Deep Salient Object Detection by Integrating Multi-level Cues

Jing Zhang[1,2] Yuchao Dai[2] and Fatih Porikli[2]
[1] School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
[2] Research School of Engineering, Australian National University, Canberra, Australia

Abstract

A key problem in salient object detection is how to effectively exploit the multi-level saliency cues in a unified and data-driven manner. In this paper, building upon the recent success of deep neural networks, we propose a fully convolutional neural network based approach empowered with multi-level fusion to salient object detection. By integrating saliency cues at different levels through fully convolutional neural networks and multi-level fusion, our approach could effectively exploit both learned semantic cues and higher-order region statistics for edge-accurate salient object detection. First, we fine-tune a fully convolutional neural network for semantic segmentation to adapt it to salient object detection to learn a suitable yet coarse per-pixel saliency prediction map. This map is often smeared across salient object boundaries since the local receptive fields in the convolutional network apply naturally on both sides of such boundaries. Second, to enhance the resolution of the learned saliency prediction and to incorporate higher-order cues that are omitted by the neural network, we propose a multi-level fusion approach where super-pixel level coherency in saliency is exploited. Our extensive experimental results on various benchmark datasets demonstrate that the proposed method outperforms the state-of-the-art approaches.

1. Introduction

Salient object detection (saliency prediction) [9][23][14] [6] aims at identifying the visually interesting object regions that are consistent with human perception. It is essential in many computer vision tasks including object-aware image retargeting [49], image cropping [38], context-aware image editing [54], recognition [40], and interactive image segmentation [31]. Even though considerable progress has been made (See [7] for a dedicated survey on salient object detection before the era of deep learning), it still exists as a challenging task and requires competent approaches to effectively handle real world scenarios.

Most of the traditional saliency detection methods are based on low-level hand-crafted features such as color and texture descriptors [24], or they compute variants of appearance uniqueness [25] and region compactness [14] based on the above primitives. Statistical priors of salient objects, e.g. contrast, boundary, and in-focus have also been investigated. These methods report acceptable results on certain datasets. However, saliency methods based on such simple hand-crafted features are often incapable of capturing semantic attributes of salient objects. As a result, their saliency maps deteriorate when the images become cluttered and complicated. By contrast, high-level semantic information plays a central role in distinguishing foreground objects from their background scenes with similar appearances. It is frequently associated with object localization, recognition, and segmentation. Numerous methods have exploited high-level information for salient object detection. To this end, deep neural network based salient object detection has achieved remarkable success and various network structures (convolutional [33] [29] [55] [11] [46], recurrent [47] [27]) have been proposed to learn competent feature representations for salient objects.

We argue that high-level semantic information could be a two-edged sword for salient object detection. On the one hand, semantic attributes provide rich and essential information in distinguishing foreground objects from difficult backgrounds [37]. On the other hand, there exist scenarios where the correlation between semantic information and saliency is low or even negative. As illustrated in Fig. 1, semantic segmentation focuses only on whether or not there exist trained categories of objects without considering whether these objects are salient. In this paper, we fine-tuned a deep learning network originally designed for semantic segmentation. By using the trained semantic segmentation model to initialize our saliency detection network, semantic information is reserved to better detect salient semantic objects.

An end-to-end neural network based saliency prediction framework has the potential to generate pixel-level saliency in a data-driven manner, but nevertheless, there exist several problems. First, the derived saliency scores from a fully convolutional neural network have low spatial reso-

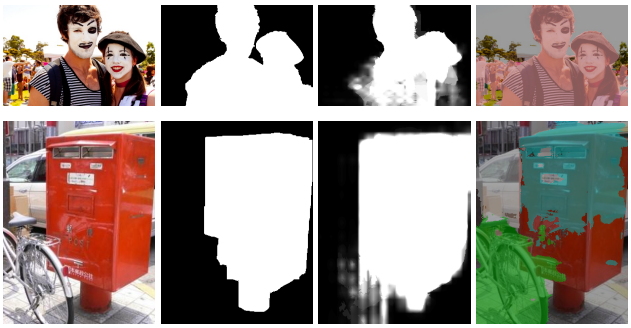


Figure 1. Difference between semantic segmentation and saliency detection. From left to right: original image, ground truth saliency map, saliency detection result and semantic segmentation result.

lution due to the pooling and stride operators. Even though the dilation or atrous convolution (see [10] [42]) could be used, significant downsampling effects may still exist in the resultant saliency maps. Second, the convolutional network can encode the local structure, yet local features may fail to capture the wider range context that is critical for salient object detection. Third, existing deep learning based methods disregard the prior knowledge (e.g. contrast prior, boundary prior and center prior) for salient object detection since they do not have a mechanism to impose such constraints.

To address the above issues, we propose to exploit the multi-level saliency cues in a unified framework. First, we use a deep fully convolutional network to learn a coarse level dense saliency map by exploiting the high-level semantic information for saliency prediction. Our network is built upon the DeepLab semantic segmentation [10]. We repurpose a very deep convolutional neural network (ResNet-101 [18]) that is originally trained for image classification, to the task of saliency prediction. We transform all fully connected layers to convolutional layers, and increase feature resolution through atrous convolutional layers [10]. In this way, we increase the resolution of the output feature four times. Similar to [10], the resultant saliency prediction map is upsampled by a factor of $8\times$ to get the original image resolution. This is done by bilinear interpolation (in essence, this upsampling operation could be achieved by deconvolution [53]). In addition, to handle the low spatial resolution of the learned saliency map and to integrate the higher-order statistics for salient object detection, we propose a multi-level fusion approach, where superpixel level coherency in saliency map is exploited to incorporate superpixel level statistics.

The main contributions of our work are summarized as follows: 1) We introduce a fully convolutional neural network based approach for salient object detection. By integrating saliency cues at different levels through our network, we effectively exploit both learned cues and higher-order region statistics for edge-accurate pixel-level detection; 2) We propose a multi-level fusion method where we employ superpixel level coherency in saliency to enhance the resolution of saliency prediction; 3) We use a small training dataset (6,000 images from the MSRA10K dataset [13]), yet achieve the state-of-the-art performance on all other datasets, which proves that our method elegantly generalizes to other datasets; 4) Our method is computationally efficient, which takes 0.2 sec to generate saliency map of a given image.

2. Related Work

Existing salient object detection methods can be roughly classified into two categories: hand-crafted feature based methods, and deep learning based methods. We refer interested readers to [7] and [8] for an in-depth survey and benchmark comparisons of conventional methods.

2.1. Hand-crafted Feature Based Methods

Many existing salient object detection methods [1] [39] [14] [43] [15] rely on hand-crafted features. They use color contrast [15] given an over-segmented image. [50] formulates saliency detection as an energy minimization problem and solves saliency assignment for each pixel. In [36] salient object detection is formulated as an image segmentation problem. By exploiting the sparsity prior for salient object, [44] solves salient object detection as a low-rank matrix decomposition problem. Objectness, which highlights the object-like regions, has also been used in [3] [25] [9].

In addition to the above low-level cues, [56] presents a robust background measure, namely "boundary connectivity", and a principle optimization framework. Along with this line, [51] ranks the similarity of image elements with foreground and background cues via graph-based manifold ranking. [17] uses the center of a convex hull notion to obtain strong background and foreground priors. Different from the above unsupervised methods, which compute pixel or superpixel saliency directly, [24] [26] and [48] regard saliency detection as a regression problem.

2.2. Deep Learning Based Methods

The above hand-crafted feature based saliency detection methods are effective for simple scenes, and they become very fragile when the scene complexity increases. Recently, deep neural network models have been adopted to salient object detection [33] [55] [11] [46] [29] [47] [27] [28] [30] [19]. Deep networks have been shown to encode high-level semantic features that capture saliency information better than hand-crafted features and reported superior performance compared with the traditional techniques.

Deep learning based methods generally train a neural network to assign saliency to each pixel or superpixel. [29] learns saliency of each superpixel by using learned features obtained from an existing CNN model instead of those

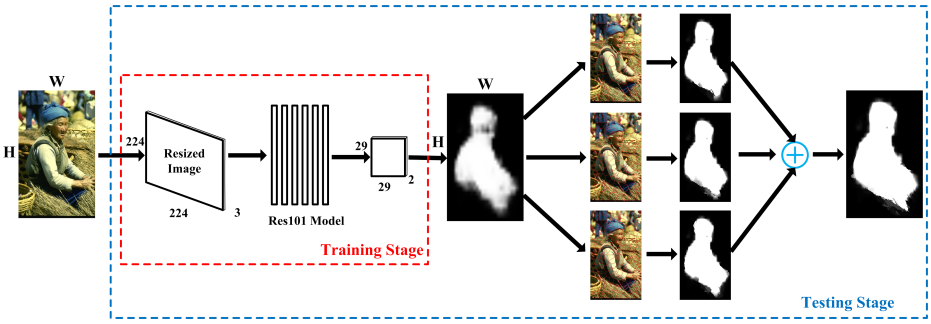


Figure 2. A nutshell of our approach to saliency detection. Given an image as input, the deep network outputs a coarse dense saliency map. Then three scale image over-segmentation are utilized to refine the low resolution saliency map and incorporate the higher-order region statistics for edge-accurate salient object detection. Finally, the three saliency maps are fused to form our final saliency map.

handcraft features. [20] aims at narrowing the semantic gap. [33] proposes a multi-task learning framework to saliency detection, where saliency detection and semantic segmentation are learned at the same time. As a reconstruction based model, [30] uses an end-to-end contrast network to produce a pixel-level saliency map. In [11], a novel Deep Image Saliency Computing (DISC) framework is presented for fine-grained image saliency computing. Two stacked DCNNs are used to get coarse-level saliency map and fine-grained saliency map respectively. [21] formulates saliency map as a generalized Bernoulli distribution. By training a deep architecture to predict saliency maps, it computes distances between probability distributions of output saliency map and the ground truth. Another noteworthy work [27] proposes a recurrent attentional convolutional-deconvolution network (RACDNN), which iteratively selects image sub-regions to perform saliency refinement. In [28] both CNN features and low-level features are integrated for saliency detection. More recently, [35] proposes an end-to-end deep hierarchical saliency network (DHSNet) based on convolutional neural networks for detecting salient objects. This work is similar to [41] where a shallow and a deep convolutional network are trained in an end-to-end architecture. By jointly modeling global and local context, [55] proposes a multi-context deep convolutional neural network for saliency detection.

As opposed to those fully convolutional neural network (FCNN) based methods [33] [35] [55], our network is much deeper. We use multiple, parallel, and dilated convolution layers that provide a wider receptive field, which improves robustness and incorporates stronger contextual information. We generate spatially consistent and boundary accurate saliency maps using histogram-based saliency affinity as well as multi-scale fusion based spatial refinement. Finally, We use a smaller training set than other methods yet achieve better results.

3. Our Approach

3.1. Overview

Targeting at effectively exploiting the multi-level saliency cues in a unified and data-driven manner, in this paper, we propose a new FCNN based approach empowered with multi-level fusion to salient object detection. By integrating saliency cues at different levels through FCNNs and multi-level fusion, our approach could effectively exploit both learned semantic cues and higher-order region statistics for edge-accurate salient object detection.

First, we fine-tune a FCNN (Deep Residual Network [18] based) for semantic segmentation to adapt it to salient object detection to learn a coarse per-pixel saliency prediction map. Due to the local receptive fields in the convolutional network, the resultant map is often smeared across salient object boundaries. Second, to enhance the resolution of the learned saliency prediction map and to incorporate higher-order cues that are omitted by the neural network, we propose a multi-level fusion approach where super-pixel level coherency in saliency is exploited. This spatial refinement for saliency detection looks wide and assigns similar saliency value to similar neighboring regions. To deal with the multi-scale effect with salient object, we conduct the spatial refinement at different levels before the multi-level refined saliency prediction maps are fused to output the final salient object detection results. A nutshell of our approach is illustrated in Fig. 2.

3.2. Saliency Prediction by Convolutional Networks

Our salient object detection network is built upon the DeepLab semantic segmentation network [10], where a deep convolutional neural network (ResNet-101 [18] in this work) trained for the task of image classification is repurposed to the task of semantic segmentation by 1) transforming all the fully connected layers to convolutional layers and 2) increasing feature resolution through atrous or dilation convolutional layers [10] [52]. In this way, the

spatial resolution of the output feature has been increased four times, which is much denser than [55] [29]. Similar to [10], the resultant saliency prediction map is upsampled by a factor of 8 to reach the same resolution as the input image by employing bi-linear interpolation. Note that, this upsampling operation could also be achieved by deconvolution [53]). Here we adopt bi-linear interpolation mainly due to its efficiency. Furthermore, as demonstrated in the following sections, our multi-scale fusion based saliency refinement could handle imperfect saliency prediction results from upsampling naturally.

Instead of learning unreferenced functions, Deep Residual Network [18] explicitly learns residual functions with reference to the layer inputs. These residual networks are easier to optimize, and can gain accuracy from considerably increased depth. Therefore, we used ResNet-101 model in our saliency map prediction network. The architecture of our network is nearly the same as the deep residual network by removing the final pooling and fully-connected layer to adapt it for dense prediction (semantic segmentation and saliency detection). To make use of the multi-scale context for final prediction, we add multiple parallel dilated convolutional layers at the end and then sum these layers to feed into the final softmax layer (in training stage) or prediction layer (in test stage). All these together form our fully convolutional neural network (FCNN), and we define our saliency map using fine-tuned deep residual network as S^I. Experimental results on 11 saliency benchmarking datasets prove that this extremely deep network ends up with more robust and better saliency map compared with the state-of-the-art methods. In Figure 3, we compare our approach with state-of-the-art methods, where (g) shows the saliency prediction map directly from the fully convolutional networks. Even with the deep FCNN only, our approach already outperforms most of the competing methods.

Training details: We trained our model using 6,000 images from the MSRA10K dataset after excluding the same images in the ASD dataset and the 3,000 testing images. We use Caffe[22] to train our network. The network parameters are fine-tuned from [10]. We use momentum-accelerated mini-batch SGD with "batch_size" 5 and "iter_size" 10 for accumulating the gradients in different iterations. The learning rate is initialized as 2.5e-4 with "poly" decay policy. The max iteration is 20000. To minimize over-fitting, we use drop-out layers in our network. We trained the model until training accuracy kept unchanged for 200 iterations. Each image is firstly scaled to the same size as $224 \times 224 \times 3$. Training takes 3 days for 20,000 iterations on a PC with one NVIDIA Quadro M4000 GPU.

3.3. Saliency Refinement by Multi-scale Fusion

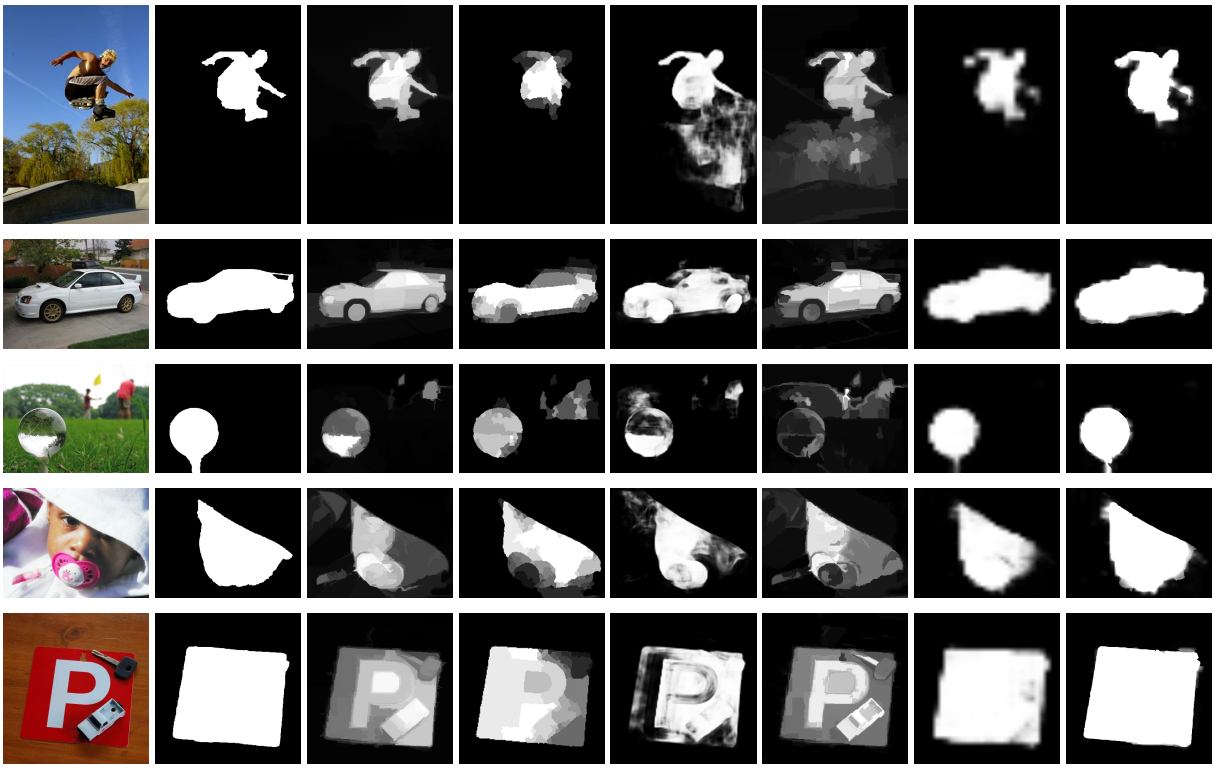
The above deep fully convolutional network learns to predict saliency map in an end-to-end manner, which already outperforms most of the state-of-the-art methods as witnessed by Fig. 3. However, due to the local receptive fields in the convolutional network, the saliency prediction map is often smeared across salient object boundaries. Additionally, the deep neural networks have not explicitly incorporated existing saliency cues at low-level or mid-level. Therefore, to enhance the resolution of the learned saliency prediction map and to incorporate higher-order cues that are omitted by the neural network, we propose a multi-level fusion approach. Under our formulation, super-pixel level coherency in saliency is exploited in spatial refinement, where saliency detection looks wide and assigns similar saliency value to similar neighboring regions. To evaluate the similarity between superpixels, color histogram distance in three color spaces are used, which include RGB color space, Lab color space and HSV color space. To deal with the multi-scale effect with salient object, we conduct the spatial refinement at different levels before the multi-level refined saliency prediction maps are fused to output the final salient object detection results.

In order to get homogeneous consistent regions, we used SLIC [2] for image over-segmentation to represent each image as a collection of superpixels. Each image X is represented as a collection of consistent elements $X = \{X_1, X_2, \cdots, X_N\}$, where the numbers of superpixels $N = \{100, 200, 300\}$ are used to achieve multi-scale image over-segmentation. This multi-scale over-segmentation strategy has been widely exploited in achieving higher resolution saliency prediction map [24]. Given saliency prediction map S^I from deep neural networks, for a specific number of N, we reach a saliency prediction score vector $S^I = \{s_1^I, s_2^I, ..., s_N^I\}$, where the per-superpixel score s_i^I is defined as the median saliency score in superpixel X_i.

Similar to [29], we formulate the spatial refinement as a multi-scale binary pixel labeling problem, and employ the following energy function (The energy function is defined on each scale separately):

$$E(S^R) = \sum_i \omega_i (s_i^R - s_i^I)^2 + \sum_{i,j} \omega_{ij} (s_i^R - s_j^R)^2, \quad (1)$$

where $S^R = \{s_1^R, s_2^R, ..., s_N^R\}$ represents the desired saliency value, s_i^R is the saliency score after refinement at superpixel X_i. The first term in Eq.-(1) encourages similarity between the refined saliency map and the input coarse saliency map, while the second term is an all-pair spatial coherence term that favors consistent saliency scores across different superpixels if they are similar to each other. ω_i is a flag, which is 1 when s_i^I does not belong to strong foreground or strong background region, and 0 when it is strong foreground or strong background region. ω_i is used to keep saliency score of those strong regions intact after spatial refinement. In our paper, we define strong region in the fol-



(c) Image (d) GT (e) MDF (f) LEGS (g) DISC (h) DRFI (i) FCNN (j) OURS

Figure 3. Salient object detection results on challenging images by different methods. From left to right: original image, ground truth saliency map, saliency maps using MDF [29], LEGS [46], DISC [11], DRFI [24], Our saliency map from Deep Residual Network and our final saliency map. It is worth noting the difference between the network output (g) and the final multi-cue fusion results (h).

lowing way: given saliency prediction map S^I from deep fully convolutional network, we find those regions with saliency value 1 or 0, and label these superpixels as strong foreground or strong background region respectively. The weight for pairwise term ω_{ij} depends on the similarity between any pair of superpixels X_i and X_j. In this paper, we define $\omega_{ij} = \exp\left(-\frac{d_{ij}^2}{\delta^2}\right)\exp\left(-\frac{e_{ij}^2}{\sigma^2}\right)$ by considering both the appearance and the spatial distance, where e_{ij} is the spatial distance between the superpixels and d_{ij} is the color histogram distance or texture histogram distance, which is defined as Chi-square distance: $d_{ij} = \sum_{k=1}^{b}\left[\frac{(h_{ik}-h_{jk})^2}{h_{ik}+h_{jk}}\right]$, where b is the number of color histogram bins in each color space, and we use $b = 512$ in this paper. h_{ik} is the histogram value of superpixel X_i in the k-th bin.

To efficiently optimize the above energy minimization problem (1), instead of solving it via global optimization, we adopt the following way. First, with the similarity matrix $\Omega = [\omega_{ij}]$ defined above, we choose a threshold th based on statistics to determine the similar superpixel pairs by verifying the similarity. If a pair of superpixels X_i and X_j satisfy the similarity constraint $\omega_{ij} \geq th$, then X_i and X_j are determined as similar superpixels for following merge. Second, we merge similar superpixels X_i and X_j, and assign median value of pixels saliency scores inside both superpixels. The superpixels are merged in the order of similarities, *i.e.*, the most similar superpixels are merged first. Taking the spatial neighbouring relation into consideration and to keep sharp boundary in salient object detection, we only merge spatially adjacent superpixels. Third, to keep the strong foreground and strong background region intact, we assign those strong superpixels their original saliency scores. To exploit the multi-scale effect in salient object detection, the above refinement is conducted in each scale individually. Finally, the refined saliency prediction maps are fused to achieve the final results. We used equal weights to combine the three scale saliency maps. These weights can be efficiently learned by using one convolutional layer.

4. Experimental Results

4.1. Experimental Setup

Data set: We have evaluated the performance of our proposed method on 11 saliency benchmark datasets. We used 3,000 images from the MSRA10K dataset [13] for testing as the remaining 6000 have been used in training (1000 images

from the MRSA1K dataset constitute the ASD [1] dataset). Most of the images in this dataset contain one salient object. The ECSSD dataset [50] contains 1,000 images of semantically meaningful but structurally complex images, which makes this dataset very challenging. The DUT dataset [51] contains 5,168 images. The SOD saliency dataset [24] contains 300 images where many images have multiple salient objects with low contrast. SED1 and SED2 [4] are both small saliency datasets, and they all contain 100 images only. Images in the SED1 dataset contain one single salient object, while those in SED2 contains two salient objects. The PASCAL-S [34] dataset is generated from the PASCAL VOC dataset [16] and contains 850 images. HKU-IS [29] is a recently released saliency dataset with 4,447 images. THUR [12] has 6232 images which contains five object classes, *i.e.* "butterfly","coffee mug","dog jump","giraffe" and "plane". Finally, ICOSEG [5] is an interactive co-segmentation dataset, which contains 643 images with single or multiple salient objects in a single image.

Compared methods: We compared our method against 10 state-of-the-art deep learning based saliency detection methods: TIP [33], RFCN [47], DISC [11], DeepMC [55], LEGS [46], MDF [29], RACDN [27], ELD [28], SPCNN [19] and DC [30], and five traditional saliency detection methods: DRFI [24], RBD [56], DSR [32], MC [23], and HS [50], which were proven in [8] as the state-of-the-art before the era of deep learning. We have three alternate ways to obtain results of these methods. Firstly, we run the original codes provided by the authors if available. Secondly, we use the saliency maps provided in the paper. Thirdly, for those methods without code or saliency maps, we use the performance as listed in other papers.

Evaluation metric: For evaluation, we use three evaluation metrics, including mean absolute error (MAE), maximum F-measure, mean F-measure, as well as PR curve. MAE can provide a better estimate of the dissimilarity between the saliency map and the ground truth. It is the average per-pixel difference between ground truth and binary saliency map, normalized to [0, 1], which is defined as:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - GT(x,y)|, \quad (2)$$

where W and H are the width and height of the respective saliency map S, GT is the ground truth saliency map.

The F-measure (F_β) is defined as the weighted harmonic mean of precision and recall:

$$F_\beta = (1+\beta^2)\frac{Precision \times Recall}{\beta^2 Precision + Recall}, \quad (3)$$

where $\beta^2 = 0.3$. $Precision$ corresponds to the percentage of salient pixels being correctly detected, while $recall$ corresponds to the fraction of detected salient pixels in relation to the ground truth number of salient pixels. The Precision-Recall (PR) curves are obtained by binarizing the saliency map in the range of [0 255]. For a given dataset, we have mean Precision and mean Recall for this dataset which are both 256 dimensional vectors.

4.2. Comparison with State-of-the-art Methods

We compared our method with other 10 deep network based methods and 5 traditional methods, and the results are shown in Table 1. First conclusion we draw from Table 1 is that, deep learning based saliency detection methods achieve better performance on almost all the datasets, except for DRFI [24], which achieve better maximum F-measure on SED1 dataset than MDF [29], DeepMC [55] and RFCN [47]. Second conclusion is that by training image segmentation and saliency detection at the same time, TIP [33] achieve almost the second best performance, which indicates that the segmentation task can improve saliency detection task to some extent. The last but the most important conclusion is that our model "OURS" achieves the best performance on almost all the 11 datasets.

In Fig. 4, we show the Precision-Recall (PR) curve on four datasets (we did not compare PR curve on all the datasets because some saliency detection methods do not have saliency maps on all the 11 datasets). Our approach consistently outperforms other methods on all the datasets. For all the four datasets, our recalls are limited in a narrow range [0.7,1]. A higher and narrower range of recall means that our saliency maps achieve consistent good results under different thresholds in the range of [0,255]. For the other methods, recalls distribute in a wide range [0,1], which shows that the corresponding saliency maps are more likely to be gray images than ours. To further prove that our saliency maps are more likely to be binary, we compute the F-measure curve on the SOD dataset in Fig. 5. Fig. 5 shows that, given threshold in the range of [0,255], our F-measure is almost the same and it's getting better with higher threshold, which proves the robustness of our method. In Fig. 6, we compare the resultant saliency maps of different methods on a challenging image, which clearly demonstrates the robustness and the binary attribute of our saliency map.

4.3. Comparison of Different Models, Depths and Spatial Refinement

In this section, we analyze the influence of network model and network depth in our saliency detection method, as well as how the spatial refinement helps our final results. Specifically, we fine-tuned two variants of deep network, namely, a 50 layers Deep Residual Network ("Res50"), and the deep VGG net [45] ("VGG"). We compared these variants with the one we used, "Res101" and the results are reported in Table 2. The three numbers inside each cell are the maximum F-measure, the mean F-measure and the

Table 1. Performance comparisons with the state-of-the-art methods on 11 benchmarking datasets. Each cell: max F-measure (higher better) / mean F-measure (higher better) / MAE (lower better). Red, blue, green: the best, the second best, and the third best.

	MSRA	ASD	ECSSD	DUT	SED1	SED2	PASCALS	ICOSEG	HKU-IS	THUR	SOD
OURS	0.9248	0.9369	0.8756	0.7747	0.9213	0.8395	0.8070	0.8628	0.8634	0.7336	0.7808
	0.9136	0.9257	0.8621	0.7631	0.9104	0.8254	0.7953	0.8493	0.8476	0.7222	0.7549
	0.0435	0.0304	0.0765	0.0652	0.0558	0.1059	0.1054	0.0672	0.0655	0.0796	0.1296
TIP [33]	0.8994	0.9380	0.8095	0.7449	-	0.8632	0.8182	-	-	0.7276	0.7807
	0.8630	0.8932	0.7589	0.6045	-	0.7778	0.7310	-	-	0.6254	0.6978
	0.0628	0.0273	0.1601	0.0758	-	0.1074	0.1695	-	-	0.0854	0.1503
RFCN [47]	-	-	0.8570	0.7379	0.8923	0.8364	0.8029	0.8432	0.8917	0.7538	0.7728
	-	-	0.8340	0.6918	0.8467	0.7616	0.7468	0.8028	0.8277	0.7062	0.7426
	-	-	0.0973	0.0945	0.1020	0.1140	0.1176	0.0948	0.0798	0.1003	0.1394
DISC [11]	0.9042	-	0.8085	0.6616	0.8857	0.7816	-	0.7999	0.7859	-	-
	0.8634	-	0.7779	0.6091	0.8667	0.7452	-	0.7609	0.7368	-	-
	0.0544	-	0.1122	0.1182	0.0772	0.1203	-	0.1147	0.1023	-	-
DeepMC [55]	0.9246	0.9301	0.7756	0.7012	0.8962	0.8141	0.7677	0.7983	0.7966	0.6858	0.7179
	0.8963	0.9067	0.7369	0.6209	0.8674	0.7766	0.7177	0.7654	0.7610	0.6189	0.6880
	0.0426	0.0281	0.1623	0.0777	0.0810	0.1223	0.1888	0.1031	0.0919	0.0924	0.1552
LEGS [46]	-	-	0.8303	0.6677	0.8897	0.8031	0.7618	0.7571	0.7662	0.6638	0.7347
	-	-	0.7855	0.6265	0.8453	0.7357	0.7176	0.7093	0.7188	0.6301	0.6870
	-	-	0.1187	0.1318	0.0997	0.1251	0.1539	0.1269	0.1186	0.1242	0.1729
MDF [29]	-	-	0.8307	0.6944	0.8916	0.8432	0.7681	0.8376	-	0.6847	0.7381
	-	-	0.8097	0.6768	0.7888	0.7658	0.7389	0.7847	-	0.6670	0.6377
	-	-	0.1081	0.0916	0.1198	0.1171	0.1420	0.1008	-	0.1029	0.1669
SPCNN [19]	-	0.9019	0.6753	-	-	-	-	-	-	-	0.6222
	-	0.7979	0.5546	-	-	-	-	-	-	-	0.5159
	-	0.0903	0.2152	-	-	-	-	-	-	-	0.2176
RACDN [27]	0.9045	-	0.8696	-	-	0.8341	-	-	0.8564	0.7160	-
	0.8997	-	0.8555	-	-	0.8165	-	-	0.8516	0.7096	-
	0.0514	-	0.0813	-	-	0.1068	-	-	0.0636	0.0866	-
ELD [28]	-	0.9310	0.8674	0.7195	-	-	0.7775	-	-	0.7312	-
	-	0.8954	0.8372	0.6651	-	-	0.7538	-	-	0.6805	-
	-	0.0349	0.0805	0.0909	-	-	0.1206	-	-	0.0952	-
DC [30]	-	-	0.8879	0.7391	0.9045	0.8567	0.8050	0.8727	0.8853	0.7441	0.8219
	-	-	0.8315	0.6902	0.8564	0.7840	0.7528	0.8291	0.8205	0.6940	0.7603
	-	-	0.0906	0.0971	0.0886	0.1014	0.1246	0.0740	0.0730	0.0959	0.1208
DRFI [24]	-	-	0.7834	0.6638	0.8731	0.8265	0.6939	0.8108	0.7771	0.6823	0.6657
	-	-	0.6440	0.5525	0.7397	0.7252	0.5596	0.6986	0.6397	0.5440	0.5613
	-	-	0.1719	0.1496	0.1454	0.1373	0.2071	0.1397	0.1445	0.2046	0.1471
RBD [56]	0.8530	0.9108	0.7164	0.6261	0.8433	0.8264	0.6611	0.7942	0.7219	0.5950	0.6383
	0.7609	0.8361	0.6195	0.5486	0.7509	0.7330	0.5745	0.7071	0.6236	0.5248	0.5418
	0.1103	0.0688	0.1739	0.1467	0.1407	0.1316	0.1985	0.1310	0.1432	0.1507	0.2069
DSR [32]	0.8371	0.8837	0.7345	0.6261	0.8299	0.7852	0.6494	0.7658	0.7414	0.6125	0.6440
	0.7357	0.8186	0.6387	0.5583	0.7277	0.7053	0.5610	0.7002	0.6438	0.5498	0.5500
	0.1230	0.0834	0.1742	0.1374	0.1614	0.1457	0.2041	0.1491	0.1404	0.1408	0.2133
MC [23]	0.8489	0.9116	0.7416	0.6273	0.8502	0.7699	0.6675	0.7857	0.7234	0.6096	0.6493
	0.7257	0.8206	0.6114	0.5289	0.7319	0.6619	0.5510	0.6790	0.5900	0.5149	0.5332
	0.1457	0.0930	0.2037	0.1863	0.1620	0.1848	0.2296	0.1729	0.1840	0.1838	0.2435

Table 2. Performance comparison between different models and network depths. Each cell: max F-measure (higher better) / mean F-measure (higher better) / MAE (lower better). Red, blue, green: the best, the second best, and the third best.

	MSRA	ASD	ECSSD	DUT	SED1	SED2	PASCALS	ICOSEG	HKU-IS	THUR	SOD
OURS	0.9248	0.9369	0.8756	0.7747	0.9213	0.8395	0.8070	0.8628	0.8634	0.7336	0.7808
	0.9136	0.9257	0.8621	0.7631	0.9104	0.8254	0.7953	0.8493	0.8476	0.7222	0.7549
	0.0435	0.0304	0.0765	0.0652	0.0558	0.1059	0.1054	0.0672	0.0655	0.0796	0.1296
Res101	0.9245	0.9208	0.8925	0.7670	0.9183	0.8382	0.8304	0.8215	0.8769	0.7478	0.8027
	0.8702	0.8673	0.8387	0.7005	0.8741	0.7748	0.7732	0.7635	0.8118	0.6852	0.7469
	0.0541	0.0486	0.0797	0.0830	0.0678	0.0997	0.1095	0.0915	0.0689	0.0920	0.1282
Res50	0.8797	0.8835	0.8218	0.6778	0.8539	0.7549	0.7488	0.7735	0.8067	0.6775	0.7062
	0.8164	0.8203	0.7471	0.5883	0.7603	0.6133	0.6676	0.7074	0.7249	0.6071	0.5921
	0.0852	0.0728	0.1269	0.1069	0.1316	0.1469	0.1537	0.1206	0.1023	0.1150	0.1779
VGG	0.9069	0.9075	0.8646	0.7162	0.9038	0.8254	0.7872	0.7895	0.8473	0.7149	0.7704
	0.8535	0.8510	0.8159	0.6651	0.8639	0.7717	0.7414	0.7335	0.7845	0.6561	0.7248
	0.0608	0.0538	0.0894	0.0894	0.0728	0.1032	0.1233	0.1030	0.0779	0.1012	0.1369

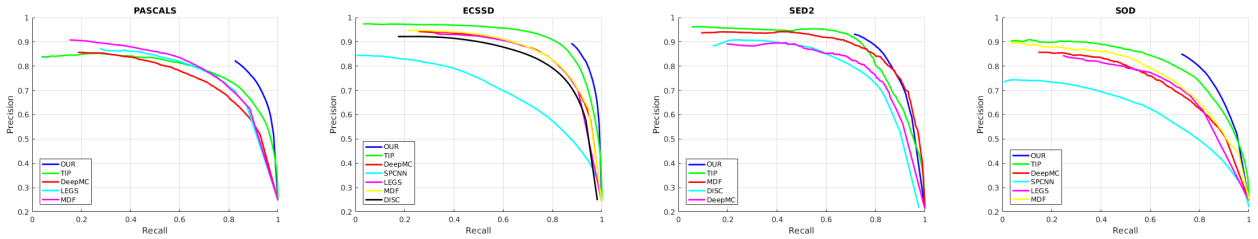


Figure 4. Comparison of Precision-Recall curves on four datasets. Our fully convolutional neural networks based multi-level fusion based approach consistently outperforms other methods on all the testing datasets.

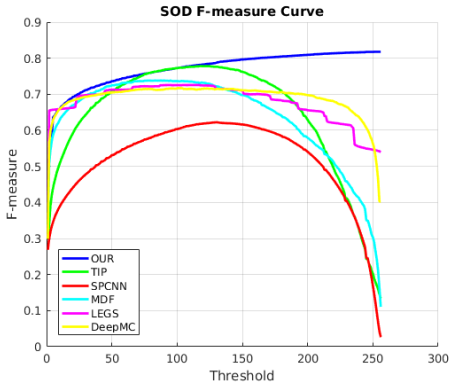


Figure 5. F-measure on the SOD dataset. It proves that our saliency maps are more likely to be binary.

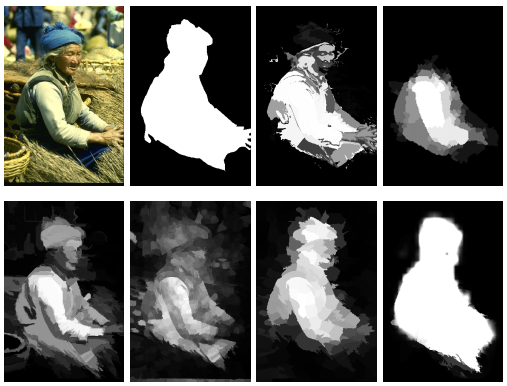


Figure 6. Saliency maps of different methods. The first row: original image, ground truth, DeepMC [55], LEGS [46]. The second row: MDF [29], SPCNN [19], TIP [33] and our saliency map.

MAE from top to bottom.

As demonstrated in Table 2, our final results ("OURS") with multi-scale fusion refinement achieve the best performance across all the 11 datasets, with almost 3% performance gain in mean F-measure and 1.5% performance improvement in MAE compared with "Res101", which is the direct coarse saliency prediction map. "Res101" achieves the second best performance, and VGG the third, Res50 the worst one. These results clearly illustrate the importance of deeper network and the contribution of our multi-level fusion in improving the saliency detection performance.

5. Conclusion

In this paper, building upon the success of deep neural networks, we propose a fully convolutional neural network based approach to salient object detection, which is empowered with multi-level fusion. We integrate saliency cues at different levels through fully convolutional neural networks and multi-level fusion, our approach could effectively exploit both learned semantic cues and higher-order region statistics for edge-accurate salient object detection. The retrained 101-layer Deep Residual Network with dilation/atrous convolution can predict much denser saliency maps. Spatial refinement process can assign consistent saliency values to regions with similar higher-order cues while keeping the strong foreground and strong background regions intact. Extensive experimental results on 11 benchmarking datasets prove that our model achieves the most accurate result with nearly binary saliency maps. Also, different from most of the existing deep learning based saliency detection approaches that often require more than 10,000 training images from more than one saliency datasets, our approach only uses 6,000 images from the MSRA10K dataset for training, and applies the same trained model for saliency prediction on other datasets, which further demonstrates the generalization ability of our approach.

Acknowledgment

This work was done when Jing Zhang was a visiting student to the Australian National University/NICTA supported by the China Scholarship Council (No: 201406290108). This work was supported in part by the Australian Research Council grants (DE140100180, DP150104645), and Natural Science Foundation of China grants (61420106007, 61671387). The first author would like to thank Prof. Mingyi He for his immeasurable support and encouragement and thank Bo Li and Jian Dong for valuable discussions.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1597–1604, June 2009. 2, 6

[2] 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.*, 34(11):2274–2282, Nov 2012. 4

[3] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2189–2202, Nov 2012. 2

[4] S. Alpert, M. Galun, A. Brandt, and R. Basri. Image segmentation by probabilistic bottom-up aggregation and cue integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):315–327, Feb 2012. 6

[5] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3169–3176, June 2010. 6

[6] A. Borji. What is a salient object? a dataset and a baseline model for salient object detection. *IEEE Trans. Image Proc.*, 24(2):742–756, Feb 2015. 1

[7] A. Borji, M. Cheng, H. Jiang, and J. Li. Salient object detection: A survey. *CoRR*, abs/1411.5878, 2014. 1, 2

[8] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *IEEE Trans. Image Proc.*, 24(12):5706–5722, 2015. 2, 6

[9] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai. Fusing generic objectness and visual saliency for salient object detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 914–921, 2011. 1, 2

[10] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014. 2, 3, 4

[11] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li. Disc: Deep image saliency computing via progressive representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 27(6):1135–1149, June 2016. 1, 2, 3, 5, 6, 7

[12] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu. Salientshape: group saliency in image collections. *The Visual Computer*, 30(4):443–453, 2014. 6

[13] 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.*, 37(3):569–582, 2015. 2, 5

[14] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook. Efficient salient region detection with soft image abstraction. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1529–1536, 2013. 1, 2

[15] M.-M. Cheng, G.-X. Zhang, N. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 409–416, 2011. 2

[16] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 6

[17] C. Gong, D. Tao, W. Liu, S. J. Maybank, M. Fang, K. Fu, and J. Yang. Saliency propagation from simple to difficult. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2531–2539, June 2015. 2

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 2, 3, 4

[19] S. He, R. W. H. Lau, W. Liu, Z. Huang, and Q. Yang. Supercnn: A superpixelwise convolutional neural network for salient object detection. *International Journal of Computer Vision*, 115(3):330–344, 2015. 2, 6, 7, 8

[20] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 262–270, Dec 2015. 3

[21] S. Jetley, N. Murray, and E. Vig. End-to-end saliency mapping via probability distribution prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 4

[23] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang. Saliency detection via absorbing markov chain. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1665–1672, Dec 2013. 1, 6, 7

[24] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2083–2090, 2013. 1, 2, 4, 5, 6, 7

[25] P. Jiang, H. Ling, J. Yu, and J. Peng. Salient region detection by UFO: Uniqueness, focusness and objectness. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1976–1983, 2013. 1, 2

[26] J. Kim, D. Han, Y.-W. Tai, and J. Kim. Salient region detection via high-dimensional color transform. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 883–890, 2014. 2

[27] J. Kuen, Z. Wang, and G. Wang. Recurrent Attentional Networks for Saliency Detection. *ArXiv e-prints*, Apr. 2016. 1, 2, 3, 6, 7

[28] G. Lee, Y. Tai, and J. Kim. Deep saliency with encoded low level distance map and high level features. *CoRR*, abs/1604.05495, 2016. 2, 3, 6, 7

[29] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5455–5463, June 2015. 1, 2, 4, 5, 6, 7, 8

[30] G. Li and Y. Yu. Deep contrast learning for salient object detection. *CoRR*, abs/1603.01976, 2016. 2, 3, 6, 7

[31] J. Li, R. Ma, and J. Ding. Saliency-seeded region merging: Automatic object segmentation. In *The First Asian Conference on Pattern Recognition*, pages 691–695, Nov 2011. 1

[32] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang. Saliency detection via dense and sparse reconstruction. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2976–2983, Dec 2013. 6, 7

[33] 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, Aug 2016. 1, 2, 3, 6, 7, 8

[34] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, June 2014. 6

[35] N. Liu and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[36] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1–8, 2007. 2

[37] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, June 2015. 1

[38] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2232–2239, 2009. 1

[39] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1139–1146, 2013. 2

[40] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2049–2056, 2006. 1

[41] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[42] G. Papandreou, L. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1742–1750, 2015. 2

[43] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 733–740, 2012. 2

[44] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 853–860, 2012. 2

[45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 6

[46] L. Wang, H. Lu, X. Ruan, and M. H. Yang. Deep networks for saliency detection via local estimation and global search. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3183–3192, June 2015. 1, 2, 5, 6, 7, 8

[47] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. *Saliency Detection with Recurrent Fully Convolutional Networks*, pages 825–841. Springer International Publishing, Cham, 2016. 1, 2, 6, 7

[48] L. Wang, J. Xue, N. Zheng, and G. Hua. Automatic salient object extraction with contextual cue. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 105–112, 2011. 2

[49] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee. Optimized scale-and-stretch for image resizing. In *ACM SIGGRAPH Asia*, pages 118:1–118:8, 2008. 1

[50] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1155–1162, 2013. 2, 6

[51] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3166–3173, 2013. 2, 6

[52] F. Yu and V. Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. *ArXiv e-prints*, Nov. 2015. 3

[53] M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 2018–2025, 2011. 2, 4

[54] G.-X. Zhang, M.-M. Cheng, S.-M. Hu, and R. R. Martin. A shape-preserving approach to image resizing. *Computer Graphics Forum*, 28(7):1897–1906, 2009. 1

[55] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1265–1274, June 2015. 1, 2, 3, 4, 6, 7, 8

[56] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2814–2821, 2014. 2, 6, 7