U-NET based Non-Local Convolutional Feature Network for Saliency Detection

Magauiya Zhussip
Ulsan National Institute of Science and Technology (UNIST), South Korea
{mzhussip}@unist.ac.kr

Abstract

Saliency detection aims to imitate human visual attention system by identifying the most distinct regions in a scene. With the emergence of CNNs, deep learning saliency detection methods become more popular over conventional solutions. Its strong representation power allows CNNs to learn local as well as deep features and hence deliver superior performance in various computer vision (CV) tasks. In this work, we propose a U-NET based non-local convolutional feature network (U-NLCF) for saliency detection problem. By integrating global and local information and using hybrid upscaling technique, our U-NLCF produces high-quality saliency map with no artifacts. Moreover, its compact architecture reduces computation time by several times, enabling near real-time salient object detection.

1. Introduction

Saliency detection aims to imitate human visual attention system by identifying the most distinct regions and predominant objects in an image. Saliency detection has been used as a preprocessing procedure for various computer vision tasks including scene classification [28], object detection [34], visual tracking [9], context aware image retargeting [17], image and video compression [7], and segmentation [25]. In spite of significant progress in this area [3, 10, 15, 24], detecting salient objects in complex real-world scenarios remains challenging task.

Usually salient objects or regions are described as regions with visually distinct features that can follow some prior criteria. Conventional saliency detection methods utilize various hand-crafted priors for local and global feature extraction. However, with the emergence of deep learning and the empirical proof of its superior performance in many computer vision tasks, data-driven saliency detection methods become more and more popular. Because of its strong representation power, deep learning based saliency detection networks are able to learn local as well as deep features and have established themselves as a *de facto* benchmark [21, 23, 31].

There have been many approaches of designing a network architecture especially tailored for saliency detection [17, 21, 22, 30, 36]. Although they have achieved exceptional performance, their complex model architecture and additional post-processing hinder computational speed-up. Moreover, CNN based methods which utilize deconvolution operation as an upsampling mechanism from low to high resolution shows checkerboard artifacts in their results [35]. Therefore, it is of great interest to develop a simpler model architecture with new upsampling methods to speed-up the computation and to remove odd artifacts. In this work, we propose a compact FCN network called U-NLCF, which uses both global and local features for saliency map prediction that can achieve near state-of-the-art performance. Our model has several unique features, as outlined below:

1. Our model is based on U-NET structure: encoder FCN followed by decoder FCN connected with skip connections. The encoder part learns global features from the input image, while the decoder tries to catch local features by upsampling encoded feature maps to the original input size. The resulting global and local features are combined into a single block giving a final saliency map.
2. We implement contrast feature blocks [23] in our network to enforce local feature maps to learn more contrast features.
3. We propose a new upsampling method based on hybrid upsampling technique [35] to deal with checkerboard artifacts resulted from traditional deconvolution operations.

2. Related works

Traditional approaches of detecting saliency objects are based on unsupervised models with some hand-crafted priors using simple edge, texture, and color features [4, 27] and more advanced features at different scales. For instance [12] proposed uniqueness, focusness, and objectness features for saliency detection, while [13, 16] operate at muti-level and utilize graph-based approach to achieve better performance. Although unsupervised methods are comparatively simple

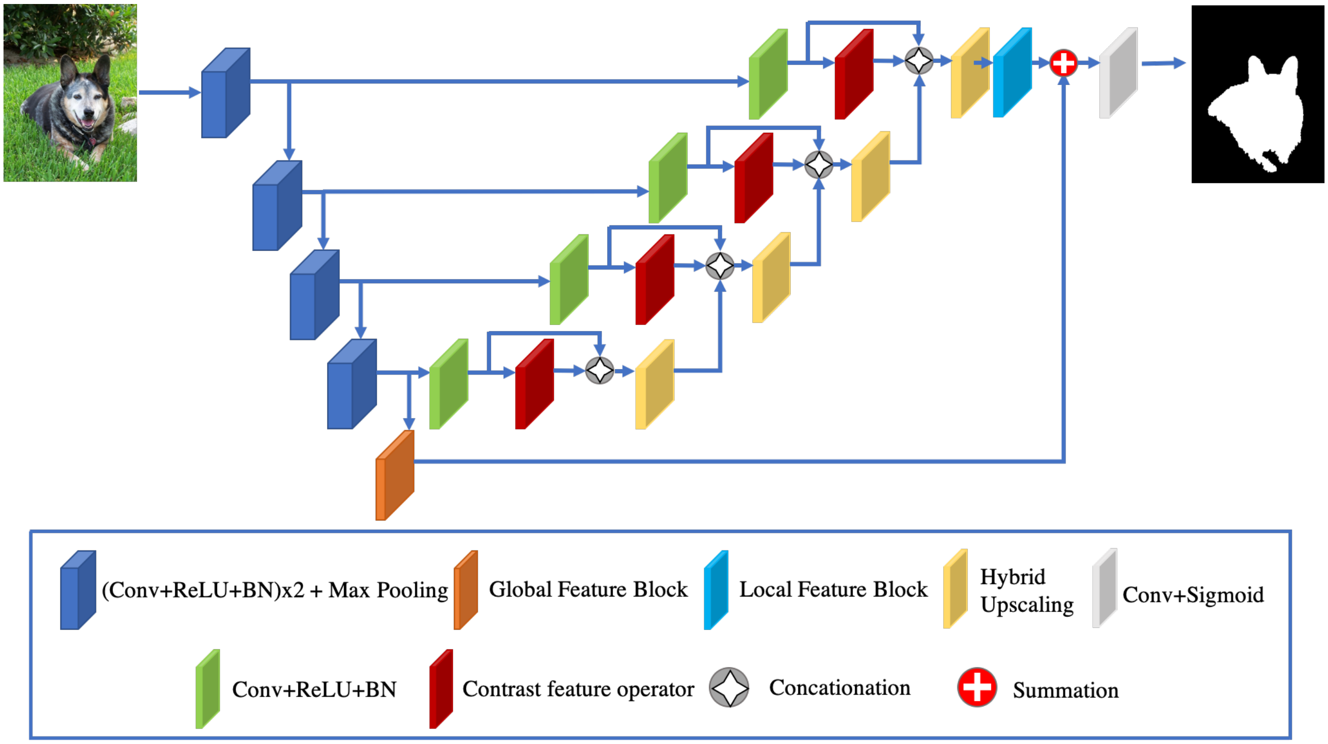

Figure 1: Architecture of the proposed U-NLCF network.

and does not require any training, deep learning based approaches have outperformed them both in terms of computational time and the quality of output saliency map.

Recently proposed method by [30] combined deep neural network for local estimation (DNN-L) with a global search network (DNN-G) to learn local patch features and predict saliency score for saliency detection. Another approach by [17] trained multiple CNNs with three fully connected layers (FCL) to learn deep features at multiple image scales. Though this complex architecture allowed to capture the saliency map at various spatial positions, contrast level, and scales, utilizing FCL in the network is computationally expensive and drops out spatial information of the input image. In order to deal with low-contrast background and integrate global and local contexts to predict saliency region, [36] proposed unified multi-context deep CNN framework. The use of global context along with local one helped to estimate saliency of the full image and its feature rich areas showing the outstanding performance.

The most recent works such as [35, 23, 21] used pretrained classification networks (e.g. VGG-16 [29], ResNet [8]) along with multi-scale features and achieved exceptional results. More precisely, [35] designed R-Dropout for convolutional layers to learn probabilistic features and developed new upscaling operator to avoid checkerboard artifacts in a predicted saliency map. On the other hand, [23] tries to learn global, local, and contrast features utilizing only CNN layers. Their proposed NLDF [23] grid network was trained using IoU loss to preserve boundaries of salient objects in the scene.

3. Proposed Method

Inspired by UCF [35] and NLDF [23], our proposed U-NLCF network integrates key features of both models into a simple U-NET architecture. Global and local feature learning techniques of NLDF along with modified hybrid upsampling operator of UCF allows our U-NCF to predict a high-quality saliency map without checkerboard artifacts.

3.1. Network Architecture

In this section we explain our proposed model architecture in details. Since, simple U-NET model have proven its effectiveness for segmentation task, we implement our network based on the U-NET architecture. However, in saliency detection problem, the primary goal of the network is to learn discriminant saliency features from the dataset. Those features should represent local and global context of an image at different scales. Therefore, U-NET model is tailored for saliency detection task by adding blocks to learn global and local feature maps. As shown on Figure 1, four convolution blocks (dark blue) are connected to the Global feature block. Convolution blocks (CB) include convolution operator with kernel size 3 × 3, ReLU, batch normalization, and max pooling with stride 2. After every max pooling number of filters doubles starting from the first CB (e.g. 64, 128, 256, 512). At global feature block, we perform three convolution operations to estimate final global feature map. While encoder part learns global features using CBs, the decoder part containes skip connected convolution layers (green block) and contrast feature blocks (red block) to catch local context. As an upsampling mechanism, we utilize hybrid upscaling blocks (yellow block) with kernel size size 4 × 4. The contrast feature blocks are designed to capture the difference between the feature and its surrounded background. At the final stage, both local (I_{Local}) and global feature maps (I_{Global}) are summed up and goes through convolution layer with sigmoid function.

3.1.1 Non-Local Features

Contrast features Salient objects in the scene usually have the quality of being distinct compared to its surrounding background, while being uniform inside. The contrast feature was formulated to various forms [20, 2]. However, we follow [23] approach, where contrast feature can be learned by the network. Therefore, according to [23], such kind of saliency features can be captured by subtracting average value from the feature map. In our network, we computed contrast feature map I_i^c for local feature map I_i at each i^{th} scale by:

$$I_i^c = I_i - AvgPool(I_i) \tag{1}$$

The average of a local feature is estimated by average pooling with stride 2.

Local features The local features are learned at different scales to create a multi-scale local feature map. In order to obtain final local feature map I_{Local}, we implement convolution with kernel size 1 × 1 to the concatinated input features:

$$I_{Local} = Conv(I_1, I_1^c, H_2) \tag{2}$$

Eventually, we have 512 feature channels with size of 256 × 256 that represents local feature map. However, in order to combine them with the global feature map, one needs to decrease the number of filters to 2 by adding convolutional layer without activation function. Thus, we have an output with 256 × 256 × 2 shape.

Global features Besides local features, in order to efficiently detect salient objects, one needs to capture global features of the scene too. Thus, we design an additional block of convolutional layers with kernel size 5 × 5 and stride 2 to learn the global context of the image. At the output, we have a 1 × 1 × 128 global feature maps, which are then convolved with 2 filters (kernel size=1×1) to obtain a final global feature map I_{Global} with shape: 1 × 1 × 2.

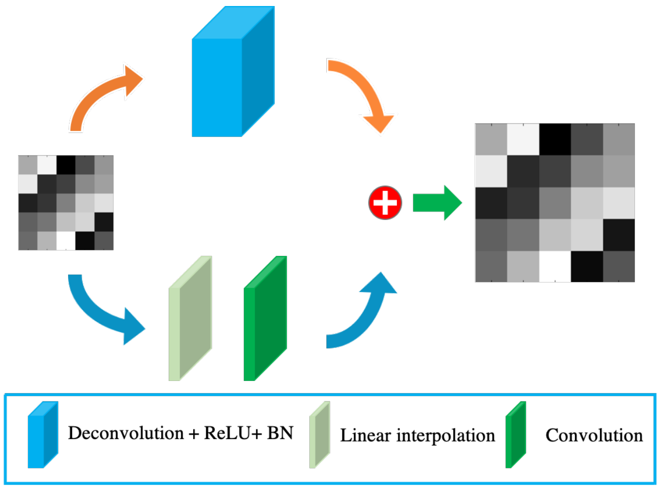

Figure 2: The hybrid upsampling.

3.1.2 Hybrid upsampling

The conventional upsampling mechanism in deep learning based methods is a simple convolution on a zero inserted input or so called deconvolution operator. Although using deconvolution filters increases representation power than utilizing fixed operators (e.g. resizing by interpolation) in the network, zero insertion to the input may cause numerical artifacts such as checkerboard [35]. Therefore, following the hybrid upsampling technique of [35], we propose a new upsampling method. Firstly, we perform deconvolution of an input followed by ReLU and batch normalization. At the same time, we resized the input to the desired size by using linear interpolation followed by convolution with kernel size 1×1 (see Figure 2). Finally, we sum up resulted outputs and obtain upsampled version of the input. This technique is implemented at each upsampling stage of the network to reduce odd artifacts at every scale.

3.1.3 Saliency map layer

The final prediction for the saliency map is estimated by adding global (I_{Global}) and local features (I_{Local}) into a single map:

$$S = Sigmoid(Conv(I_{Global} + I_{Local})) \tag{3}$$

Convolution layer is used to decrease the number of filters from 2 to 1, while sigmoid function ensures that output is in $[0 - 1]$ range.

4. Experiments

4.1. Datasets

Our proposed U-NLCF is tested on two different public benchmark datasets: ECSSD [32] and DUT-OMRON [33]. As a training dataset, we have used images from MSRA10K

[5], which are resized to 256 × 256 and augmented (mirror reflection, rotation) to produce 80,000 training images.

DUT-OMRON: The dataset includes 5,168 high quality images. Since images have complex background and many salient objects in a single scene, this testset is considered challenging for saliency detection algorithms.

ECSSD: This testset of 1,000 images is distinguished by its complex structure and meaningful semantic objects. The ground truth segmentations were performed by five subjects.

4.2. Performance Metrics

We adopt two evaluation metrics to measure the performance of our U-NLCF model. The first one is the F-measure score. It is a balanced mean of both average precision and average recall:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{4}$$

where β^2 is set to 0.3 as suggested by [2]. Precision and recall values for (4) are obtained by binarizing the output saliency map with adaptive threshold. The threshold value is image dependent and for output saliency map (S) of an input image, it is equal to:

$$\lambda = 2 \times mean(S) \tag{5}$$

The second performance metrics is a mean absolute error (MAE), which is calculated in the following way:

$$MAE = mean(|G - S|) \tag{6}$$

where G is binarized ground-truth mask of the input image.

4.3. Implementation Details

We have used Tensorflow framework [1] to implement our U-NLCF model. Since we did not use any pretrained model (e.g. VGG-16 [29] , ResNet [8]), all the parameters are initialized using *Xavier* [6] method and trained from scratch by minimizing cross-entropy loss function. U-NLCF model is trained for 75 epochs using Adam optimizer [14] with initial learning rate of 10^{-4}, which drops 10 times every 20 epochs. The batch size was set to 4 and training process took approximately 52 hours with NVIDIA Titan X (Pascal).

4.4. Comparison with State-of-the-arts

We compare our saliency network with several state-of-the-art methods: Multiscale Deep Features (MDF) [17], Deep Contrast Learning (DCL) [18], Cellular Automata (BSCA) [26], Uncertain Convolutional Features (UCF) [35], Learning Pixel-wise Contextual Attention (PiCANet) [21], DRFI [11], Deep Saliency (DS) [19], and Non-local Deep Features (NLDF) [23]. In Table 1, quantitative comparison results are presented. We observe that our model (U-NLCF) consistently performs better than most of the methods in terms of MAE metrics for ECSSD dataset. Although our U-NLCF could not outperform PiCANet [21], it shows considerable performance gain over conventional methods (BSCA [26] and DRFI [11]) and some deep learning methods (DCL [18] and DS [19]), which indicates the competitiveness of the proposed method. However, proposed method shows one of the worst results among deep learning methods on challenging DUT-OMRON dataset.

We also made a computation time comparison with four state-of-the-art deep learning methods, namely MDF [17], DCL [18], PiCANet [21], and NLDF [23]. As shown in Table 2, our proposed model is 2 to 84 times faster than all methods except NLDF. U-NLCF attains slightly faster computational speed compared to NLDF, which indicates that real-time salient object detection is possible with our method.

Table 1: The F-measure and MAE results of different saliency detection methods on two benchmark datasets.

Method	ECSSD		DUT-OMRON	
	F_β	MAE	F_β	MAE
U-NLCF	0.804	0.102	0.591	0.129
UCF [35]	0.852	0.069	0.628	0.120
MDF [17]	0.807	0.105	0.644	0.092
DCL [18]	0.829	0.150	0.684	0.157
PiCANet [21]	-	**0.035**	-	**0.068**
DS [19]	0.826	0.122	0.603	0.120
NLDF [23]	-	0.063	-	0.080
BSCA [26]	0.705	0.182	0.509	0.190
DRFI [11]	0.733	0.164	0.550	0.138

Table 2: Inference time of state-of-the-art deep learning based methods

	MDF	DCL	PiCANet	NLDF	U-NLCF
sec/img	8.00	1.50	0.18	0.08	**0.07**

In terms of visual comparison, our methods shows high-quality results (see Figure 3). Saliency map predicted by our method reliably highlights the most salient objects in many challenging scenes like salient objects near the image boundary or disconnected from each other (separate salient regions). However, we still see some minor checkerboard artifacts inside the salient object, which indicates that hybrid upsampling mechanism should be further developed.

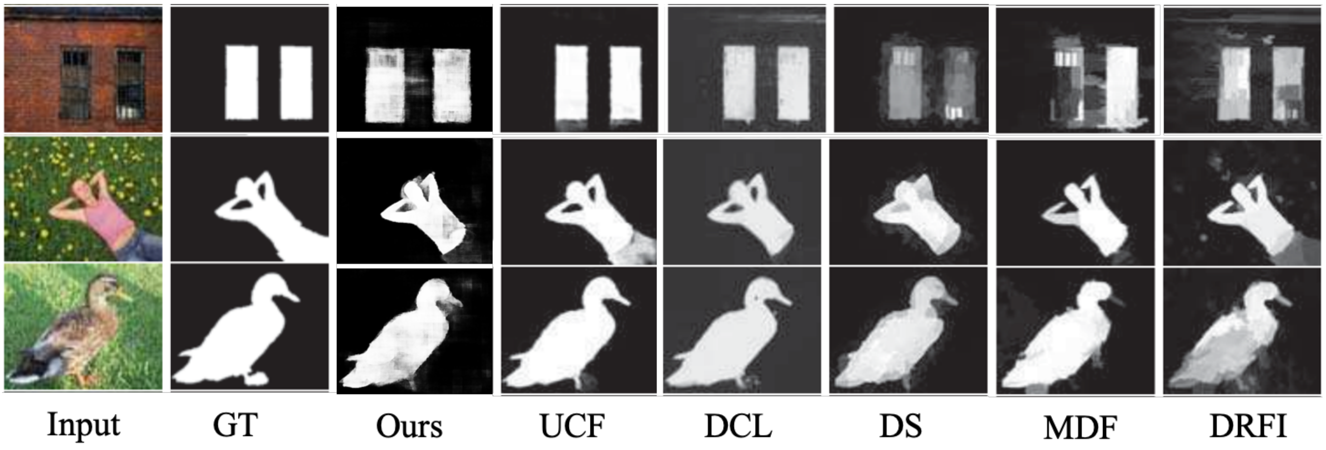

Figure 3: Comparison of saliency maps on the ECSSD dataset.

5. Conclusion

In this paper, we propose a U-NET based model for saliency detection problem. Our network integrates global and local features to efficiently detect salient regions in the scene. Moreover, a new upsampling mechanism is implemented to reduce the checkerboard artifacts and to better preserve boundaries of the salient object. Qualitative and quantitave comparison with other state-of-the-art methods shows that U-NLCF is faster than existing deep learning methods and provides comparably better results than most of the methods on ECSSD dataset.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016. 4

[2] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*, pages 1597–1604. IEEE, 2009. 3, 4

[3] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 73–80. IEEE, 2010. 1

[4] Z.-h. Chen, Y. Liu, B. Sheng, J.-n. Liang, J. Zhang, and Y.-b. Yuan. Image saliency detection using gabor texture cues. *Multimedia Tools and Applications*, 75(24):16943–16958, 2016. 1

[5] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015. 4

[6] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010. 4

[7] C. Guo and L. Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans. Image Processing*, 19(1):185–198, 2010. 1

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 4

[9] S. Hong, T. You, S. Kwak, and B. Han. Online tracking by learning discriminative saliency map with convolutional neural network. In *International Conference on Machine Learning*, pages 597–606, 2015. 1

[10] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 1

[11] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2083–2090, 2013. 4

[12] P. Jiang, H. Ling, J. Yu, and J. Peng. Salient region detection by ufo: Uniqueness, focusness and objectness. In *Proceedings of the IEEE international conference on computer vision*, pages 1976–1983, 2013. 1

[13] J.-S. Kim, J.-Y. Sim, and C.-S. Kim. Multiscale saliency detection using random walk with restart. *IEEE transactions on circuits and systems for video technology*, 24(2):198–210, 2014. 1

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[15] D. A. Klein and S. Frintrop. Center-surround divergence of feature statistics for salient object detection. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2214–2219. IEEE, 2011. 1

[16] C. Li, Y. Yuan, W. Cai, Y. Xia, and D. Dagan Feng. Robust saliency detection via regularized random walks ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2710–2717, 2015. 1

[17] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5455–5463, 2015. 1, 2, 4

[18] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 478–487, 2016. 4

[19] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang. Deepsaliency: Multi-task deep neural network model for salient object detection. *IEEE Transactions on Image Processing*, 25(8):3919–3930, 2016. 4

[20] F. Liu and M. Gleicher. Region enhanced scale-invariant saliency detection. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 1477–1480. IEEE, 2006. 3

[21] N. Liu, J. Han, and M.-H. Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, 2018. 1, 2, 4

[22] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu. Predicting eye fixations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 362–370, 2015. 1

[23] Z. Luo, A. K. Mishra, A. Achkar, J. A. Eichel, S. Li, and P.-M. Jodoin. Non-local deep features for salient object detection. In *CVPR*, volume 2, page 7, 2017. 1, 2, 3, 4

[24] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2232–2239. IEEE, 2009. 1

[25] P. Mehrani and O. Veksler. Saliency segmentation based on learning and graph cut refinement. In *BMVC*, pages 1–12, 2010. 1

[26] Y. Qin, H. Lu, Y. Xu, and H. Wang. Saliency detection via cellular automata. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 110–119, 2015. 4

[27] P. L. Rosin. A simple method for detecting salient regions. *Pattern Recognition*, 42(11):2363–2371, 2009. 1

[28] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE transactions on pattern analysis and machine intelligence*, 29(2):300–312, 2007. 1

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 4

[30] L. Wang, H. Lu, X. Ruan, and M.-H. Yang. Deep networks for saliency detection via local estimation and global search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3183–3192, 2015. 1, 2

[31] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu. A stagewise refinement model for detecting salient objects in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4019–4028, 2017. 1

[32] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1162, 2013. 3

[33] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013. 3

[34] H. Zha, J. Feng, G. Zeng, J. Wang, P. Wang, and S. Li. Salient object detection for searched web images via global saliency. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3194–3201. IEEE, 2012. 1

[35] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin. Learning uncertain convolutional features for accurate saliency detection. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 212–221. IEEE, 2017. 1, 2, 3, 4

[36] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1265–1274, 2015. 1, 2