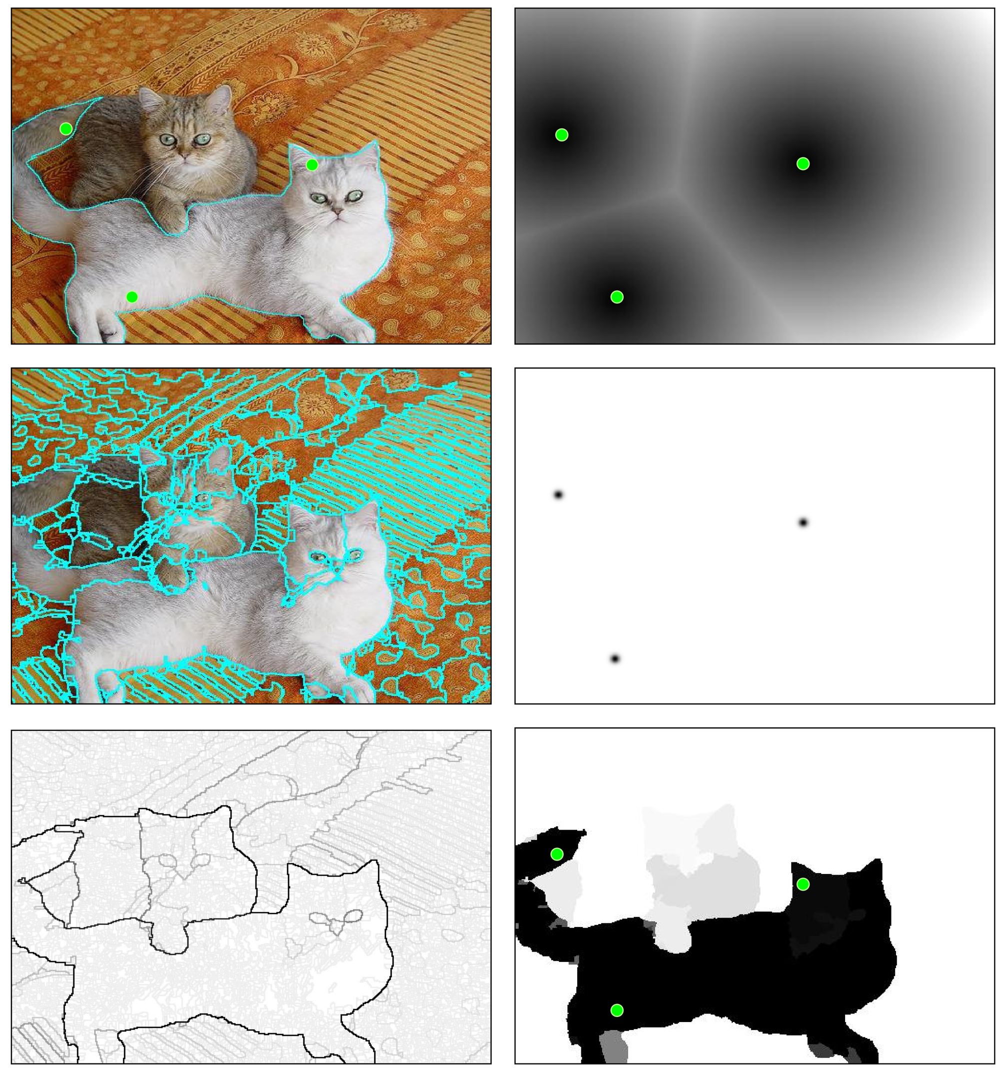
Content-Aware Multi-Level Guidance for Interactive Instance Segmentation

Soumajit Majumder[1], Angela Yao[2]

[1]University of Bonn and [2]National University of Singapore

Motivation



Current interactive instance segmentation ignores the structures in the input image when generating *guidance* maps from user clicks. Our work proposes :

- ▶ Novel transformation of clicks based on superpixels and object proposals.
- ▶ Framework which accounts for the scale of an object.

With our guidance maps, a basic FCN outperforms existing approaches with state-of-the-art segmentation networks (DeepLabv3+) !

Guidance Maps

- ▶ $\{\mathcal{S}\}$ - set of superpixels [5], $\{\mathbf{s}_t = f_{SP}(\mathbf{p}_t)\}$ - set of positive and negative superpixels for user-provided positive and negative clicks.
- ▶ $d_c(s_i, s_j)$ - Euclidean distance between the centers of superpixels s_i and s_j.
- ▶ **Superpixel guidance map**

$$\mathcal{G}_t^{sp}(\mathbf{p}) = \min_{s\in\{\mathbf{s}_t\}} d_c\left(s, f_{SP}(\mathbf{p})\right), \quad \text{where} \quad t=\{0,1\}, \tag{1}$$

- ▶ $\{\mathcal{L}_p\}$ - set of category-independent object proposals [5] for an image with support of pixel location $\mathbf{p}$.
- ▶ **Object-based guidance map**

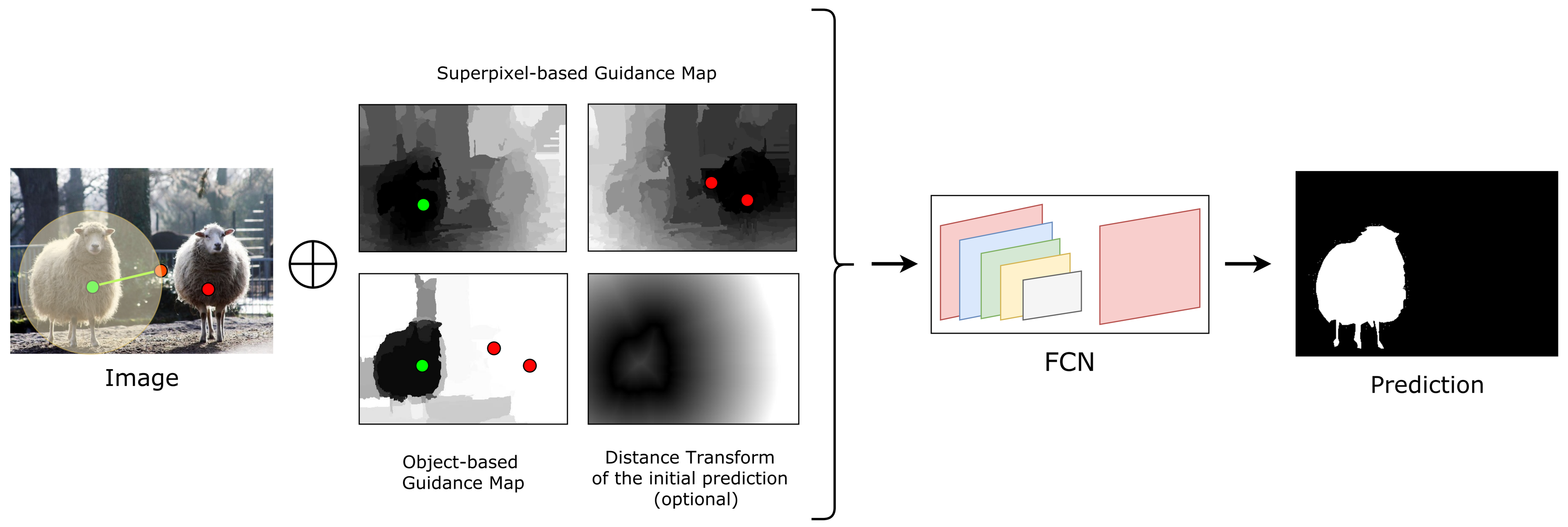
$$\mathcal{G}^{o}(\mathbf{p}) = \sum_{\mathbf{p}'\in\{\mathbf{p}_0\}} \sum_{\mathcal{L}\in\{\mathcal{L}_{p'}\}} \mathbf{1}[\mathbf{p}\subset\mathcal{L}] \tag{2}$$

- ▶ s - estimated scale, f_1, f_2 - tolerance factors.
- ▶ **Scale-aware guidance map**

$$\mathcal{G}^{o\text{-}sc}(\mathbf{p}) = \sum_{\mathbf{p}'\in\{\mathbf{p}_0\}} \sum_{\mathcal{L}\in\{\mathcal{L}_{p'}\}} \mathbf{1}[\mathbf{p}\subset\mathcal{L}]\cdot\mathbf{1}[f_1 \le |\mathcal{L}|/s^2 \le f_2]. \tag{3}$$

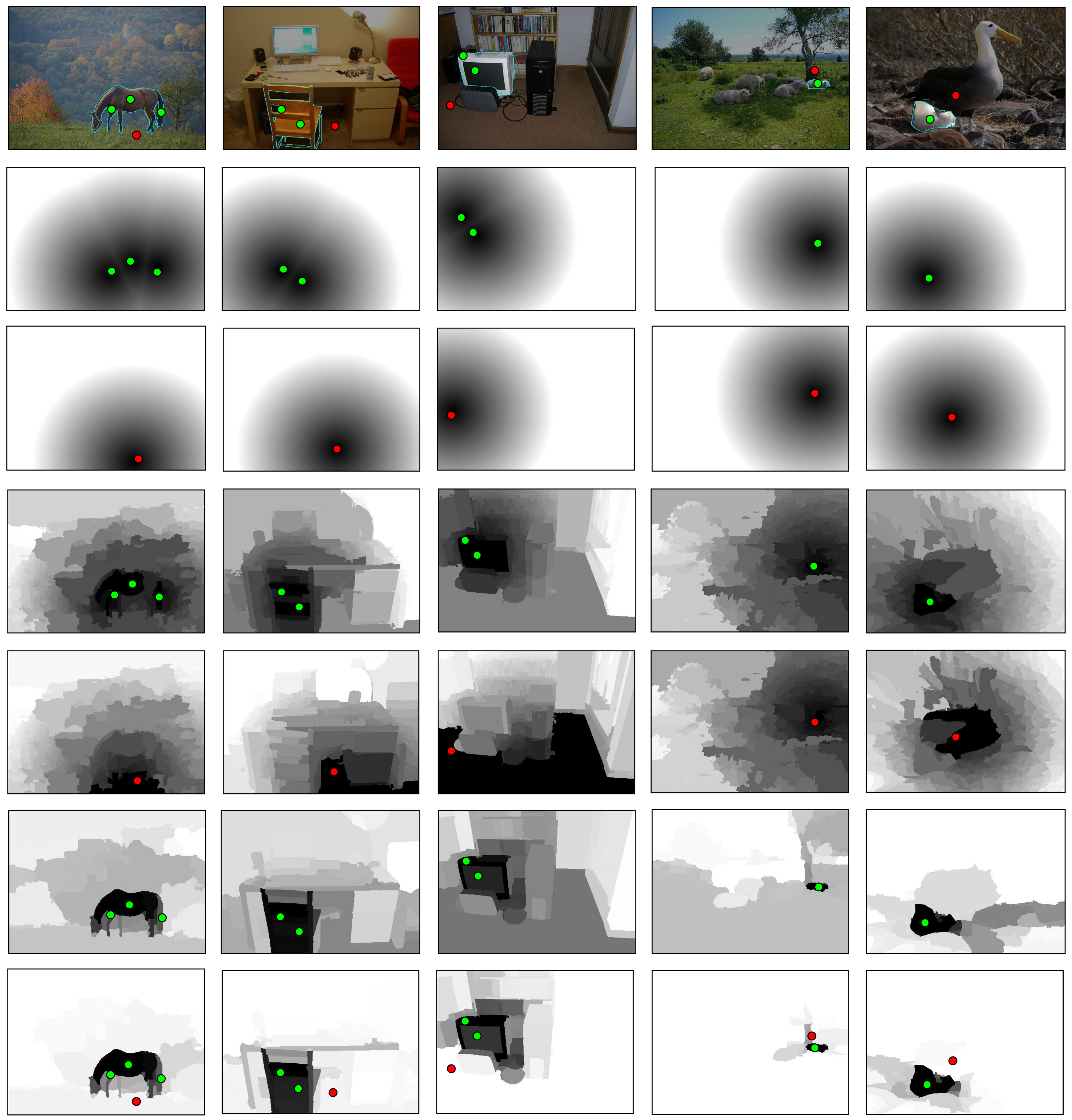
- ▶ 621 objects (from PASCAL VOC 2012) smaller than 32×32.
- ▶ 2% improvement over the scale agnostic version.

Outline



Outline The generated guidance maps are concatenated (denoted as $\oplus$) with the 3-channel image and is fed to the segmentation network.

Example of Guidance Maps



Row 1 : Original image with object of interest highlighted. **Rows 2-3** : positive and negative euclidean distance map. **Rows 4-5** : positive and negative superpixel based guidance map. **Row 6** : Object proposal based guidance map. **Final row** : Scale-aware guidance map.

Impact of Structure-Based Guidance

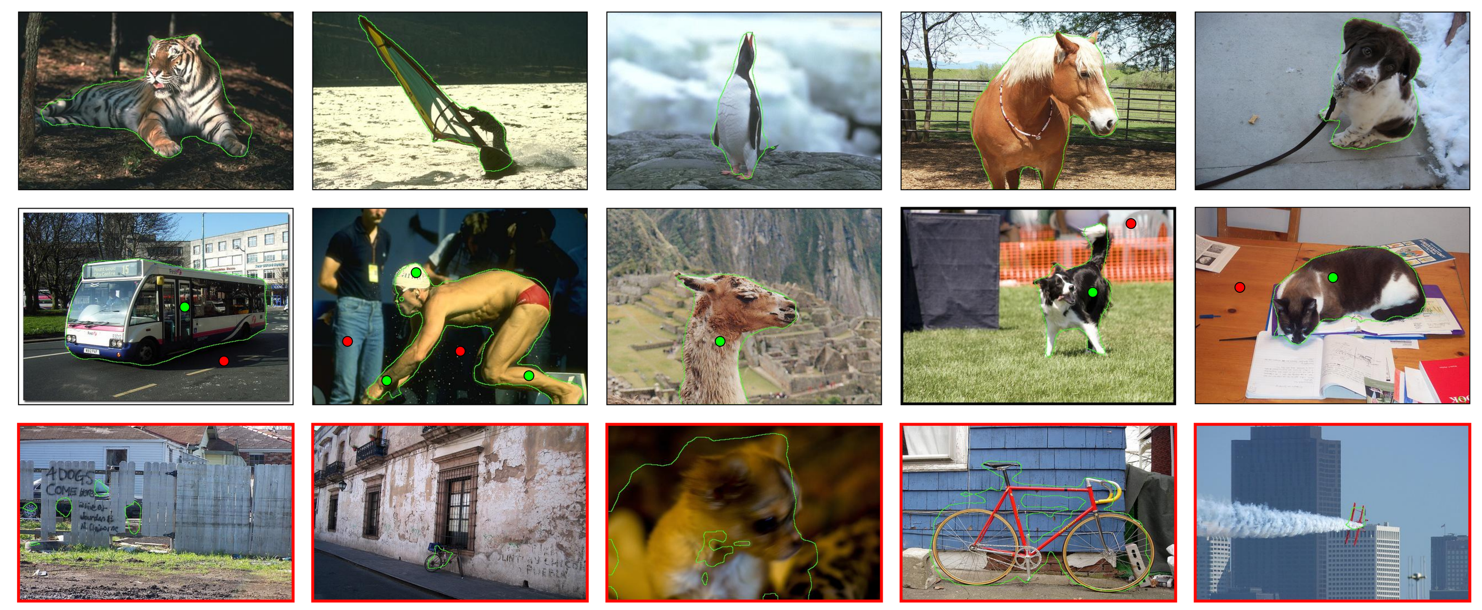
	GrabCut @90%	Berkeley @90%	VOC 2012 @85%
Euclidean [6]	6.04	8.65	6.88
Superpixel	4.44	6.67	4.23
Superpixel + Object	3.82	6.05	4.02
Superpixel + Object + Iterative[3]	**3.58**	**5.60**	**3.62**

Clicks required to segment instance.Guidance maps leveraging structural information require significantly less clicks than Euclidean distance-based guidance, especially .

Results

Comparison to State-of-the-Art : Average clicks required.

Method	Base Network	GrabCut @90%	Berkeley @90%	VOC 12 @85%	MS-COCO seen@85%	MS-COCO unseen@85%
Graph cut [2]	-	11.10	14.33	15.06	18.67	17.80
iFCN [6]	FCN-8s	6.04	8.65	6.88	8.31	7.82
ITIS [3]	DeepLabv3+	5.60	-	3.80	-	-
DEXTR [4]	DeepLabv2	4.00	-	4.00	-	-
VOS-Wild [1]	ResNet-101	3.80	-	5.60	-	-
Ours	FCN-8s	**3.58**	**5.60**	**3.62**	**5.40**	**6.10**



First Row: 'acceptable' segmentations without any user guidance. **Second row**: a few clicks removes background and undesired objects. **Third row**: Representative failures include small objects, occlusion, motion blur and objects with fine structures.

Discussion

- ▶ Does encoding user clicks with superpixels and object proposals simplify learning ?
- ▶ Too easy ? Base network meets the mIoU criteria without any clicks: VOC 2012 (433 of 697), Grabcut (13 of 50), Berkeley (15 of 100).
- ▶ Too hard ? For objects with very fine detailing, e.g. bike wheel spokes, partially occluded chairs our algorithm exhausted the 20 click budget.

References

[1] Bénard et al., Interactive video object segmentation in the wild, arXiv 2017.
[2] Boykov et al., Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images, ICCV 2001.
[3] Mahadevan et al., Iteratively trained interactive segmentation, BMVC 2018.
[4] Maninis et al., Deep extreme cut: From extreme points to object segmentation, CVPR 2018.
[5] Pont-Tuset et al., Multiscale combinatorial grouping for image segmentation and object proposal generation, TPAMI 2017.
[6] Xu et al., Deep interactive object selection, CVPR 2016.