Background-Driven Salient Object Detection

Zilei Wang, *Member, IEEE*, Dao Xiang, Saihui Hou, and Feng Wu, *Fellow, IEEE*

Abstract—**The background information is a significant prior for salient object detection, especially when images contain cluttered background and diverse object parts. In this paper, we propose a background-driven salient object detection (BD-SOD) method to more comprehensively exploit the background prior, aiming at generating more accurate and robust salient maps. To be specific, we first exploit the background prior to conduct the saliency estimation, i.e., computing the regional saliency values. In this stage, the background prior is utilized in threefold: restricting the reference regions to only the background regions, weighting the contribution of reference regions, and leveraging the importance of different features. Benefiting from such an explicit utilization, the proposed model can greatly mitigate the negative interference of the cluttered background and diverse object parts. We then embed the background prior into the optimization graph for saliency refinement. Specifically, two virtual supernodes (representing the background and foreground, respectively) are introduced with extra connections, and the nonlocal feature connections between similar regions are also set up. These connections enhance the power of optimization graph to alleviate the perturbations from diverse parts, and thus help to achieve the uniformity of saliency values. Finally, we provide systematical studies to investigate the effectiveness of the proposed BD-SOD in exploiting the valuable background prior. Experimental results on multiple public benchmark datasets, including MSRA-1000, THUS-10000, PASCAL-S, and ECSSD, clearly show that BD-SOD consistently outperforms the well-established baselines and achieves state-of-the-art performance.**

Index Terms—**Salient object detection, background prior, saliency estimation, graph-based optimization.**

I. Introduction

SALIENT object detection aims to identify the spatial locations and scales of the most attention-grabbing objects in a given image [1], [2]. It is an important problem in the multimedia research and has shown to be helpful for various tasks, such as image retrieval [3], [4], adaptive image display [5], and content-aware image editing [6]. Different from eye-fixation saliency prediction [7], salient object detection emphasizes *saliency* and *integrity* of detected objects. Thus how to deal with the cluttered background and diversity of object parts within an image is always one of the major challenges for salient object detection [8].

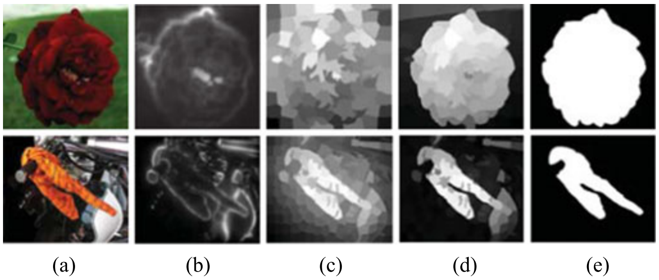

Fig. 1. Visual comparison of different spatial reference strategies for contrast computation. From left to right: (a) the original image; (b) local contrast maps [1]; (c) global contrast maps [10]; (d) our produced saliency maps by restricting the reference regions to identified background; and (e) ground truth. Better viewed in the color version.

The visual saliency conceptually originates from the uniqueness, rarity, or unpredictability of salient objects [9]. Thus it is commonly characterized by the *contrast* of primitive image features (e.g., color, intensity, or orientation) in the pixel or superpixel level [1], [8]–[17], as contrast is one of the most predominant factors in human cognition [18]. In the literature, the local [15] or global [9], [10] contrast is usually adopted to derive the saliency map.

Specifically, the local contrast methods compare the features of each region with those of its neighbors to compute the saliency value. Consequently, only the edge pixels with high contrasts are predicted to be salient and the inner foreground pixels belonging to salient objects are missed [see Fig. 1(b)]. On the opposite, the global contrast methods refer to all other pixels, and thus can achieve better internal consistency. But some background regions may also be assigned high saliency values [see Fig. 1(c)]. To produce more reasonable saliency maps, the recently proposed methods [17], [19], [20] usually combine the local and global contrasts. The computed contrasts in these methods essentially represent the peculiarity of a region within one image, which is reasonable to salient object detection for simple images. However, a critical gap would occur if the background becomes cluttered or the parts of a foreground object are very diverse, as partial background regions may be considered salient or different parts of one object may get highly varying saliency values. To tackle this important issue, in this work, we particularly take into account the disparities of regional contents other than the spatial relationship between image regions, i.e., adopting different processing strategies for different contents when deriving the saliency value of a certain region. More specifically, the background prior, which has been shown helpful for

Manuscript received July 1, 2016; revised October 13, 2016 and December 2, 2016; accepted December 2, 2016. Date of publication December 7, 2016; date of current version March 15, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61673362 and Grant 61233003, and in part by the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jing-Ming Guo.

Z. Wang, D. Xiang, and S. Hou are with the Department of Automation, University of Science and Technology of China, Hefei 230027, China (e-mail: zlwang@ustc.edu.cn; xiangdao@mail.ustc.edu.cn; saihui@mail.ustc.edu.cn).

F. Wu is with the School of Information Science and Technology, West Campus, University of Science and Technology of China (USTC), Hefei 230027, China (e-mail: fengwu@ustc.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2636739

improving the detection performance [8], [12], [14], would be comprehensively exploited [see Fig. 1(d)].

We propose a background driven salient object detection (BD-SOD) method in this paper. Our purpose is to improve the quality of saliency maps, i.e., producing more accurate and robust saliency values, such that the foreground objects are easier to detect as a whole. In particular, we use a background map, consisting of the probability of each region belonging to the background, to represent the background prior of an image. Such background maps can be obtained in practice by employing some heuristic method [8] or inverting the saliency map produced by some handy detection method [10], [19]. Using the background map, we can further identify the foreground and background regions of an image (e.g., via a thresholding strategy). The background map, background regions and foreground regions together form the background clues we would exploit in the subsequent saliency computations, including saliency estimation and saliency refinement.

The first stage of BD-SOD, saliency estimation, is to predict the contrasts of image regions independently and then produce an initial saliency map for each image. Here we propose to conduct such saliency estimation in the light of background clues, which are specifically exploited in three folds. First, we restrict the reference regions to only background regions when computing the contrast of a target region. Such a restriction directly eliminates the negative interference of diverse object parts by significantly suppressing their effects. Second, we adaptively weight the contributions of different reference regions according to their degree of correlation with the target region in content. Such a weighting strategy predisposes the background regions to get smaller saliency values owing to their higher probability of being similar to reference regions. Third, we utilize the identified background and foreground to leverage the importance of different features (e.g., color, texture) such that the predominant feature that determines the regional saliency can be automatically captured according to the specific characteristics of an image.

The second stage of BD-SOD, saliency refinement, is to optimize the saliency map by computing the saliency correlations between image regions and then conducting saliency diffusion. It aims to produce more robust saliency maps, i.e., making the same type of regions gain nearly equal saliency values. In practice, a graph-based optimization model is usually adopted [13], [21], where the nodes represent image regions and the weighted edges represent their connection strengths in saliency diffusion. In this paper we propose to embed the background prior into the optimization graph, i.e., building an enhanced graph. Specifically, we first introduce two virtual supernodes that represent the background and foreground respectively, and each regular node is required to connect both of them. Such a process can mitigate the perturbations from cluttered background and diverse object parts, since the supernodes represent the collection of diverse image contents and the connections to them are more reliable. Moreover, we set up extra nonlocal connections between similar regions in the feature and saliency spaces. These connections encourage the same type of regions to get closer saliency values even if they are spatially far from each other. This is important especially for the objects containing similar but far-between parts, e.g., ones with an elongated shape.

Benefiting from the explicit and comprehensive exploitation of background prior, the proposed BD-SOD can produce more accurate and robust saliency maps. We systematically evaluate the effectiveness of BD-SOD through the experiments on multiple benchmark datasets, including *MSRA-1000* [2], *THUS-10000* [9], *PASCAL-S* [22], and *ECSSD* [23]. The results show that BD-SOD is always able to boost the detection performance no matter which background prior is fed, and achieves state-of-the-art overall performance on these benchmarks.

The main contributions of this work are summarized as follows:

1) We propose a background driven salient object detection method, which exploits the identified background prior to compute saliency maps in multiple folds. Consequently, more accurate and robust saliency maps can be generated.
2) We propose a saliency refinement model by embedding the background prior into the optimization graph, which can effectively mitigate the interference of cluttered background and diverse object parts. Thus the object can be finally assigned more consistent saliency values.
3) We experimentally compare and analyze the representative saliency detection approaches with different background priors, and provide convincing evidence for the importance of background prior in salient object detection.

The rest of this paper is organized as follows. We review related works in Section II. We provide the brief skeleton of BD-SOD in Section III, and then elaborate on the details of two main components in Section IV and Section V. In Section VI, we report the empirical results on public benchmarks. Finally, the conclusions are provided in Section VII.

II. Related Works

This paper focuses on the data-driven bottom-up salient object detection. Most existing methods are inspired by the biological visual attention mechanism, which compute such bottom-up saliency using the contrast of image features [24]. To be specific, Itti *et al.* [15] first propose the fundamental framework of the contrast-based saliency model, which particularly uses the center-surrounded differences across multi-scale low-level features. Motivated by this pioneer work, many approaches [1], [9], [10], [14], [17], [25]–[28] are then proposed along a similar route. These methods mainly differ in one or more of the following aspects: the used image features, spatial scope of surroundings, level of operations (pixels or regions), and contrast metrics. On the other hand, deep neural network (DNN) is recently introduced to salient object detection due to its powerful learning ability, and shows impressive detection performance.

A. Contrast-Based Saliency Estimation

Among the consideration factors in designing the contrast based methods, the spatial scope of surroundings is a crucial one that determines the capability of detecting salient objects. Thus we review this type of works along such a major route. According to the spatial scope of referred surroundings, we can roughly divide them into two broad types, i.e., the local contrast methods and global contrast methods.

Local contrast methods only use the immediate neighborhoods of a pixel or region to compute contrasts. For example, Liu *et al.* [1] propose a set of local features including multi-scale contrasts, center-surround histogram, and color spatial distribution to describe a salient object, and choose conditional random field to combine them for saliency estimation. Similarly, Ma *et al.* [26] employ the local color dissimilarities computed by Gaussian distance to measure the pixel-level contrasts. Goferman *et al.* [29] utilize multiple context-aware factors (local low-level color contrasts, global considerations, visual organizational rules, *etc.*) to conduct saliency detection. For the results, the local contrast methods tend to focus on the object boundaries rather than uniformly highlighting the entire object, or react to small conspicuous background regions [9].

Global contrast methods refer to the entire image to compute contrasts. For example, Fang *et al.* [28] use the amplitude spectrum of quaternion Fourier transform (QFT) to represent the color, intensity, and orientation distributions of image patches. Then the saliency value of one patch is obtained by calculating the global differences between the QFT amplitude spectrum of the patch and that of all other patches. Margolin *et al.* [30] integrate the pattern and color distinctness in a unique manner, in which principal component analysis (PCA) is employed in representing the set of image patches. Shi *et al.* [23] estimate the regional saliency by combining local contrast and location heuristic in different scales using hierarchical inference. Generally, the global contrast models can achieve more consistent results than the local ones. But the resulting saliency maps may be obscuring, especially for the images with large-sized salient objects or visually varied background.

The plain local or global contrast methods mainly consider the spatial relationship when merging the contrasts of reference regions into the overall saliency of a region. And the contents of image regions are ignored completely although they are indeed important according to the semantics of *saliency*. In this work, we particularly take into account the type of regional contents in saliency computation, i.e., exploiting the background prior. Actually, several previous works [12], [14] have attempted to use the background prior though they just consider the pre-defined image boundaries or manually selected seeds in a heuristic manner. Here directly taking the image boundaries is from an empirical observation that the boundary regions of images are mostly likely to belong to background. For example, Wei *et al.* [12] define the saliency as the shortest-path distance to image boundary. Jiang *et al.* [14] compute the contrast with image boundary as the backgroundness feature. Yang *et al.* [13] use the boundary patches as background queries in evaluating saliency. Zhu *et al.* [8] employ the boundary connectivity to measure saliency, which essentially characterizes the spatial layout of image regions with respect to image boundaries. Similarly, Li *et al.* [31] introduce the distances to the heuristically selected foreground/background seeds to define saliency, where two Mahalanobis distances and a superpixel-wise fisher vector representation are specially proposed. Compared with the plain contrast-based methods, these methods can work better owing to utilizing the background prior. However, they strongly depend on the selection of background prior.[1] For example, the heuristic boundary prior may fail to highlight the salient objects when the object touches the boundary, or the boundary is too flat to represent the visually varied background (e.g., the image has a border), as shown in Fig. 2.

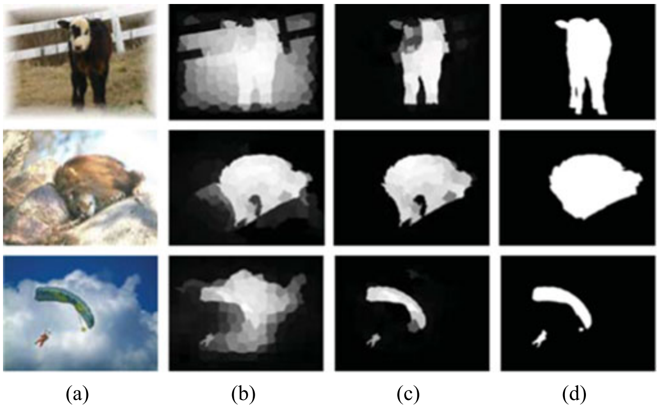

Fig. 2. Visual comparison of different background priors on helping saliency computation. From left to right: (a) the original image; (b) the generated saliency map by [8] with the boundary prior; (c) our detection results with the predicted background map; and (d) the ground truth. Obviously, the heuristic boundary prior would fail when the boundary cannot represent the visually varied background. Best viewed in the color version.

B. Saliency Computation With Deep Neural Network

Recently, some researchers [19], [20], [32], [33] introduce deep neural network to salient object detection due to its powerful learning ability [34]. Specifically, Han *et al.* [35] use stacked denoising autoencoders under deep learning architectures to model the background of an image, and then formulate the separation of salient objects from the background as a problem of measuring reconstruction residuals of deep autoencoders. In addition, Li *et al.* [19] propose to employ three deep convolutional neural networks (CNN) to extract multi-scale features from nested windows and then adopt a deep neural network with two fully connected layers to estimate saliency values. Similarly, Zhao *et al.* [32] propose to detect salient objects in multi-contexts, where the global and local context of a superpixel are modeled under a unified deep learning framework. Essentially, the two works utilize DNN to extract richer features with respect to saliency, and then establish an end-to-end saliency predictor. On the other hand, Wang *et al.* [20] propose to not only learn patch features with DNN-L for local saliency detection, but also conduct global search with DNN-G to directly predict regional saliency scores. He *et al.* [33] propose a SuperCNN to learn the contrast features from the superpixel-level color uniqueness and color distribution sequences. These two works put their emphasis on the utilization of the powerful modeling capacity of DNN for computing the saliency values. It can be seen that DNN is mainly applied to learn more discriminative features or model the saliency prediction.

[1] We provide more comparative analyses in the experiments.

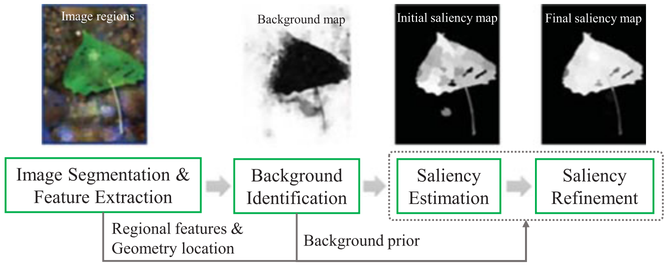

Fig. 3. Pipeline of the proposed BD-SOD. Firstly, an image is segmented into visually homogeneous regions, and the regional features are extracted. Secondly, the background regions are identified, resulting in a background map for each image. Thirdly, the contrast-based saliency is independently estimated for each region, forming an initial saliency map. Finally, the saliency map is refined via the graph-based saliency diffusion. Particularly, both saliency estimation and saliency refinement explicitly exploit the background prior besides the traditional contrast and geometry location. Best viewed in the color version.

In this paper, we particularly focus on how to better predict the saliency values given regional features. Meanwhile, different from the learning-based methods that build a fixed saliency prediction model by training, we aim to build a novel adaptive approach to obtain more accurate and uniform saliency maps, which can derive the saliency values according to the regional contents and specific characteristics of an image.

III. Approach Skeleton

Aiming at obtaining more robust and accurate saliency maps, we propose a Background Driven Salient Object Detection (BD-SOD) approach in this paper. Particularly, BD-SOD is designed to achieve such a goal by mitigating the negative interference from diverse regions and uniforming the saliency values of foreground objects. Here the background prior would be deeply exploited as one of major clues. Fig. 3 illustrates the pipeline of the proposed BD-SOD. We first give a brief introduction to each component as follows.

1) *Image segmentation*: Image segmentation is to decompose images into visually homogeneous superpixels, which can produce more robust saliency detection results and improve processing efficiency compared to the pixel-level saliency detection [9], [10]. In this work, we adopt the SLIC method [36] to segment images into compact and edge-preserving superpixels. For a given image $\mathcal{I}$, let $R = \{r_i\}_{i=1}^n$ denote the set of produced superpixels, where n is the number of regions. Then we extract the features of all superpixels, denoted by $\{\mathbf{f_i}\}_{i=1}^n$. Since the superpixels produced by SLIC usually have relatively regular shapes, the geometric position of one image region is described by its central point $\mathbf{p}_i$, which is generated by averaging the x/y coordinates of pixels contained in the region.

2) *Background identification*: Background identification is to identify the potential background regions of a new image. We adopt a background map $\mathcal{B} = \{b_i\}_{i=1}^n$ to represent the identified background, where the normalized b_i represents the probability of the region r_i belonging to the background. Such a background map can be produced heuristically [8] or by some learning model [19], and we would elaborate on the details of the used methods in the experimental settings. Besides such a soft background map, we generate the corresponding background regions (denoted by R_b) and foreground regions (denoted by R_f) to indicate the background/foreground subset of image regions. In our implementation, the thresholding strategy is adopted due to its high efficiency. In practice, two thresholds T_b and T_f are preset with $T_b > T_f$, and then a region is considered to belong to R_b if $b_i \geq T_b$ or R_f if $b_i \leq T_f$. Specially, the regions satisfying $T_f < b_i < T_b$ are considered unreliable and thus would not be used in saliency computation as the background clue.

3) *Saliency estimation*: Saliency estimation is to predict the saliency values of image regions separately from the extracted features and obtained background clues (i.e., the identified background and foreground). To yield an accurate saliency map, we particularly focus on two important aspects of utilizing the background clues. Firstly, we adaptively leverage the importance of different features via the weights derived from R_b and R_f. Secondly, we strictly restrict the reference regions to R_b, and meanwhile weight their contributions according to the disparities of their features and feature of the target region when pooling the pair-wise contrasts. As a result, an initial saliency map $\tilde{S}$ consisting of the regional saliency values $\{\tilde{s}_i\}_{i=1}^n$ would be generated.

4) *Saliency refinement*: Saliency refinement is to re-evaluate the saliency values of image regions by utilizing their pair-wise saliency correlations, and the purpose is to improve saliency consistency, which is to let the same type of regions get nearly equal saliency values no matter how diverse their features are. This process is important for retaining the integrity of salient objects. In this work, we conduct saliency refinement through embedding the background prior into the optimization graph, where the obtained background map would mainly be utilized. Specifically, two virtual supernodes, i.e., the background node V_b and foreground node V_f, are introduced with extra connections, and multiple types of connection edges are set up. The saliency refinement would produce the final saliency map $\mathcal{S} = \{s_i\}_{i=1}^n$.

As shown in Fig. 3, image segmentation provides the operands of saliency computation (superpixels with features), and background identification provides the background prior we need to exploit in this paper. According to such a pipeline, the key of BD-SOD actually lies in the last two components, i.e., saliency estimation and saliency refinement, which truly exploit the provided information to produce the saliency maps. We elaborate on each of them in the following two sections.

IV. Background-Driven Saliency Estimation

In this section, we provide the details of saliency estimation, i.e., how to better compute an initial saliency map $\tilde{S}$ using the regional features and background clues. Here the contrast-based model is taken owing to its good adaptability to the image-specific characteristics, and we would exploit the background clues in two major aspects: feature leverage and contrast estimation.

A. Features

In this work, we adopt two typical visual features, i.e., the color and texture, due to their predominance in determining saliency [18]. Specifically, we take the average value of a region in the CIE-Lab color space as its color descriptor, which can eliminate the slight noise within a homogeneous region. Let $\mathbf{c}_j$ denote the color vector of the j-th pixel. Then the color feature $\mathbf{f}_i^c$ of the region r_i is

$$\mathbf{f}_i^c = \frac{1}{m_i} \sum_{j \in r_i} \mathbf{c}_j \tag{1}$$

where m_i is the number of pixels contained by the region r_i. For the texture feature, a 2-level Haar wavelet decomposition [37] is employed due to its effectiveness in saliency computation. Specifically, the Haar wavelet decomposition, which covers the 3 color channels, 2 levels, and 6 feature images (i.e., the mean of the coefficients over a 3×3 neighborhood for the horizontal, vertical, diagonal and approximation sub-images, and the gradient and variance over the same neighborhood in the approximation sub-image), generates the pixel-level textures and then the texture feature of a region is obtained by averaging on the pixels belonging to the region. As a result, a 36-dimensional texture feature $\mathbf{f}_i^t$ is generated for the region r_i.

We produce the final regional features by concatenating all extracted features in a weighting manner, where the weights indicate the importance of different features. For our used color and texture features, the final feature $\mathbf{f}_i$ of the region r_i is denoted by

$$\mathbf{f}_i = [w_c \mathbf{f}_i^c ; w_t \mathbf{f}_i^t] \tag{2}$$

where w_c and w_t correspond to the weights of the color and texture features.

Usually, color is considered as a major factor in saliency prediction, and texture plays a complementary role. However, it is not always right for all natural images, and their relative importance may vary a lot. To automatically leverage the importance of different features according to the image-specific characteristics, we introduce an adaptive strategy guided by the background clues. Intuitively, the weight should represent the relative power of a certain feature in describing the saliency. For example, w_t needs to be larger if the foreground and background mainly differ in texture. Hence, we heuristically compute the importance weight of a feature according to its dissimilarity between the identified foreground and background regions. Here the overlap of the feature histograms is specifically adopted. Formally, for one dimension of a certain feature, let $\mathbf{h}_f$ and $\mathbf{h}_b$ denote the normalized histograms of R_f and R_b. Then the overlap is defined as

$$o(\mathbf{h}_f, \mathbf{h}_b) = \frac{2 \sum_{i=1}^{l} h_f(i) \cdot h_b(i)}{\sum_{i=1}^{l} (h_f(i)^2 + h_b(i)^2)} \tag{3}$$

where $h_f(i)$ and $h_b(i)$ are the i-th elements of $\mathbf{h}_f$ and $\mathbf{h}_b$, and l is the number of histogram bins. Specially, $l = 25$ is set for the color feature and $l = 60$ for the texture feature. It is evident that $o \in [0, 1]$. Finally, the overlap of a feature is generated by averaging on all the one-dimensional overlaps.

The feature with a smaller overlap is considered more discriminative and thus needs to be assigned a larger weight. Let o_c and o_t denote the overlap values of color and texture, respectively. Then the corresponding color weight w_c and texture weight w_t are defined as follows:

$$w_c = \exp\left\{-\frac{o_c}{o_c + o_t}\right\}, w_t = \exp\left\{-\frac{3 \times o_t}{o_c + o_t}\right\}. \tag{4}$$

Here a fixed penalty factor of 3 is set for w_t due to the observation that color is generally more reliable than texture in saliency detection. That is, the feature weight is determined by both the image-specific characteristics and feature discrimination.

B. Contrast Estimation

The contrast of a region refers to its overall difference from other regions within the same image, and it is generally believed that the region with a higher contrast is apter to catch human attention [10]. In this work, we measure the difference between image regions using the Euclidean distance of their features as in the previous works [8]–[11]. Formally, the difference between the regions r_i and r_j is

$$d_{ij} = \|\mathbf{f}_i - \mathbf{f}_j\|_2^2 = w_c^2 \|\mathbf{f}_i^c - \mathbf{f}_j^c\|_2^2 + w_t^2 \|\mathbf{f}_i^t - \mathbf{f}_j^t\|_2^2. \tag{5}$$

In saliency estimation, however, each region is expected to only assign a scalar value $\tilde{s}_i$ to represent its overall contrast. Thus we need to integrate the pair-wise differences $\{d_{ij}\}_{j \neq i}$ into $\tilde{s}_i$. Indeed, there are several typical pooling strategies in computer vision, e.g., max-pooling and average-pooling [38]. Considering the scenario of saliency detection, we adopt the weighted sum-pooling strategy here, i.e.,

$$\tilde{s}_i = \sum_{j \neq i} \omega_{ij} d_{ij} = \sum_{j \neq i} \omega_{ij} \|\mathbf{f}_i - \mathbf{f}_j\|_2^2 \tag{6}$$

where the weight ω_{ij} is used to control the contribution of d_{ij} to the overall contrast $\tilde{s}_i$. Obviously, how to define ω_{ij} would determine the quality of the saliency map $\tilde{\mathcal{S}}$.

In this paper, we particularly take into consideration the contents of involved regions besides the traditional geometric factor when determining such important weights. That is, ω_{ij} would strongly depend on the contents of the regions r_i and r_j, which are represented by the features $\mathbf{f}_i$ and $\mathbf{f}_j$. Before elaborating on the details, we first give the geometric factor that expresses the effect of different spatial locations on determining saliency values. Here a Gaussian metric is employed since a farther reference region usually has less impact on the overall contrast [8]–[10]. Assume the central points of the regions r_i and r_j are $\mathbf{p}_i$ and $\mathbf{p}_j$. Then the geometric factor is defined as $\mathcal{G}_i(\mathbf{p}_j) = \exp\{-\|\mathbf{p}_i - \mathbf{p}_j\|_2^2 / 2\sigma_p^2\}$, which represents the spatial effect of the reference region r_j to $\tilde{s}_i$. Here σ_p is a bandwidth parameter to control the spatial range of pooling the pair-wise contrasts.

Now we discuss how to effectively utilize the content information of the reference regions with the goal of more accurately estimating the saliency value of a certain region. In salient

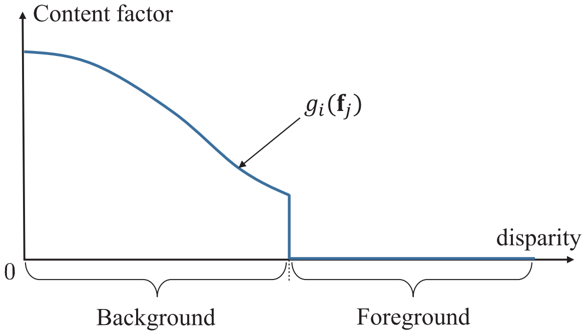

Fig. 4. Illustration of content factor for different reference regions. Specifically, the factor is set as 0 for the foreground references. As for the background references, the factor is monotonically decreasing as the disparity between the reference region and target region increases. That is, it is expected that the saliency value of a region is determined mainly by the most similar background references.

object detection, the obstacles to achieving accurate saliency maps mainly lie in the interference of cluttered background and diverse object parts, especially for complicated images. Meanwhile, salient objects in semantics are opposite to background. Therefore, we propose to treat the reference regions in a content-specific manner rather than blindly process them. In this work, we particularly propose two specific strategies to mitigate the interferences within an image. Firstly, we restrict the reference regions of contrast estimation to only the background regions. Due to explicitly utilizing the content types (background or foreground), such a restriction almost completely eliminates the interference of diverse foreground parts. Secondly, we weight the contributions of the reference regions according to their degrees of content correlation to the target region, aiming at further alleviating the interference of cluttered background. To be specific, we propose to reduce the effects of the reference regions that are too different from the target region, which means more similar reference regions would have larger contributions. Consequently, the background regions tend to result in lower saliency values since they are more likely to be similar to reference regions.

Let $g_i(\mathbf{f}_j)$ denote the content factor of the reference region r_j to $\tilde{s}_i$. Now we investigate the concrete form of $g_i(\mathbf{f}_i)$, where the background map $\mathcal{B}$ together with R_b is used as the background clue. According to the above analysis, $g_i(\mathbf{f}_j)$ should be piecewise for distinguishing the types of reference regions, and strongly depend on $\mathbf{f}_j$ for reflecting the regional contents. Fig. 4 illustrates such an expression curve. More specifically, we adopt the Gaussian function to define $g_i(\mathbf{f}_j)$, i.e.,

$$g_i(\mathbf{f}_j) = \begin{cases} \exp\{-d_{ij}/2\sigma_f^2\}, & \text{if } r_j \in R_b \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where σ_f is a bandwidth parameter to control the decay rate of the content factor.

By now we can obtain the overall weight ω_{ij} by multiplying the geometry and content factors, and further compute the saliency value $\tilde{s}_i$ of the region r_i by replacing ω_{ij} in (6), i.e.,

$$\begin{aligned} \tilde{s}_i &= \sum_{j \neq i} g_i(\mathbf{f}_j) \mathcal{G}_i(\mathbf{p}_j) d_{ij} \\ &= \sum_{j \in R_b} w_{ij} \, \|\mathbf{f}_i - \mathbf{f}_j\|_2^2 \end{aligned} \tag{8}$$

where

$$w_{ij} = \exp\left\{ -\frac{\|\mathbf{f}_i - \mathbf{f}_j\|_2^2}{2\sigma_f^2} - \frac{\|\mathbf{p}_i - \mathbf{p}_j\|_2^2}{2\sigma_p^2} \right\} \tag{9}$$

represents the connection strength between the regions r_i and r_j. It is noticeable that from (8) the reference regions of computing a saliency value have been strictly restricted to the background regions, i.e., the identified background R_b. Compared with referring to the entire image in the previous works, such a restriction would result in more robust and accurate saliency maps. Our extensive experiments verified its effectiveness, and Fig. 5(h) provides a visual example.

V. Background Prior-Embedded Saliency Refinement

In saliency estimation, the saliency values of different image regions are computed independently, and their saliency correlations are not exploited at all. Consequently, the predicted saliency map $\tilde{S}$ may be highly sensitive to slight variations of regional contents [39]. For salient object detection, such instability and intra-object inconsistency would damage the final detection performance. To tackle this issue, saliency optimization is usually conducted which essentially diffuses the saliency values among image regions according to their pair-wise relationship. It has been proven that saliency optimization can achieve better performance compared to directly adopting the estimated contrasts [10], [12], [30]. Saliency optimization is generally implemented by a graph-based model, where the nodes denote the regions of an image and the weighted undirected edges represent their pair-wise connections [13], [21], [40]. To generate more robust saliency maps, we propose to build an enhanced optimization graph in this work, where the background prior would be embedded.

Formally, let $G = (V, E)$ represent a regular graph, where V is the set of regional nodes with the size $|V| = n$, and E is the set of undirected edges. Concretely, the node v_i corresponds to the region r_i, and the edge e_{ij} (equivalent to e_{ji}) corresponds to the connection between v_i and v_j with a weight w_{ij} to denote the correlation degree in diffusing saliency. Here the weight defined in (9) is used for each pair of nodes. It can be seen that the connection strength of two regions may be low if their features are very different no matter whether they belong to the same type (foreground or background). In saliency refinement, however, a larger weight is actually expected if two involved regions belong to the same type, such that their saliency values are closer. Such a negative situation probably occurs for the complicated images consisting of diverse regions. In this work, we plan to address this issue by introducing extra connections rather than directly rectifying the weights in (9). Specifically, we introduce two virtual supermodes into the optimization graph

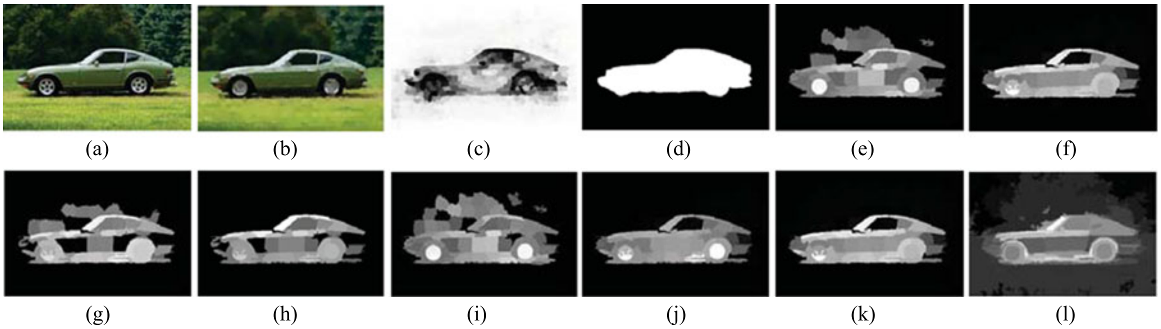

Fig. 5. Component-wise visual example of BD-SOD. Here the notations are as follows: Clr and Txt denote the color and texture features; Ide, Est, and Ref represent our main components, i.e., the background identification, saliency estimation, and graph-based saliency refinement. Then we can show the resulting saliency maps for their different combinations: (a) the raw image; (b) segmented image, where the average color in RGB is used for each superpixel; (c) learned background map by single CNN; (d) ground truth; (e) $\mathrm{Txt} + \mathrm{Ide} + \mathrm{Est} + \mathrm{Ref}$; (f) $\mathrm{Clr} + \mathrm{Ide} + \mathrm{Est} + \mathrm{Ref}$; (g) $\mathrm{Clr} + \mathrm{Txt} + \mathrm{Est} + \mathrm{Ref}$; (h) $\mathrm{Clr} + \mathrm{Txt} + \mathrm{Ide} + \mathrm{Est}$; (i) $\mathrm{Txt} + \mathrm{Est} + \mathrm{Ref}$; (j) $\mathrm{Clr} + \mathrm{Txt} + \mathrm{Ide} + \mathrm{Est} + \mathrm{Ref}$ (i.e., the complete BD-SOD); (k) optimized saliency map by [13]; and (l) global contrast map generated by [9]. Evidently, each component of BD-SOD contributes to the final saliency map. Best viewed in the color version.

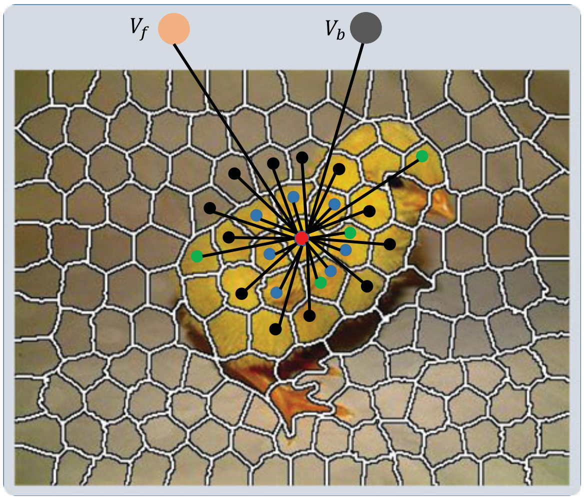

Fig. 6. Illustration of building the enhanced optimization graph in BD-SOD. Here each regular node corresponds to an image superpixel, and two virtual supernodes V_f and V_b represent the foreground and background, respectively. The edges of a node (red point) include the connections to its spatial neighbors (blue points), ones to "neighbors of neighbors" (black points), ones to V_f and V_b, and nonlocal feature connections to similar nodes (green points). Best viewed in the color version.

that represent the background and foreground, and also set up the nonlocal feature connections. Fig. 6 provides an illustration, and we explain the details as follows.

The purpose of introducing supernodes is to strengthen the connection of each region to its belonging type so that more robust saliency maps are generated. Conceptually, the supernodes V_b and V_f represent the collections of background and foreground regions of an image, respectively. Thus V_b and V_f contain richer information than the regular nodes, which implies that the connections to supernodes would be more reliable. Let $\hat{G} = (\hat{V}, \hat{E})$ denote the enhanced optimization graph possessing two supernodes. Then each regular node v_i is required to connect to both V_b and V_f, which actually makes the saliency value of a region be refined towards its belonging type, i.e., 1 for the foreground and 0 for the background. Let w_{if} and w_{ib} denote the connection weights of the node v_i to V_f and V_b, and then we define them as

$$w_{if} = \alpha_f(1 - b_i), w_{ib} = \alpha_b b_i \tag{10}$$

where α_f and α_b represent the impact scaling factors of the foreground and background supernodes, which essentially control the effect of the prior background map.

In addition, we introduce the nonlocal feature connections in optimization graph besides the traditional spatially local connections. These nonlocal connections can help achieve the intra-object consistency of saliency values, especially for complicated objects. To be specific, we adopt two kinds of similarity metrics to set up such connections. The first is from the estimated saliency map $\tilde{S}$. That is, we enforce the regions belonging to the same type with high confidence to connect with each other. Specifically, we produce the foreground and background sets as $N_f = \{v_k | \tilde{s}_k \geq 0.9, k = 1, 2, \cdots, n\}$ and $N_b = \{v_k | \tilde{s}_k \leq 0.1, k = 1, 2, \cdots, n\}$, and then set up the pairwise connections in $N_f(N_b)$. The second is from the features representing regional contents, i.e., requiring similar regions to be interconnected. Specifically, each regular node connects its k-nearest neighbors (e.g., $k = 4$) in the feature space, and the similarity is measured by the Euclidean distance.

Now we elaborate on the procedure of saliency refinement with the enhanced graph $\tilde{G} = (\tilde{V}, \tilde{E})$. We first give the definition of the cost function as

$$g(\mathcal{S}) = \underbrace{\sum_{i=1}^{n} \lambda_i (s_i - \tilde{s}_i)^2}_{\text{Fitting}} + \underbrace{\sum_{(i,j)\in\hat{E}} w_{ij}(s_i - s_j)^2}_{\text{Smoothness}}. \tag{11}$$

Here $\mathcal{S}$ is the refined saliency map of an image, consisting of all regional saliency values $\{s_i\}_{i=1}^{n}$. Particularly, $s_0 = 0$ is fixed for the background supernode, and $s_{n+1} = 1$ for the foreground supernode. The trade-off parameter λ_i is set as 1 for each node in N_f and N_b due to their high confidence, and 0.1 for other nodes. Obviously, it is expected that the produced saliency map does not change too much from its initial saliency assignment

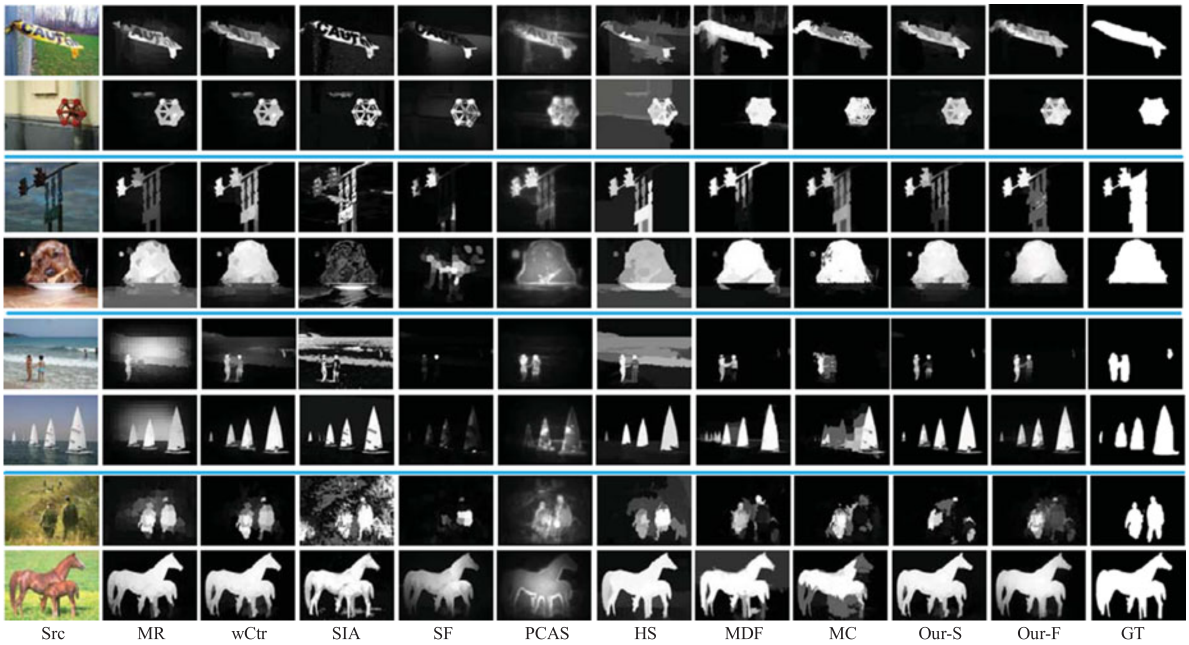

Fig. 7. Detection results on some example images for different methods. The images are from the datasets MSRA-1000 (rows $1 \sim 2$), THUS-10000 (rows $3 \sim 4$), Pascal-S (rows $5 \sim 6$), and ECSSD (rows $7 \sim 8$). Specifically, Src and GT represent the raw images and their ground truth, and the remaining columns represent different salient object detection methods. Here the baselines include MR [13], wCtr [8], SIA [11], SF [10], PCAS [30], HS [23], MDF [19], and MC [32]. Best viewed in the color version.

(through the fitting term), and differ too much between similar nodes (through the smoothness term).

We obtain the saliency map $\mathcal{S}$ by minimizing the cost function in (11). Before solving such an optimization problem, we define the following notations:

$$\begin{cases} \mathbf{\Psi} = \mathrm{diag}\{\psi_1, \psi_2, \cdots, \psi_n\}, & \psi_i = \lambda_i + w_{if} \\ \theta = [\theta_1, \theta_2, \cdots, \theta_n]^\top, & \theta_i = \frac{\lambda_i \tilde{s}_i + w_{if}}{\lambda_i + w_{if}} \\ \mathbf{B} = \mathrm{diag}\{b_1, b_2, \cdots, b_n\}, & b_i = w_{ib} \\ \mathbf{D} = \mathrm{diag}\{d_1, d_2, \cdots, d_n\}, & d_i = \sum_j w_{ij} \end{cases} \tag{12}$$

where $\mathbf{D}$ is the degree matrix of the graph. We decompose the smoothness term in (11) into the following three terms: connections to V_f, ones to V_b, and regular connections E. Then the cost function can be reformulated as

$$g(\mathcal{S}) = (\mathcal{S} - \theta)^\top \mathbf{\Psi} (\mathcal{S} - \theta) + \mathcal{S}^\top (\mathbf{B} + \mathbf{D} - \mathbf{W}) \mathcal{S} + c. \tag{13}$$

Here c is a constant value, and $\mathbf{W} \in \mathbb{R}^{n \times n}$ is the affinity matrix consisting of the edge weights $\{w_{ij}\}$. The objective function in (13) is quadratic on $\mathcal{S}$, so we can derive a closed-form solution as

$$\mathcal{S} = (\mathbf{\Psi} + \mathbf{B} + \mathbf{D} - \mathbf{W})^{-1} \mathbf{\Psi} \theta \tag{14}$$

which is exactly the final saliency map of an image.

VI. Experimental Results

A. Dataset and Setup

1) Datasets: We evaluate the performance of BD-SOD on four widely used benchmark datasets, i.e., MSRA-1000 [2], THUS-10000 [9], PASCAL-S [22], and ECSSD [23]. Among these datasets, 1) MSRA-1000 contains 1000 images with pixel-level saliency annotations; 2) THUS-10000 is one of the largest benchmark datasets for saliency detection, which consists of 10000 images with pixel-level annotations for salient object regions; 3) PASCAL-S is a collection of 850 natural images built on the validation set of the PASVAL VOC 2010 segmentation challenge; and 4) ECSSD consists of 1000 semantically meaningful but structurally complex images acquired from the Internet. Fig. 7 provides some example images from these datasets.

2) Evaluation criteria: We adopt two types of standard metrics to measure the overall performance of different methods by following most previous works [8], [10], [11].

a) Precision and recall: Precision (also called *positive predictive value*) represents the ratio of the correctly assigned salient pixels to all the pixels of extracted regions, while Recall (also known as *sensitivity*) measures the percentage of the detected salient pixels w.r.t. the ground truth. Given a saliency map normalized to $[0, 255]$, a binary foreground mask M can be generated by setting a threshold in $[0, 255]$. Then a threshold would result in a precision value and a recall value by comparing the binary M against the ground truth G, i.e.,

$$\mathrm{Precision} = \frac{|M \cap G|}{|M|}, \mathrm{Recall} = \frac{|M \cap G|}{|G|}. \tag{15}$$

Moreover, we also adopt an adaptive thresholding strategy proposed in [2] to binarize the saliency map. Specifically, the threshold is defined as $T_a = \frac{2}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \mathcal{S}(x, y)$, where W and H are the width and height of an image, and $\mathcal{S}(x, y)$ denotes the saliency value in the position (x, y). Then we can compute the weighted harmonic mean measure [2], [10] (or F-measure) other than the precision and recall. F-measure actually reflects the overall quality of detected objects (higher is better), which is defined as

$$F_\beta = \frac{(1+\beta^2)\cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}. \tag{16}$$

Here β is a control parameter to compromise the importance between the precision and recall. Similar to [2], [10], we set $\beta^2 = 0.3$ to emphasize the precision.

b) Mean absolute error (MAE) The aforementioned metrics do not explicitly consider the true negative saliency assignments, which means the non-salient pixels are marked to be salient. So they are more favorable for the methods that only successfully assign the salient pixels but fail for the non-salient ones [10], [11]. MAE is introduced as a supplement that is defined as

$$\text{MAE} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\mathcal{S}(x,y) - G(x,y)|. \tag{17}$$

Comparatively, F-measure is usually adopted as the major metric since the salient objects need to be explicitly segmented according to the generated saliency maps in salient object detection. For a given dataset, the overall metric values are obtained by averaging on all involved images with the optional standard deviation.

3) *Experimental Settings*

To fully evaluate the effectiveness of BD-SOD, we adopt two types of background identification approaches to produce different background priors. The first is the heuristic method, and the image boundary is used as the background prior, as in wCtr [8]. Such a way is highly efficient but may suffer from poor performance especially for complicated images. The second is the learning-based method, and we concretely explore two CNN-based models (i.e., MDF and CNN-1) that possess different complexity and accuracy. Specifically, MDF [19] employs three deep CNNs to extract multi-scale features from nested windows and then adopts a deep neural network with two fully connected layers to estimate saliency values. For MDF, we first use the provided model [19] to produce a saliency map for each image, and then revert the predicted saliency map into a background map via $b_i = 1 - \hat{s}_i$, where $\hat{s}_i$ is the normalized saliency value of the i-th image region. MDF can result in a relatively accurate background map due to employing three CNNs to extract richer features. For sufficient comparison, we also propose a simple model by employing only one CNN to directly learn the background maps, denoted by CNN-1. CNN-1 has lower complexity than MDF due to adopting single network, and meanwhile produces better background prior than the heuristic boundary. Through these experiments, it is expected to show that BD-SOD can always produce good saliency maps for various background priors.

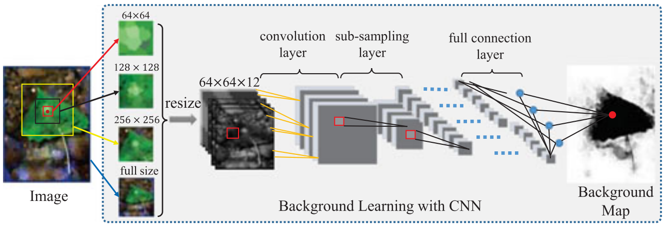

Fig. 8. Illustration of learning the background map with CNN-1. For each superpixel of an image, the four spatial context patches are produced, including three cropped subimages and one raw full-size image. These subimages are down-sampled to the same size (e.g., 64×64) to fit for the CNN model, and consequently a $64 \times 64 \times 12$ patch is formed for each image region. Here an existing network is adopted except necessary modification of the input and output sizes. Particularly, different spatial parts of the input are divided into four scale groups to first extract multiple content features, and then they are fused to represent the overall contrast information. In the training phase, the background superpixels are labeled by 1, otherwise 0. Best viewed in the color version.

Here we briefly explain the key details of CNN-1, which is illustrated in Fig. 8. CNN-1 uses four different sizes of contexts for an image region, including three center-surrounding rectangle subimages and one raw full-size image. Specifically, the rectangle subimages are produced by cropping a patch of a given size around the central point of the target region. In our experiments, the three sizes of subimages are set as 64×64, 128×128, and 256×256. To fit the CNN model, each context patch is down-sampled into a 64×64 standard subimage. Then a $64 \times 64 \times 12$ patch is formed to represent a region, where the third dimension (12) is derived from stacking four different contexts with three channels (R, G, B) per context. For the CNN model, we directly adopt an existing network using *cuda-convnet*, which achieves 11% error on CIFAR-10 in 75 minutes,[2] and image translations and horizontal reflections are applied. In our implementation, we modify the input size of *conv1* from 3 into 12 according to the format of input data and the output size of *fc10* from 10 into 2 to fit for the binary classification. Here the input images of multiple scales are stacked to feed into the first convolution layer with four scale groups. Through such a design, the features for different scales of contexts [32] are first extracted separately, and then fused to represent the overall contrast information of the target region (as in the integration of multiple modal features [41]). That is, the final prediction is yielded by hierarchically representing the contrast features, which is in agreement with the intension of saliency. The training dataset for CNN-1 is built by randomly choosing 5000 images from the dataset in [13], and has no overlap with the evaluated datasets. We follow the standard strategy [42] to train CNN-1, and no data augmentation is adopted. More specifically, the training set is divided into five subsets. The classifier is trained on the first four subsets with the initial learning rate of 0.001, and is validated on the fifth subset. When the validation error reaches a plateau, we further conduct two epochs on the whole training set with the lower learning rates (i.e., 0.0001 and 0.00001).

The control parameters of BD-SOD are preset and fixed throughout the experiments. They are given as follows. First, two thresholds for producing the background and foreground are

[2] [Online]. Available: https://github.com/dnouri/cuda-convnet/tree/master/example-layers for details

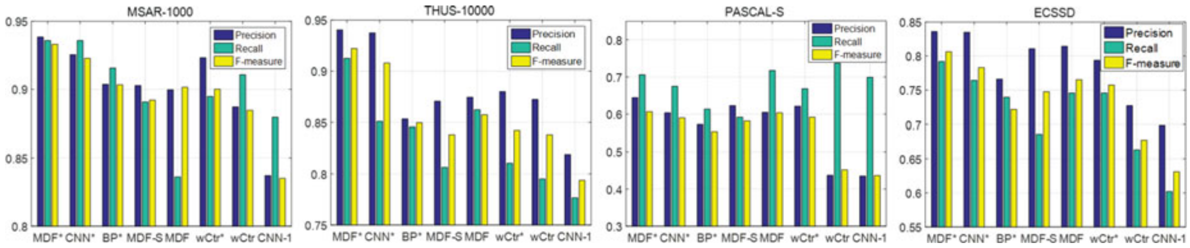

Fig. 9. Resulting precision, recall, and F-measure for different combinations of components. Here *MDF**, *CNN**, and *BP** correspond to the proposed BD-SOD with the MDF-S learned background prior, CNN-1 learned background prior, and boundary prior, respectively. *MDF* [19] and *wCtr* [8] are two baselines. *MDF-S* is the finest-scale version of *MDF*, and *wCtr** is a variant of *wCtr* using the CNN-1 learned background prior. *CNN-1* represents the proposed learning model with single CNN. Best viewed in the color version.

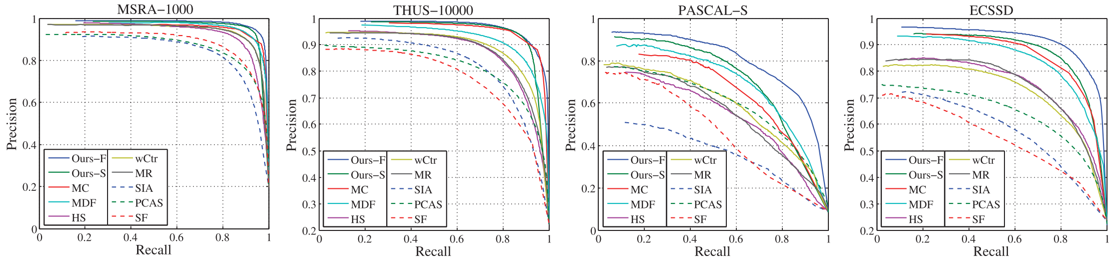

Fig. 10. Precision and recall curves of all used methods on MSRA-1000, THUS-10000, PASCAL-S, and ECSSD. Best viewed in the color version.

set as $T_b = 0.8$ and $T_f = 0.4$, which are empirically determined by tuning on MSAR-1000. Second, the bandwidth parameters in saliency estimation are set as $\sigma_f = 15$ and $\sigma_p = 0.25$ by referring to the similar works [8], [10]. Third, the impact scaling factors in (10) are set as $\alpha_f = 0.5$ and $\alpha_b = 5$ to emphasize that the identified background is more important than the foreground for saliency refinement.

B. Performance Analysis

In this subsection, we evaluate the effectiveness of the proposed BD-SOD through different combinations of main components. To be specific, two representative approaches, i.e., robust background detection (wCtr) [8] and multi-scale deep features (MDF) [19], are adopted as the baselines due to their relative superiority. wCtr belongs to the traditional saliency computation model and utilizes the background prior, where the backgroundness of an image patch is measured by its connectivity to image boundary. MDF belongs to the emerging learning-based approaches, and it does not explicitly utilize the background prior. We adopt the precision, recall, and F-measure with adaptive threshold to present the detection performance, and four benchmark datasets are all examined.

According to the framework of BD-SOD, we mainly consider two components: background identification and saliency computation (including saliency estimation and saliency refinement). For background identification, the aforementioned three background priors are all investigated. For saliency computation, we compare the proposed BD-SOD to both established baselines. To fully show the effects of different components, we specifically consider the following combinations: a) the computation model in [8] with the boundary prior and CNN-1 learned background prior, which are denoted by *wCtr* and *wCtr**; b) the proposed BD-SOD with three background priors, which are denoted by *BP** for the boundary prior, *CNN** for the CNN-1 learned prior, and *MDF** for the MDF learned prior (only the finest scale is adopted); and c) the plain learning-based model, including *CNN-1*, *MDF* [19] and its finest-scale version *MDF-S*.

The results on four datasets are reported in Fig. 9. We give the concrete analysis as follows. Firstly, the overall performance (F-measure) of *wCtr** is superior to *wCtr* on all used datasets as more accurate background prior from CNN-1 is fed. Such results imply that better background prior is more beneficial, even for the stereotyped model. In particular, *wCtr** is better in precision but worse in recall on MSRA-1000 and PASCAL-S, which implies that *wCtr* tends to focus on both the salient regions and background regions. Secondly, comparing *BP** and *wCtr* using the same background prior, it can be seen that the proposed BD-SOD outperforms *wCtr*. Thirdly, comparing *MDF** and *MDF-S*, we can observe that BD-SOD is able to further boost the performance even for a relatively accurate background prior, which implies that the adaptive strategies in BD-SOD are effective. Finally, from the results of *BP**, *CNN**, and *MDF**, we can see that BD-SOD can consistently improve the performance as more accurate background prior is fed.

C. Performance Comparison

In this subsection, we compare the proposed BD-SOD with eight representative saliency detection approaches, including the hierarchical saliency (HS) [23], multiscale CNN features learning (MDF) [19], multiple-context saliency detection (MC) [32], soft image abstraction (SIA) [11], robust background detection (wCtr) [8], manifold ranking (MR) [13], saliency filter (SF) [10], and patch distinctness (PCAS) [30]. These methods cover the mainstream and emerging models of saliency detection. Specifically, we adopt the implementation from [8] for SF, MR, and wCtr, and the original implementation provided by the authors

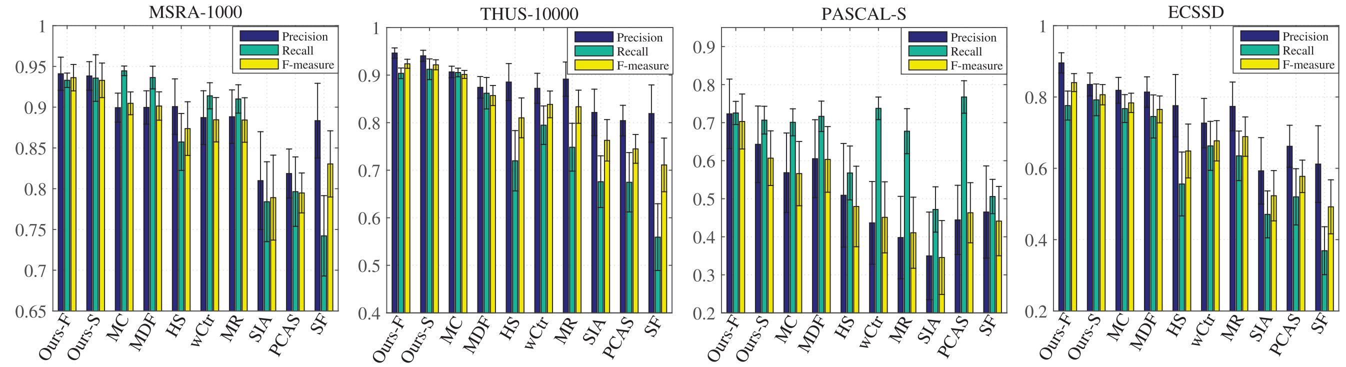

Fig. 11. Precision, recall, and F-measure of all used methods on MSRA-1000, THUS-10000, PASCAL-S, and ECSSD, where an adaptive threshold is adopted. For these metrics, a larger value is better. Best viewed in the color version.

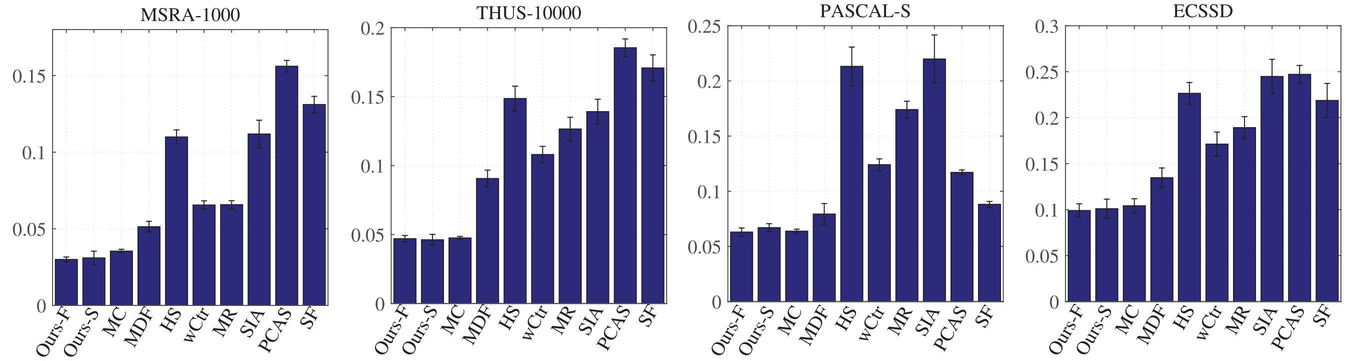

Fig. 12. Mean absolute error (MAE) of all used methods on MSRA-1000, THUS-10000, PASCAL-S, and ECSSD. For this metric, a smaller value is better.

for SIA, HS, PCAS, MC, and MDF. For sufficient comparison, we evaluate the multi-scale version of BD-SOD (*Our-F*) besides the single-scale one (*Our-S*), in which five scales (including 200, 400, 600, 800, and 1000 superpixels per image) are particularly used [23], [43], [44], and the final saliency map of an image is generated by averaging on the results of five scales. Here the MDF background prior is chosen due to its superiority. An intuitive visual comparison of these methods is provided in Fig. 7, which covers some good results for simple images and some hard examples in Pascal-S and ECSSD. It can be observed that our proposed method can more uniformly highlight the entire salient objects.

Fig. 10 reports the resulting precision-recall curves on four benchmark datasets, Fig. 11 reports the corresponding precision, recall, and F-measure, and Fig. 12 shows their averaged MAE scores. From the results, it can be seen that the multi-scale version of BD-SOD performs better than the single-scale one overall, especially for the complicated images (e.g. in PASCAL-S and ECSSD), which is mainly benefited from the increase of the accuracy performance. In addition, our proposed method (*Our-F*) outperforms all baselines on the overall performance (F-measure and MAE), and achieves state-of-the-art on these datasets. In particular, compared with the advanced *MDF* and *MC*, the proposed BD-SOD can always get higher *precision* on all datasets, but may decrease *recall* (e.g., for MSRA-1000). This is because that BD-SOD emphasizes the saliency of detected regions, while *MDF* and *MC* focus on the coverage rate to salient objects. Fortunately, BD-SOD can achieve a good trade-off for complicated images, e.g., both *precision* and *recall* improved on PASCAL-S.

D. *Component-wise Analysis*

In this subsection, we conduct extra experiments to investigate the contributions of different components of BD-SOD. Since different settings of BD-SOD possess similar tendency for the detection performance, the single-scale version (600 superpixels per image) is adopted here for simplicity, and the CNN-1 learned background prior and MSRA-1000 dataset are used. Fig. 5 gives a component-wise visual example, and here we particularly analyze the used image features and connections of optimization graph. For the performance presentation, the precision-recall curve and three metrics (precision, recall, and F-measure) of the adaptive threshold are presented.

1) Image Feature Evaluation: In our method, both the color and texture features are adopted. In order to make clear the contributions of different features to the final performance, we evaluate different settings by restricting the used feature to one of the color and texture features. The results are reported in Fig. 13. As predicted, color works much better than texture because it is usually predominant in grabbing the visual attention, and combining both of them achieves the best performance owing to their mutual supplement.

2) Graph Connection Examination: The graph connections mainly determine the performance of saliency refinement. Here we analyze the effects of three kinds of connections in our

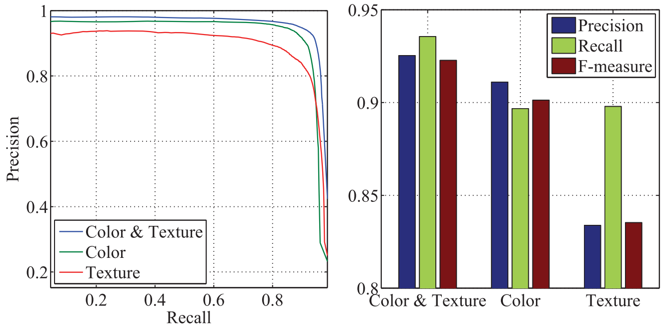

Fig. 13. Precision-recall curves (left), and precision, recall, and F-measure values of the adaptive threshold (right) for the saliency maps with different features on MSRA-1000. Best viewed in the color version.

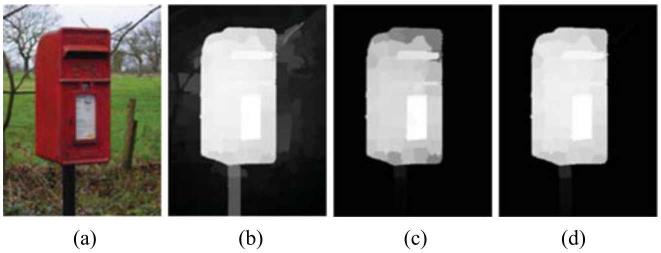

Fig. 14. Saliency maps generated by BD-SOD with different combinations of graph connections: (a) source image, (b) saliency map produced by only local connections, (c) that by both local and prior connections, and (d) that by full connections consisting of local, nonlocal, and prior connections.

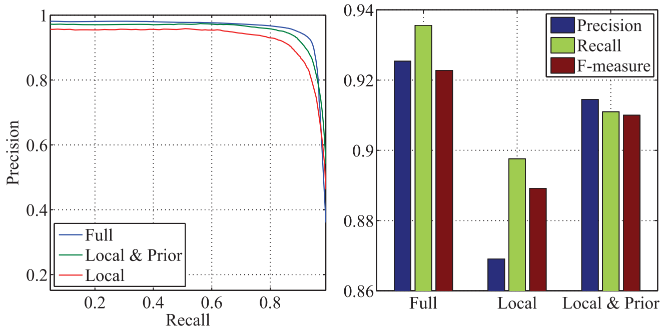

Fig. 15. Precision-recall curves (left), and precision, recall, and F-measure values of the adaptive threshold (right) for the saliency maps with different combinations of graph connections on MSRA-1000.

graph model, i.e., the spatially local connections, prior connections to both supernodes, and non-local feature connections, aiming at illustrating that each of them is beneficial to the final performance. Specifically, we evaluate the detection performance by successively accumulating different connections. Fig. 14 gives an intuitive visual example and Fig. 15 presents the performance results. As exhibited, setting up the prior and non-local connections can gradually improve the quality of saliency maps with more consistent saliency assignments.

VII. Conclusion

In this work, we proposed a background driven salient object detection method (BD-SOD) to produce more accurate and robust saliency maps. The key idea of BD-SOD is to mitigate the negative interference of cluttered background and diverse object parts within an image. To this end, we explicitly and comprehensively exploited the background prior. Specifically, we proposed a novel saliency estimation model to deeply utilize the background clues, i.e., restricting the reference regions to only background regions, weighting the contribution of reference regions, and leveraging the importance of different features. Furthermore, we proposed an enhanced optimization graph for saliency refinement by embedding the background prior. Particularly, two virtual supernodes are introduced with extra connections and the nonlocal feature connections are also set up. Finally, we verified the effectiveness of BD-SOD through the systematical experiments. The results show that the background prior is critically useful for salient object detection, and our method achieves state-of-the-art performance on multiple benchmark datasets. In the future, we plan to develop some more advanced method to exploit the background prior. For example, the detection method can keep superior even if a very rough background prior is provided.

References

[1] T. Liu *et al.*, “Learning to detect a salient object,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[2] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, “Frequency-tuned salient region detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1597–1604.

[3] Y. Gao, M. Shi, D. Tao, and C. Xu, “Database saliency for fast image retrieval,” *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 359–369, Mar. 2015.

[4] J. Huang, X. Yang, X. Fang, W. Lin, and R. Zhang, “Integrating visual saliency and consistency for re-ranking image search results,” *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 653–661, Aug. 2011.

[5] L. Chen *et al.*, “A visual attention model for adapting images on small displays,” Microsoft Res., Redmond, WA, USA, Tech. Rep. MSR-TR-2002-125, 2002.

[6] M. Ding and R. F. Tong, “Content-aware copying and pasting in images,” *Vis. Comput.*, vol. 26, nos. 6–8, pp. 721–729, 2010.

[7] X. Sun, H. Yao, and R. Ji, “What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 1552–1559.

[8] W. Zhu, S. Liang, Y. Wei, and J. Sun, “Saliency optimization from robust background detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2814–2821.

[9] M. M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. M. Hu, “Global contrast based salient region detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[10] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, “Saliency filters: Contrast based filtering for salient region detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 733–740.

[11] M. M. Cheng *et al.*, “Efficient salient region detection with soft image abstraction,” in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 1529–1536.

[12] Y. Wei, F. Wen, W. Zhu, and J. Sun, “Geodesic saliency using background priors,” in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 29–42.

[13] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. H. Yang, “Saliency detection via graph-based manifold ranking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3166–3173.

[14] H. Jiang *et al.*, “Salient object detection: A discriminative regional feature integration approach,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2083–2090.

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

[16] V. Gopalakrishnan, Y. Hu, and D. Rajan, “Salient region detection by modeling distributions of color and orientation,” *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 892–905, Aug. 2009.

[17] N. İmamoğlu, W. Lin, and Y. Fang, “A saliency detection model using low-level features based on wavelet transform,” *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 96–105, Jan. 2013.

[18] W. Einhauser and P. Konig, “Does luminance-contrast contribute to a saliency map for overt visual attention?,” *Eur. J. Neurosci.*, vol. 17, pp. 1089–1097, 2003.

[19] G. Li and Y. Yu, “Visual saliency based on multiscale deep features,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 5455–5463.

[20] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, “Deep networks for saliency detection via local estimation and global search,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 3183–3192.

[21] K. Y. Chang, T. L. Liu, H. T. Chen, and S. H. Lai, “Fusing generic objectness and visual saliency for salient object detection,” in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 914–921.

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

[23] J. Shi, Q. Yan, L. Xu, and J. Jia, “Hierarchical image saliency detection on extended CSSD,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 717–729, Apr. 2015.

[24] C. Koch and S. Ullman, “Shifts in selective visual attention: Towards the underlying neural circuitry,” *Human Neurbiol.*, vol. 4, pp. 219–227, 1985.

[25] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu, “Visual saliency detection by spatially weighted dissimilarity,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 473–480.

[26] Y. F. Ma and H. J. Zhang, “Contrast-based image attention analysis by using fuzzy growing,” in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 374–381.

[27] Z. Liu *et al.*, “Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut,” *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1275–1289, Aug. 2012.

[28] Y. Fang *et al.*, “Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum,” *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 187–198, Feb. 2012.

[29] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-aware saliency detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[30] R. Margolin, A. Tal, and L. Zelnik-Manor, “What makes a patch distinct?,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 1139–1146.

[31] S. Li, H. Lu, Z. Lin, X. Shen, and B. Price, “Adaptive metric learning for saliency detection,” *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3321–3331, Nov. 2015.

[32] R. Zhao, W. Ouyang, H. Li, and X. Wang, “Saliency detection by multicontext deep learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 1265–1274.

[33] S. He, R. W. Lau, W. Liu, Z. Huang, and Q. Yang, “Supercnn: A superpixelwise convolutional neural network for salient object detection,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 1–15, 2015.

[34] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang, “Multimodal deep autoencoder for human pose recovery,” *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5659–5670, Dec. 2015.

[35] J. Han *et al.*, “Background prior based salient object detection via deep reconstruction residual,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1309–1321, Aug. 2014.

[36] R. Achanta, *et al.*, “Slic superpixels compared to state-of-the-art superpixel methods,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[37] E. Shahrian and D. Rajan, “Weighted color and texture sample selection for image matting,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 718–725.

[38] Z. Wang, J. Feng, S. Yan, and H. Xi, “Linear distance coding for image classification,” *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 537–548, Feb. 2013.

[39] A. Borji, D. N. Sihite, and L. Itti, “Salient object detection: A benchmark,” in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 414–429.

[40] L. Zhou, Z. Yang, Q. Yuan, Z. Zhou, and D. Hu, “Salient region detection via integrating diffusion-based compactness and local contrast,” *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3308–3320, Nov. 2015.

[41] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, “Multispectral deep neural networks for pedestrian detection,” in *Proc. Brit. Mach. Vis. Conf.*, 2016.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[43] Z. Liu, W. Zou, and O. Le Meur, “Saliency tree: A novel saliency detection framework,” *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937–1952, May 2014.

[44] L. Zhu, D. A. Klein, S. Frintrop, Z. Cao, and A. B. Cremers, “A multisize superpixel approach for salient object detection based on multivariate normal distribution estimation,” *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5094–5107, Dec. 2014.

Zilei Wang (M’13) received the B.S. and Ph.D. degrees in control science and engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2002 and 2007, respectively.

He is currently an Associate Professor with the Department of Automation, USTC, and the Founding Lead of the Vision and Multimedia Research Group (http://vim.ustc.edu.cn). His research interests include computer vision, multimedia, and deep learning.

Prof. Wang is a Member of the Youth Innovation Promotion Association, Chinese Academy of Sciences.

Dao Xiang received the B.S. and Ph.D. degrees in control science and engineering from the University of Science and Technology of China, Hefei, China, in 2011 and 2016, respectively.

His research interests are image classification, salient object detection, and object segmentation.

Saihui Hou received the B.S. degree in control science and engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2014, and is currently working toward the Ph.D. degree in control science and engineering at the USTC.

His current research interests include object recognition and deep learning.

Feng Wu (M’00–SM’06–F’13) received the B.S. degree in electrical engineering from Xidian University, Xi’an, China, in 1992, and the M.S. and Ph.D. degrees in computer science from Harbin Institute of Technology, Harbin, China, in 1996 and 1999, respectively.

He is currently a Professor with the University of Science and Technology of China (USTC), Hefei, China, and the Dean of the School of Information Science and Technology (SIST), USTC. Before that, he was a Principle Researcher and a Research Manager with Microsoft Research Asia, Beijing, China. He has authored or coauthored more than 200 papers (including several dozens IEEE transaction papers) and top conference papers at MOBICOM, SIGIR, CVPR, and ACM MM. He has 77 granted U.S. patents. His 15 techniques have been adopted into international video coding standards. His research interests include image and video compression, media communication, and media analysis and synthesis.

Prof. Wu serves as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEM FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, and several other international journals. He also served as a TPC Chair for MMSP 2011, VCIP 2010, and PCM 2009, and Special Sessions Chair for ICME 2010 and ISCAS 2013. He was the recipient of the Best Paper Award as a coauthor for IEEE T-CSVT 2009, PCM 2008, and SPIE VCIP 2007. He was also the recipient of the IEEE Circuits and Systems Society 2012 Best Associate Editor Award.