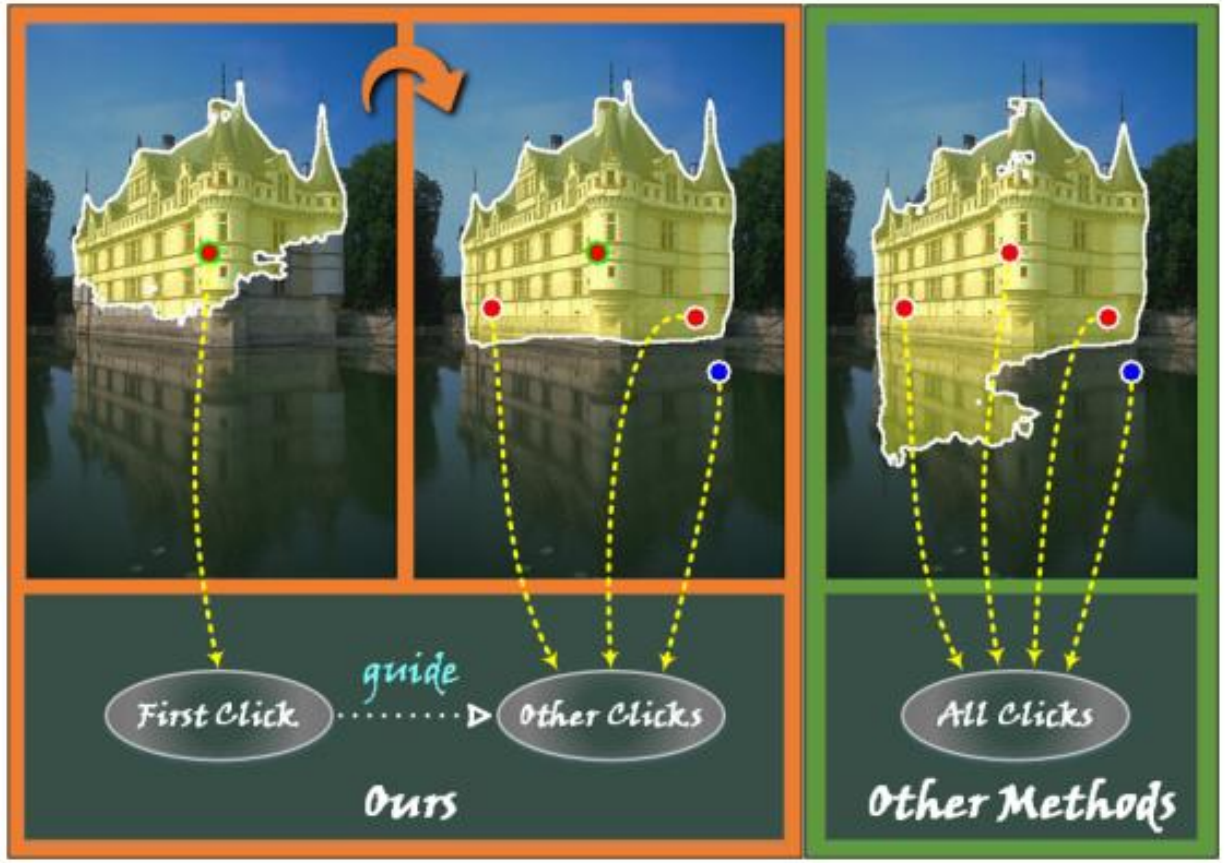


Figure 1. The crucial role of the first click in our method. We utilize the first click as a segmentation anchor to guide other clicks for a precise segmentation, while the conventional click-based interactive segmentation methods treat all clicks indiscriminately.

No.	1	2	3	4	5	6	7	8	9	10
PI	.751	.076	.045	.027	.020	.017	.015	.015	.009	.010
CD	.769	.312	.243	.207	.201	.211	.189	.188	.178	.186

Table 1. Statistics of user interactions. PI: Performance improvement (mean IoU) by adding different interaction points. CD: Centrality degree for describing how close the point is to the center of the object (only for positive points). Higher CD means closer to the center. The computational details are mentioned in Sec. 3.5.

- Focus on first click
- Click loss and a structural integrity strategy
- SOTA of 5 datasets

The overall architecture : FCA Net

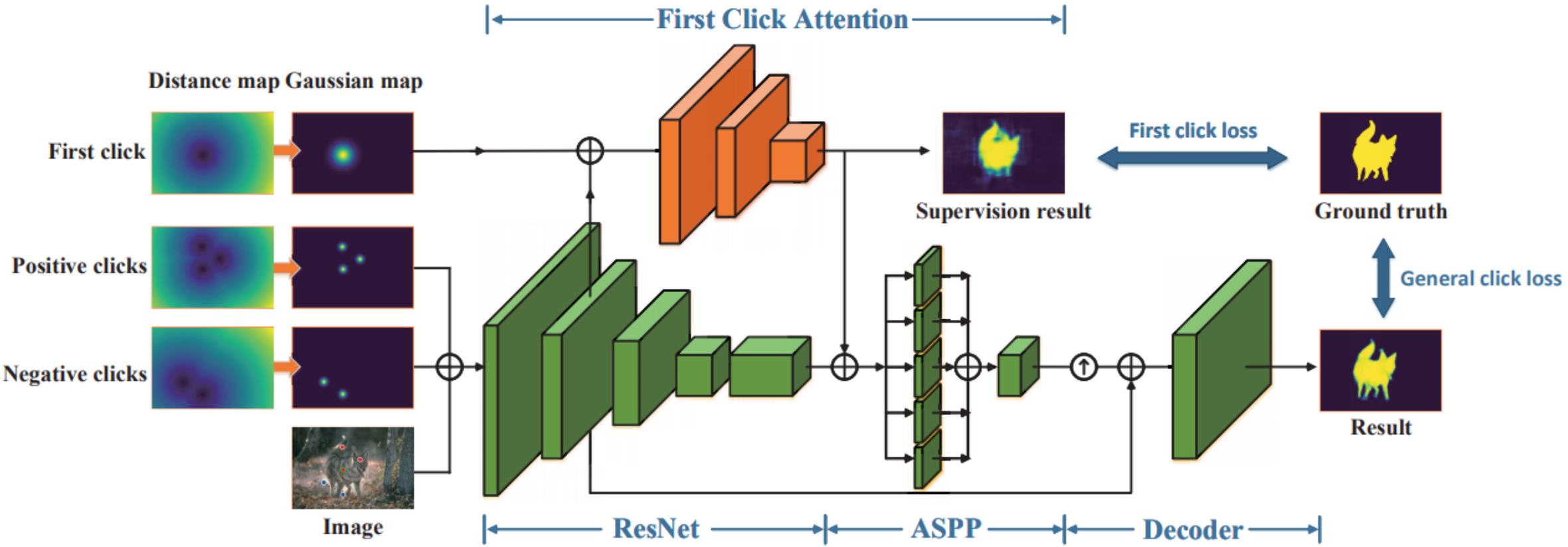
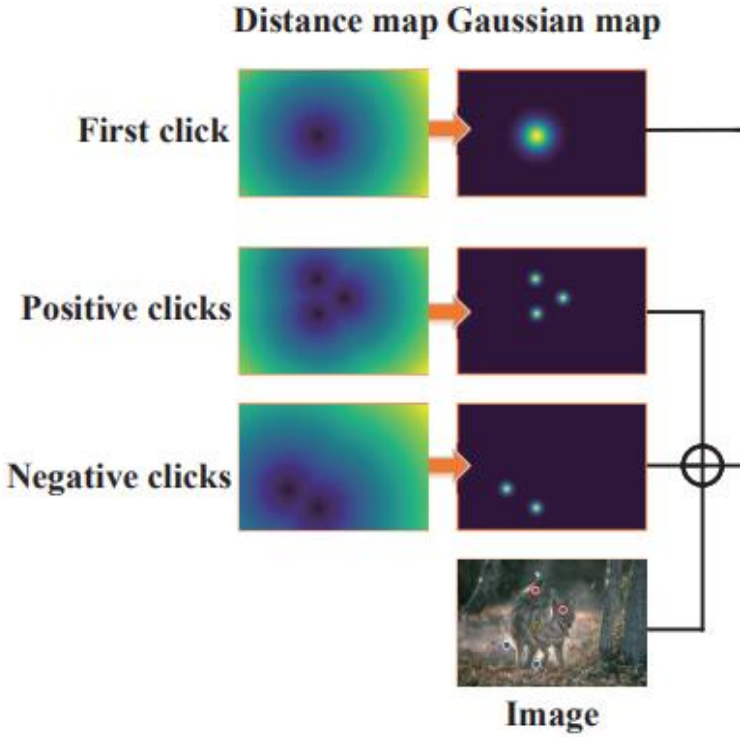
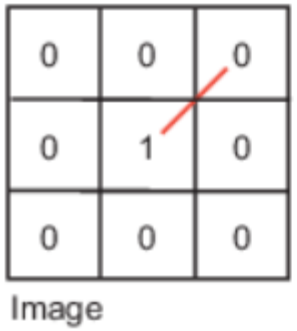


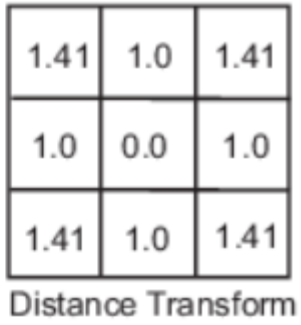
Figure 2. The overall architecture of FCA-Net. The green part shows the basic segmentation network with backbone, ASPP, and decoder modules. The orange part shows the first click attention module. Symbols "⊕" and "↑" mean concatenation and up-sampling operations, respectively. Consult Sec. 3.1 for more details.

Network Architecture

- Input : First click distance map(CDM) + Positive CDM + Negative CDM + image







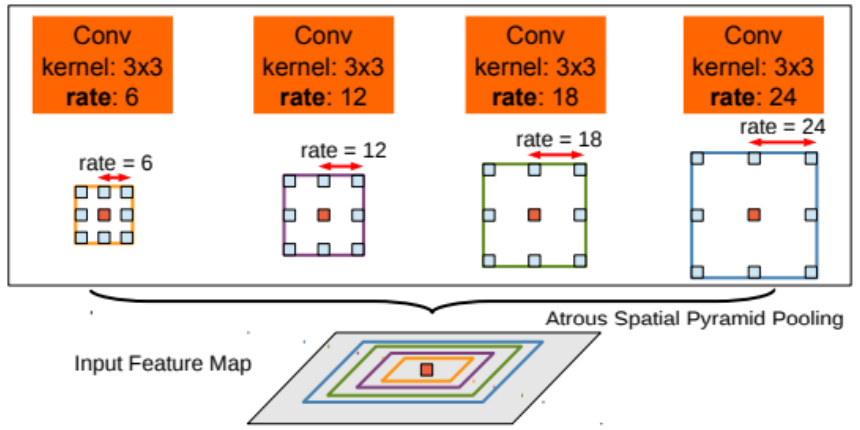
Euclidean distance map : distance between positive value pixel and background (value : 0) pixel

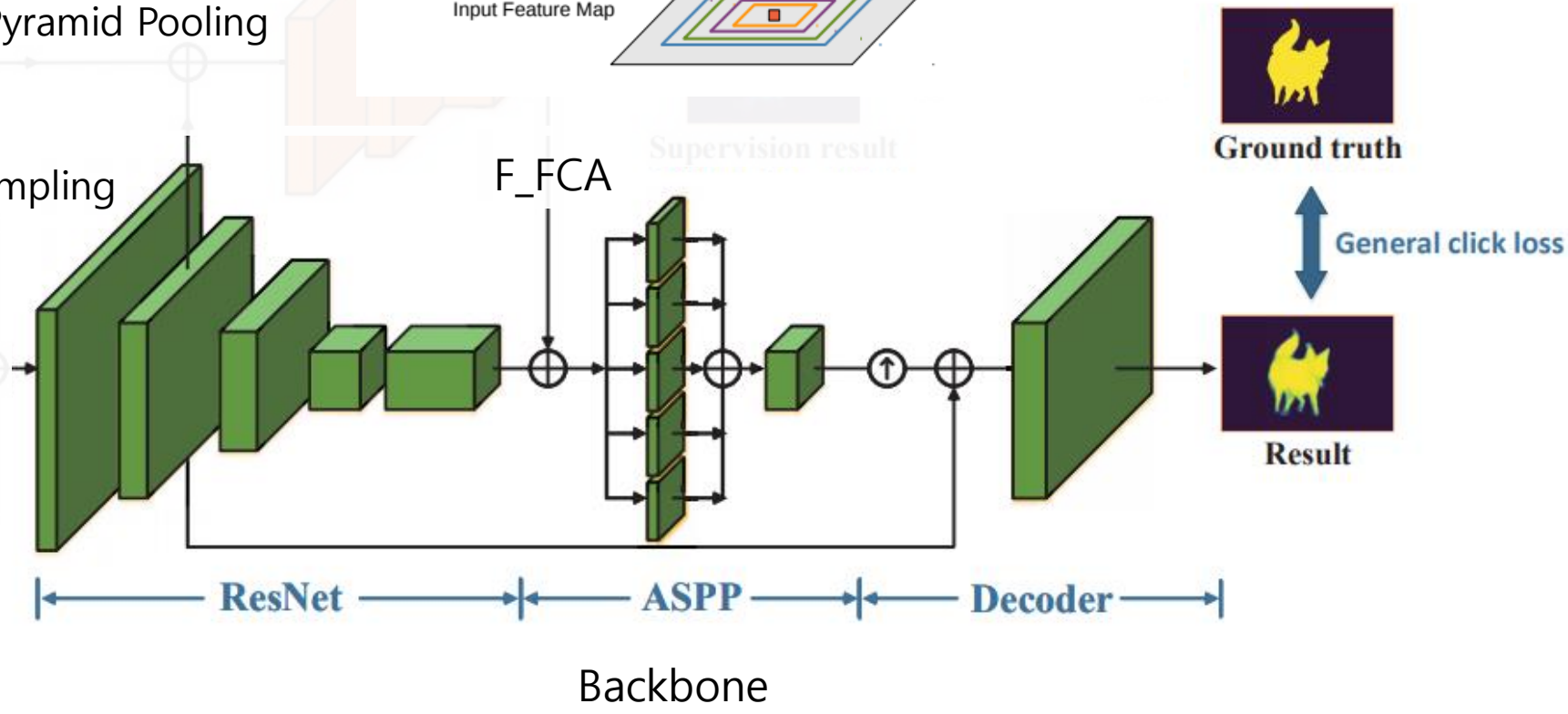
Distance map to Gaussian map : multiply by a negative value and divide by a value(sigma)

Network Architecture

- Base Network
  - FCN similar to DeepLab v3+
  - Backbone : ResNet101
  - ASPP : Atrous Spatial Pyramid Pooling
  - Decoder
  - ⊕ : concat , ↑ : up-sampling
  - Last four stage :
    {F1, F2, F3, F4}

Fig. ASPP

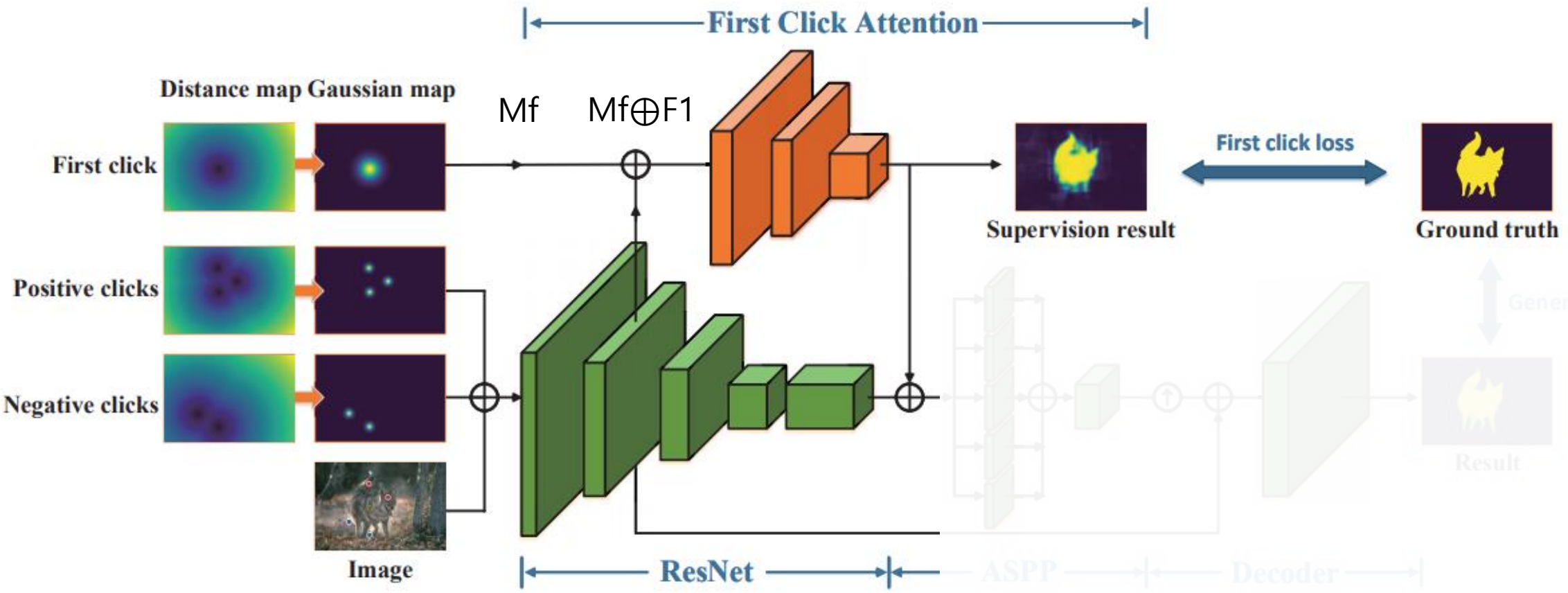




ASPP figure : https://arxiv.org/pdf/1606.00915.pdf

Network Architecture

- First click attention module : six 3 × 3 convolution layers.



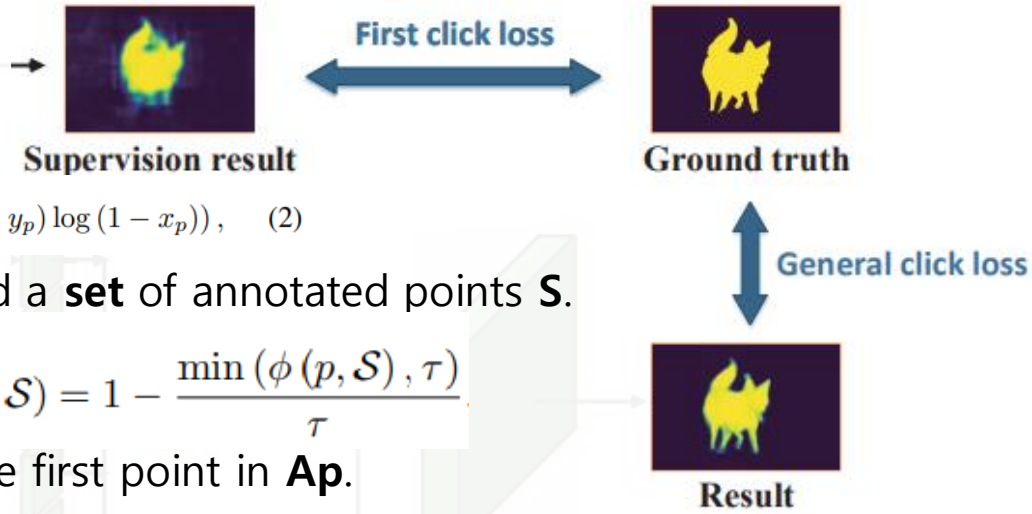
Click Loss

- Notation
  - Gts all pixels are represented as **G**.
  - **Gp** and **Gn** to represent the sets of pixels in foreground and background according to the ground truth.
  - **A** represents all the annotated points.
  - **Ap** and **An** represent positive points and negative points respectively.
  - **d(p1, p2)** to represent the Euclidean distance between point **p1** and point **p2**
  - **φ(p, S)** to represent the shortest distance from one point **p** to another region **S**. $\phi(p, \mathcal{S}) = \min_{\forall p_s \in \mathcal{S}} d(p, p_s)$.

Click Loss

- Loss with notation
  - Gts all pixels are represented as **G**. **Gp** and **Gn**
  - **A** represents all the annotated points. **Ap** and **An**
  - **d :** Euclidean distance
  - **φ(p, S) :** shortest distance $\phi(p, \mathcal{S}) = \min_{\forall p_s \in \mathcal{S}} d(p, p_s)$.
  - **Weighted Binary Cross Entropy(BCE)** $\ell(p) = -(y_p \log(x_p) + (1 - y_p) \log(1 - x_p)), \quad (2)$
  - **ψ** to represent the distance weight between a point **p** and a **set** of annotated points **S**.
  - **τ** is the influence range of each annotated point. $\psi(p, \mathcal{S}) = 1 - \frac{\min(\phi(p, \mathcal{S}), \tau)}{\tau}$
  - **α** and **β** are used to adjust the range of loss. **af** means the first point in **Ap**.
  - In experiments, we choose τ at 100, α at 0.8, β at 2.0.

$$\mathcal{L}_f = \frac{1}{N} \sum_{p \in \mathcal{G}} (\widetilde{w}_p \cdot \ell(p)).$$



$$\mathcal{L}_g = \frac{1}{N} \sum_{p \in \mathcal{G}} (\widehat{w}_p \cdot \ell(p)).$$

$$\widehat{w}_p = \begin{cases} \alpha + \psi(p, \mathcal{A}_p)(\beta - \alpha), & y_p = 1 \\ \alpha + \psi(p, \mathcal{A}_n)(\beta - \alpha), & y_p = 0 \end{cases}$$

$$\widetilde{w}_p = \alpha + \psi(p, \{a_f\})(\beta - \alpha) y_p$$

Structural Integrity Strategy

- The prediction masks of neural networks may contain some scattered regions of wrong results.
- To maintain the structural integrity of the segmentation based on interaction points.
- **P** represent points which are predicted as foreground.
- **Postprocess** these prediction areas according to the interaction points and get new **P'** , which is formulated as follows:

$$\mathcal{P}' = \{p \in \mathcal{P} | \exists_{a \in \mathcal{A}_p} \sigma(p, a) = 1\}$$

where σ(p1, p2) = 1 when there is an eight-connected path from point p1 to point p2.

- Code : CV2.floodFill

Strength Analysis

- **Focus Invariance.**
  - If the neural network treats these points equally as the first point, it will often result in a wrong segmentation.
- **Location Guidance.**
  - With the first click attention, the prediction will focus on the location of the first click and get a better result.
- **Error-Tolerant Ability.**
  - We want to segment the penguin. A positive point on the right near the boundary of the target object accidentally falls into the background area. With the guidance of the first point, the influence of these error points will be greatly reduced.

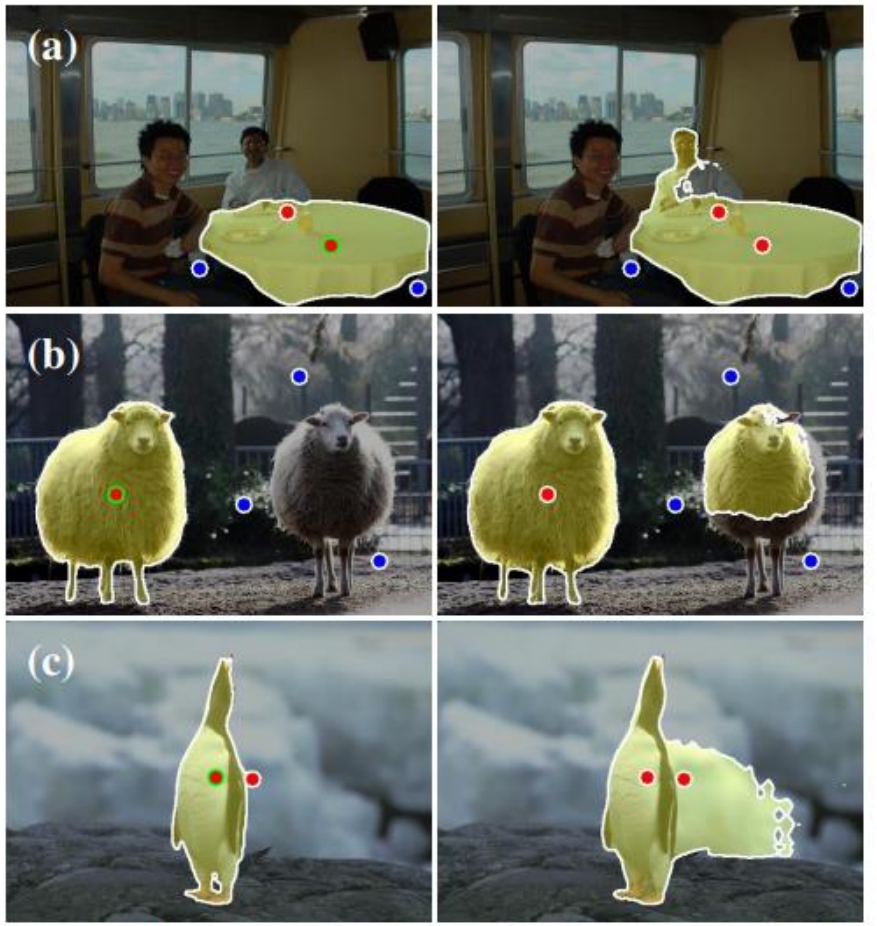


Figure 4. Illustration for benefits of first click attention. The left and right columns show the prediction masks with and without the FCA module, respectively.

Implementation Details

- **General Click Simulation.**
  - The foreground and background are determined randomly within [1, 10] and [0, 10], respectively.

$$\mathcal{C}_p = \{p \in \mathcal{G}_p | \phi(p, \mathcal{G}_n) > P_1, \phi(p, \mathcal{A}^*) > P_2\}. \qquad \mathcal{C}_n = \{p \in \mathcal{G}_n | \phi(p, \mathcal{G}_p) \in (N_1, N_2), \phi(p, \mathcal{A}^*) > N_3\}.$$

- **First Click Simulation.**
  - E(p) closer to 1 means that the first click point locates at a more central position of the object.
  - Choose the point whose E(p) equals 1 in cropped training images as the first point.

$$\mathcal{E}(p) = \frac{\phi(p, \mathcal{G}_n)}{\max_{\forall p_0 \in \mathcal{G}_p} \phi(p_0, \mathcal{G}_n)}.$$

- **Others..**
  - **Iterative training strategy.** 512 × 512 cropped image, batch size is 8, learning rate to 0.007 for ResNet and 0.07 for other parts and take stochastic gradient descent with 0.9 momentum for optimization, the polynomial learning rate decay for 30 epochs and constant learning rate for additional 3 epochs in the end.

Ablation study

#	FCANet	PASCAL	Berkeley
1	BS	4.21	5.74
2	BS + FCA	3.66	5.22
3	BS + FCA + CL	3.33	4.94
4	BS + FCA + CL + Iter	2.98	4.23
5	BS2 + FCA + CL + Iter	2.79	3.92

Table 3. Ablation study of proposed methods. BS: baseline; BS2: baseline implemented by Res2Net; FCA: first click attention module; CL: click loss; Iter: iterative training.

Iterative training : https://arxiv.org/pdf/1805.04398.pdf

Result

Method	GrabCut @90%	Berkeley @90%	PASCAL VOC @85%	DAVIS @90%	MSCOCO (seen)@85%	MSCOCO (unseen)@85%
GC [6]$_{ICCV01}$	11.10	14.33	15.06	17.41	18.67	17.80
GRC [45]$_{POG05}$	16.74	18.25	14.56	N/A	17.40	17.34
RW [18]$_{PAMI06}$	12.30	14.02	11.37	18.31	13.91	11.53
GM [3]$_{IJCV09}$	12.44	15.96	14.75	19.50	17.32	14.86
ESC [19]$_{CVPR10}$	8.52	12.11	11.79	17.70	13.90	11.63
GSC [19]$_{CVPR10}$	8.38	12.57	11.73	17.52	14.37	12.45
DOS [47]$_{CVPR16}$	6.04	8.65	6.88	12.58	8.31	7.82
RIS [31]$_{ICCV17}$	5.00	6.03	5.12	N/A	5.98	6.44
LD [30]$_{CVPR18}$	4.79	N/A	N/A	9.57	N/A	N/A
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

Table 2. Comparison of the mean number of clicks (mNoC) on 6 sets over 5 datasets. SIS means the proposed structural integrity strategy for post-process. FCA-Net* indicates our model with Res2Net [16] as a backbone.

f-BRS-B	ResNet-34	**2.46**	4.65	8.21
	ResNet-50	2.98	**4.34**	7.81
	ResNet-101	2.72	4.57	**7.41**

Limitation

- (a), FCA-Net is not good at segmenting multiple instances in an image at the same time. Fortunately, in real-world applications, the limitation can be alleviated by annotating for each instance object with its own first click.
- (b-c), we show two interesting scenes, where the center of these instances may not be clicked by users due to the structure or occlusion.

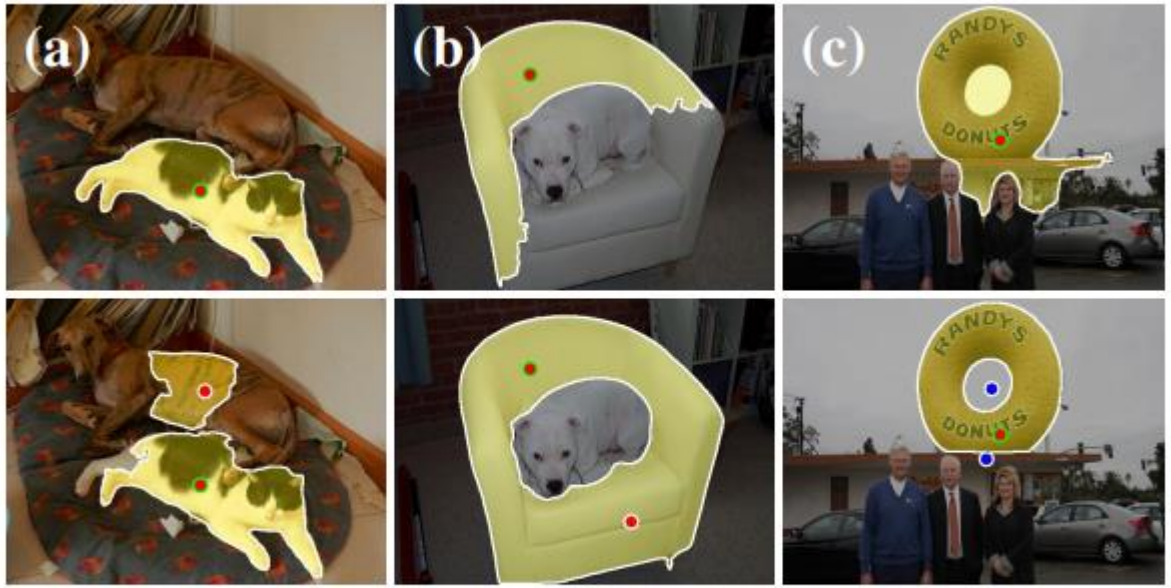


Figure 6. Illustration the possible limitations of the proposed FCA-Net. Points with green ring represent the first click.

Another good works

- **[1] Efficient Full Image Interactive Segmentation by Leveraging Within-image Appearance Similarity / google R /** https://arxiv.org/pdf/2007.08173.pdf
- **[2] F-brs : https://arxiv.org/abs/2001.10331**

[1]

[2]

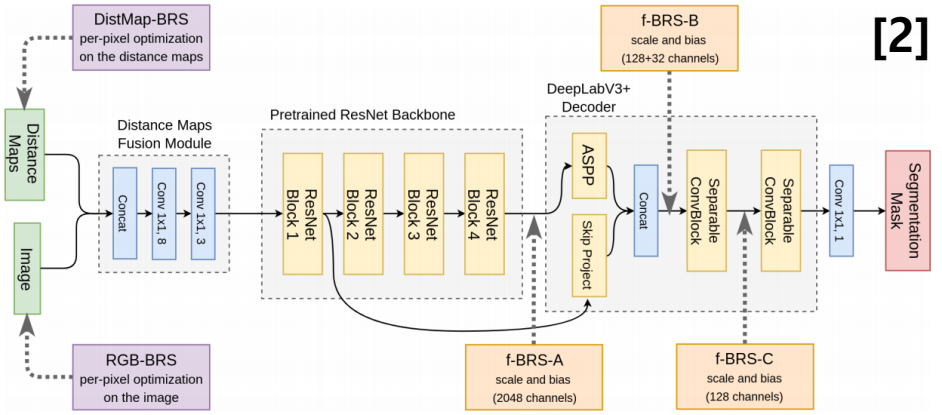


Figure 2. Illustration of the proposed method described in Section 3. f-BRS-A optimizes scale and bias for the features after pre-trained backbone, f-BRS-B optimizes scale and bias for the features after ASPP, f-BRS-C optimizes scale and bias for the features after the first separable convblock. The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

FCANet, Reference

- Paper : http://mftp.mmcheng.net/Papers/20CvprFirstClick.pdf
- Code : https://github.com/frazerlin/fcanet