Efficient Salient Object Detection Model with Dilated Convolutional Networks

Fei GUO†, Yuan YANG†, Yong GAO†, Nonmembers, and Ningmei YU†, Member

SUMMARY Introduction of Fully Convolutional Networks (FCNs) has made record progress in salient object detection models. However, in order to retain the input resolutions, deconvolutional networks with unpooling are applied on top of FCNs. This will cause the increase of the computation and network model size in segmentation task. In addition, most deep learning based methods always discard effective saliency prior knowledge completely, which are shown effective. Therefore, an efficient salient object detection method based on deep learning is proposed in our work. In this model, dilated convolutions are exploited in the networks to produce the output with high resolution without pooling and adding deconvolutional networks. In this way, the parameters and depth of the network are decreased sharply compared with the traditional FCNs. Furthermore, manifold ranking model is explored for the saliency refinement to keep the spatial consistency and contour preserving. Experimental results verify that performance of our method is superior with other state-of-art methods. Meanwhile, the proposed model occupies the less model size and fastest processing speed, which is more suitable for the wearable processing systems.

key words: saliency detection model, deep learning, dilated convolution

1. Introduction

Saliency detection models focus on identifying and segmenting the most conspicuous objects or regions in images. It is utilized as an important preprocessing step in many vision tasks including image segmentation [1], object recognition [2], [3], visual tracking [4], image retrieval [5], [6]. Although many effective methods are developed, the saliency detection in complex scenes is still a challenging task. Conventional saliency detection models [7]–[10] are mainly based on heuristic saliency priors. According to the visual attention study [11], [12], contrast priors are the main roles in saliency detection. Hence, global or local contrasts are exploited as the hand-craft features to characterize and formulate the saliency detection [8], [13], [14]. Besides, boundary prior is utilized to detect foreground objects [9], [15], [16]. Though these saliency priors are demonstrated effective in many works, the saliency detection in cluttered scenes is a still challenging task.

Recently, deep convolutional neural networks (CNNs) have made much progress in many vision applications, such as image classification [17], [18], object detection [19], [20], object tracking [21], [22], semantic segmentation [23]–[26]. Current methods [27]–[30] suggest that deep CNNs have promotions in saliency detection, especially in handling complex scenes. Though improvements on substantial performance have been made, major issues are still existed in deep CNNs based methods. Firstly, most deep CNNs methods discard completely the saliency priors, which are demonstrated effective in many works. The integration of the saliency priors into CNNs will have benefits in the saliency detection. Secondly, CNNs have greatly reduced resolutions in their final layer activations due to pooling or strided convolution operations throughout the network [31]. In order to atone for the resolution loss, segmentation networks either introduces additional connections to make use of localization power of low-middle layer, or adds a deconvolutional network on top of the CNNs with unspooling layers. Both approaches results into an even more increase in the number of parameters used in segmentation network.

For this purpose, we propose an efficient saliency detection method based on deep learning. We firstly utilize a deep learning network to capture the semantic features and segment the salient objects in a data-driven manner. In this network, dilated convolutions are exploited to increase the output resolution and retain the receptive field. In this way, we can decrease the parameters sharply compared with the fully convolutional networks (FCNs) methods without adding the deconvolutional networks for the salient object detection. Furthermore, the detection map predicted by network is encoded to refine the saliency detection. The manifold ranking model is explored to rank the encoded detection map, and get the final saliency with the spatial consistency and contour preserving.

2. Related Work

In current, the main saliency detection model are included the hand-crafted models and deep learning methods. For the hand-crafted models, the first saliency computational framework is proposed by Itti et al. [32], in which the framework is generated on the basis of center-surround mechanisms. After that, a framework based on isophote is proposed by valenti et al. [33]. The framework linearly combines the saliency maps estimated by isocenters, curvedness, and color boosting to compute the final saliency map. Next, a frequency-tuned method is explored by Achanta et al. [14]. This method measures the pixel-wise color difference to estimate the saliency maps with full resolution. Klein et al. [34] proposed a saliency maps computational
model based on center-surround contrasts, which leverages the Kullback-Leibler Divergence of image features to compute the contrasts in an information-theoretic way. In another study [35], the center-surround contrast is explored by regarding it as a cost sensitive max-margin classification problem. Despite of the center-surround contrast, contrast of the global region is also explored in many works. Cheng et al. [36] measure the global contrast between the object region and background regions in an image to compute the saliency. Pirazzi et al. [8] define the Euclidean distance of colors and leverage Gaussian blurring kernel to efficiently compute the global contrast. Cheng et al. [37] propose a soft abstraction that generating the homogeneous regions by exploiting histogram quantization and a global Gaussian Mixture Model. Later in [38], a multi-scale contrast of local region based method is studied. In this method, saliency values are computed across multiple segmentations, and combined to get the saliency maps with pixel-wise. Besides, more and more saliency prior knowledge is incorporated in recent saliency detection models. Background priors are exploited in many salient object detection methods [9], [15], [39], [40]. Yang et al. [9] propose a two-stage saliency detection model that leveraging the manifold ranking algorithms on the graph with undirected weights. Li et al. [39] propose to formulate the saliency detection as the dense and sparse reconstruction errors with respect to the pseudo background. In [40], saliency computation is formulated via absorbing Markov Chain. Wei et al. [15] study the geodesic distance of the background regions to refine the initial saliency prediction.

Recently, deep learning based saliency detection models have shown superior performance than hand-crafted methods in handling complex scenes. He et al. [41] proposed a superpixel-wise CNN architecture based on hierarchical contrast features. In order to build more advanced features, two contrast sequences are fed into the networks for each scale of superpixels. After that, multi-scale saliency maps are combined to produce the more accurate results. Li et al. [27] present a CNN architecture with fully connected layers that responsible for the extraction of different-scale features, and then refine the saliency map to yield more accurate results. Wang et al. [30] propose a saliency detection model with local estimation and global search. For the local estimation, features of local patch are learned by a local deep neural network to compute the local saliency values. For the global search, global features are formed by integrating the estimated local saliency maps, global contrasts and geometric information. Another global network is explored to generate the global salient object region with the global features. Finally, the weighted sum of salient regions generates the final saliency map. A unified deep learning framework for the saliency detection is proposed in [42]. In the framework, VGG-net is utilized to extract the high level features, and the low level features are used to form a distance map. Li et al. [28] present a deep saliency model with multi-tasks based on FCN. In this model, the saliency prior knowledge is encoded, and a multi-task learning scheme is set up to perform the tasks of saliency detection and semantic segmentation. Hou et al. [43] present a novel saliency detection model based on the holistically nested edge detector (HED) architecture. In this model, short connections are introduced into the skip-layer structures of the HED. This model can provide more representations at each layers through leveraging features of multi-level and multi-scale extracted from FCNs. Wang et al. [44] propose a novel saliency detection method. In this model, the recurrent fully convolutional network (RFCNs) is introduced. More accurate inference is predicted by incorporating saliency priors and iteratively learning to refine the saliency maps in the recurrent architecture. However, the recurrent networks structure leads to a computational overhead and the final results depend heavily on the initial estimations. Li et al. [45] propose an an deep contrast network to detect the salient object. The network consists of two complementary components to generate the saliency map, and a fully CRF model is incorporated in to refine the results. Liu et al. [46] propose an deep hierarchical saliency network, the network first learns various global structured saliency cues to make a coarse prediction. Then, a novel hierarchical recurrent CNN is adopted to refine the saliency. Luo et al. [47] propose a simplified convolutional neural network which combines local and global information through a multi-resolution 4 × 5 grid structure, and then refine the saliency map using a loss function inspired by the Mumford-Shah functional which penalizes errors on the boundary.

3. The Proposed Framework

As demonstrated in Fig.1, our saliency detection model includes two parts: Dilated-FCN (DFCN) and graphical model for saliency refinement. For the DFCN, the task of salient object detection is performed instead of the traditional FCNs network. For the traditional FCNs, to compensate for the resolution loss due to pooling or strided convolution operators in CNN, it introduces the additional connections to leverage the location power of low-middle layers or add deconvolutional network on top of FCN with unpooling layers. These solutions result in an even more increase in the model size. Thus, we propose the DFCN network to overcome the increasing model size. In DFCN, we simply remove the pooling layers to increase the resolutions of feature maps instead of adding deconvolutional network in traditional FCNs. However, removing subsampling will lead to the reduction of receptive field. Thus, dilated convolutions are introduced in the higher layers to enlarge the corresponding receptive field. Based on the saliency maps of the DFCN, we further explore the manifold ranking model to refine the saliency map. For an input image, the DFCN network is exploited to predict the salient foreground map. Meanwhile, the superpixels on the predicted map are treated as the seed nodes, and the saliency map can be refined by the manifold ranking model.
3.1 The Saliency Detection Network

The construction of DFCN is based on the architecture of dilated Residual network (DRN) presented by Yu et al. [48]. We replace the 1000-way linear classification layer of DRN with a 1 × 1 kernel linear convolutional layer and two output channels to construct the DFCN into a dense saliency map detection network. Moreover, dilation convolutions are utilized to the 4th and 5th groups of ResNet to generate high resolution output. The original ResNet is divided into five groups of convolutional layers, the resolution of final feature maps is 1/32 of the input image [49]. These five layers are called as downsampling layers in this work. The output in the original model is described as

\[
(g_l^i \ast f_l^i)(p) = \sum_{a+b=p} g_l^i(a) \ast f_l^i(b)
\]  

where \(g_l^i\) is the \(i\)th layer in group \(l\), for \(l = 1, \ldots, 5\). The \(f_l^i\) denotes the filter associated with layer \(g_l^i\). The domain of \(p\) is the feature map in \(g_l^i\).

In each group of the original Resnet, feature maps share the same resolution in different layers. To make the final saliency map denser, two final groups of convolutional layers (\(g_4^i\) and \(g_5^i\)) are explored. As shown in Fig. 2, subsampling in the last two downsampling layers are skipped by setting the stride to 1, and dilation rates of subsequent layers are increased using the dilation algorithm to increase the receptive fields as [50]. The first step in the conversation is to remove the striding in \(g_4^i\) and \(g_5^i\). This will cause that the output resolution of \(g_4^i\) doubled without reducing the receptive field. However, this will reduce the receptive fields of the subsequent layers by a factor of 2 in each dimension. Hence, we utilize 2-dilated convolutions to replace the convolution operators in those layers:

\[
(g_l^i \ast 2 f_l^i)(p) = \sum_{a+2b=p} g_l^i(a) f_l^i(b)
\]  

where \(i > 2\). The same transformation is applied to \(g_1^5\):

\[
(g_1^5 \ast 2 f_1^5)(p) = \sum_{a+2b=p} g_1^5(a) f_1^5(b)
\]  

Two striding layers have been eliminated after the subsequent layers of \(g_1^5\). This elimination will result in the reduction of receptive fields by a factor of 4 in each dimension. In order to compensate for the loss, the convolution operators in those layers are replaced by 4-dilated convolutions:

\[
(g_l^i \ast 4 f_l^i)(p) = \sum_{a+4b=p} g_l^i(a) f_l^i(b)
\]
Fig. 3 The detection results of the network

where $i > 2$.

The output resolution in DFCN is increased to the $1/8$ resolution of the input compared with the $1/32$ resolution in Resnet. Moreover, the DFCN share the same model size with the Resnet. After that, bilinear interpolation is adopted to upsample the feature map to full resolution, the detection results are shown in Fig. 3.

3.2 Manifold Ranking Model

The saliency maps generated from DFCN are usually with common defects in spatial consistency and contour preserving. In order to keep the spatial consistency and preserve the contour for the saliency maps, a manifold ranking method is studied. The manifold ranking theory is referred in [9]. Given an image, it is segmented as superpixels $\{X_i\}_{i=1}^N$ by SLIC [51]. Next, some nodes are set as queries, the remaining are ranked in terms of their relevances with the queries. We set up a ranking function $f$ to measure the relevances. This function computes corresponding ranking score $f_i$ for each $x_i$ and can be denoted as a vector $f = [f_1, \ldots, f_n]^T$.

Let $y = [y_1, \ldots, y_n]$ be an indicator vector, in which if the node $x_i$ is set as query, $y_i = 1$, if not $y_i = 0$. After that, graph model $G(V,E)$ based on superpixels is constructed. In the graph model, $V$ are the superpixel nodes, $E$ are the edges of adjacent nodes with the weight $w_{ij}$. Given the graph, the degree matrix is defined as $D = \text{diag}(d_1, \ldots, d_m)$, in which $d = \sum_j w_{ij}$. Moreover, the edge weight $w$ is defined as

$$w_{ij} = e^{-\frac{||c_i - c_j||^2}{\sigma^2}}$$

where $c_i$ and $c_j$ are the mean value of two super-pixels.

We can compute the optimal query ranking score by optimizing the following problem:

$$f_{ran} = \text{argmin} \frac{1}{2} \left( \sum_{ij} w_{ij} \left\| \frac{f_i}{\sqrt{d_{ii}}} - \frac{f_j}{\sqrt{d_{jj}}} \right\|^2 + \mu \sum_{i=1}^n ||f_i - y_i||^2 \right)$$

where $\mu$ balances the smoothness and fitting constraints. The minimum optimization is performed by setting the derivative to 0, the final ranking function can be denoted as:

$$f_{ran} = (D - \alpha W)^{-1} y$$

where $\alpha = (1 + \mu)^{-1}$, $W$ is the unnormalized Laplacian matrix.

3.3 Saliency Map Refinement

As illustrated in Fig. 4, the saliency map is estimated in four stages for an image: object prediction by DFCN, image segmentation with superpixels, estimation of the query nodes and saliency map refinement by manifold ranking-based model propagation.

In DFCN, semantic structural information on object perception is adaptively captured by the trained network to predict salient object maps with pixel-wise. The predicted salient object maps are referred as Deepmap.

In the stage2, we segment the input image with superpixels. The superpixel segmentation is implemented by SLIC.

In the stage3, we segment the Deepmap with adaptive threshold. The pixels with high values in the binary map are set as foreground query nodes.

In stage4, the query nodes are fed into Eq. (7) to model the manifold ranking, generating the final refined saliency map $S_{ref}$, which is denoted as:
where $\bar{f}_{ran}$ is the normalized version of $f_{ran}$, and $i$ indexes the nodes on the graph.

4. Results and Discussions

4.1 Datasets

The proposed method is evaluated on three benchmark dataset: MSRA-B [52], PASCALS [53] and ESCCD [54]. The ESCCD dataset has 1000 selected images with complex scenes. The MSRA dataset covers various image contents and is a common dataset in saliency detection. Most of the images are in simple scenes with high contrast. The PASCAL-S dataset consists of 850 images with multiple complex objects and cluttered background. It is more challenging for the saliency detection task.

4.2 Implementation Details

The DFCN is implemented on the basis of tensor-flow framework. All evaluations are performed using Amax server with 4 NVidia TESLA-K80 and Ubuntu 16.04 in Intel Xeon E5 CPU with 128GB RAM. We training our network using 5K images from MSRA dataset. All images are resized to $320 \times 320$ pixels for training and testing. The momentum parameter is set as 0.99, the learning rate is chosen as $10^{-8}$, and the weight decay is 0.0005. The SGD learning procedure is accelerated using NVidia TESLA-K80 GPU. For the graphical model process, the super-pixel segment is carried out by the SLIC. The super-pixel number $N$ is set as 200. The refinement scheme is implemented in Matlab platform.

4.3 Evaluation Metrics

In order to evaluate the performance of the proposed models, the metrics such as Precision-recall (PR) curve, F-measure and mean absolute error (MAE) are adopted. The precision and recall metrics are computed by binarizing the saliency map, and comparing with ground truth. The PR curves represent the mean precision and recall at different thresholds. The F-measure is defined as

$$F\text{- measure} = \frac{(1 + \beta^2) \times \text{Precision} \times \text{recall}}{\beta^2 \times \text{Precision} + \text{recall}}$$

where $\beta^2 = 0.3$. In addition, MAE denotes the difference of mean absolute pixel, and measures the numerical distance, which is demonstrated by

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x, y) - G(x, y)|$$

Where $S$ is the saliency map, $G$ is the ground truth.

The performance of the proposed method is compared with 9 state-of-art methods, including DRFI [54], HS [10], PCA [55], RC [36], SVO [56], DSR [39], LEGS [57], DS [28], MDF [27], DCL [45], RFCN [44] and DSS [43]. Note that LEGS, DS, MDF, DCL, RFCN and DSS are methods based on deep learning.

4.4 Qualitative Performance Comparison

In order to demonstrate the qualitative performance, we provide the results of our approach and the other state-of-the-art methods in several sample images, which are shown in Fig. 5. These sample images have single objects, multiple objects, and small objects with complex background clutters. It shows that hand-crafted methods fail to detect the accurate salient objects from the cluttered backgrounds in almost scenes. For the deep learning based methods, the LEGS and DS methods fail to detect the multiple objects and small objects. The MDF method cannot detect the large object and multiple objects in low contrast scenes. Meanwhile, the DCL method fails to detect the salient object in low contrast scenes. These detected results are not with spatial consistency and contour preserving. The results detected by RFCN, DSS methods are the most precise and close to the ground-truth. It is obvious that our method can highlight the salient objects uniformly and keep the object contours in the background clutters. Owing to the fusion of boundary prior knowledge and deep feature maps in our method,
the detected saliency maps are comparable with the methods such as RFCN and DSS and outperform other methods in almost scenarios. However, our method has a poor performance of our algorithm in the small objects. Because the salient object tend to be the large and high contrast object in a scene. Meanwhile, the training dataset has less images with small objects, which lead that the deep learning network hard to learn the salient features.

4.5 Quantitative Performance Comparison

Figure 6 and Table 1 show the quantitative comparison results. The processing speed and the number of parameters are also demonstrated in Table 2. In Fig. 5, the performance of our proposed model is substantially superior to the other approaches with a considerable margin, and comparable with RFCN in terms of the PR curves and F-measure. Moreover, as shown in Table 1, our method achieves the considerable superior performance with a relatively high F-measure and low MAE in three challenging datasets.

In Table 2, the number of parameters in our model is the least with other deep learning models. This can effectively save the memory in hardware system. Besides, the processing speed is the fastest over other salient detection method. In the inference stage of our method, the inference time is only 0.02s for an image of 320*320, which is extremely faster than other method. Given the refinement stage, it also needs 0.04s. The overall processing time is 0.06s, which is an advantage over other methods. According to Table 1 and Table 2, the average F-measure of our method in three datasets accounts for 99.1% and 97.2% of RFCN and DSS. However, the model size of our method accounts for 10.7% and 14.4% of RFCN and DSS, respectively. These results demonstrate that our method has better overall performance with other deep learning based methods.
Table 1  Quantitative comparison of different saliency detection methods on F-measure and MAE (the best two results are denoted as bold and underline respectively)

| Method  | MSRAB     | ESCCD     | PASCALS  |
|---------|-----------|-----------|-----------|
|         | F-measure | MAE       | F-measure | MAE       | F-measure | MAE       |
| Ours    | 0.918     | 0.046     | 0.906     | 0.069     | 0.794     | 0.096     |
| DSS     | 0.927     | 0.038     |           |           |           |           |
| RFCN    | 0.946     | 0.022     | 0.898     | 0.097     | 0.798     | 0.092     |
| DCL     | 0.916     | 0.047     | 0.879     | 0.071     | 0.822     | 0.108     |
| MDF     | 0.885     | 0.104     | 0.765     | 0.105     | 0.656     | 0.146     |
| DS      | 0.863     | 0.063     | 0.810     | 0.160     | 0.818     | 0.170     |
| LEGS    | 0.870     | 0.081     | 0.785     | 0.118     | 0.695     | 0.157     |
| DRFI    | 0.838     | 0.118     | 0.733     | 0.166     | 0.616     | 0.207     |
| DSR     | 0.824     | 0.121     | 0.717     | 0.173     | 0.551     | 0.215     |
| PCA     | 0.782     | 0.185     | 0.627     | 0.248     | 0.530     | 0.249     |
| RC      | 0.820     | 0.137     | 0.701     | 0.187     | 0.640     | 0.225     |
| SVO     | 0.585     | 0.331     | 0.357     | 0.404     | 0.266     | 0.318     |

Table 2  Complexity of the deep learning based models and the runtime of different methods (The best results are denoted as bold)

| Methods | DRFI | HS | PCA | RC | SVO | DSR | LEGS | MDF | DS | RFCN | DCL | DSS | Ours |
|---------|------|----|-----|----|-----|-----|------|-----|----|------|-----|-----|------|
| Speed (FPS) | 0.7  | 0.53 | 4.34 | 6.49 | 0.01 | 10.2 | 0.52 | 0.04 | 8.33 | 0.65 | 2.18 | 12.5 | 16.67 |
| Parameters (M) | 73.6 | 116 | 128 | 186 | 64  | 138 | 20   |

Fig. 7  Performance comparison of our methods with and without refinement on PASCALS dataset: (a) the precision and recall curves, (b) the F-measure curves under different thresholds.

4.6 Effectiveness of Refinement Scheme

We quantitatively compare our approach with and without graph-based manifold ranking refinement on PASCALS benchmark dataset. Figure 7 shows that our approach with refinement will lead to a better saliency detection performance in most cases compared with the one without refinement. This can be attributed to the fact that graph-based manifold ranking refinement can capture more topological information among super-pixels that is helpful to the boundary preservation of salient objects.

5. Conclusions

In this work, we propose an efficient and effective saliency detection method. For this method, a dilated fully convolutional network is introduced to predict the saliency map, which preserves the respective field without adding extra networks. Meanwhile, in order to yield more accurate saliency results, a refinement scheme based on manifold ranking is proposed, which aims to propagate the predicted salient information for further improvement of saliency performance. Extensive evaluations in three widely database verify that our proposed model has superior performance of saliency detection compared with state-of-art methods. Moreover, this method has much less parameters with other FCNs methods, which is easy to deploy to the embedded processing systems.

Acknowledgments

This work is supported in part by projects of the Natural Science Foundation of China (Grant no.51477138), the Key Research and Development Program of Shaanxi Province (Grant no.2017ZDXM-GY-130), Xi’an City Science and Technology Project (Grant no. 2017080CG/RC043XALG009), and Industrial research project of Science and Technology Department of Shaanxi Province (Grant no. 2017GY-083).
References

[1] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, “Saliency driven total variation segmentation,” Int. Conf. Comput. Vis., pp. 817–824, 2009.

[2] U. Rutishauser, D. Walther, C. Koch, and P. Perona, “Is bottom up attention useful for object recognition?” IEEE Conf. Comput. Vis. Pattern Recog., pp. II–II, 2004.

[3] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, “Stc: A simple to complex framework for weakly-supervised semantic segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2314–2320, 2017.

[4] A. Borji, S. Frintrop, D.N. Sihite, and L. Itti, “Adaptive object tracking by learning background context,” IEEE Conf. Comput. Vis. Pattern Recog. Worksh., 2012.

[5] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, “3-D object retrieval and recognition with hypergraph analysis,” IEEE Trans. Image Process., vol. 21, no. 9, pp. 4290–4303, 2012.

[6] M.M. Cheng, Q.-B. Hou, S.-H. Zhang, and P.L. Rosin, “Intelligent visual media processing: When graphics meets vision,” J. Comput. Sci. Tech., vol. 32, pp. 110–121, 2017.

[7] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, “Automatic salient object segmentation based on context and shape prior,” Proc. British Machine Vision Conference, pp. 110.1–110.12, 2011.

[8] F. Perazzi, P. Krähenbühl, Y. Priich, and A. Hornung, “Saliency filters: Contrast based filtering for salient region detection,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 733–740, 2012.

[9] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, “Saliency detection via graph-based manifold ranking,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3166–3173, 2013.

[10] Q. Yan, L. Xu, J. Shi, and J. Jia, “Hierarchical saliency detection,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1155–1162, 2013.

[11] L. Itti and C. Koch, “Computational modeling of visual attention,” Nature reviews neuroscience, vol. 2, no. 3, pp. 194–203, 2001.

[12] D. Parkhurst, K. Law, and E. Niebur, “Modeling the role of salience in the allocation of overt visual attention,” Vision research, vol. 42, no. 1, pp. 107–123, 2002.

[13] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 11, pp. 1254–1259, 1998.

[14] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency tuned salient region detection,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1597–1604, 2009.

[15] Y. Wei, F. Wen, W. Zhu, and J. Sun, “Geodesic saliency using background priors,” Proc. European Conference on Computer Vision, pp. 29–42, 2012.

[16] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, “Real-time salient object detection with a minimum spanning tree,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2334–2342, 2016.

[17] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in Neural Processing Systems, pp. 1097–1105, 2012.

[18] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587, 2014.

[20] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.

[21] L. Wang, W. Ouyang, X. Wang, and H. Lu, “Visual tracking with fully convolutional networks,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3119–3127, 2015.

[22] L.J. Wang, W. Ouyang, X. Wang, and H. Lu, “Stct: Sequentially training convolutional networks for visual tracking,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1373–1381, 2016.

[23] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440, 2015.

[24] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A.L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” arXiv preprint arXiv: 1412.7062, 2014.

[25] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” Proc. IEEE International Conference on Computer Vision, pp. 1520–1528, 2015.

[26] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P.H. Torr, “Conditional random fields as recurrent neural networks,” Proc. IEEE International Conference on Computer Vision, pp. 1529–1537, 2015.

[27] G. Li and Y. Yu, “Visual saliency based on multiscale deep features,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 5455–5463, 2015.

[28] X. Li, L. Zhao, H. Mei, M. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, “Deepsaliency: Multi-task deep neural network model for salient object detection,” IEEE Trans. Image Process., vol. 25, no. 8, pp. 3919–3930, 2016.

[29] G. Lee, Y.-W. Tai, and J. Kim, “Deep saliency with encoded low level distance map and high level features,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 660–668, 2016.

[30] A. Achanta, M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Snavely, “Saliency detection via local estimation and global search,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3183–3192, 2015.

[31] A. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[32] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 11, pp. 1254–1259, 1998.

[33] R. Valenti, N. Sebe, and T. Gevers, “Image saliency by isocentric curvatures and color,” ICCV, pp. 2185–2192, 2009.

[34] D.A. Klein and S. Frintrop, “Center-surround divergence of feature statistics for salient object detection,” ICCV, 2011.

[35] X. Li, Y. Li, C. Shen, A.R. Dick, and A. van den Hengel, “Contextual hypergraph modeling for salient object detection,” ICCV, pp. 3328–3335, 2015.

[36] M.M. Cheng, N.J. Mitra, X. Huang, P.H.S. Torr, and S.-M. Hu, “Global contrast based salient region detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 569–582, 2014.

[37] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, “Efficient salient region detection with soft image abstraction,” ICCV, pp. 1529–1536, 2013.

[38] Z. Jiang and L.S. Davis, “Submodular salient region detection,” CVPR, pp. 2043–2050, 2013.

[39] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, “Saliency detection via dense and sparse reconstruction,” ICCV, 2013.

[40] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, “Saliency detection via absorbing markov chain,” ICCV, 2013.

[41] S. He, R. Lau, W. Liu, Z. Huang, and Q. Yang, “Supercnn: A superpixel-wise convolutional neural network for salient object detection,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 330–344, 2015.

[42] E. Crook, “Efficient salient object detection,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3119–3127, 2015.
[44] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, “Salient object detection with recurrent fully convolutional networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol.41, no.7, pp.1734–1746, 2019.

[45] G. Li and Y. Yu, “Deep contrast learning for salient object detection,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.478–487, 2016.

[46] N. Liu and J.W. Han, “DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016.

[47] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, “Non-local Deep Features for Salient Object Detection,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017.

[48] F. Yu, V. Koltun, and T. Funkhouser, “Dilated Residual Networks,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2017.

[49] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” CVPR, pp.1, 2, 5, 2016.

[50] F. Yu and V. Koltun, “Multiscale context aggregation by dilated convolutions,” ICLR2016, arXiv: 1511.07122v3, 2016.

[51] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, “SLIC superpixels compared to state-of-the-art superpixel methods,” IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.11, pp.2274–2282, Nov. 2012.

[52] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, “Learning to detect a salient object,” IEEE Trans. Pattern Anal. Mach. Intell., vol.33, no.2, pp.353–367, 2011.

[53] Y. Li, X. Hou, C. Koch, J.M. Rehg, and A.L. Yuille, “The secrets of salient object segmentation,” IEEE Conf. Comput. Vis. Pattern Recog., pp.280–287, 2014.

[54] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, “Salient object detection: A discriminative regional feature integration approach,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.2083–2090, 2013.

[55] R. Margolin, A. Tal, and L. Zelnik-Manor, “What makes a patch distinct?” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1139–1146, 2013.

[56] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, “Fusing generic objectness and visual saliency for salient object detection,” International Conference on Computer Vision, IEEE Computer Society, 2011.

[57] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, “Deep networks for saliency detection via local estimation and global search,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.3183–3192, 2015.

Fei Guo received his B.S. degree and M.S. degree in electronics engineering from Xi'an University of Technology, Xi'an, China, in 2011 and 2014, respectively. Now he is a Ph.D. candidate in Xi'an University of Technology. His main research interests include image processing, computer vision and VLSI Design.

Yuan Yang received the B.S. degree in power electronics from Xi'an University of Technology, Xi'an, China, in 1997, and the M.S. and Ph.D. degrees in electronics engineering from Xi'an University of Technology, Xi'an, China, in 2000 and 2004, respectively. Since 2000, she has been a member of the faculty of School of Automation and Information Engineering, Xi'an University of Technology, Xi'an, China, where she is currently a professor. From March to August 2004, she was in the center of VLSI, Kyushu University, Fukuoka, Japan, as a visiting scholar. She then came back to Xi'an University of Technology, and engaged in teaching and researches in electronics engineering. She is the author of three books, more than 80 articles, and more than 10 inventions. Her main research interests include VLSI design, circuit and system design. Prof. Yang’s awards and honors include the first prize in science and technology of Shaanxi Province, the first prize in science and technology of Shaanxi colleges and Universities, the first prize in science and technology of Xi'an, excellent doctoral dissertations in Shaanxi.

Yong Gao received the B.S. degree in applied physics from Xi'an University of Technology in 1982. He received the M.S. and Ph.D. degrees in Microelectronics and solid-state electronics from Xi'an Jiaotong University, Xi'an, China, in 1988 and 1995, respectively. From 1996 to 2009, he has been a Professor with the Electronics Department, Xi'an University of Technology, China. Since 2010, he has been a professor in the department of electronics in Xi'an University of Technology and the department of electrical engineering at Xi'an Polytechnic University in China. His research mainly include power electronic devices, computer vision and very large scale integration design.

Ningmei Yu received the B.S. degree in electronics engineering from Xi'an University of Technology, Xi'an, China, in 1986. She received the M.S. and Ph.D. degrees from TOHOKU University, Sendai, Japan, in 1996 and 1999, respectively. Since 2001, she has been a member of the faculty of School of Automation and Information Engineering, Xi'an University of Technology, Xi'an, China, where she is currently a professor. Her main research mainly include wireless communication, image compression chip and digital-analog hybrid integrated circuit technology.