Saliency Detection by Multi-Context Deep Learning

Rui Zhao[1,2] Wanli Ouyang [2] Hongsheng Li [2,3] Xiaogang Wang[1,2]
[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]Department of Electronic Engineering, The Chinese University of Hong Kong
[3]School of Electronic Scicence, University of Electronic Science and Technology of China
{rzhao, wlouyang, hsli, xgwang}@ee.cuhk.edu.hk

Abstract

Low-level saliency cues or priors do not produce good enough saliency detection results especially when the salient object presents in a low-contrast background with confusing visual appearance. This issue raises a serious problem for conventional approaches. In this paper, we tackle this problem by proposing a multi-context deep learning framework for salient object detection. We employ deep Convolutional Neural Networks to model saliency of objects in images. Global context and local context are both taken into account, and are jointly modeled in a unified multi-context deep learning framework.

To provide a better initialization for training the deep neural networks, we investigate different pre-training strategies, and a task-specific pre-training scheme is designed to make the multi-context modeling suited for saliency detection. Furthermore, recently proposed contemporary deep models in the ImageNet Image Classification Challenge are tested, and their effectiveness in saliency detection are investigated. Our approach is extensively evaluated on five public datasets, and experimental results show significant and consistent improvements over the state-of-the-art methods.

1. Introduction

Saliency detection, aiming at highlighting visually salient regions or objects in an image, has been a fundamental problem drawing extensive attentions in recent years. It has a wide range of applications in computer vision and image processing tasks, such as image/video compression and summarization [38], content-aware image resizing [6], and photo collage [53]. Saliency information has also been exploited in high-level vision tasks, such as object detection [37], and person re-identification [62, 61]. A large number of approaches [63, 52, 40, 39, 32, 35, 60, 57, 56, 47, 41, 31, 27, 25, 24, 23, 11, 44, 17, 8, 13, 1, 21] are proposed to capture different saliency cues.

Many conventional saliency detection methods focus on design of low-level saliency cues, or modeling background priors. There are noticeable problems in these methods. 1) Computational saliency models need effective feature representations to estimate saliency, but sometimes the contrast between hand-crafted low-level features cannot help salient objects stand out from context. 2) Moreover, contrast is not only in terms of difference between visual cues, but also relates to high-level cognition and understanding. For example in Figure 1, a dark gray house appears in dark yellow bush. Objects like the house cannot be classified as salient objects from the low-contrast background either based on low-level saliency cues or background priors, but they are semantically salient in high-level cognition, *i.e.* they are distinct in object categories. Therefore, ***saliency detection is considered as a high-level task in our work***.

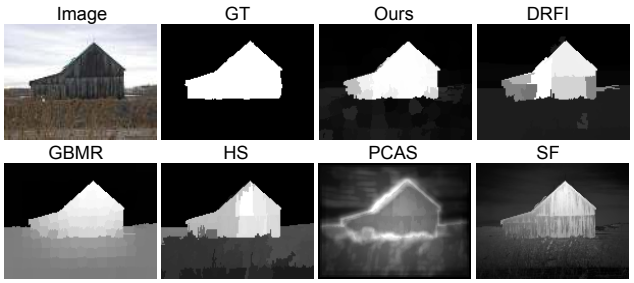

Figure 1. Examples to show problems in conventional approaches. From top left to bottom right: image, groundtruth mask, our saliency maps, and saliency maps of other five latest approaches, including DRFI [25], HS [56], GBMR [57], PCAS [41], and SF[44].

The deep Convolutional Neural Network (CNN) [30], which recently showed its powerfulness in extracting high-level feature representations [16], can well solve aforementioned problems. From another perspective, saliency detection is a task to simulate the mechanism of human attention, which is a neurocognitive reaction controlled by human brains. Deep CNN aims to mimic the functions of neocortex in human brain as a hierarchy of filters and non-linear operations. For better detecting semantically salient objects, high-level knowledge on object categories becomes important. Suppose that if the deep model can recognize the gray house, then the problems in Figure 1 can be easily solved. As indicated in [49], pre-training can provide

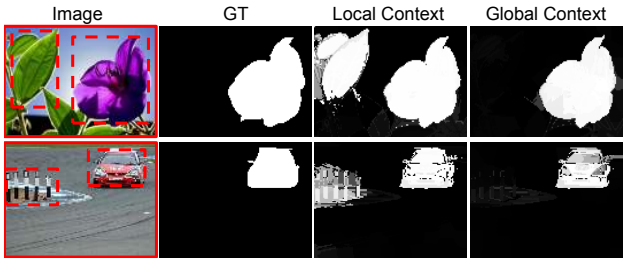

Figure 2. Examples to show importance of global context. From left to right: image, groundtruth saliency mask, our saliency map predicted with local context, and our saliency map predicted with global context.

a good initialization for training deep models, and is able to preliminarily memorize some related high-level information. Therefore, it is desirable to see the influence of pre-training in modeling saliency.

An appropriate scope of context is also very important to help a salient object stand out from its context meanwhile keep those non-salient objects suppressed in background. In Figure 2, high-level knowledge tells us information about flowers, leaves, cars and guard fences, but cannot answer which are salient objects. If a local context (*e.g*. the red dashed boxes in Figure 2(a)) is adopted to determine the saliency, then all these object are highlighted as salient objects, as shown in Figure 2(c). This becomes general object detection and segmentation problems. Due to the extremely large variation in positions and scales of objects of different categories, a global context (including the full image) is more suitable to determine object saliency. Because a global context takes all objects in an image into account, and only with a global context we can model the contrast between all objects. As shown in Figure 2(d), if the flower and the leaf are considered together, then only the flower is classified as the salient object; if the car and the guard fence are considered as in the same picture, then only the car is detected as the salient object. In addition, it is known that deep models are also powerful in learning global contextual feature representations.

Based on the above motivations, a new multi-context deep learning framework for saliency detection is proposed. Our work has two major **contributions**:

First, a deep model with multiple contexts is designed to capture object saliency. The global context is utilized to model saliency in full image, while the local context is used for saliency prediction in meticulous areas. The global and local context are integrated into the multi-context deep learning framework for saliency detection, and the global- and local-context modeling are jointly optimized.

Second, we explore the influence of different pre-training strategies, and introduce a task-specific pre-training scheme to pre-train the deep models using the ImageNet image classification dataset. In addition, several contemporary deep architectures in ImageNet Image Classification Challenge are tested, and their effectiveness in saliency detection are investigated.

2. Related Work

2.1. Salient Object Segmentation

Salient object segmentation approaches can be roughly categorized into two groups: bottom-up methods and top-down methods.

Bottom-up methods can be further divided into two categories, *i.e*. local and global. Local approaches (*e.g*. [22, 17, 36]) design saliency cues by considering the contrast between each image element (pixel, region, or patch) and its locally surrounding neighborhood. Global approaches estimate saliency scores by calculating the holistic statistics on uniqueness of each image element over the whole image. Cheng *et al*. [13, 10] used 3D color histograms as regional features to compute global contrast with all image regions. Perazzi *et al*. [44] applied two measures of contrast that rate the uniqueness and the spatial distribution to derive image saliency. However, these global features are weak in capturing semantic information.

Top-down methods take advantages of high-level category-specific information as prior knowledge, and are usually task-dependent. Judd *et al*. [28] learned a top-down saliency model object detectors such as faces, humans, animals, and text. Borji *et al*. [7] combine bottom-up saliency cues with top-down features learned via multiple object detectors. Yang *et al*. [58] proposed a top-down saliency model by jointly learning a Conditional Random Field and a dictionary. These methods explore high-level information from $3 \sim 5$ object categories. However, our deep models encodes prior knowledge on $1,000$ object classes from ImageNet, and has much stronger generalization capability.

In addition, some other interesting priors were also proposed to assist saliency detection, such as flash cues [18], boundary and background priors [55, 57, 25, 63].

2.2. Objectness and Object Proposal

Objectness was introduced to measure how likely a region contains an object regardless of object categories. Alexie [3, 4] proposed to combine local appearance contrast and boundary characteristics to measure the objectness score of a bounding box. Based on such measures, some *object proposal* methods [9, 64] further generated candidate object regions as a preprocessing step for object detection, which can effectively speed up the process comparing to the classical sliding-window detection paradigm. A recent work [54] proposed to extract generic objects by jointly handling localization and segmentation tasks.

Different than *object proposal*, which enumerates preliminarily likely candidates for object detection regardless of their contrastive relations, saliency detection requires to take the context of full images and the contrast between objects into account. Also, objectness score in [3, 4] or

the ranking score in [9] was measured over a candidate bounding box, which can only provide a rough score map highlighting all possible objects. Contrarily, saliency detection aims to produce accurate segmentation over the salient ones. Despite the difference, objectness score can be used as high-level prior knowledge [23], which could be further combined with with low-level saliency cues for saliency detection.

2.3. Hierarchical Structure for Saliency Detection

Latest works on saliency detection have showed the trend of using deep/hierachical architectures to model visual saliency. Yan *et al*. [56] presented a hierarchical framework to reduce the influence of small-scale structures in saliency detection. Lin *et al*. [34] proposed to unsupervisedly learn a set of mid-level filters to capture local contrast, and to fuse multi-level saliency calculation by convolution. Unlike their methods where mid-level filters are handcrafted, filters of CNNs in our framework are automatically and jointly learned in a discriminative manner. Jiang *et al*. [26] introduced successive Markov random fields (sMRF) to model visual saliency, which shared the similar spirit as this work of mimicing the deep propagation (a chain of synaptic communications) along visual cortex. However, the sMRF is a hierachical graphical model, which is a generative model optimized by belief propagation, while visual saliency in our work is computed in a discriminative model optimized by stochastic gradient descent.

2.4. Deep Convolutional Neural Networks

Since the introduction by LeCun Yann [30], deep CNN has been applied to a wide range of computer vision tasks such as hand-written digit classification and face detection. Recently, the latest generation of CNNs have substantially outperformed handcrafted approaches in computer vision field. Notably, best performing entries [29, 46, 59, 50, 42, 43] on ImageNet ILSVRC [14] and PASCAL VOC [15] benchmarks are all variants of deep CNNs since 2012.

Some recent approaches close to our work included cascaded stages in deep learning to solve problems that need meticulous refinement, such as in facial landmark detection [48] and human pose estimation [51]. Saliency detection also need such refinement since global-context model cannot well capture the very detailed information in local neighborhoods. However, we propose a multi-context deep model to consolidate both global context and local context in a unified framework.

3. Our Approach

In this paper, we propose a multi-context deep learning framework for saliency detection, and focus on modeling saliency with global context and local context simultaneously. Furthermore, different pre-training strategies are investigated, and an effective task-specific pre-training scheme is introduced. Figure 3 shows an overview of our approach.

3.1. Global-context Modeling by Deep CNN

As shown in Figure 3, the upper branch (global-context modeling) of our saliency detection pipeline is a deep CNN architecture with global and coarse context. Superpixel segmentation is firstly performed on images using the SLIC [2] method, and the input of global-context CNN is a superpixel-centered large context window including the full image. Regions exceeding image boundaries are padded with mean pixel value of the training dataset. The padded image are then warped to $227 \times 227 \times 3$ as input, where the three dimensions represent width, height, and number of channels. With this proposed normalization and padding scheme, the superpixel to be classified is always located at the center of the image, and the spatial distribution of the global context is normalized in this way. Moreover, it ensures the input covers the whole range of the original image. The last layer of the network structure has 2 neurons followed by a softmax function as output, indicating the probabilities of centered superpixel whether being in background or belonging to a salient object.

The winning model in the classification task of ImageNet 2013, *i.e*. the Clarifai model [59], is adopted as our baseline model. The Clarifai model contains 5 convolutional layers and 2 fully connected layers, as shown in Figure 3. Denote by $conv\#$ a convolutional layer, by $lrn\#$ a local response normalization layer, $pool\#$ a pooling layer and by $fc\#$ a fully connected layer. $conv\#$ and $fc\#$ layers consist of a linear transformation followed by a nonlinear rectified linear unit function denoted by $relu\#$, and only $conv\#$ and $fc\#$ layers have learnable parameters. The structure of the network can be described by the size of feature maps at each layer as $conv1$ $(111 \times 111 \times 96)$ $-relu1$ $-pool1$ $-lrn1$ $-conv2$ $(27 \times 27 \times 256)$ $-relu2$ $-pool2$ $-lrn2$ $-conv3$ $(13 \times 13 \times 384)$ $-relu3$ $-conv4$ $(13 \times 13 \times 384)$ $-relu4$ $-conv5$ $(13 \times 13 \times 256)$ $-relu5$ $-pool5$ $-fc6(4096)$ $-relu6$ $-dropout6$ $-fc7(4096)$ $-relu7$ $-dropout7$ $-fc8(2)$. For $conv$ layers, the size of feature maps is defined as width$\times$height$\times$depth, where the first two dimensions describe the spatial size and the depth defines the number of channels. Pooling is applied after three layers. The total number of parameters in the above model is about 58 million. We refer readers to [59] for further details.

Apart from the Clarifai model, there are also other contemporary models such as AlexNet [29], NIN [33], OverFeat [46], DeepID-Net [42], and GoogLeNet [50]. It is flexible to incorporate any of these contemporary deep models into our framework, and in the experimental section we investigate the performance of saliency detection using some of these contemporary architectures.

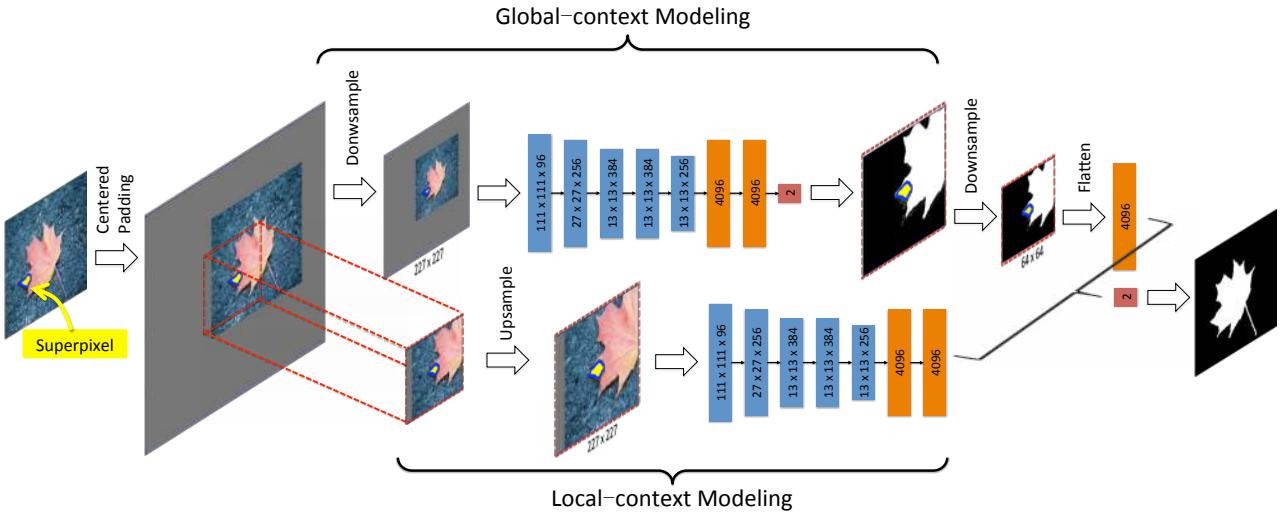

Figure 3. Upper branch: Deep CNN-based global-context modeling for saliency detection with a superpixel-centered window padded with mean pixel value. Lower branch: local-context modeling with a closer-focused superpixel-centered window, and global-context saliency detection results are combined into finally fully-connected layer in the local-context model. We visualize the network layers with their corresponding dimensions, where convolutional layers are in blue, fully connected layers (with parameters initialized using pre-trained model parameters) in orange, and fully connected layers (with parameters randomly initialized) in red. Layers without parameters are omitted in this figure.

3.2. Integrated Multi-context Model

While the CNN at the upper branch (global-context model) aims to robustly model saliency with few large errors, CNN at the lower branch are designed to look at details - it focuses on a smaller context to refine the saliency prediction of the centered superpixel. In this work, the local-context model takes an input with a similar form as in the global-contex model, but with one third of the scope of context, and then normalized to $227 \times 227 \times 3$. The local-context model shares the same deep structure with the global-context model, but with independent parameters. Other deep structures can also be flexibly incorporated in the local-context model. Overall, prediction of a superpixel-centered input window is performed by estimating the saliency probability

$$\mathbf{score}(\mathbf{x}_{gc},\ \mathbf{x}_{lc}) = P(y = 1 \mid \mathbf{x}_{gc},\ \mathbf{x}_{lc};\ \theta_1), \tag{1}$$

where $\mathbf{x}_{gc}$ and $\mathbf{x}_{lc}$ are output of the penultimate layer of the global context model and the local context model respectively. y is the prediction of saliency for the centered superpixel, where $y = 1$ for salient superpixel and $y = 0$ for background.

We train a binary classifier on top of the last network layer to classify background and saliency by minimizing a unified softmax loss between the classification result and the groundtruth label.

$$L(\ \theta;\ \{\mathbf{x}_{gc}^{(i)}, \mathbf{x}_{lc}^{(i)}, y^{(i)}\}_{i=1}^{m}) = \\ -\frac{1}{m} \sum_{\substack{i\in\{1,\ldots,m\} \\ j\in\{0,1\}}} \mathbf{1}_{\{y^{(i)}=j\}} \log P(y^{(i)} = j \mid \mathbf{x}_{gc}^{(i)},\ \mathbf{x}_{lc}^{(i)}; \theta_j), \tag{2}$$

In our approach, the parameters in our framework can be decomposed to several parts, *i.e.* $\theta_j = \{\ \mathbf{w}_{gc,j},\ \mathbf{w}_{lc,j},\ \alpha,\ \beta\}$, where $\mathbf{w}_{gc,j}$ are last-layer parameters in the neural network for global-context modeling, $\mathbf{w}_{lc,j}$ are last-layer parameters for local-context modeling, and α, β are parameters of an ambiguity modeling function controlling the need of local-context modeling. Thus, the posterior probability in Eq.(2) is factorized into product of experts [19], *i.e.* we aim to infer the label probability via two components simultaneously:

$$P(y = j \mid \mathbf{x}_{gc}, \mathbf{x}_{lc};\ \theta_j) \propto \Phi(\mathbf{x}_{gc};\ \theta_j^{\Phi}) \cdot \Psi(\mathbf{x}_{gc}, \mathbf{x}_{lc}; \theta_j^{\Psi}), \tag{3}$$

$$\theta_j^{\Phi} = \mathbf{w}_{gc,j},\ \theta_j^{\Psi} = \{\ \mathbf{w}_{gc,j},\ \mathbf{w}_{lc,j},\ \alpha,\ \beta\},\ j \in \{0,\ 1\}. \tag{4}$$

Specifically, Φ tries to estimate the saliency probability based on global-context modeling,

$$\Phi(\mathbf{x}_{gc};\ \theta_j^{\Phi}) \propto e^{\mathbf{w}_{gc,j}^{\mathrm{T}} \mathbf{x}_{gc}}, \tag{5}$$

and Ψ is based on both the global context and local context,

$$\Psi(\mathbf{x}_{gc}, \mathbf{x}_{lc}; \theta_j^{\Psi}) \propto e^{f_u(\alpha \mathbf{w}_{gc,j}^{\mathrm{T}} \mathbf{x}_{gc} + \beta) \cdot \mathbf{w}_{lc,j}^{\mathrm{T}} x_{lc}}. \tag{6}$$

Then, the corresponding unnormalized saliency prediction score function is formulated as

$$f(\mathbf{x}_{gc}, \mathbf{x}_{lc};\ \theta_1^{\Psi}) = \mathbf{w}_{gc,1}^{\mathrm{T}} \mathbf{x}_{gc} + f_u(\alpha \mathbf{w}_{gc,1}^{\mathrm{T}} \mathbf{x}_{gc} + \beta) \mathbf{w}_{lc,1}^{\mathrm{T}} \mathbf{x}_{lc}, \tag{7}$$

where $f_u(\cdot)$ is defined as

$$f_u(t) = \begin{cases} t & \text{for} \quad 0 \le t \le 1 \\ 0 & \text{for} \quad \text{otherwise} \end{cases} \tag{8}$$

$f_u(\alpha \mathbf{w}_{gc,1}^{\mathrm{T}} \mathbf{x}_{gc} + \beta)$ models ambiguity of the saliency prediction of the global context model, and α and β can integrate multiple contexts in modeling to perform saliency detection from a joint model. Intuitively, $\alpha \mathbf{w}_{gc,1}^{\mathrm{T}} \mathbf{x}_{gc} + \beta$ in range $(-\infty, 0] \cup [1, +\infty)$ leads to a zero $f_u(\alpha \mathbf{w}_{gc,1}^{\mathrm{T}} \mathbf{x}_{gc} + \beta)$, which

means it has a high-confidence prediction in the global-context model so that $f(\mathbf{x}_{gc}, \mathbf{x}_{lc};\ \theta_1^{\Psi})$ only relies on the global context information. Properly setting α and β incorporate a non-zero $f_u(\alpha \mathbf{w}_{gc,1}^{\mathrm{T}} \mathbf{x}_{gc} + \beta) \cdot \mathbf{w}_{lc,1}^{\mathrm{T}} \mathbf{x}_{lc}$ (weighted local-context modeling) to handle ambiguous predictions with low confidence in global-context modeling.

To this end, our problem can be formulated as minimizing the following loss function:

$$\begin{aligned}
&\underset{\substack{\{\mathbf{w}_{gc,j}, \mathbf{w}_{lc,j}\}_{j=0}^{1}, \\ \alpha, \beta}}{\operatorname{argmin}} L(\{\mathbf{w}_{gc,j}, \mathbf{w}_{lc,j}\}_{j=0}^{1}, \alpha, \beta;\ \{\mathbf{x}_{gc}^{(i)}, \mathbf{x}_{lc}^{(i)}, y^{(i)}\}_{i=1}^{m}) = \\
&-\frac{1}{m} \sum_{\substack{i \in \{1,\ldots,m\} \\ j \in \{0,1\}}} \mathbf{1}_{\{y^{(i)}=j\}} \log \frac{e^{\mathbf{w}_{gc,j}^{\mathrm{T}} \mathbf{x}_{gc}^{(i)} + f_u(\alpha \mathbf{w}_{gc,j}^{\mathrm{T}} \mathbf{x}_{gc}^{(i)} + \beta) \mathbf{w}_{lc,j}^{\mathrm{T}} \mathbf{x}_{lc}^{(i)}}}{\sum_l e^{\mathbf{w}_{gc,l}^{\mathrm{T}} \mathbf{x}_{gc}^{(i)} + f_u(\alpha \mathbf{w}_{gc,l}^{\mathrm{T}} \mathbf{x}_{gc}^{(i)} + \beta) \mathbf{w}_{lc,l}^{\mathrm{T}} \mathbf{x}_{lc}^{(i)}}} \\
&+\lambda_1 \sum_{j \in \{0,1\}} \|\mathbf{w}_{gc,j}\|_2^2 + \lambda_2 \sum_{j \in \{0,1\}} \|\mathbf{w}_{lc,j}\|_2^2,
\end{aligned} \tag{9}$$

where the parameters are simultaneously optimized with other layers' parameters by backpropagating the loss.

3.3. Task-specific Pre-training and Fine-tuning

It was demonstrated in [16] that fine-tuning a Deep CNN model pre-trained for image classification with the target task (*e.g.* object detection) data can significantly improve the performance of target task. Particularly, deep model structures at the pre-training and fine-tuning stages are only different in the last fully connected layer for predicting labels. Except for the last fully connected layers for classification, the parameters learned at the pre-training phase are directly used as initial values for the fine-tuning stage.

Similar strategy in [16] can be directly used to fine-tune the contemporary CNN models for saliency detection. However, the pre-training task and fine-tuning task have disparity in following aspects. 1) **Input data**. Image classification task takes full images as inputs, while our global-context model requires superpixel-centered windows padded with mean pixel value, and the local-context model takes a cropped input, which serves to provide local context for finer prediction. Both the input of the global- and local-context models have changed scales and translations compared to the original images, which leads our multi-context model to learning different feature representations. 2) **Class labels**. Dataset in ImageNet for Image classification has $1,000$ classes, while saliency detection solves a binary classification problem. Despite the disparity in class labels, it is shown that deep CNN pre-trained for $1,000$-class classification can be generalized for fine-tuning the classification problem with fewer classes [16]. 3) **Loss function**. The loss function in image classification task aims to differentiate $1,000$ classes, while the loss function for saliency detection in our approach is defined as in Eq. (9) to perform binary classification.

To apply the contemporary models like the Clarifai model to our problem, the disparities mentioned above need to be considered. Therefore, we explore several pre-training strategies, and propose task-specific initialization for fine-tuning our deep saliency detection models. Task-specific pre-training has been proved to be very effective on object detection [42].

Pre-training We pre-train our models using image data from ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014 [45]. This challenge contains two different datasets: 1) the classification and localization dataset and 2) the detection dataset. The classification and localization dataset is divided into three subsets, *train*, *validation*, and *test* data. We use the *train* dataset for pre-training, which contains 1.2 million images with labels of $1,000$ categories. In addition, labels are provided at the image-level and the object-level, *i.e.* category labels are available for both full image and object bounding boxes.

Based on the object-level annotations, we can easily generate another type of annotations to suit the input format in saliency detection, which we call the superpixel-level annotation. Specifically, we randomly sample a superpixel within a object bounding box (where the superpixel is most likely located within the object region), and produce a superpixel-centered window including full image, also padded with mean pixel value in ImageNet training data. The label of each window is determined by thresholding the overlap ratio between the centered superpixel and corresponding groundtruth salient object mask. In our experiments, the threshold is set to 0.5.

We investigate following strategies for pre-training:

Strategy 1. No pre-training, *i.e.* randomly initializing model parameters for fine-tuning.

Strategy 2. Pre-training the deep models using training images with image-level annotations of $1,000$ classes.

Strategy 3. Pre-training the deep models using training images with object-level annotations of $1,000$ classes.

Strategy 4. Pre-training the deep models using training images with superpixel-level annotations of $1,000$ classes.

In Table 1, we show two settings of pre-training schemes. The R-CNN [16] for object detection and segmentation adopted **strategy 2** in training (denoted by $S1$). Different from R-CNN, a new task-specific scheme (denoted by $S2$) with pre-training strategies based on superpixel-level and object-level annotation are proposed for the global- and local-context modeling. Superpixel-level annotation aligns spatial location of objects when pre-training the global-context model, and it is consistent with the input format of the global-context model. Features pre-trained with object-level annotation are sensitive to the location of objects, and they provide more appropriate pre-training information for the local-context model.

In Section 4.3, we quantitatively investigate the influences of different pre-training strategies in saliency detection, and validate our hypothesis that task-specific pre-

training strategy provides a better initialization for fine-tuning.

Pre-training strategy	global context	local context
S1: R-CNN [16]	*strategy 2*	*strategy 2*
S2: Task-specific	*strategy 4*	*strategy 3*

Table 1. Pre-training strategies used for comparison.

Fine-tuning. We use the MSRA10k dataset [10] for fine-tuning our deep saliency detection models. The MSRA10k dataset is a subset of the MSRA Salient Object Dataset [36], which originally provides salient object annotation in terms of bounding boxes provided by 3-9 users. Cheng *et al*. [10] selected $10,000$ images with consistent bounding box labeling in MSRA dataset, and provided pixel-level saliency annotations. We randomly select $8,000$ images for training, and $2,000$ for validation. From each image, we select an average $200 \sim 300$ of superpixels, and in total about 2.1 million input windows for training and 0.6 million for validation are generated. Fine-tuning for $100,000$ iterations costs 31 hours on a PC with Intel I7 3.6GHz GPU, 32GB RAM and a GTX TITAN GPU. Testing an image with 200 superpixel takes about 0.8 seconds only using global-context model, and 1.6 seconds using the unified multi-context model.

4. Experimental Results

4.1. Benchmark Datasets

ASD [1] includes $1,000$ images sampled from the MSRA Salient Object Database [36]. Although our training data originates from the same dataset, we separate images in ASD dataset from our training set to avoid overlap.

SED1 [5] contains 100 images of a single salient object annotated manually by three users.

SED2 [5] contains 100 images of two salient objects annotated manually by three users.

ECSSD [56] contains $1,000$ structurally complex images acquired from the Internet, and the groundtruth masks were annotated by five labelers.

PASCAL-S [32] was built on the validation set of the PASCAL VOC 2010 segmentation challenge. It contains 850 natural images with both saliency segmentation groundtruth and eye fixation groundtruth. Saliency groundtruth masks were labeled by 12 subjects.

Evaluation Metrics[1]. We follow the evaluation protocol as in [1, 13, 32], where saliency maps are binarized at every threshold within range $[0, 255]$, and all saliency maps are evaluated by the F-measure score [1], which is obtained as a harmonic mean of average precision and average recall, *i.e.* $F_\beta = \frac{(1+\beta^2)\times precision \times recall}{\beta^2 \times precision + recall}$, where β^2 is set to 0.3 following the convention as in [1, 56, 32].

[1]We use the code provided by [32] at http://cbi.gatech.edu/salobj/ for evaluation of our results on all the five datasets.

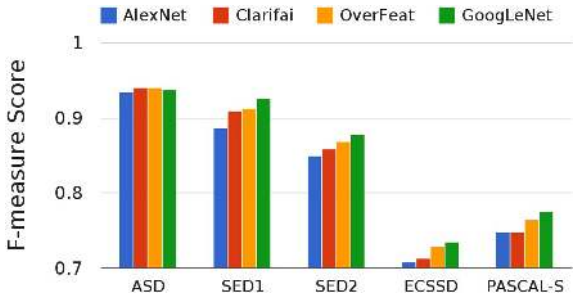

Figure 6. Saliency detection performance (F-measure score) using contemporary deep models, including AlexNet[29], Clarifai [59], OverFeat [46], and GoogLeNet [50].

4.2. Evaluation on the Multi-context Model

We separate the global-context branch in our framework as a baseline model, and we call it as the single-context model. It is also trained and tested under the same experimental settings for comparison. As shown in Figure 4(a), our proposed multi-context model consistently outperforms the single-context model on all the five datasets. Especially on the PASCAL dataset, our multi-context model increases the F-measure score by around 5%. Some examples of the saliency maps are shown in Figures 4(d-e), and it is clearly shown that the multi-context model refines the erroneous predictions by the single-context model since it combines both global context and local context.

4.3. Evaluation on Task-specific Pre-training

We evaluate the performance of the single-context model with different pre-training strategies on the ECSSD dataset and PASCAL-S dataset since evaluations on larger datasets show more robust statistics. As shown in Figures 5(a-b), evaluation results on both datasets have similar characteristics: 1) random initialization (**strat.1**) of network parameters leads to the worst performance; 2) image-level pre-training strategy (**strat.2**) and object-level pre-training strategy (**strat.3**) obtain similar results on the ECSSD dataset, while image-level pre-training slightly outperforms object-level pre-training on the PASCAL-S dataset; 3) superpixel-level pre-training strategy (**strat.4**) outperforms other pre-training strategies on both datasets.

Then we evaluate the performance of task-specific pre-training strategies for the whole framework. As shown in Table 1, we test two settings, the $S1$ follows the scheme used in R-CNN [16], while the second one $S2$ is our task-specific pre-training scheme introduced in Section 3.3. From the results on five datasets in Figure 5(c), we can conclude that our task-specific pre-training scheme consistently outperforms the conventional pre-training method adopted in R-CNN, which validates the effectiveness of the proposed task-specific pre-training approach.

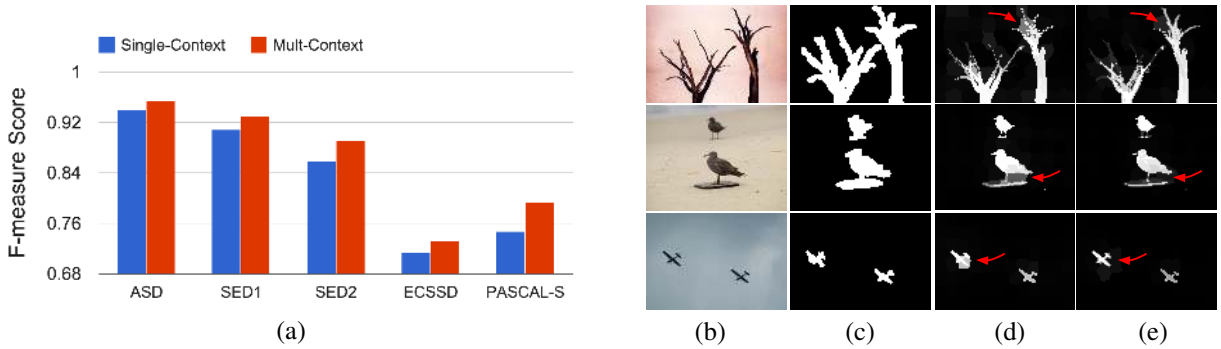

Figure 4. (a): F-measure scores on the five saliency detection datasets for evaluation of the single-context model and the multi-context model. (b-e): Qualitative comparison between the single-context model and the multi-context model. (b) is input image. (c) is ground-truth saliency map. (d) is our saliency map from the single-context model. (e) is our saliency map from the multi-context model. Red arrows indicate that regions with erroneous prediction by single-context model are refined by the multi-context model.

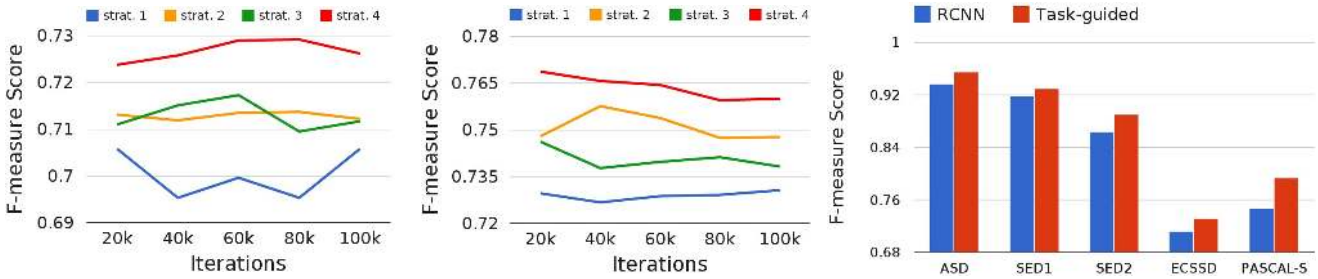

Figure 5. Evaluation of the single-context model with different pre-training strategies on (a) ECSSD dataset and (b) PASCAL-S dataset. The abbreviations "strat. #" in (a-b) represent the strategies introduced in Section 3.3. (c) Evaluation of multi-context model learned using task-specific pre-training scheme on five datasets.

4.4. Evaluation on Contemporary Deep Structures

Our framework is flexible to incorporate other contemporary deep models, and for simplicity we replace the model structure in the global-context model with other contemporary model structures for evaluation. Evaluated structures include AlexNet [29], Clarifai [59], OverFeat [46], and GoogLeNet [50]. As shown in Figure 6, GoogLeNet[2] slightly outperforms other deep models on the five dataset, but performances of these contemporary deep models do not vary very much. We expect extra performance gain if GoogleNet is used in our multi-context model with task-specific pre-training scheme, as introduced in Section 3.3.

4.5. Evaluation on Overall Performance

In Table 2, we compare our approach with nine latest state-of-the-art methods, including IS [20], GBVS [17], SF [44], GC [12], CEOS [40], PCAS [41], GBMR [57], HS [56], and DRFI [25]. Our approach significantly outperforms all the state-of-the-art salient object segmentation algorithms. The PASCAL-S dataset was proposed in CPMC-GBVS [32], but we do not include their method in Table 2 for a fair comparison because they used eye fixation label in training. Our approach obtains a higher F-measure score than theirs (0.7930 vs. 0.7457) on PASCAL-S dataset. Also, we qualitatively compare our saliency maps with those by other methods in Figure 7. It is obvious that our approach is able to highlight the salient object parts more coherently, and has a better prediction especially in complex scene with confusing background, such as the cases in the 6^{th} and 7^{th} rows in Figure 7. More comparisons can be found at our project website.

	ASD	SED1	SED2	ECSSD	PASCAL-S
IS [20]	0.5943	0.5540	0.5682	0.4731	0.4901
GBVS [17]	0.6499	0.7125	0.5862	0.5528	0.5929
SF [44]	0.8879	0.7533	0.7961	0.5448	0.5740
GC [12]	0.8811	0.8066	0.7728	0.5821	0.6184
CEOS [40]	0.9020	0.7935	0.6198	0.6465	0.6557
PCAS [41]	0.8613	0.7586	0.7791	0.5800	0.6332
GBMR [57]	0.9100	0.9062	0.7974	0.6570	0.7055
HS [56]	0.9307	0.8744	0.8150	0.6391	0.6819
DRFI [25]	0.9448	0.9018	0.8725	0.6909	0.7447
Ours	**0.9548**	**0.9295**	**0.8903**	**0.7322**	**0.7930**

Table 2. The F-measure scores of benchmarking approaches on five public datasets.

5. Conclusion

In this paper, we propose a multi-context deep learning framework for saliency detection. Firstly, we intro-

[2]Our implementation of GoogLeNet is pre-trained with less extensive data augmentation, and gets 67% top-1 ILSVRC accuracy.

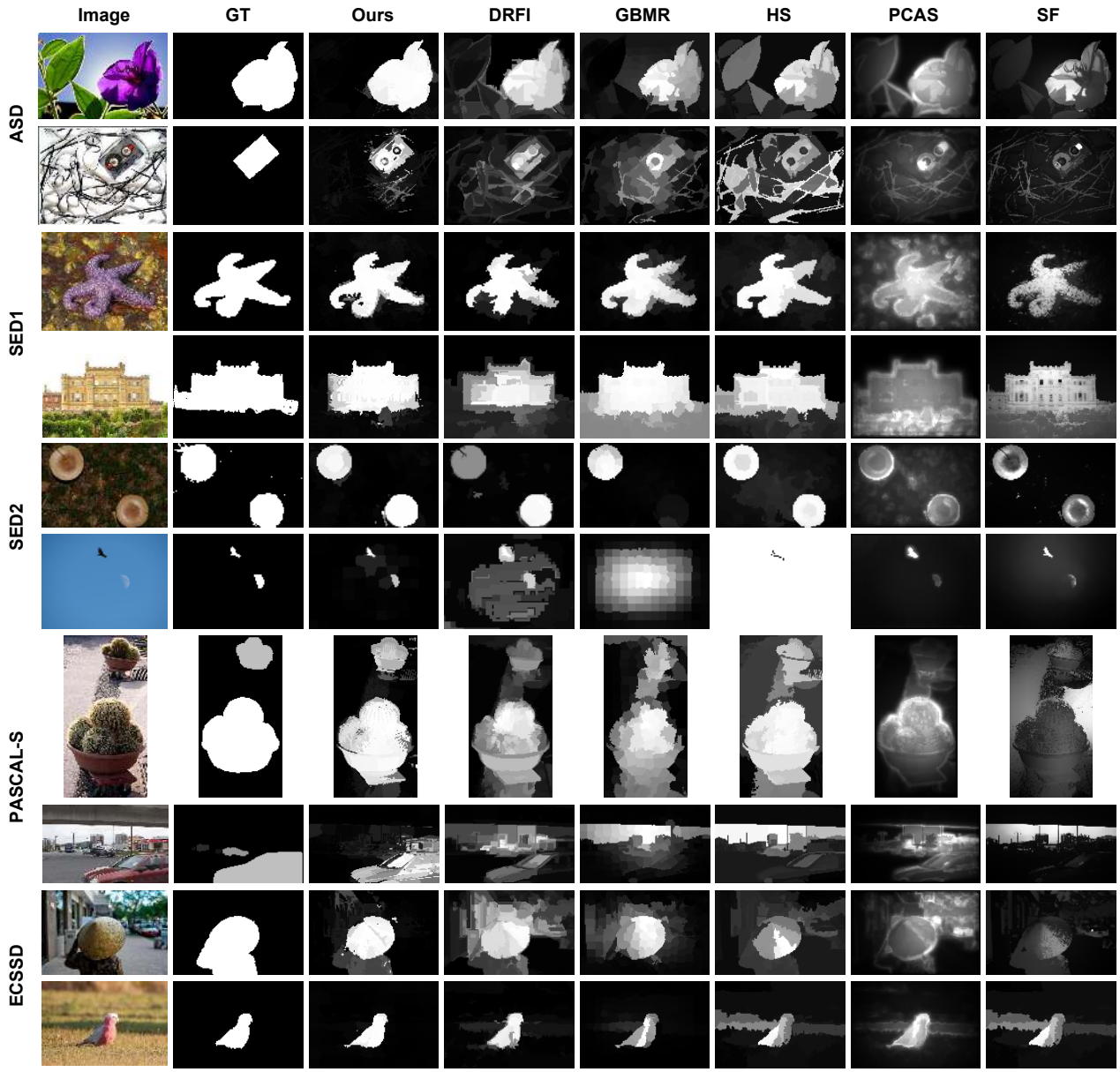

Figure 7. Example images from five datasets and the saliency maps by compared methods. Methods for comparison include DRFI [25], HS [56], GBMR [57], PCAS [41], CEOS [40], GC [12], and SF [44].

duce multi-context saliency modeling using deep Convolutional Neural Networks. Global context and local context are utilized and integrated into a unified multi-context deep learning framework for saliency detection. Global- and local-context models are jointly optimized. Secondly, different pre-training strategies are investigated to learn the deep model for saliency detection, and a task-specific pre-training scheme designed for our multi-context deep model is proposed. Moreover, recently proposed contemporary deep models in ImageNet Image Classification Challenge are tested, and their effectiveness in saliency detection are investigated. Experiments validate each component in our framework, and show our approach significantly and consistently outperforms all the state-of-the-art methods.

6. Acknowledgement

This work is supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project Nos. CUHK419412, CUHK417011, CUHK14206114 and CUHK14207814), Hong Kong Innovation and Technology Support Programme (Project reference ITS/221/13FP), Shenzhen Basic Research Program (JCYJ20130402113127496), NSFC (Project Nos. 61301269, 91320101, 61472410) and Sichuan High Tech R&D Program (No.2014GZX0009).

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009. 1, 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. on PAMI*, 34(11):2274–2282, 2012. 3

[3] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010. 2

[4] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Trans. on PAMI*, 34(11):2189–2202, 2012. 2

[5] S. Alpert, M. Galun, R. Basri, and A. Brandt. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *CVPR*, 2007. 6

[6] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. In *ACM TOG*, volume 26, page 10, 2007. 1

[7] A. Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *CVPR*, 2012. 2

[8] A. Borji, D. N. Sihite, and L. Itti. Salient object detection: A benchmark. In *ECCV*. 2012. 1

[9] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, 2010. 2, 3

[10] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu. Global contrast based salient region detection. *IEEE Trans. on PAMI*, 2014. 2, 6

[11] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook. Efficient salient region detection with soft image abstraction. In *ICCV*, 2013. 1

[12] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook. Efficient salient region detection with soft image abstraction. In *ICCV*, 2013. 7, 8

[13] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, 2011. 1, 2, 6

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

[15] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 3

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013. 1, 5, 6

[17] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, 2006. 1, 2, 7

[18] S. He and R. W. Lau. Saliency detection with flash and no-flash image pairs. In *ECCV*, 2014. 2

[19] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002. 4

[20] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Trans. on PAMI*, 34(1):194–201, 2012. 7

[21] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007. 1

[22] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on PAMI*, 20(11):1254–1259, 1998. 2

[23] Y. Jia and M. Han. Category-independent object-level saliency detection. In *ICCV*, 2013. 1, 3

[24] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang. Saliency detection via absorbing markov chain. In *ICCV*, 2013. 1

[25] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013. 1, 2, 7, 8

[26] R. Jiang and D. Crookes. Deep salience: Visual salience modeling via deep belief propagation. In *AAAI*, 2014. 3

[27] Z. Jiang and L. S. Davis. Submodular salient region detection. In *CVPR*, 2013. 1

[28] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009. 2

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3, 6, 7

[30] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 1, 3

[31] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang. Saliency detection via dense and sparse reconstruction. In *ICCV*, 2013. 1

[32] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014. 1, 6, 7

[33] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2014. 3

[34] Y. Lin, S. Kong, D. Wang, and Y. Zhuang. Saliency detection within a deep convolutional architecture. In *AAAI Workshop*, 2014. 3

[35] R. Liu, J. Cao, Z. Lin, and S. Shan. Adaptive partial differential equation learning for visual saliency detection. In *CVPR*, 2014. 1

[36] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Trans. on PAMI*, 33(2):353–367, 2011. 2, 6

[37] P. Luo, Y. Tian, X. Wang, and X. Tang. Switchable deep network for pedestrian detection. In *CVPR*, 2014. 1

[38] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *ACM Multimedia*, 2002. 1

[39] L. Mai and F. Liu. Comparing salient object detection results without ground truth. In *ECCV*. 2014. 1

[40] R. Mairon and O. Ben-Shahar. A closer look at context: From coxels to the contextual emergence of object saliency. In *ECCV*. 2014. 1, 7, 8

[41] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *CVPR*, 2013. 1, 7, 8

[42] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, Y. Xiong, C. Qian, et al. Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection. *arXiv preprint arXiv:1409.3505*, 2014. 3, 5

[43] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy, and X. Tang. Deepid-net: Deformable deep convolutional neural networks for object detection. In *CVPR*, 2015. 3

[44] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012. 1, 2, 7, 8

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014. 5

[46] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. 3, 6, 7

[47] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *CVPR*, 2013. 1

[48] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013. 3

[49] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, 2013. 1

[50] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014. 3, 6, 7

[51] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. *arXiv preprint arXiv:1312.4659*, 2013. 3

[52] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *CVPR*, 2014. 1

[53] J. Wang, L. Quan, J. Sun, X. Tang, and H.-Y. Shum. Picture collage. In *CVPR*, 2006. 1

[54] X. Wang, L. Zhang, L. Lin, Z. Liang, and W. Zuo. Joint task learning via deep neural networks with application to generic object extraction. In *NIPS*, 2014. 2

[55] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *ECCV*, 2012. 2

[56] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *CVPR*, 2013. 1, 3, 6, 7, 8

[57] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 1, 2, 7, 8

[58] J. Yang and M.-H. Yang. Top-down visual saliency via joint crf and dictionary learning. In *CVPR*, 2012. 2

[59] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*. 2014. 3, 6, 7

[60] J. Zhang and S. Sclaroff. Saliency detection: a boolean map approach. In *ICCV*, 2013. 1

[61] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013. 1

[62] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 1

[63] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *CVPR*, 2014. 1, 2

[64] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 2