Journal of Visual Communication and Image Representation 54 (2018) 1–9

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Salient object detection via a local and global method based on deep residual network☆

Dandan Zhu[a], Ye Luo[a,*], Lei Dai[b], Xuan Shao[a], Qiangqiang Zhou[c], Laurent Itti[d], Jianwei Lu[a,e,*]

[a] School of Software Engineering, Tongji University, China
[b] School of Automotive and Traffic Engineering, Jiangsu University, China
[c] School of Electronics and Information, Tongji University, China
[d] Dept. of Computer Science and Neuroscience Program, University of Southern California, United States
[e] Institute of Translational Medicine, Tongji University, China

ARTICLE INFO

Keywords:
Salient object detection
Deep residual network
Local and global features

ABSTRACT

Salient object detection is a fundamental problem in both pattern recognition and image processing tasks. Previous salient object detection algorithms usually involve various features based on priors/assumptions about the properties of the objects. Inspired by the effectiveness of recently developed deep feature learning, we propose a novel Salient Object Detection via a Local and Global method based on Deep Residual Network model (SOD-LGDRN) for saliency computation. In particular, we train a deep residual network (ResNet-G) to measure the prominence of the salient object globally and extract multiple level local features via another deep residual network (ResNet-L) to capture the local property of the salient object. The final saliency map is obtained by combining the local-level and global-level saliency via Bayesian fusion. Quantitative and qualitative experiments on six benchmark datasets demonstrate that our SOD-LGDRN method outperforms eight state-of-the-art methods in the salient object detection.

1. Introduction

Saliency detection attempts to identify the most important and conspicuous object regions in an image by the human visual and cognitive system. It is a fundamental problem in neural science, psychology and computer vision. Many computer vision researchers propose computational models to simulate the process of human visual attention or identify salient objects. Recently, salient object detection had drawn a large amount of attention in variety of computer vision tasks, such as object detection [1], person re-identification [2], object retargeting [3], image retrieval [4,5], video summarization [6] and image deblurring [7], etc.

Visual saliency can be viewed into different perspectives and contrast is one of them. Based on the observation that salient object is always distinguishing itself from its surroundings, contrast as a prior has been widely used to detect salient object. According to the range of the context that the contrast is computed to, it can be further categorized into local contrast and global contrast methods. The local contrast based methods usually compute center-surround difference to obtain the object-of-interest region standing out from their surroundings [8]. Due to the lack of the global information, methods of this category tend to highlight the boundaries of salient objects and neglect the interior content of the object. Meanwhile, the global contrast based methods take the entire image into consideration to estimate the saliency of every pixel or every image segment, thus the whole salient object is detected but the details of the object structure are always missing. There are also some methods proposed to improve the performance of salient object detection via integrating both local and global-level cues [9]. The aforementioned methods may work well for low-level saliency, but they are neither sufficient nor necessary, especially in the cases when the saliency is also related to the human perception or is task-dependent.

In order to obtain more accurate semantic features for salient object detection, deep neural networks have recently been widely used. These methods include Multisacle Deep Feature (MDF) [10], Deep Image Saliency Computing via Progressive Representation Learning (DISC) [11], Local Estimation and Global Search (LEGS) [12], and Encoded Low level Distance map (ELD) [13]. They utilized high-level features from the deep convolution neural network (CNN) and demonstrate superior results over previous works that utilizes only low-level features. CNN based methods did show their superiority on deep feature extraction, but compared to the methods using deep residual network

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding authors at: School of Software Engineering, Tongji University, China (Y. Luo).
E-mail addresses: yeluo@tongji.edu.cn (Y. Luo), jwlu33@tongji.edu.cn (J. Lu).

https://doi.org/10.1016/j.jvcir.2018.03.017
Received 27 August 2017; Received in revised form 23 January 2018; Accepted 28 March 2018
Available online 24 April 2018

(ResNet), there are still some limitations. Specifically for MDF, the author treats each region as an independent unit in feature extraction without any shared computation. In addition, this method uses not only high-level features but also low-level features for complement and enhancement when doing salient detection. Due to the fact that CNNs fail to extract distinguishing features when the textures of salient object regions are similar to the background. Moreover, since CNN is generally performed on image patches, only features at local level are extracted, thus it fails to capture the global relationship of image regions and can't maintain the label consistency in a relative large region.

In order to solve these problems, in this paper, we propose a novel image salient object detection model named Salient Object Detection via a Local and Global Method Based on Deep Residual Network (i.e. SOD-LGDRN). Instead of CNN, we apply the deep residual network (ResNet) to salient object detection, from which more distinctive features of the salient objects can be obtained. Different from previous methods, high-level semantic features are extracted from both local and global levels via two ResNets, respectively. Features extracted from an entire image via a global ResNet (ResNet-G) roughly identify the global concepts of the salient object(e.g. the location, the scale and the size of the salient object). However, the residual network at the global level considers the entire image but pays less attention to the local context information, and this may misinterpret the background as salient regions or lack the subtle structure of a salient object. Hence, local features of the image are extracted via local ResNets (ResNet-L) as complements to the global information. By considering deep features at the local and the global levels simultaneously, we can expect to obtain a salient object with homogeneously highlighted region and accurate object boundary.

We briefly describe the implementation of our approach as shown in Fig. 1. At first, the entire image as input is sent to ResNet-G to extract global level features thus to generate a saliency map in a global context. Secondly, a multilevel image segmentation method is employed to segment the target image into multiple segments, and the segments at each level are warped and then fed to a ResNet-L to extract local features. The obtained local features at multiple level are further used to estimate the local saliency map. In order to deal with the noise caused by the image segmentation, we also introduce a spatial coherence refinement method to enhance the smoothness of the local saliency map. At last, the global and the local saliency maps are fused to obtain the final saliency map via a Bayesian integration method. In summary, our major contribution is threefold:

(1) A novel salient object detection model is proposed based on the global and the local semantic features extracted from two deep residual networks: ResNet-G and ResNet-L.
(2) Features learned from deep residual network are at the first time extracted and applied to salient object detection.
(3) We proposed the model only uses high-level features for saliency detection, and the performance can be significantly improved without using low-level features for complement or refinement.

The rest of the paper is organized as follows. In Section 2, we first review and evaluate the related work. Then we introduce our proposed SOD-LGDRN model in Section 3. In Section 4, we conduct experiments on six public datasets: MSRA-B, PASCAL-S and ECSSD, ASD, SED1 and THUR15K and then make comparisons with eight state-of-the-art methods. In the last Section, we present the conclusion and discussion of the future work.

2. Related work

In this section, we discuss the related work on salient object detection. In addition, we also briefly review deep neural network that are closely related to this work.

To estimate visual saliency, various methods are proposed and most of the saliency detection approaches can be generally categorized as global and local schemes. Local methods measure saliency by computing local contrast and rarity. Itti et al. [14] propose the center-surround scheme to extract low-level features such as color, intensity, orientation and texture, and use a linear and non-linear combination of multi-scale saliency map. Hereafter, Ma and Zhang [15] utilize color contrast in a local neighborhood as a measure of saliency. Goferman et al. [9] provide a context-aware (CA) method that used three principles including local low-level cues, global consideration and visual representation rules to highlight salient objects along with their contexts. However, these methods still have some problems to be solved, for example, they tend to be more sensitive to the boundaries of salient objects, but it can't highlight the object interior uniformly. On the other hand, global methods generally detect saliency by using holistic contrast and color statistics of the entire image. Achanta et al. [16] propose a frequency tuned method to calculate the image pixel saliency by subtracting the average color of the image. Cheng et al. [17] compute image saliency on the basis of color histogram contrast and region contrast. Xu et al. [18] utilize histograms based global contrast and spatial coherence to detect saliency. Most of the global contrast methods depend on color uniqueness when doing global statistics. However, these methods can be insufficient to capture semantic features in a natural image. Besides, they usually ignore the spatial relation (i.e. spatial weight and distance) between different parts in the images. Liu et al. [19] propose a set of features from both global and local perspectives, which are integrated by conditional random field (CRF) to generate final saliency map. Fareed et al. [20] proposed a salient region detection algorithm based on sparse representation and graph ranking, which combined the Gaussian and Bayesian procedures to produce smooth and precise saliency map. While the combination of global and

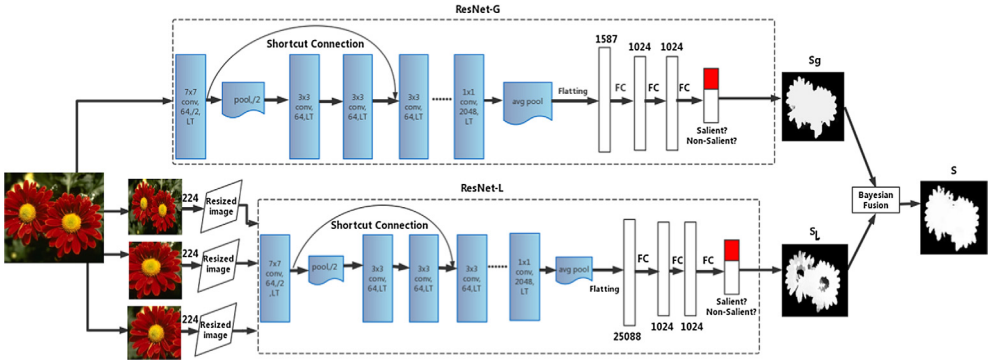

Fig. 1. Architecture overview of our proposed deep salient object detection model. The ResNet-G takes the entire image as input and generates global saliency map S_g. The ResNet-L takes image segments at multiple levels as input and produces the local-level saliency map S_l. The aggregated saliency map S is obtained by fusing S_g and S_l via a Bayesian integration method. LT denotes to the linear transformation operation.

local models [21,22] is technically sound, these approaches have two major problems. First, these methods mainly rely on handcrafted low-level features which may fail to describe image with clutter background and object structures. Second, these methods can detect salient objects, but it can't highlight semantic features.

Deep neural networks have recently achieved many successes in computer vision tasks, including image classification [23,24], object detection [25–27] and semantic segmentation [28,29], etc. The succeed stem from the expressibility and capacity of deep architectures that facilitates learning complex features and models to account for interacted relationship directly from training examples. Due to the deep neural network mainly takes image patches/regions as inputs, they tend to fail to capture long range label dependencies in the scene parsing (i.e. saliency detection). In order to solve this problems, Zhao et al. [30] model the semantic features of salient objects and propose a multi-task deep saliency detection model based on a convolutional neural network. Wang et al. [12] propose a novel saliency detection algorithm using deep neural network combined with local estimation and global search. We propose to utilize deep ResNet in both global and local perspectives for salient objects detection, where the ResNet-G takes the entire image as input and generates global saliency map and the ResNet-L estimates local saliency of each pixel. Compared with previous work, our method outperforms previous methods that use only low-level features.

3. The framework of our SOD-LGDRN method

In this section, we present the proposed SOD-LGDRN method in detail. The pipeline of SOD-LGDRN is illustrated in Fig. 1. We model the complete framework of salient object detection with two deep ResNets: ResNet-G and ResNet-L. The ResNet-G takes the original image as input and generates semantic features globally. Since features extracted at the global level can not accurately describe the detailed structure of the salient object, we consider using the ResNet-L to extract local features as a complementary to the global information. Besides, in order to deal with the noise caused by the image segmentation, a spatial coherence method is integrated in the ResNet-L as a smoothing step.

3.1. Global-level saliency map

Global-level Feature Extraction In order to extract the global level semantic features of the image, we take the entire image as the input of the deep ResNet-G. Specifically, for each pixel in image I, a global-level feature is obtained via a feature extractor ϕ_g which is implemented by ResNet-G. Examples of the learned convolutional features are illustrated in Fig. 2. Fig. 2(a) is an input image and Fig. 2(b) illustrates the learned convolutional filters in the first layer, which capture color, contrast, edge and corner information in a local region. Fig. 2(c), Fig. 2(d) show the outputs of the 50_{th} and the 150_{th} layers, from which we can see that different levels of salient parts are highlighted at different levels of feature maps. (Better viewed at high resolution.) Then, for each feature, we apply a linear transformation that assigns a corresponding saliency score to the pixel, which can be expressed as:

$$SG(I_i) = w_{g,i}^T \phi_g(I_i) + b_{g,i}, \qquad i = 1,2,....,N, \tag{1}$$

where $w_{g,i}$ and $b_{g,i}$ represent the weights and biases of detector, and the values of these parameters are obtained by performing a linear transformation. Here, N is the pixel number of the original image I and I_i is the i_{th} pixel. By this way, we can obtain a saliency map SG. Although the saliency map generated via the deep ResNet-G can highlight the entire salient regions, it suffers two problems: (1) the generated saliency map on global level may be difficult to distinguish between the foreground and background; (2) when the background region is heavily cluttered, it can't preserve the subtle structure of salient objects. The reasons directly leading to the above two problems are the lack of consideration of the neighboring local information when extracting global-level features.

SLCI Smoothing In order to solve these problems, we use the superpixel-based local context information (SLCI) method [11] to preserve the boundary of the salient object when generating a global-level saliency map. Given an image I, we first use the Simple Linear Iterative Clustering segmentation method (SLIC) [31] to segment I into M superpixels as $\{H_i\}_{i=1\cdots M}$. Once obtaining the global-level saliency map SG, for each pixel, the saliency value is assigned the same as the ones with similar appearance in the local region. Denotes H_i the superpixel containing the pixel $p(x,y)$. Then the formula of SLCI smoothing for the pixel $p(x,y)$ is as follows:

$$R_i^g = \frac{1}{|H_i|} \sum_{p \in H_i} SG(p), \tag{2}$$

where $|H_i|$ denotes the cardinality of the set and R_i^g is the saliency value for the superpixel H_i. The direct meaning of Eq. (2) is that the saliency value of a pixel can be replaced with the average saliency value over all pixels in the superpixel it belongs to. For each pixel in H_i, the saliency value is set to the same as the saliency value of the superpixel. That's:

$$S_g(p) = R_i^g, \qquad \forall\, p(x,y) \in H_i. \tag{3}$$

3.2. Local-level saliency map

In order to obtain the local-level saliency map, an image is first segmented into multiple sub-regions, and then the multiple sub-regions are fed into the ResNet-L to generate the local-level semantic features. At last, saliency is estimated for each sub-region via the extracted local level features. We also introduce a method to refine the local level saliency map by considering the spatial coherence among multiple sub-regions.

3.2.1. Multi-level image segmentation

There are two steps to perform multi-level image segmentation: first, we employ the graph-based image segmentation method [32] to segment the image into multi-level sub-regions; secondly, in order to produce accurate segmentation and reduce the number of small meaningless segments, we apply region merging method [33]. Specifically, for each image, we segment it into $L = 10$ different levels. From the coarsest level (level 1) to the finest level (level 10), the number of

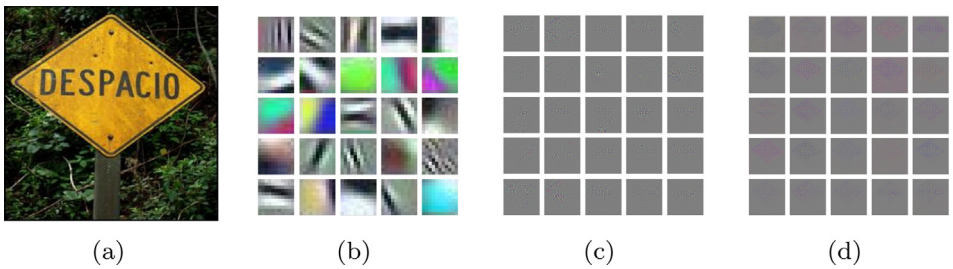

Fig. 2. Visualization of features learned by ResNet-G on PASCAL-S dataset. (a) Input image (b) 25 convolutional filters in the first layer (c) the learned features at the 50_{th} layer (d) the learned features at the 150_{th} layer.

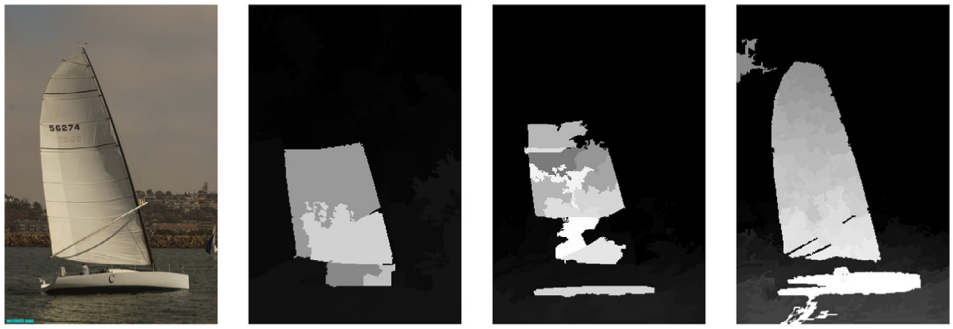

Fig. 3. Visual comparison of saliency maps at different scales of segmentation. (a) Original image (b) $S^{(1)}$ at scale 1 (c) $S^{(5)}$ at scale 5 (d) $S^{(10)}$ at scale 10.

segments for the image is set from 20 to 200, and the number of image segments in the intermediate levels follows a geometric series.

3.2.2. Multiscale local feature extraction

We extract features for each image segment with a deep residual network which has been trained on the ImageNet datasets using the Caffe framework. Due to the segments obtained by the multi-level image segmentation having variant size, we first wrap each of them in a bounding box size of 200×200. Then, at each level, all the warped segments are fed into the ResNet-L to obtain the saliency map at level k as: $S^{(k)}(H_i) = w_{l,i}^T \phi_l(H_i) + b_{l,i}, k = 1,2,....L$. Here, H_i is the i_{th} warped segment at k_{th} level of image segmentation, and $w_{l,i}$ and $b_{l,i}$ are obtained similarly as in Eq. (1). An example of the obtained $\{S^(k)\}$ is shown in Fig. 3. From this figure, we can see that as the segmentation performed from coarse level to fine level, the boundary of the salient object becomes accurate at the cost of the inhomogeneous salient object regions. Once obtaining the saliency maps at all levels, we aim to further fuse these saliency maps together to get an aggregated saliency map as $SL = \sum_{k=1}^{L} a_k S^{(k)}$, and the weights a_k are learned by performing a least-squares regression method on the training dataset:

$$\arg\min_{\{a_k\}} \sum_{i \in I_u} \|GT_i - \sum_k a_k S_i^{(k)}\|^2, \tag{4}$$

where I_u represents the set of indices of the images in the training dataset and GT_i is the groundtruth of the i_{th} image in the training set. Here, we define $\|\cdot\|$ to calculate the squared sum of all elements.

3.2.3. Spatial coherence refinement

Due to the fact that image segmentation method can't get the perfect segmentation results and our model assigns saliency values to each segment individually, problems by lacking of the spatial coherence between segments will inevitably appear in the resulting saliency map. Therefore, in order to enhance spatial coherence, we use a superpixel-based saliency refinement method. As we did before, the saliency value of the superpixel is set as the averaged saliency value of all pixels in it as $R_i = \frac{1}{|H_i|} \sum_{p \in H_i} SL(p)$ and $S_l(p) = R_i, \forall\, p \in H_i$, where R_i is the saliency value of the superpixel H_i at local saliency map SL. We enhance the saliency map by setting an objective function, which can be reduced to solving a linear system as:

$$\arg\min_{\{R_i^l\}} \sum_i (R_i^l - R_i)^2 + \sum_{i,j} \omega_{i,j} (R_i^l - R_j^l)^2, \tag{5}$$

where R_i^l is the saliency value of the superpixel H_i after refinement. The first term in the formula denotes the difference between the original saliency map and the refined saliency map, and the second term indicates the difference of spatial consistency between different superpixels. Here, $\omega_{i,j}$ refers to the weight between superpixels R_i^l and R_j^l aiming to enhance the spatial consistency. The $\omega_{i,j}$ is thus defined as follows:

$$\omega_{i,j} = exp\left(-\frac{d^2(H_i,H_j)}{2\sigma_1^2}\right), \tag{6}$$

where $d(H_i,H_j)$ is the normalized Euclidean distance between the color histogram of H_i and H_j in $CIEL*a*b$. In our experiment, σ_1 is the standard deviation of the distance between pairwise superpixels and set to 0.1. After refinement, the refined local saliency map S_l becomes:

$$S_l(p) = R_i^l, \qquad \forall\, p(x,y) \in H_i. \tag{7}$$

3.3. Saliency map fusion

Given two saliency maps S_g and S_l, inspired by the fact that both S_g and S_l are the prior information of the locations of the salient object, we can use a Bayesian integration method [20,34,35] to achieve a better fusion result. Specifically, we utilize one of them as the prior to compute the posterior probability of the other one, and combine the two computed posterior linearly to obtain the final fused saliency map S.

Take S_l as a prior and computation of the posterior probability for S_g for example. Firstly, we threshold the saliency map S_g by its mean saliency value and obtain its foreground and background regions described by F and B, respectively. Then, in each region, we compute the likelihoods by comparing S_g in terms of the foreground and background bins at pixel z:

$$\begin{aligned} p(S_g(z)|F) &= \frac{N_{b_F(S_g(z))}}{N_F}, \\ P(S_g(z)|B) &= \frac{N_{b_B(S_g(z))}}{N_B}, \end{aligned} \tag{8}$$

where N_F and N_B are the number of pixels in the foreground F and the background B, respectively. Denote $N_{b_F(S(z))}$ and $N_{b_B(S(z))}$ the number of pixels whose color features belong to the foreground bin $b_F(S_g(z))$ and background bin $b_B(S_g(z))$, respectively. Consequently, the posterior probability $p(F|S_g)$ with regards to S_g can be computed with S_l as the prior as:

$$p(F|S_g(z)) = \frac{S_l(z)p(S_g(z)|F)}{S_l(z)p(S_g(z)|F) + (1-S_l(z))p(S_g(z)|B)}. \tag{9}$$

Similarly, the posterior probability $p(F|S_l(z))$ with S_g as the prior can be computed as well. Once obtained $p(F|S_g(z))$ and $p(F|S_l(z))$, we can combine them together to get fused result S based on Bayesian integration:

$$S = p(F|S_l) + p(F|S_g). \tag{10}$$

4. Experiments and discussion

In this section, experimental results are reported to validate the proposed salient object detection model. Firstly, we compare our method with eight state-of-the-art methods on three public datasets qualitatively and quantitatively. Besides, in order to show the

effectiveness of our SOD-LGDRN method by using deep residual network, we perform comparisons by using different deep neural network under our SOD-LGDRN framework. Moreover, to show the superiority of using 152-layers deep ResNet for salient object detection, we experiment on deep ResNets with different layers. Meanwhile, in order to show the effectiveness of salient object detection by ResNet-L and ResNet-G in our proposed SOD-LGDRN method, we perform the performance analysis of them on ECSSD dataset. Finally, performance generalization of our model is provided.

4.1. Dataset

To evaluate our model, six public benchmark datasets are used: ECSSD, MSRA-B, PASCAL-S, SED1, THUR15K and ASD.

ECSSD [36]: This dataset contains 1000 structurally complex images acquired from the Internet with pixel-wise ground truth masks.

MSRA-B [37]: This dataset has 5000 images, and is widely used for salient object detection. Most of the images contain only one salient object. Pixel-wise annotation was provided by [38].

PASCAL-S [39]: This dataset was bulit using the validation set of the PASCAL VOC 2010 segmentation challenge. It contains 850 images with pixelwise salient object annotation. The ground truth saliency masks were labeled by 12 subjects.

SED1 [36]: This dataset is exploited recently which contains 100 images of single object.

THUR15K [40]: It consists of 15000 images, with 6233 pixel-accurate ground truth annotations for five specific categories: plane, dog jump, coffe mug, butterfly and giraffe.

ASD [16]: It consists of 1000 images labeled with pixle-wise ground truth.

All the experiments are run on the MATLAB R2011b platform on a workstation with Intel Xeon(R) CPU(2.60 GHz) and 500 GB RAM.

4.2. Evaluation criteria

In order to objectively compare our method with other methods, we use Precision-Recall curve, F-measure, MAE (Mean Absolute Error) three criteria for evaluation.

Precision-Recall Curve Since Precision-Recall (P-R) curve [44] provides rich information compared to ROC [9], we use P-R curve to evaluate the proposed algorithm.

To obtain a P-R curve, for each saliency map, we binarized it with 256 threholds, ranging from 0 to 255. Each binary mask (D) is compared with the ground-truth (G) which is also a binary mask. We compute the averaged precision and recall for each employed datasets.

$$Precision = \frac{|G \bigcap D|}{|D|}, Recall = \frac{|G \bigcap D|}{|G|}. \tag{11}$$

F-measure Compared to P-R curve, F-measure provides balanced evaluation of precision and recall. Due to precision is always more important than recall, F-measure is calculated as:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}. \tag{12}$$

We set β^2 as 0.3 empirically. In order to obtain good precision and recall values, every single saliency map is binarized via an adaptive threshold as:

$$T_\delta = \frac{2}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} s(x,y), \tag{13}$$

where w and h are the width and the height of saliency map respectively. Denotes $s(x,y)$ the saliency value of pixel (x,y). We then acquire averaged precision, recall and F-measure in each dataset.

MAE Since precision and recall values sometimes fail to reflect True Postive Rate (TPR) and Ture Negative Rate (TNR). MAE is exploited to reduce the average errors when fitting relations between TPR and TNR. MAE defines average pixel-wise absolute difference between ground truth and saliency map. The corresponding formula is as follows:

$$MAE = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} |s(x,y) - g(x,y)|. \tag{14}$$

4.3. Performance comparisons with state-of-the-art methods

In this subsection, we evaluate the proposed model on ECSSD, MSRA-B and PASCAL-S datasets. We trained our model using the 5000 images from the MSRA-B dataset and divide MSRA-B into two subsets, one subset of 4000 images for training and the other subset of 1000 images for test. The results of our experiments compare with eight state-of-the-art methods: Visual Saliency Based on Scale-Space Analysis in the Frequency Domain (HFT [41]), Background and Foreground Seed (BFS [42]), Bootstrap Learning (BL [43]), Global and Local cues (GL [22]), Deep Image Saliency Computing via Progressive Representation Learning (DISC [11]), Local Estimation and Global Search (LEGS [12]), Multiscale Deep CNN Features (MDF [10]) and Encoded Low level Distance map (ELD [13]). We compare with them quantitatively and qualitatively.

Quantitative Comparison Fig. 4 shows the saliency maps generated by our method and eight other methods. Experimental results have shown that our method not only highlights entire salient objects, but also preserve the details very well. In particular, our method can accurately detect the salient object and clearly display the object. As can be seen, our method performs well in a variety of challenging cases, such as the low contrast between salient object and background (row 1, row 3 and row 6), cluttered background (row 2), multiple salient objects (4-th and 7-th rows) and objects touching the image boundary (3-th and 4-th rows).

Qualitative Comparison As shown in the Fig. 5, our method achieves the highest precision in the entire recall range in three datasets. The results of P-R curves demonstrate that the proposed method outperform the state-of-the-art methods. On the ECSSD dataset, the proposed method obtains the highest precision value of 98.5%, which is 7.5% higher than the best one (91% in the ELD method). On the other hand, the minimal recall value of Ours method is 80%, significantly higher than those of the other methods.

Fig. 6 shows the F-measure values on the ECSSD, MSRA-B and PASCAL-S datasets, respectively. Among all these methods, our approach achieves the best performance on the overall F-measure as well as significant improvement in both precision and recall. On the ECSSD dataset, our method achieves 89.0% precision and 80% recall while the second best (ELD) achieves 86.8% precision and 72.3% recall. Performance improvement becomes more obvious on MSRA-B dataset. Compared with the second best (ELD), our method increases the F-measure from 0.72 to 0.78 and achieves an increase of 5% in precision while at the same time improving the recall by 15.1%.

Fig. 7 demonstrates that our method significantly outperforms other existing methods in terms of the MAE measurement, which provides a better estimation of the visual distance between the predicted saliency map and the ground truth. Our method successfully lowers the MAE by 5.0% with respect to the second best method (ELD) on the ECSSD dataset. On two other datasets, MSRA-B and PASCAL-S, our algorithm lowers the MAE by 2% and 3% respectively with respect to the second best methods (ELD). All these further show that our predicted saliency map is the best estimation of the ground truth.

4.4. Comparison with other deep neural networks

To further demonstrate the effectiveness of the proposed framework, we replace our SOD-LGDRN architecture with CNN, VGG16 and AlexNet, and compare their performance, respectively. It is worth

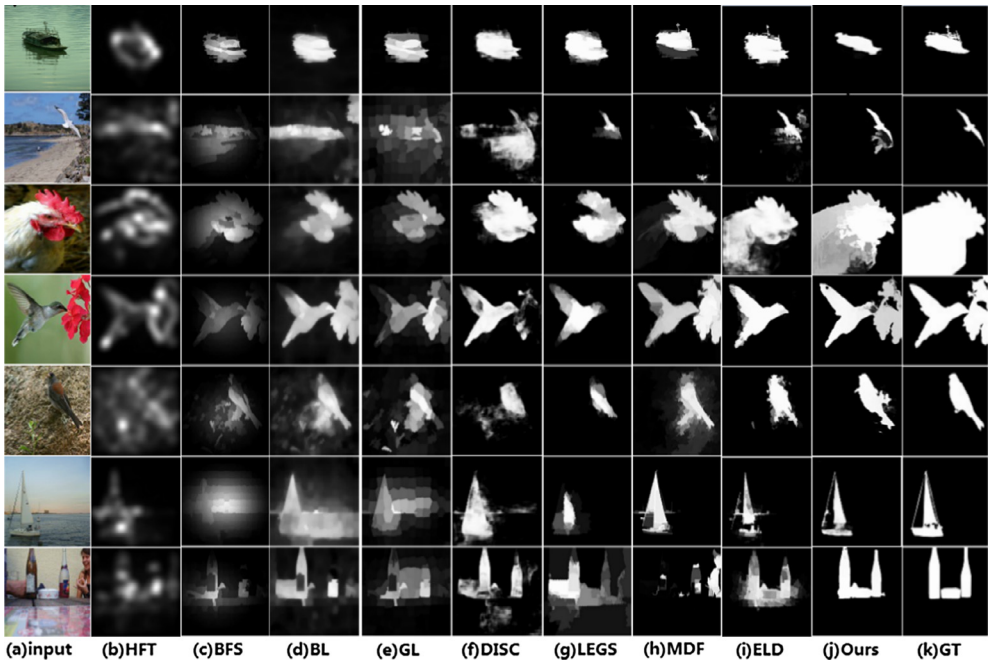

Fig. 4. Visual comparison of our results with the state-of-the-art methods on ECSSD, MSRA-B and PASCAL-S datasets. Top and bottom two rows are images from the ECSSD, MSRA-B datasets, middle three rows are images from the PASCAL-S dataset. The ground truth (GT) is shown in the last column. Our method produces saliency maps closest to the ground truth. We compare our method against saliency model HFT [41], BFS [42], BL [43], GL [22], DISC [11], LEGS [12], MDF [10], ELD [13].

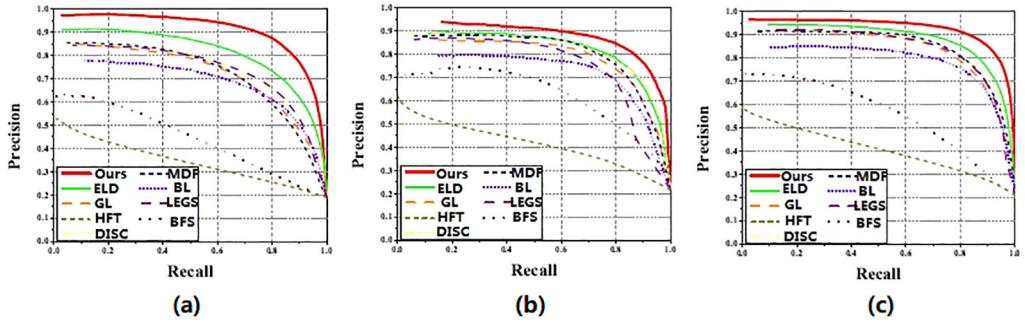

Fig. 5. Quantitative comparison the precision-recall curves of different methods on three datasets: (a) ECSSD dataset (b) MSRA-B dataset (c) PASCAL-S dataset.

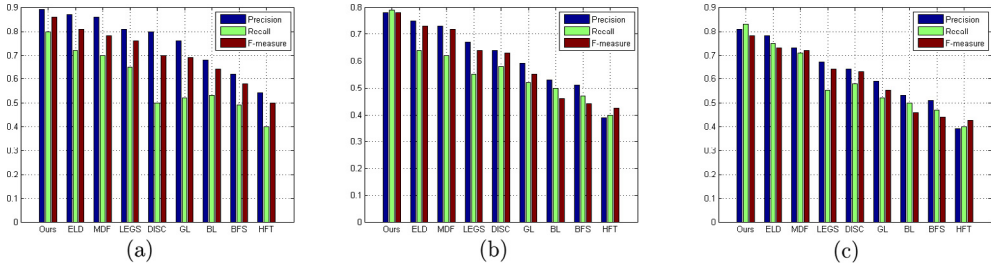

Fig. 6. Quantitative comparison our method with different methods from F-measure, Precision, Recall on three datasets: (a) ECSSD dataset (b) MSRA-B dataset (c) PASCAL-S dataset.

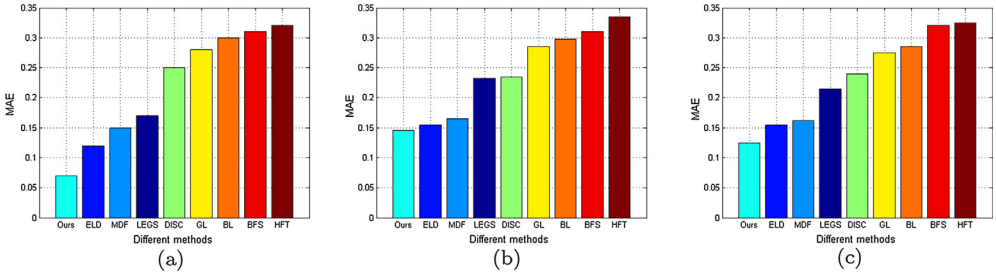

Fig. 7. Quantitative comparison the MAE values of different methods on three datasets: (a) ECSSD dataset (b) MSRA-B dataset (c) PASCAL-S dataset. Better viewed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

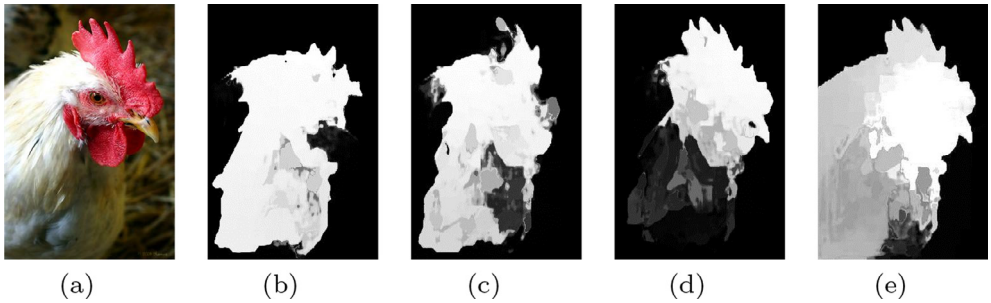

Fig. 8. Visual comparison of our results with other different types of deep neural networks. The samples are taken from the MSRA-B dataset: (a) Original image (b) CNN (c) AlexNet (d) VGG (e) Ours.

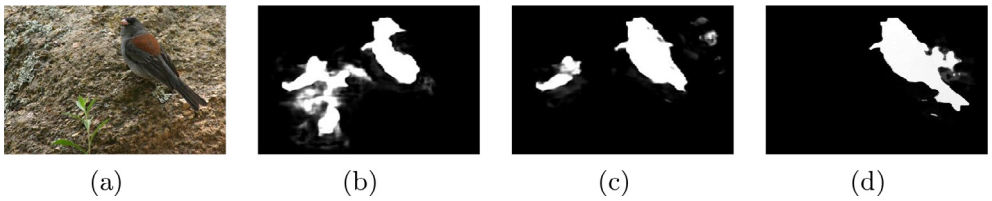

Fig. 9. Visual comparison of saliency map of the ResNet with 51-layers, 101-layers and 152-layers. The samples are taken from the MSAR-B dataset: (a) Original image (b) 51-layers RseNet (c) 101-layers ResNet (d) 152-layers ResNet.

Fig. 10. Visual comparison of the salient object detection result by ResNet-L and ResNet-G in our SOD-LGDRN method. (a) Original image (b) Global saliency map by ResNet-G (c) Local saliency map by ResNet-L (d) The fused result.

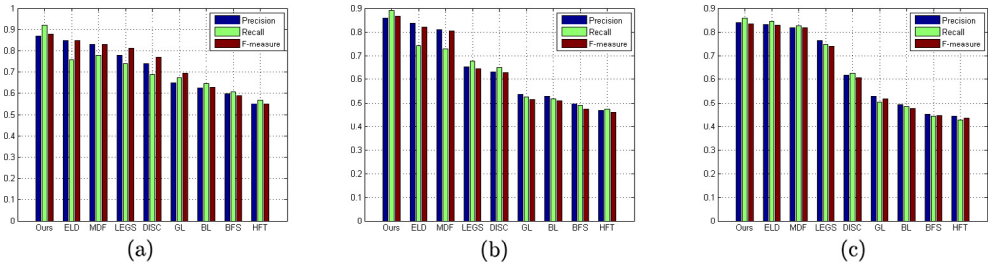

Fig. 11. Visual comparison the generalization ability of our method with different methods on three datasets: (a) ASD dataset (b) SED1 dataset (c) THUR15K dataset. From this figure, we can see that SOD-LGDRN framework has well generalization ability.

noting that both VGG16 and AlexNet networks are improved and enhanced on the CNN network structure. In order to make a fair comparison, we put the SOD-LGDRN framework of different networks on the same dataset (PASCAL-S). The experiments are all carried out on a workstation with NVIDIA GeForce GTX TITAN Black and 1 TB RAM. The results are shown in Fig. 8. The SOD-LGDRN architecture with 152-layers ResNet performs consistently better than SOD-LGDRN framework with VGG16 and AlexNet. As can be seen from Fig. 8, our SOD-LGDRN framework with 152-layers Resnet is significantly superior to SOD-LGDRN framework with VGG16 and AlexNet. Therefore, it can demonstrate the effectiveness of our SOD-LGDRN framework with ResNet.

4.5. Comparison with the different layers of deep ResNet

In order to demonstrate the superiority of the SOD-LGDRN framework with 152-layers ResNet, we compared the SOD-LGDRN framework with 152-layers ResNet to the SOD-LGDRN model with 51-layers, 101-layers ResNet. Except for the different layers of ResNet, all other settings of the SOD-LGDRN framework are kept untouched. Sample results of salient object detection by SOD-LGDRN with 152-layers and SOD-LGDRN with 51-layers, 101-layers are shown in Fig. 9. We can see that, SOD-LGDRN architecture with 152-layers not only highlights the overall objects but preserves the boundary and structure details. While the boundary detected by the SOD-LGDRN model with 51-layers and 101-layers are more background are detected. Therefore, SOD-LGDRN framework with 152-layers achieved better performance in detecting salient object.

4.6. ResNet-L v.s. ResNet-G on salient object detection

In order to show the effectiveness of salient object detection by ResNet-L and ResNet-G in our proposed SOD-LGDRN method, we

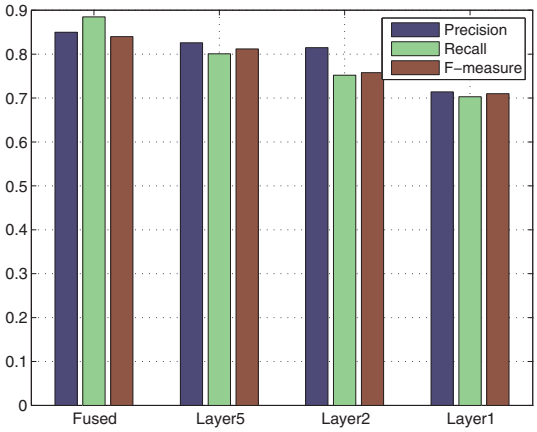

Fig. 12. The effectiveness of multi-level segmentation. Layer1, Layer2 and Layer5 denotes three segmentation levels.

Table 1
Comparison of per image of execution time (seconds) on the ECSSD dataset.

Method	HFT	BFS	BL	GL	DISC	LEGS	MDF	ELD	Ours
Times(s)	0.036	0.423	0.251	0.252	0.341	0.438	0.201	0.183	**0.125**

perform the performance analysis of them on ECSSD dataset. The sampled detection results are shown in Fig. 10. From this figure, we can see that the global saliency map obtained via ResNet-G (i.e. Fig. 10(b)) cannot keep the complete object boundary (e.g. the hindquarters of the yak) and some of the background on the right bottom of the image are wrongly detected, while the local saliency map generated via ResNet-L (i.e. Fig. 10(c)) can obtain complete object boundary but the interior of the salient object cannot be highlighted homogeneously. Only the fused result by our proposed SOD-LGDRN method can not only highlight the overall object but preserve the object boundary and the details of the structure.

4.7. Ablation study

4.7.1. Multi-level image segmentation

As discussed in Section 3.2, our method employs multi-level image segmentation algorithm when extracting local-level features. As shown in Fig. 12, the performance of the multi-level segmentation is significantly better than the performance of single-level segmentation. The aggregated saliency map from 10 levels of image segmentation is significantly improved, and the average accuracy is improved by 2.4% and the recall rate is 8.4% when it is compared with the result obtained by single-level segmentation.

4.7.2. Execution time

In the experiment, we implement our SOD-LGDRN model under the caffe framework and train them using stochastic gradient descent (SGD) with momentum of 0.99, and weight decay of 0.005. The learning rate for the training the SOD-LGDRN is initialized as 10^{-9} with a batch size of 64. All the experiments are run on the MATLAB R2011b platform on a workstation with Intel Xeon(R) CPU(2.60 GHz) and 500 GB RAM.

In fact, more accurate results are achieved at the cost of longer run-time. However, this is not what we expect, we hope to achieve better performance, but also can maintain high efficiency. As shown in Table 1, we can see that proposed method has achieved good performance and run-time is the shortest.

4.8. Performance of generalization on other datasets

In this section, we evaluate and analyze the generalization of SOD-LGDRN. In order to learn a particular model, we need to spend a lot of time and effort to train on a large number of labeled data. Therefore, we need a model with generalization function. To evaluate the effect of SOD-LGDRN on other datasets, we evaluated SOD-LGDRN on three datasets (i.e. ASD, SED1 and THUR15K). In our experiment, except for the different datasets, all other settings of the SOD-LGDRN model is keep untouched. We test the performance of these three datasets by using the model we learned on the ECSSD dataset. The test results are shown in Fig. 11. As can be seen from Fig. 11, although the model is trained on other datasets, its performance still outperforms some of the state-of-the-art approaches. Thus, we can demonstrate that our SOD-LGDRN model has generalization ability.

5. Conclusion

In this paper, we propose a novel salient object detection framework. Compared with existing image saliency computing methods, our framework achieves superior performance without relying on various features (priors/assumptions). We model the image saliency from both local and global observations. Specifically, our salient object detection model is built upon deep ResNets. Firstly, ResNet-G generates a global level saliency map by taking the overall image as the input. Secondly, we decompose the input image into a series of regions, then put theses regions into the ResNet-L to produce local level saliency map while preserving object details. Our method outperforms eight state-of-the-art approaches in term of visual qualitative and quantitative results on pubilc datasets. Meanwhile, our method has superior generalization ability across datasets. As a future work, we are planning to combine our model with object location or object recognition to explore the efficient algorithms.

Acknowledgment

This work was supported by the General Program of National Natural Science Foundation of China (NSFC) under Grant No. 61572362. This research was also partially supported by the General Program of National Natural Science Foundation of China (NSFC) under Grant No. 81571347, and the Fundamental Research Funds for the Central Universities under Grant No. 22120180012.

References

[1] Z. Ren, S. Gao, L.T. Chia, W.H. Tsang, Region-based saliency detection and its application in object recognition, IEEE Trans. Circ. Syst. Video Technol. 24 (5) (2014) 769–779.
[2] L. Wu, C. Shen, A.V.D. Hengel, Personnet: Person re-identification with deep convolutional neural networks, Comput. Vision Pattern Recog. (2016) 3908–3915.
[3] Y.H. Sung, W.Y. Tseng, P.H. Lin, L.W. Kang, C.Y. Lin, C.H. Yeh, Significance-Preserving-Guided Content-Aware Image Retargeting, Springer Berlin Heidelberg, 2013.
[4] J. Johnson, R. Krishna, M. Stark, L.J. Li, Image retrieval using scene graphs, in: Computer Vision and Pattern Recognition, 2015, pp. 3668–3678.
[5] J. Huang, X. Yang, X. Fang, W. Lin, R. Zhang, Integrating visual saliency and consistency for re-ranking image search results, IEEE Trans. Multimedia 13 (4) (2011) 653–661.
[6] S. Marat, M. Guironnet, D. Pellerin, Video summarization using a visual attention model, in: Signal Processing Conference, 2007 European, 2015, pp. 1784–1788.
[7] C. Zhang, W. Lin, W. Li, B. Zhou, J. Xie, J. Li, Improved image deblurring based on salient-region segmentation, Signal Process. Image Commun. 28 (9) (2013) 1171–1186.
[8] L. Itti, C. Koch, A saliency-based search mechanism for overt and covert shifts of visual attention, Vision. Res. 40 (10-12) (2000) 1489–1506, http://dx.doi.org/10.1016/S0042-6989(99)00163-7.
[9] S. Goferman, L. Zelnikmanor, A. Tal, Context-aware saliency detection, IEEE Trans. Pattern Anal. Mach. Intell. 34 (10) (2012) 1915.
[10] G. Li, Y. Yu, Visual saliency based on multiscale deep features, in: Computer Vision and Pattern Recognition, 2015, pp. 5455–5463.
[11] T. Chen, L. Lin, L. Liu, X. Luo, X. Li, Deep image saliency computing via progressive representation learning, IEEE Trans. Neural Networks Learn. Syst. 27 (6) (2016) 1135.
[12] L. Wang, H. Lu, R. Xiang, M.H. Yang, Deep networks for saliency detection via local estimation and global search, in: Computer Vision and Pattern Recognition, 2015, pp. 3183–3192.

[13] G. Lee, Y.W. Tai, J. Kim, Deep saliency with encoded low level distance map and high level features, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 660–668.
[14] L. Itti, C. Koch, Comparison of feature combination strategies for saliency-based visual attention systems, Proc. SPIE – Int. Soc. Opt. Eng. 3644 (1) (1999) 473–482.
[15] Y.F. Ma, H.J. Zhang, Contrast-based image attention analysis by using fuzzy growing, in: Eleventh ACM International Conference on Multimedia, 2003, pp. 374–381.
[16] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: Computer Vision and Pattern Recognition, 2009, pp. 1597–1604.
[17] M.M. Cheng, G.X. Zhang, N.J. Mitra, X. Huang, S.M. Hu, Global contrast based salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 409–416.
[18] L. Xu, L. Zeng, H. Duan, An effective vector model for global-contrast-based saliency detection, J. Vis. Commun. Image Represent. 30 (2015) 64–74.
[19] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.Y. Shum, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2010) 353–367.
[20] M.M.S. Fareed, G. Ahmed, C. Qi, Salient Region Detection Through Sparse Reconstruction and Graph-based Ranking, Academic Press, Inc., 2015.
[21] K. Shi, K. Wang, J. Lu, L. Lin, Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2115–2122.
[22] N. Tong, H. Lu, Y. Zhang, X. Ruan, Salient object detection via global and local cues, Pattern Recogn. 48 (10) (2015) 3258–3267.
[23] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: International Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.
[24] C. Da, H. Zhang, Y. Sang, Brain CT Image Classification with Deep Neural Networks, Springer International Publishing, 2015, pp. 653–662.
[25] D. Erhan, C. Szegedy, A. Toshev, D. Anguelov, Scalable object detection using deep neural networks, in: Computer Vision and Pattern Recognition, 2014, pp. 2155–2162.
[26] J. Kung, D. Zhang, G.V.D. Wal, S. Chai, S. Mukhopadhyay, Efficient object detection using embedded binarized neural networks, J. Signal Process. Syst. (11) (2017) 1–14.
[27] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, F. Wu, Background prior-based salient object detection via deep reconstruction residual, IEEE Trans. Circ. Syst. Video Technol. 25 (8) (2015) 1309–1321.
[28] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 640–651.
[29] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, Comput. Sci. (4) (2014) 357–361.
[30] X. Li, L. Zhao, L. Wei, M.H. Yang, F. Wu, Y. Zhuang, H. Ling, J. Wang, Deepsaliency: Multi-task deep neural network model for salient object detection, IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. 25 (8) (2015) 3919.
[31] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, Slic superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274.
[32] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comput. Vision 59 (2) (2004) 167–181, http://dx.doi.org/10.1023/B:VISI.0000022288.19776.77.
[33] C. Zhou, D. Wu, W. Qin, C. Liu, An efficient two-stage region merging method for interactive image segmentation, Comput. Electrical Eng. 54 (15) (2016) 220–229.
[34] Y. Xie, H. Lu, M.H. Yang, Bayesian saliency via low and mid level cues, IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. 22 (5) (2013) 1689–1698.
[35] X. Li, H. Lu, L. Zhang, R. Xiang, M.H. Yang, Saliency detection via dense and sparse reconstruction, in: IEEE International Conference on Computer Vision, 2013, pp. 2976–2983.
[36] Q. Yan, L. Xu, J. Shi, J. Jia, Hierarchical saliency detection, in: Computer Vision and Pattern Recognition, 2013, pp. 1155–1162.
[37] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.Y. Shum, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2011) 353–367.
[38] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, S. Li, Salient object detection: A discriminative regional feature integration approach, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 2083–2090.
[39] Y. Li, X. Hou, C. Koch, J.M. Rehg, A.L. Yuille, The secrets of salient object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 280–287.
[40] M.M. Cheng, N.J. Mitra, X. Huang, S.M. Hu, Salientshape: group saliency in image collections, Visual Comput. 30 (4) (2014) 443–453.
[41] J. Li, M.D. Levine, X. An, X. Xu, H. He, Visual saliency based on scale-space analysis in the frequency domain, IEEE Trans. Pattern Anal. Mach. Intell. 35 (4) (2013) 996–1010.
[42] J. Wang, H. Lu, X. Li, N. Tong, W. Liu, Saliency detection via background and foreground seed selection, Neurocomputing 152 (2015) 359–368.
[43] N. Tong, H. Lu, X. Ruan, M.-H. Yang, Salient object detection via bootstrap learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1884–1892.
[44] C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang, Saliency detection via graph-based manifold ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3166–3173.