Context Proposals for Saliency Detection[☆]

Aymen Azaza[a,b,*], Joost van de Weijer[b,*], Ali Douik[a], Marc Masana[b]

[a]*Noccs, National Engineering School of Sousse, Tunisia.*
[b]*Computer Vision Center Barcelona, Spain.*

Abstract

One of the fundamental properties of a salient object region is its contrast with the immediate context. The problem is that numerous object regions exist which potentially can all be salient. One way to prevent an exhaustive search over all object regions is by using object proposal algorithms. These return a limited set of regions which are most likely to contain an object. Several saliency estimation methods have used object proposals. However, they focus on the saliency of the proposal only, and the importance of its immediate context has not been evaluated.

In this paper, we aim to improve salient object detection. Therefore, we extend object proposal methods with context proposals, which allow to incorporate the immediate context in the saliency computation. We propose several saliency features which are computed from the context proposals. In the experiments, we evaluate five object proposal methods for the task of saliency segmentation, and find that Multiscale Combinatorial Grouping outperforms the others. Furthermore, experiments show that the proposed context features improve performance, and that our method matches results on the FT datasets and obtains competitive results on three other datasets (PASCAL-S, MSRA-B and ECSSD).

Keywords: Computational Saliency, Object Segmentation, Object Proposals

*Corresponding author

Email addresses: aymen.azaza@cvc.uab.es (Aymen Azaza), joost@cvc.uab.es (Joost van de Weijer), ali.douik@enim.rnu.tn (Ali Douik), mmasana@cvc.uab.es (Marc Masana)

1. Introduction

To rapidly extract important information from a scene, the human visual system allocates more attention to salient regions. Research on computational saliency focuses on designing algorithms which, similarly to human vision, predict which regions in a scene are salient. In computer vision, saliency is used both to refer to eye-fixation prediction [60, 62] as well as to salient object segmentation [25, 34]. It is the latter which is the focus of this article. Computational saliency has been used in applications such as image thumbnailing [39], compression [51], and image retrieval [56].

Object proposal methods have recently been introduced in saliency detection methods [34]. They were first proposed for object recognition, which was long dominated by sliding window approaches (see e.g. [11]). Object proposal methods reduce the number of candidate regions when compared to sliding window approaches [55]. They propose either a set of bounding boxes or image segments, which have a high probability of containing an object [19, 27]. Recently, these methods have been applied in saliency detection [12, 34, 57]. Object proposals especially help in obtaining exact boundaries of the salient objects [34]. In addition, they can reduce the computational costs of evaluating saliency based on a sliding window [36].

The saliency of an object is dependent on its context, i.e. an object is salient (or not) with respect to its context. If a visual feature, e.g. color, textures or orientation, of an object differs from that of its context it is considered salient. Traditionally, this has been modeled in saliency computation with the center-surround mechanism [13, 16], which approximates visual neurons. This mechanism divides the receptive field of neurons into two regions, namely the center and surround, thereby modeling the two primary types of ganglion cells in the retina. The first type is excited by a region in the center, and inhibited by a surround. The second type has the opposite arrangement and is excited from the surround and inhibited by a center. In computational saliency the center-surround mechanism has been implemented in different ways. For example, [24] model this by taking the difference between fine (center) and coarse scale (surround) representations of image features. Even though this has been shown to successfully model eye fixation data, for the task of salient object detection this approach is limited to the shapes of the filters used. It can only consider the differences between circle regions of different radii. This led [36] to consider center-surround between arbitrary rectangles in the images for salient object detection. In this work we will fur-

ther generalize the concept of center-surround but now to arbitrarily shaped object proposals.

To generalize the concept of center-surround to arbitrary shaped object proposals we extend object proposals with context proposals. We consider any object proposal method which computes segmentation masks. For each object proposal we compute a context proposal which encompasses the object proposal and indicates the part of the image which describes its direct surrounding. To compute the saliency with respect to the context proposals, we use a similar approach as [38]. For an object to be salient, it should be so with respect to the region described by the context proposal. In addition, because typically an object is occluding a background, it is expected that the features in the context proposal do not vary significantly. As a consequence, the saliency of the object proposal is increased if the corresponding context-proposal is homogeneous in itself, and different with respect to the object segment. In [38] these observations on context-based saliency led to an iterative multi-scale accumulation procedure to compute the saliency maps. Here, however, we circumvent this iterative process by directly computing context proposals derived from object proposals, and subsequently computing the context saliency between the proposal and its context proposal.

The main contribution of our paper is that we propose several context based features for saliency estimation. These are computed from context proposals which are computed from object proposals. To validate our approach we perform experiments on a number of benchmark datasets. We show that our method matches state-of-the-art on the FT dataset and improves state-of-the-art results on three benchmark (PASCAL-S, MSRA-B and ECSSD datasets). In addition, we evaluate several off-the-shelf deep features and object proposal methods for saliency detection and find that VGG-19 features and multiscale combinatorial grouping (MCG) obtain the best performance.

This paper is organized as follows. In Section 2 we discuss the related work. In Section 3 we provide an overview of our approach to saliency detection. In Section 4 the computation of context proposals is outlined. Next we provide details on the experimental setup in Section 5 and give results in Section 6. Conclusions are provided in Section 7.

2. Related work

In this section we provide an overview of salient object detection methods and their connection with object proposal methods. More complete reviews

on saliency can be found in [4, 66, 65].

Saliency detection One of the first methods for computational saliency was proposed by [24]. Their model based on the feature integration theory of [54] and the work of [26] decomposes the input image into low level feature maps including color, intensity and orientation. These maps are subsequently merged together using linear filtering and center surround structures to form a final saliency map. Their seminal work initiated much research in biologically inspired saliency models [13, 42, 49] as well as more mathematical models for computational saliency [1, 17, 21, 32]. The central surround allows to measure contrast with the context, however it is confined to predefined shapes; normally the circle shape of the Gaussian filters [24] or rectangle shapes in the work of [36]. In this paper we will propose a method for arbitrary shaped contexts.

Local and global approaches for visual saliency can be classified in the category of bottom-up approaches. Local approaches compute local center-surround contrast and rarity of a region over its neighborhoods. [24] derive a bottom-up visual saliency based on center surround difference through multiscale image features. [36] propose a binary saliency estimation method by training a CRF to combine a set of local, regional, and global features. [17] propose the GBVS method which is a bottom-up saliency approach that consists of two steps: the generation of feature channels as in Itti's approach, and their normalization using a graph based approach. A saliency model that computes local descriptors from a given image in order to measure the similarity of a pixel to its neighborhoods was proposed by [48]. [14] propose a AWS method which is based on the decorrelation and the distinctiveness of local responses.

Another class of features for saliency are based on global context or rarity; the saliency of a feature is based on its rarity with respect to the whole image. [15] consider the difference of patches with all other patches in the image to compute global saliency. [59] compute saliency by considering the reconstruction error which is left after reconstructing a patch from other patches (other patches can be from the same image or from the whole dataset). [25] compute the rarity of a feature by comparing the contrast between a 15 pixel border around the image and the object proposal histogram. Other than these methods we propose a method to compute the saliency with respect to the direct context of the object. Finally, to compute saliency [23] combined local and global objectness cues with a set of candidates location.

Our work has been inspired by a recent paper [38] which demonstrates the importance of visual context for saliency computation. The work is based on the observation that an object is salient with respect to its context. And since context is an integral part of saliency of an object, it should therefore be assigned a prominent role in its computation. The final saliency map is computed by alternating between two steps: 1. the fusing of image regions based on their color distance into larger and larger context segments, and 2. the accumulation of saliency votes by the context segments (votes are casted to the region which is enclosed by the context segments). The steps are alternated until the whole image is clustered together into a single context segment. The procedure is elegant in its simplicity and was shown to obtain excellent results. However, the iterative nature of the computation renders it computationally very demanding.

Deep convolutional neural networks have revolutionized computer vision over the last few years. This has recently led to several papers on deep learning for saliency detection [57, 29, 67, 43, 8]. Both [29] and [67] consider parallel networks which evaluate the image at various scales. [57] use two networks to describe local and global saliency. [53] combine a local and global model to compute saliency. The main challenge for saliency detection with deep networks is the amount of training data which is not always available. This is solved in [57, 29, 67] by training on the largest available saliency dataset, namely MSRA-B [36], and testing on the other datasets (both [29, 30] also use pretrained network weights trained on the 1M Imagenet dataset). Like these method, we will use a pretrained network for the extraction of features for saliency detection.

Object Proposal methods Object detection based on object proposals methods has won in popularity in recent years [55]. The main advantages of these methods is that they are not restricted to fixed aspect ratios as most sliding window methods are, and more importantly, they allow to evaluate a limited number of windows. As a consequence more complicated features and classifiers can be applied, resulting in state-of-the-art object detection results. The generation of object hypotheses can be divided into methods whose output is an image window and those that generate object or segment proposals. The latter are of importance for salient object detection since we aim to segment the salient objects from the background.

Among the first object proposal methods the work of [6], named the Constrained Parametric Min-Cuts (CPMC) method, uses graph cuts with

different random seeds to obtain multiple binary foreground and background segments. [2] proposes to measure the objectness of an image window, where they rank randomly sampled image windows based on their likelihood of containing the object by using multiple cues among which edges density, multiscale saliency, superpixels straddling and color contrast. [10] proposed an object proposal method similar to the CPMC method by generating multiple foreground and background segmentations. A very fast method for object proposals was proposed by [7], which generates box proposals at 300 images per second.

An extensive comparison of object proposal methods was performed by [19]. Among the best evaluated object proposal methods (which generate object segmentation) are the selective search [55], the geodesic object proposals [27] and the multiscale combinatorial grouping method [3]. Selective search proposes a set of segments based on hierarchical segmentations of the image where the underlying distance measures and color spaces are varied to yield a large variety of segmentations. [27], propose the geodesic object proposals method, which applies a geodesic distance transfer to compute object proposals. Finally, Multiscale Combinatorial Grouping [3] is based on a bottom-up hierarchical image segmentation. Object candidates are generated by a grouping procedure which is based on edge strength.

Several methods have applied object proposals to saliency detection [12, 34, 57]. The main advantage of saliency detection methods based on object proposals over methods based on superpixels [64] is that they do not require an additional grouping phase, since the object proposals are expected to encompass the whole object. Other than general object detection, salient object segmentation aims at detecting objects which are salient in the scene. Direct surrounding of objects is of importance to determine the object's saliency. Therefore, in this paper we extend the usage of object proposals for saliency detection with context proposals, which allow us to directly assess the saliency of the object with respect to its context.

3. Method Overview

The main novelty of our method is the computation of context features from context proposals. To illustrate the advantage of this idea consider Fig. 1. In this figure several implementation of the center surround idea for saliency detection are shown. The circular surround was used in the original work by [24]. This concept was later generalized to arbitrary rectangles [36].

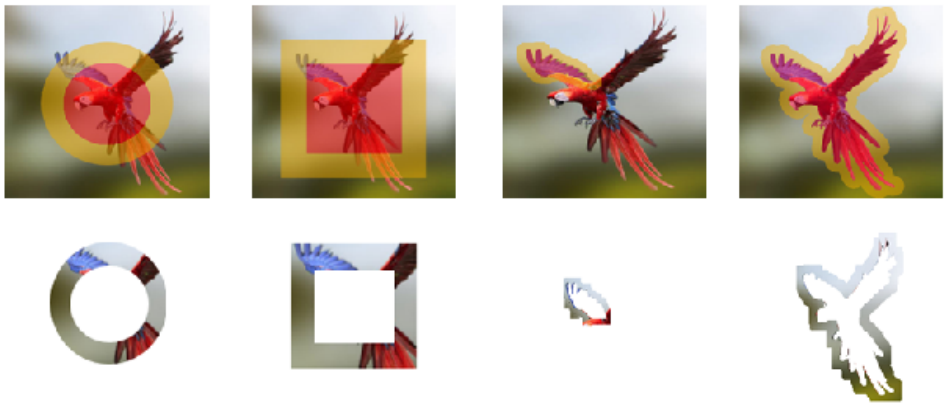

Figure 1: Top row: examples of different center surround approaches (a) circular surround (b) rectangular surround (c) superpixels surround, and (d) the context proposals. Bottom row: the surround for each of the methods. It can be seen that only the object proposal based surround correctly separates object from background.

Both these approaches have the drawback that they only are a rough approximation of the real object shape and the contrast between the (circle or rectangular) object and its surround does not very well represent the real saliency of the object. This is caused by the fact that when we approximate the object by either a square or circle, part of the object is in the surround, and part of the surround is in the object.

In principle the center surround idea could be extended to superpixels which are often used in saliency detection [64], see Fig. 1. However, superpixels generally only cover a part of the object, and therefore their surround is often not homogeneous, complicating the analysis of the saliency of the center. Finally, [38] show that a surround which can adapt to the shape of the object (center) is an excellent saliency predictor. For its computation they propose an iterative procedure. In this paper we propose to use object proposals methods [3], which are designed to directly provide a segmentation of the object, for the computation of context-based saliency. Since object proposals have the potential to correctly separate object from surround (see final column on the right in Fig. 1), we hypothesize that considering their contrast can lead to a better saliency assessment than with the other methods.

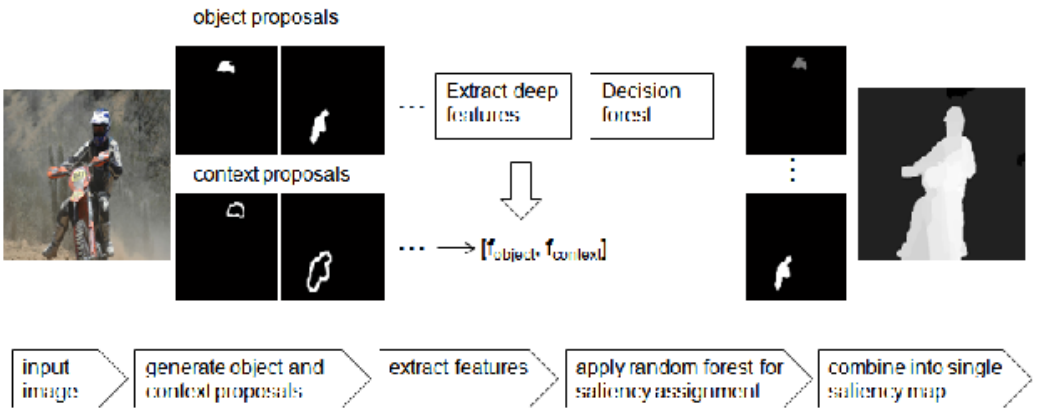

Figure 2: Overview of our method at test time. A set of object proposals is computed. From these a set of accompanying context proposals is derived. We extract deep convolutional features from both object and context ($\mathbf{f}_{object}$ and $\mathbf{f}_{context}$). At training for each object proposal its saliency is computed based on the ground truth, and a random forest is trained to regress to the saliency. At testing this random forest is applied to predict the saliency of all proposals, which are combined in the final saliency map

An overview of the saliency detection algorithm is provided in Fig. 2. Next, any object proposal algorithm can be used here that provides pixel-precise object contours, such as [3, 55, 27, 22, 44]. We extend each object proposal with a context proposal which is its immediate surround (see Section 4.1). We then proceed by computing deep features for both the object proposal and its context proposal from which we derive several context features (see Section 4.2).

Given the feature vector of the object and context for each of the proposals in the training set we train a random forest classifier. As the saliency score for each object proposal we use the average saliency of the pixels in the proposal: pixels have a saliency of one if they are on the ground truth salient object or zero elsewhere (this procedure is further explained in Section 4.4). At testing time we infer the saliency for all the object proposals by applying the random forest regressor. The final saliency map is computed by taking for each pixel the average of the saliency of all the proposals that contain that pixel.

The overall method is similar to several previous papers on saliency. A similar approach was proposed by [25] and later used by [34]. In [25] they use a random forest classifier to score each region in the image instead of

every object proposal in our method. [34] use the CPMC method for object proposals [6] and similar as [25] they apply a random forest to predict region saliency based on regional features. In contrast to these methods we investigate the usage of context proposal for object saliency detection.

4. Context Proposals for Saliency Computation

In this section we start by describing our approach for context proposal generation. Then in Section 4.2 we describe how to compute the context features from the context proposals. Next in Section 4.3 we describe the deep features we directly use as features for the object proposals, and which we also use to compute the context features. Finally, in Section 4.4 we explain how we arrive at the final saliency estimation by using a random forest regressor on both the object and context features.

4.1. Context Proposal Generation

Recently, several saliency methods have applied object proposal algorithms to generate proposals for salient objects [34, 25]. Consider an object proposal, represented by the mask M which is equal to one for all pixels within the object proposal and zero otherwise. Then we define the context of the proposal to be

$$\begin{aligned} &C = \left(M \oplus B^{(n)}\right) \setminus M \\ &\text{smallest } n \text{ for which } |C| \geq |M| \end{aligned} \tag{1}$$

where B is a structural element and $\oplus$ is the dilation operator. We used the notation

$$B^{(n)} = \overbrace{B^{(1)} \oplus B^{(1)} \oplus B^{(1)} \oplus B^{(1)}}^{n \text{ times}} \tag{2}$$

to indicate multiple dilations. In our work we choose $B = N_8$ which is the eight connected set (a 3x3 structural element with all ones). We use $|C|$ to indicate the number of non-zero values in C. If we would consider arbitrary n in the first part of this equation, this equation could be interpreted as generating a border for the object proposal M which thickness is equal to n. We define the context to be the smallest border which has equal or more pixels than M. In practice, the context is computed by iteratively dilating with B until we reach a border which contains more pixels than the object proposal M. In Fig. 3 we provide examples of context borders for several

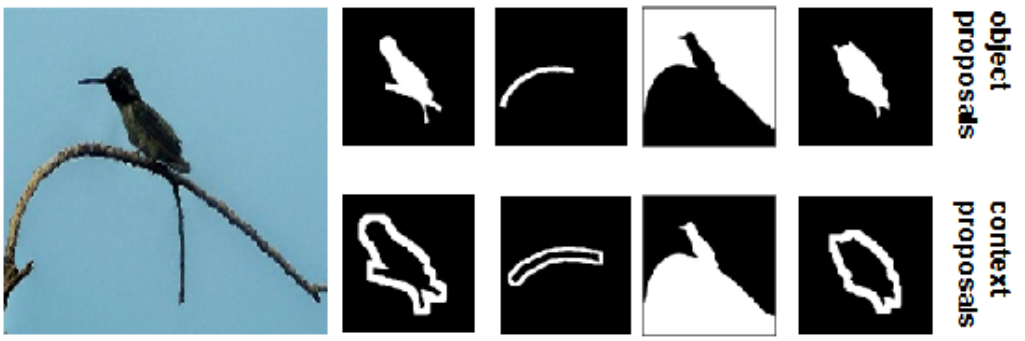

Figure 3: Input image and (top row) examples of object proposals and (bottom row) examples of context proposals.

object proposals. Note that the context border is wider for larger object proposals. The idea is to verify if the object proposal is salient with respect to its context.

4.2. Context Feature Computation

Next we outline the computation of the context features. We consider two properties which define a good context proposal. *Context contrast* which measures the contrast between the features which make up the salient object and the features which describe its context. Secondly *Context continuity* which is based on the observation that the salient object is often occluding a background which continues behind it. As a consequence, we expect the features which describe the context on opposite sides of the salient object to be similar. In human vision research it was verified that salient objects (targets) are faster found on a homogeneous background than when surrounded by a heterogeneous background (distractor) [9]. Context continuity is an indicator of background homogeneity, since homogeneous backgrounds lead to higher context continuity, and heterogeneous ones would lead to low context continuity.

The first context saliency feature which we consider combines both described properties, context contrast and context continuity, into a single measure. Consider a pixel m_i in the object proposal M. Then we define two related coordinates d_i^{φ} and u_i^{φ} which are coordinates of the points on the context when considering a line with orientation φ through point m_i (see Fig. 4). The saliency of a point m_i is larger when the feature representation at m_i is more different from the feature representation on its context at d_i

and u_i. In addition, we would like the distance between the points on the context (d_i and u_i) to be similar. Combining these two factors in one saliency measures yields:

$$c_1^{\varphi}(m_i) = \arctan\left(\frac{\min\left(s_i^{d,\varphi}, s_i^{u,\varphi}\right)}{s_i^{du,\varphi} + \lambda}\right). \tag{3}$$

where the numerator contains the context contrast and the denominator the context continuity. The *arctan* and the constant λ are used to prevent large fluctuations in saliency for small values of $s_i^{du,\varphi}$. The distances are defined with

$$s_i^{u,\varphi} = \left\|\mathbf{f}\left(u_i^{\varphi}\right) - \mathbf{f}\left(m_i\right)\right\|, \tag{4}$$

$$s_i^{d,\varphi} = \left\|\mathbf{f}\left(d_i^{\varphi}\right) - \mathbf{f}\left(m_i\right)\right\|, \tag{5}$$

$$s_i^{du,\varphi} = \left\|\mathbf{f}\left(d_i^{\varphi}\right) - \mathbf{f}\left(u_i^{\varphi}\right)\right\|. \tag{6}$$

Here $\mathbf{f}(m_i)$ denotes a feature representation of the image at spatial location m_i, and $\|.\|$ is the L2 norm. This feature representation could for example be the RGB value at that spatial location, but also any other feature representation such as for example a deep convolutional feature representation as we will use in this article. Now that we have defined the saliency for a single point considering its context points along a line with orientation φ, we define the overall saliency for a context proposal as the summation over all pixels m_i in the object proposal considering all possible lines:

$$C^1 = \frac{1}{|M|}\sum_{m_i \in M}\int_0^{\pi} c_1^{\varphi}(m_i)\, d\varphi. \tag{7}$$

It should be noted that we exclude lines which do not have context on both sides of the object. This happens for example for objects on the border of the image.

Considering all orientations is computationally unrealistic and in practice we approximate this equation with

$$C^1 = \frac{1}{|M|}\sum_{m_i \in M}\sum_{\varphi \in \Phi} c_1^{\varphi}(m_i), \tag{8}$$

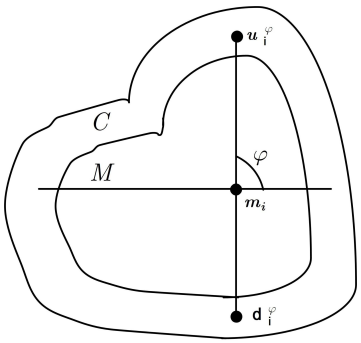

Figure 4: Graphical representation of variables involved in context feature computation.

where Φ is a set of chosen orientations between $[0, \pi)$. In this paper we have considered four orientations

$$\Phi = \left\{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\right\}. \tag{9}$$

The saliency of one point in the object proposal is hence computed by considering its context along four orientations. To be less sensitive to noise on the context both $\mathbf{f}\left(d_i^{\varphi}\right)$ and $\mathbf{f}\left(u_i^{\varphi}\right)$ are extracted from a Gaussian smoothed context proposal.

As a second context feature we ignore the object proposal and only consider the values on the context proposal to compute the saliency. This feature solely focuses on context continuity. In this case we would like the saliency to be larger when the values on the context have a smaller distance. We propose to use the following measure:

$$c_2^{\varphi}\left(m_i\right) = \arctan\left(\frac{1}{s_i^{du,\varphi} + \lambda}\right) \tag{10}$$

again λ prevents large fluctuations for low values of s_i.

Similarly we compute the $C^2\left(m_i\right)$ for the object proposal with

$$C^2 = \frac{1}{|M|} \sum_{m_i \in M} \int_0^{\pi} c_2^{\varphi}\left(m_i\right) d\varphi. \tag{11}$$

and its approximation

$$C^2 = \frac{1}{|M|} \sum_{m_i \in M} \sum_{\varphi \in \Phi} c_2^{\varphi}\left(m_i\right). \tag{12}$$

Figure 5: Context continuity: features on opposites sides of the object proposal are expected to be similar. Examples of (left) omni-directional context continuity and (right) horizontal context continuity.

In addition to C^1 and C^2 which measure context saliency based on comparing features on all sides of the object proposal, we introduce also a measure for horizontal context continuity C^3 where we use $\Phi^H = \{0\}$, and we compute

$$C^3 = \frac{1}{|M|} \sum_{m_i \in M} \sum_{\varphi \in \Phi^H} c_1^{\varphi}(m_i). \tag{13}$$

The motivation for a special measure for horizontal context continuity is provided in Fig. 5. Natural scenes contain more horizontal elongation than other orientations; the C^3 measure is designed to detect horizontal clutter.

The context measures proposed here are motivated by the work of [38]. They propose an iterative procedure to compute context based saliency. We prevent the iterative procedure by directly computing the context from the object proposals. In addition, we propose a measure of horizontal context which is not present in [38]. Also instead of RGB features we use deep features to compute the context saliency.

4.3. Off-the-Shelf Deep Features

Here we explain the computation of the deep features, which we use as the feature $\mathbf{f}$ in Eq. 4-6 to compute the three context features Eq. 8, Eq. 12- 13. These are combined into one context feature

$$\mathbf{f}_{context} = \{C^1, C^2, C^3\} \tag{14}$$

net. / bl.	AlexNet	VGG-19	ResNet-152
1.	$[11 \times 11, 96]$	$[3 \times 3, 64] \times 2$	$[7 \times 7, 64]$
2.	$[5 \times 5, 256]$	$[3 \times 3, 128] \times 2$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
3.	$[3 \times 3, 384]$	$[3 \times 3, 256] \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 8$
4.	$[3 \times 3, 384]$	$[3 \times 3, 512] \times 4$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 36$
5.	$[3 \times 3, 256]$	$[3 \times 3, 512] \times 4$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$

Table 1: Overview of the *convolutional* layers of different networks. The convolutional part can be divided in 5 blocks (bl.) for all three networks. For each block we show the convolutional size, the number of features, and how many times this layer pattern is repeated. The non-linear activation layers are omitted. In our evaluation we will use the last layer of each block to extract convolutional features.

for each context proposal. The deep feature is also used directly as a descriptor for the object proposal by pooling the deep feature over all pixels in the object proposal with

$$\mathbf{f}_{object} = \frac{1}{|M|} \sum_{m_i \in M} \mathbf{f}\left(m_i\right).. \tag{15}$$

Deep convolutional features have shown excellent results in recent papers on saliency [57, 29, 67, 31]. A straight-forward way to use deep features is by using a pre-trained network, for example trained for the task of image classification on ImageNet [28], to extract features. These so called off-the-shelf features can then be used as local features. A good overview of this approach is given by [45], who successfully apply this technique to a variety of tasks including object image classification, scene recognition, fine grained recognition, attribute detection and image retrieval.

To choose the best deep features for saliency detection we evaluate three popular networks, namely AlexNet [28], VGG-19 [50] and ResNet [18]. The configuration of the convolutional layers of the networks is given in Table 1.

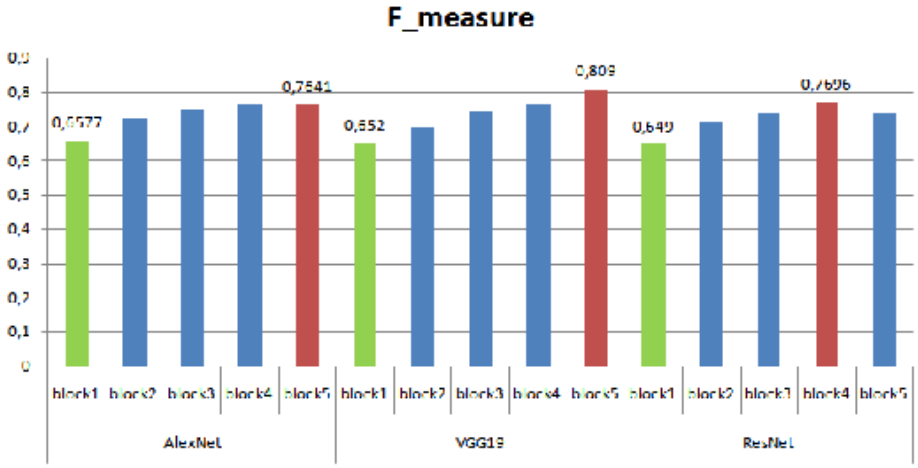

Figure 6: Evaluation on 5 convolutional layers for the three architectures used in our framework.

We evaluate the performance of the different blocks for saliency estimation. The results using both object features $\mathbf{f}_{object}$ and context features $\mathbf{f}_{context}$ are summarized in Fig. 6. We found the best results, similar to the ResNet, were obtained with block 5 of VGG-19 (which layer name is *conv5_4*). Based on these results we choose to extract block 5 deep features with VGG-19 for all images. We spatially pool the features within each object to form a 512-dimensional $\mathbf{f}_{object}$ and the 3-dimensional $\mathbf{f}_{context}$ according to Eq. 14-15. In addition, we found that applying a standard whitening, where we set the variance over all features of $\mathbf{f}_{object}$ to 1, prior to applying the classifiers improved results.

4.4. Saliency Score of Object Proposals

Based on the features which are extracted from the object proposal and its context we train a random forest regressor to estimate the saliency of the object proposal. To compute the saliency score, sal^{object}, for object proposals we use the following equation:

$$sal^{object} = \frac{|M \cap S|}{|M|} \tag{16}$$

here M is the set of all pixels in the object proposal and S is the set of all pixels which are considered salient in the ground truth. A $sal = 0.8$ means that 80% of the pixels in the object proposal are considered salient.

We found that this score is not optimal when considering context proposals, and we propose to use the following equation

$$sal^{context} = \max\left(\frac{|M \cap S|}{|M|} - \frac{|C \cap S|}{|C|}, 0\right) \tag{17}$$

where C is the set of pixels in the context. The $sal^{context}$ measure lowers the score if salient pixels are in the context.

We train two separate random forest regressors, one based on the deep features of the object proposal regressing to sal^{object} and one based on the context features regressing to $sal^{context}$. The final saliency score at testing time is computed by adding results of the two regressors. The final saliency map is computed by averaging the saliency of all the object proposals which are considered in the image. We have also considered to assign to each pixel the maximum saliency of all object proposals which include the pixel, but found this to yield inferior results.

5. Experimental Setup

In this section we describe the features on which we base the saliency computation, the datasets on which the experiments are performed, and the evaluation protocol we use.

5.1. Datasets

We evaluate our proposed algorithm on several benchmark datasets that are widely used.

Pascal-S [34]: This dataset was built on the validation set of the Pascal VOC 2010 segmentation challenge. It contains 850 images with both saliency segmentation ground truth and eye fixation ground truth. Saliency ground truth masks were labeled by 12 subjects. Many of the images in this dataset contain multiple salient objects.

MSRA-B [36]: This dataset contains 5,000 images and is one of the most used datasets for visual saliency estimation. Most of the images contain only one salient object.

FT [1]: This dataset contains 1,000 images, most of the images contain one salient object. It provides only salient object ground truth which is derived from [61] and is obtained using user-drawn rectangles around salient objects.

ECSSD [63]: It contains 1,000 images acquired from the PASCAL VOC dataset and the internet and the ground truth masks were annotated by 5 subjects.

5.2. Evaluation

We evaluate the performance using PR (precision-recall) curve and F-measure. Precision measures the percentage of salient pixels correctly assigned, and recall the section of detected salient pixels which belongs to the salient object in the ground truth.

We compute precision and recall of saliency maps by segmenting the salient object with a threshold T and comparing the binary map with the ground truth. All saliency maps are also evaluated using the F-measure score which is defined as:

$$F_\beta \;=\; \frac{(1+\beta^2)\cdot \text{precision}\cdot \text{recall}}{\beta^2\cdot \text{precision} + \text{recall}} \tag{18}$$

where β^2 is set to 0.3 following [34, 57, 52, 67]. As a threshold we use the one which leads to the best F_β. This was proposed in [5, 41] as a good summary of the precision-recall curve. We compare our method against 8 recent CNN methods: Deeply supervised salient object (DSS)[20], Deep contrast learning (DCL) [30], Reccurent fully convolutional networks (RFCN) [58], Deep hierarchical salieny (DHS) [35], Multi-task deep saliency (MTDS) [33], Multiscale deep features (MDF) [29], Local and global estimation (LEGS) [57], Multi context (MC) [67] and we compare also against 8 classical methods including Discriminative regional feature integration (DRFI) [25], Hierarchical saliency (HS) [63], Frequency tuned saliency (FT) [1], Regional principal color based saliency detection (RPC) [37], (CPMC-GBVS) [34], Graph-based manifold ranking (GBMR) [64], Principal component analysis saliency (PCAS) [40], Textural distinctiveness (TD) [47] and a Context aware method [15] (GOF). For a fair comparison we did not include (CPMC-GBVS) method[34] because they use eye fixation label in training.

Based on crossvalidation experiments on PASCAL-S training set we set the number of trees in the random forest to 200, we set $\lambda = 40$ in Eq. 3 and Eq. 10 and we set the minimum area of object proposals to be considered at 4,500 pixels. We use these settings for all datasets.

6. Experimental Results

In this section we provide our experimental results. We provide an evaluation of five popular object proposal approaches. Next we evaluate the relative gain which is obtained by adding the features based on context proposals. We evaluate also our context features with different context shapes including the conventional circular or rectangular neighborhood. Finally, we compare to state-of-the-art methods on several benchmark datasets.

6.1. Object Proposal based Saliency Detection

Object proposal method evaluation: In recent years several methods have proposed to use object proposals for salient object segmentation. However, to the best of our knowledge, there is no work which evaluates the different object proposal approaches to saliency detection. [19] have provided an extensive evaluation of object proposals for object detection. Based on their analysis we have selected the three best object proposal methods which output segments based on their criteria, namely repeatability, recall, and detection results. The object proposal methods we compare to are selective search (SS) [55], the geodesic object proposals (GOP) [27], and the multiscale combinatorial grouping (MCG) method [3]. We have added two recent object proposals to this list which are based on deep learning, namely FastMask [22] and SharpMask [44]. We do these experiments on the PASCAL-S dataset because it is considered one of the most challenging saliency datasets; also it is labeled by multiple subjects without restriction on the number of salient objects [34].

We evaluate the performance of object proposal methods as a function of proposals. Results are provided in Table. 2. Results of MCG are remarkable already for as few as 16 proposals per image, and they stay above the other methods when increasing the number of proposals. The results of SS can be explained by the fact that the ranking of their proposals is inferior to the other methods. The inferior ranking is not that relevant for object detection where typically thousands of proposals are considered per image[1]. The results of the two methods based on deep learning, namely FastMask and SharpMask, are somewhat surprising because they are known to obtain better results for object detection [22, 44]. In a closer analysis we found that MCG obtains

[1]Selective search applies a pseudo random sorting which combines random ranking with a ranking based on the hierarchical merging process.

Number of proposals	8	16	32	64	128	256
SS	59.00	64.60	70.20	74.20	77.50	78.40
GOP	66.20	71.50	73.30	76.30	77.70	79.60
MCG	77.20	77.50	78.60	79.30	80.20	**80.90**
SharpMask	73.79	74.07	73.34	73.15	73.70	74.01
FastMask	75.87	75.03	74.42	74.04	–	–

Table 2: The F-measure performance as the number of proposals evaluated on the PASCAL-S dataset for selective search (SS), geodesic object proposals (GOP), multiscale combinatorial grouping (MCG), SharpMask and FastMask

higher overlap (as defined by IoU) with the salient object groundtruth. In addition, deep learning approaches typically extract the salient object among the first 8-16 proposals, and therefore do not improve, and sometimes even deteriorate, when considering more proposals. Based on the results we select MCG to be applied on all further experiments, and we set the number of object proposals to 256.

Context proposals: The proposed context features are motivated by the work of [38]. Different from it, our paper does not use an iterative procedure but is based on object proposals. We add a comparison in Table 3 of the performance of our context features against their method on the PASCAL-S dataset. Note that here we only consider our context feature for a fair comparison, and do not use the object feature. We have also included results when only using RGB features, which are the features used by [38]. Our context features clearly outperform the context features based on both RGB and deep features. We have also included timings of our algorithm. Since most of the time was spend by the MCG algorithm (35.3s) we have also included results with the FastMask object proposals (using 8 proposals). In this case the computation of the context features takes (5.4s). Note that this is based on an unoptimized matlab implementation. Also we add a visual comparison between our method and [38] in Fig. 7.

Next we compare our context proposals, which follow the object proposal boundary, with different context shapes. We consider rectangular and circular context, which are derived from the bounding boxes based on the object proposals [2]. For the three different context shapes we extract the same con-

[2]The context of the rectangular bounding box is computed by considering its difference

	feature	proposals	PASCAL-S	Time(s)
Mairon	RGB	-	65.57	140
Our context	RGB	MCG	69.06	40.9
Our context	DF	MCG	74.90	49.0
Our context	DF	FastMask	73.65	6.7

Table 3: Comparison between our context features and the context method proposed by [38] in terms of F-measure and computational speed in seconds. We provide results for our method based on RGB and deep features (DF), and with MCG or FastMask as a object proposal method.

.

Method	PASCAL-S
Our context features	74.90
Rectangular center surround	67.64
Circular center surround	63.71

Table 4: Comparison between our context shape and the conventional circular or rectangular neighborhood in terms of F-measure.

.

text features. The results are summarized in Table 4 and show that our approach clearly outperforms the rectangular and circular shaped contexts. Thereby showing that accurate context masks result in more precise saliency estimations.

In the following experiment we evaluate the additional performance gain of the saliency features based on context proposals. The results are presented in Table 5 for four datasets. We can see that a consistent performance gain is obtained by the usage of context proposals. The absolute gain varies from 0.7 on FT to 1.6 on PASCAL-S. This is good considering that the context feature only has a dimensionality of 3 compared to 512 for the object feature.

with a rectangle which is $\sqrt{2}$ larger. In case of the circular context we consider the circle center to have a radius of $r = \frac{w+h}{4}$ and its context is computed by considering the difference with a radius larger by a factor of $\sqrt{2}$. Like this the context for both the rectangle and the circle has again the same surface area as the object (center).

	object	context	object & context
PASCAL-S	80.64	74.90	82.31
MSRA-B	89.90	89.24	90.90
FT	89.80	87.96	91.5
ECSSD	86	82.64	86.90

Table 5: The results on four datasets in F-measure for saliency based only on object proposals, only context proposals and a combination of the two.

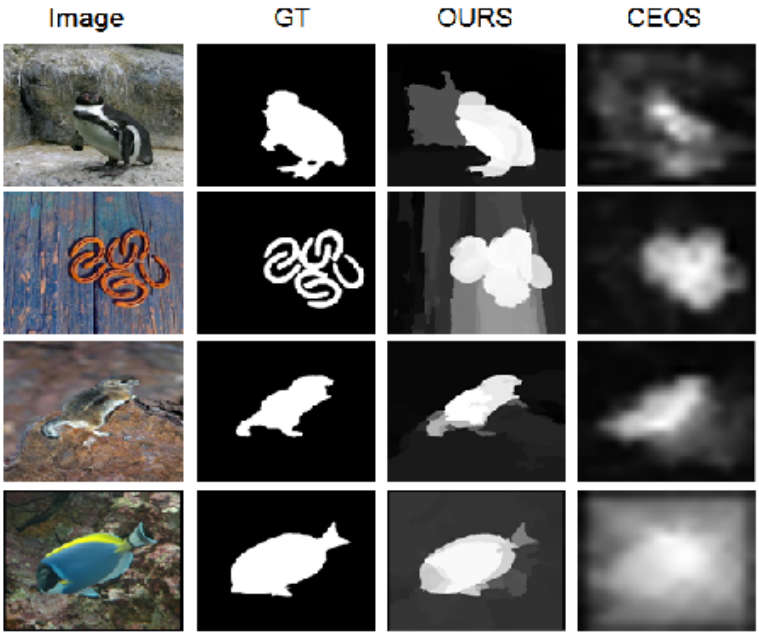

Figure 7: Visual comparison between our method and the method of [38]. Our method results in clearer edges since saliency is assigned to whole object proposals.

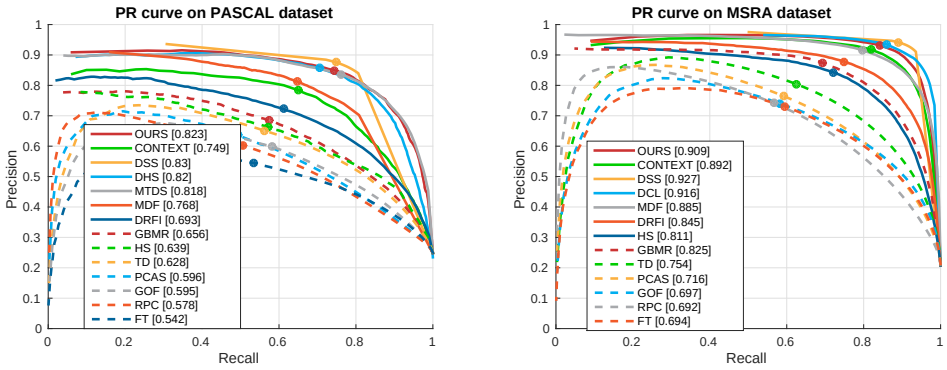

Figure 8: Precision-Recall curves on (left) Pascal-S dataset and (right) on MSRA-B dataset

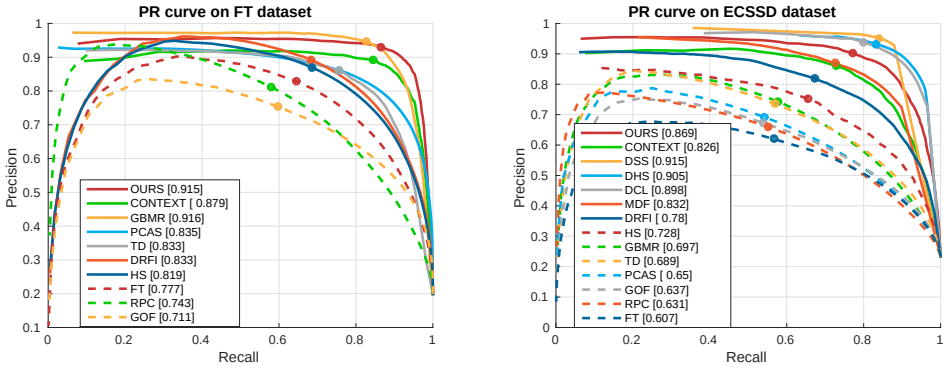

Figure 9: Precision-Recall curves on (left) FT dataset and (right) ECSSD dataset.

6.2. Comparison with the state-of-the-art

Experiments have been conducted on the PASCAL-S, MSRA-B, FT and ECSSD datasets. Traditionally these datasets proposed an original train and testset split [34]. However, several of these datasets are too small to train deep neural networks. Therefore, methods based on deep learning generally train on the MSRA-B trainset which is the largest available dataset [25, 29, 30]. To be able to compare with all results reported in the literature, we report in Table 6 both results; the results trained on the original training set and those based on training on the MSRA-B training set (these results are indicated by an asterix). As an evaluation metric we use the F-measure. We report both qualitative and quantitative comparison of our methods with state-of-the-art methods. We also report our results in Figs. 8-9. Note that these are based on training on the original training set of each dataset. Furthermore, we have only included the curves of the methods in Figs. 8-9 when this data is made available by the authors.

On the challenging PASCAL-S dataset our method trained on the original dataset obtains an F-measure of 82.3, and is the third method. On the MSRA-B dataset we are outperformed by several recent end-to-end trained saliency methods but still obtain competitive results of 90.9. On the FT dataset we obtain similar to state-of-the-art results when trained on the original dataset, and slightly better than state-of-the-art when trained on the MSRA-B dataset. Finally, on the ECSSD dataset we obtain the best results when considering only those which are trained on the ECSSD training dataset, but are outperformed by recent end-to-end trained networks trained on MSRA-B.

We added a qualitative comparison in Fig. 10. We tested our method in different challenging cases, multiple disconnected salient objects (first two rows), and low contrast between object and background (third and fourth row). Notice that our method correctly manages to assign saliency to most parts of the spider legs. Finally, results of objects touching the image boundary are shown where our method successfully includes the parts that touch the border (last two rows).

7. Conclusions

The direct context of an object is believed to be important for the saliency humans attribute to it. To model this directly in object proposal based saliency detection, we pair each object proposal with a context proposal. We

	Pascal-S	MSRA-B	FT	ECSSD
FT[1]	54.2	69.4	77.7	60.7
PRC[37]	57.8	69.2	74.3	63.1
GOF[15]	59.5	69.7	71.1	63.7
PCAS [40]	59.6	71.6	83.5	65
TD [47]	62.8	75.4	83.3	68.9
HS[63]	63.9	81.1	81.9	72.8
GBMR [64]	65.6	82.5	91.6	69.7
DRFI[25]	69.3	84.5	83.3	78
LEGS[57]	75.2	87	–	82.5
MC[67]	79.3	–	–	73.2
MDF[29]	76.8*	88.5	–	83.2*
MTDS [33]	81.8*	–	–	80.9*
DHS[35]	82*	–	–	**90.5***
DCL [30]	82.2*	91.6	–	89.8*
RFCN[58]	**82.7***	**92.6**	–	89.8*
DSS[20]	**83***	**92.7**	–	**91.5***
Ours (trained on original trainset)	82.3	90.9	91.5	86.9
Ours (trained on MSRA-B)	78.1*	90.9	**91.8***	85.4*

Table 6: Comparison of our method and context features against state-of-the-art methods. The results are based on training on the original trainset of each datasets. The methods which use the MSRA-B dataset to train are indicated with a $*$.

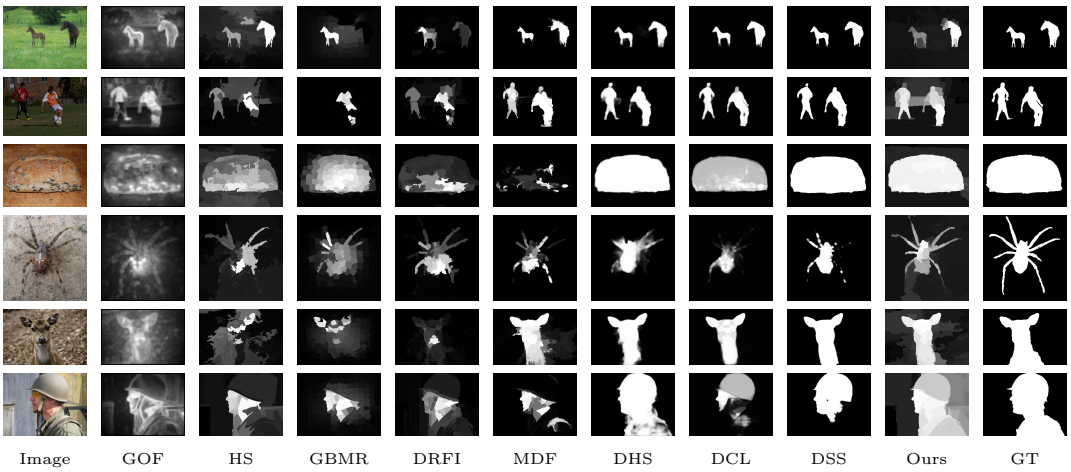

Figure 10: Visual comparison of saliency maps generated from 9 different methods, including our method. Methods for comparison includes DSS [20], DCL [30], DHS [35], MDF[29], DRFI [25], GOF [15], HS [63], and GBMR [64].

propose several features to compute the saliency of the object based on its context; including features based on omni-directional and horizontal context continuity.

We evaluate several object proposal methods for the task of saliency segmentation and find that multiscale combinatorial grouping outperforms selective search, geodesic object, SharpMask and Fastmask. We evaluate three off-the-shelf deep features networks and found that VGG-19 obtained the best results for saliency estimation. In the evaluation on four benchmark datasets we match results on the FT datasets and obtain competitive results on three datasets (PASCAL-S, MSRA-B and ECSSD). When only considering methods which are trained on the training set provided with the dataset, we obtain state-of-the-art on PASCAL-S and ECSSD.

For future research, we are interested in designing an end-to-end network which can predict both object and context proposals and extract their features. We are also interested in evaluating the usage of context proposals for other fields where object proposals are used, notably in semantic image segmentation. Finally, extending the theory to object proposals and saliency detection in video would be interesting [46].

Acknowledgements

This work has been supported by the project TIN2016-79717-R of the Spanish Ministry of Science, CHIST-ERA project PCIN-2015-226 and Masana acknowledges 2017FI-B-00218 grant of Generalitat de Catalunya. We also acknowledge the generous GPU support from NVIDIA corporation.

References

[1] Achanta, R., Hemami, S., Estrada, F., Susstrunk, S., 2009. Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1597–1604.

[2] Alexe, B., Deselaers, T., Ferrari, V., 2012. Measuring the objectness of image windows. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. IEEE, pp. 2189–2202.

[3] Arbelaez, P., Pont-Tuset, J., Barron, J., Marques, F., Malik, J., 2014. Multiscale combinatorial grouping. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 328–335.

[4] Borji, A., Itti, L., 2013. State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (1), 185–207.

[5] Borji, A., Sihite, D. N., Itti, L., 2012. Salient object detection: A benchmark. In: European Conference on Computer Vision. Springer, pp. 414–429.

[6] Carreira, J., Sminchisescu, C., 2010. Constrained parametric min-cuts for automatic object segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3241–3248.

[7] Cheng, M.-M., Zhang, Z., Lin, W.-Y., Torr, P., 2014. Bing: Binarized normed gradients for objectness estimation at 300fps. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3286–3293.

[8] de San Roman, P. P., Benois-Pineau, J., Domenger, J.-P., De Rugy, A., Paclet, F., Cataert, D., 2017. Saliency driven object recognition in

egocentric videos with deep cnn: toward application in assistance to neuroprostheses. Computer Vision and Image Understanding.

[9] Duncan, J., Humphreys, G. W., 1989. Visual search and stimulus similarity. Psychological review 96 (3), 433.

[10] Endres, I., Hoiem, D., 2014. Category-independent object proposals with diverse ranking. IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2), 222–234.

[11] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9), 1627–1645.

[12] Frintrop, S., Werner, T., García, G. M., 2015. Traditional saliency reloaded: A good old model in new shape. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 82–90.

[13] Gaborski, R., Vaingankar, V. S., Canosa, R., 2003. Goal directed visual search based on color cues: Cooperative effects of top-down & bottom-up visual attention. Proceedings of the Artificial Neural Networks in Engineering, Rolla, Missouri 13, 613–618.

[14] Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., Dosil, R., 2009. Decorrelation and distinctiveness provide with human-like saliency. In: International Conference on Advanced Concepts for Intelligent Vision Systems. Springer, pp. 343–354.

[15] Goferman, S., Zelnik-Manor, L., Tal, A., 2012. Context-aware saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (10), 1915–1926.

[16] Han, B., Li, X., Gao, X., Tao, D., 2012. A biological inspired features based saliency map. In: Computing, Networking and Communications (ICNC), 2012 International Conference on. IEEE, pp. 371–375.

[17] Harel, J., Koch, C., Perona, P., 2006. Graph-based visual saliency. In: Advances in neural information processing systems. pp. 545–552.

[18] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

[19] Hosang, J., Benenson, R., Dollár, P., Schiele, B., 2015. What makes for effective detection proposals? IEEE Transactions on Pattern Analysis & Machine Intelligence.

[20] Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z., Torr, P., 2017. Deeply supervised salient object detection with short connections. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, pp. 5300–5309.

[21] Hou, X., Zhang, L., 2007. Saliency detection: A spectral residual approach. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–8.

[22] Hu, H., Lan, S., Jiang, Y., Cao, Z., Sha, F., 2016. Fastmask: Segment multi-scale object candidates in one shot. arXiv preprint arXiv:1612.08843.

[23] Huo, L., Jiao, L., Wang, S., Yang, S., 2016. Object-level saliency detection with color attributes. Pattern Recognition 49, 162–173.

[24] Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis & Machine Intelligence 20 (11), 1254–1259.

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

[26] Koch, C., Ullman, S., 1987. Shifts in selective visual attention: towards the underlying neural circuitry. In: Matters of intelligence. Springer, pp. 115–141.

[27] Krähenbühl, P., Koltun, V., 2014. Geodesic object proposals. In: European Conference on Computer Vision. Springer, pp. 725–739.

[28] Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105.

[29] Li, G., Yu, Y., 2015. Visual saliency based on multiscale deep features. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 5455–5463.

[30] Li, G., Yu, Y., June 2016. Deep contrast learning for salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 478–487.

[31] Li, G., Yu, Y., 2016. Deep contrast learning for salient object detection. arXiv preprint arXiv:1603.01976.

[32] Li, S., Lu, H., Lin, Z., Shen, X., Price, B., 2015. Adaptive metric learning for saliency detection. IEEE Transactions on Image Processing 24 (11), 3321–3331.

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

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

[35] Liu, N., Han, J., 2016. Dhsnet: Deep hierarchical saliency network for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 678–686.

[36] Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.-Y., 2011. Learning to detect a salient object. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 33. IEEE, pp. 353–367.

[37] Lou, J., Ren, M., Wang, H., 2014. Regional principal color based saliency detection. PLoS ONE 9 (11), e112475.

[38] Mairon, R., Ben-Shahar, O., 2014. A closer look at context: From coxels to the contextual emergence of object saliency. In: The European Conference on Computer Vision. Springer, pp. 708–724.

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

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

[41] Martin, D. R., Fowlkes, C. C., Malik, J., 2004. Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (5), 530–549.

[42] Murray, N., Vanrell, M., Otazu, X., Parraga, C. A., 2013. Low-level spatiochromatic grouping for saliency estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (11), 2810–2816.

[43] Pan, J., Sayrol, E., Giro-i Nieto, X., McGuinness, K., O'Connor, N. E., 2016. Shallow and deep convolutional networks for saliency prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 598–606.

[44] Pinheiro, P. O., Lin, T.-Y., Collobert, R., Dollár, P., 2016. Learning to refine object segments. In: European Conference on Computer Vision. Springer, pp. 75–91.

[45] Razavian, S., Azizpour, A., H. Sullivan, J., Carlsson, S., 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 806–813.

[46] Rudoy, D., Goldman, D. B., Shechtman, E., Zelnik-Manor, L., 2013. Learning video saliency from human gaze using candidate selection. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1147–1154.

[47] Scharfenberger, C., Wong, A., Fergani, K., Zelek, J. S., Clausi, D., et al., 2013. statistical textural distinctiveness for salient region detection in natural images. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 979–986.

[48] Seo, H. J., Milanfar, P., 2009. Static and space-time visual saliency detection by self-resemblance. Journal of vision 9 (12), 15–15.

[49] Siagian, C., Itti, L., 2007. Rapid biologically-inspired scene classification using features shared with visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2), 300–312.

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

[51] Stella, X. Y., Lisin, D. A., 2009. Image compression based on visual saliency at individual scales. In: Advances in Visual Computing. Springer, pp. 157–166.

[52] Sun, J., Lu, H., Liu, X., 2015. Saliency region detection based on markov absorption probabilities. IEEE Transactions on Image Processing 24 (5), 1639–1649.

[53] Tong, N., Lu, H., Zhang, Y., Ruan, X., 2015. Salient object detection via global and local cues. Pattern Recognition 48 (10), 3258–3267.

[54] Treisman, A. M., Gelade, G., 1980. A feature-integration theory of attention. Cognitive psychology 12 (1), 97–136.

[55] Uijlings, J. R., van de Sande, K. E., Gevers, T., Smeulders, A. W., 2013. Selective search for object recognition. International journal of computer vision 104 (2), 154–171.

[56] Wan, S., Jin, P., Yue, L., 2009. An approach for image retrieval based on visual saliency. In: International Conference on Image Analysis and Signal Processing. IEEE, pp. 172–175.

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

[58] Wang, L., Wang, L., Lu, H., Zhang, P., Ruan, X., 2016. Saliency detection with recurrent fully convolutional networks. In: European Conference on Computer Vision. Springer, pp. 825–841.

[59] Wang, M., Konrad, J., Ishwar, P., Jing, K., Rowley, H., 2011. Image saliency: From intrinsic to extrinsic context. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 417–424.

[60] Wang, S., Jiang, M., Duchesne, X. M., Laugeson, E. A., Kennedy, D. P., Adolphs, R., Zhao, Q., 2015. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. Neuron 88 (3), 604–616.

[61] Wang, Z., Li, B., 2008. A two-stage approach to saliency detection in images. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 965–968.

[62] Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., Xiao, J., 2015. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. arXiv preprint arXiv:1504.06755.

[63] Yan, Q., Xu, L., Shi, J., Jia, J., 2013. Hierarchical saliency detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1155–1162.

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

[65] Zhang, D., Fu, H., Han, J., Wu, F., 2016. A review of co-saliency detection technique: Fundamentals, applications, and challenges. CoRR abs/1604.07090.
URL `http://arxiv.org/abs/1604.07090`

[66] Zhao, Q., Koch, C., 2013. Learning saliency-based visual attention: A review. Signal Processing 6, 1401–1407.

[67] Zhao, R., Ouyang, W., Li, H., Wang, X., 2015. Saliency detection by multi-context deep learning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1265–1274.