Depth-Aware Salient Object Detection and Segmentation via Multiscale Discriminative Saliency Fusion and Bootstrap Learning

Hangke Song, Zhi Liu, *Senior Member, IEEE*, Huan Du, Guangling Sun, Olivier Le Meur, and Tongwei Ren, *Member, IEEE*

***Abstract*— This paper proposes a novel depth-aware salient object detection and segmentation framework via multiscale discriminative saliency fusion (MDSF) and bootstrap learning for RGBD images (RGB color images with corresponding Depth maps) and stereoscopic images. By exploiting low-level feature contrasts, mid-level feature weighted factors and high-level location priors, various saliency measures on four classes of features are calculated based on multiscale region segmentation. A random forest regressor is learned to perform the discriminative saliency fusion (DSF) and generate the DSF saliency map at each scale, and DSF saliency maps across multiple scales are combined to produce the MDSF saliency map. Furthermore, we propose an effective bootstrap learning-based salient object segmentation method, which is bootstrapped with samples based on the MDSF saliency map and learns multiple kernel support vector machines. Experimental results on two large datasets show how various categories of features contribute to the saliency detection performance and demonstrate that the proposed framework achieves the better performance on both saliency detection and salient object segmentation.**

***Index Terms*— Depth information, discriminative saliency fusion, random forest, saliency detection, salient object segmentation.**

I. Introduction

TO COPE with the complex natural scene which contains various objects at different scales, human visual system (HVS) has developed the visual attention mechanism to select the most salient parts that stand out from the other parts in the scene [1]. The research on saliency models was stimulated by modeling human visual attention, and since the seminal work by Itti *et al.* [2], a number of computational saliency models have been proposed for either salient object detection [3]–[12] or human fixation prediction [13]–[16]. In this paper, we mainly focus on the former class of saliency models, i.e., for salient object detection. The output of a saliency model is a gray-scale saliency map, which is also related with the visual importance map as discussed in [17]. Saliency maps, which can highlight salient objects and suppress background regions, have been successfully applied in a variety of applications including salient object segmentation [18]–[20], salient object detection [21], [22], content aware image/video retargeting [23], [24], content-based image/video compression [25], [26], object recognition [27] and visual scanpath prediction [28].

Saliency models for salient object detection are usually based on an unsupervised computational framework, which is generally built on the following structure: (a) extract some basic visual features such as color, texture and orientation from the input image; (b) investigate feature channels in parallel and compute a saliency map for each feature; and (c) integrate these different feature based saliency maps to generate the final saliency map. Here, the last step, which is typically carried out by taking the weighted average, i.e., linear summation, is important for generating the final saliency map with the better quality [29]. In this regard, a few researches try to address the integration problem by finding the optimal values for the weights in the linear summation. In [3], the linear fusion weights for different features are learned under the conditional random field (CRF) framework, and in [12], the large-margin framework is adopted to learn the weights. Due to the highly nonlinear essence of visual attention mechanism, the above linear mapping might not perfectly capture the characteristics of feature integration [30] of HVS. Therefore, some nonlinear methods have been used for saliency fusion. In [31], a series of nonlinear classifiers are learned via AdaBoost to combine different feature based saliency maps. In [32], the region covariance is employed to encode local structure information and nonlinearly integrate different feature maps. Similar ideas of nonlinear saliency fusion are also used for aggregation of saliency maps generated by different existing saliency models in [29], [33], and [34]. Recently, owing to the emergence of several large datasets, supervised saliency models, which are learned from the training samples to generate saliency maps,

Manuscript received June 13, 2016; revised February 13, 2017 and March 29, 2017; accepted May 24, 2017. Date of publication June 2, 2017; date of current version June 23, 2017. This work was supported by the National Natural Science Foundation of China under Grants 61471230 and 61171144, and in part by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Janusz Konrad. *(Corresponding author: Zhi Liu.)*

H. Song, Z. Liu, and G. Sun are with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: hksong0209@163.com; liuzhisjtu@163.com; sunguangling@shu.edu.cn).

H. Du is with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China, and also with the Third Research Institute of Ministry of Public Security, Shanghai 201204, China (e-mail: huan_du@163.com).

O. Le Meur is with IRISA, University of Rennes 1, Rennes 35042, France (e-mail: olemeur@irisa.fr).

T. Ren is with the Software Institute, Nanjing University, Nanjing 210008, China (e-mail: rentw@nju.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/TIP.2017.2711277

show the power to detect salient objects. For example, in [35], rich feature representations are learned using convolutional neural networks for object detection and segmentation.

The traditional saliency models only exploit features from color images, but neglect the additional depth information, considering that human observers perceive the real world in the 3D space. Depth information can be directly captured or estimated for a single image [36], and can be also represented by using the disparity map estimated from stereoscopic images [37]. However, the potential of depth information for saliency detection has not been adequately exploited. Several studies were reported on how human attention may be affected by depth cues. In [38], the effect of depth on saliency analysis is investigated and the depth cue is found to be powerful in predicting human fixations. The correlation and influence of depth cues in modeling saliency were studied and the experiments demonstrated that depth continues to be a significant contributor to human fixation prediction [39], [40] and to salient object detection [41].

The features extracted from the depth map can be directly used for measuring saliency, and an integration of depth-induced saliency with the saliency estimated from color image is a common paradigm of depth-aware saliency models for salient object detection. For example, the disparity map estimated from binocular images in [42] and the anisotropic center-surround difference on the depth map in [43] are utilized to identify salient objects. In [41], the depth saliency estimated from point cloud data is integrated with the saliency of color image using nonlinear regression. In [44], both depth weighted color contrast and depth contrast are exploited to measure saliency. In [45], the primitive depth and color contrasts are refined by depth based object probability and region merging for saliency measurement.

In summary, most of the existing depth-aware saliency models for salient object detection can be classified into two categories, i.e., the depth-based saliency model and the depth-weighting saliency model. The depth feature alone may be not a reliable cue when the depth map is not accurate or object regions show weak depth contrasts with background regions. On the other hand, although the depth-weighting saliency model can easily adopt the existing 2D saliency models as summarized in [46], it may miss certain salient regions signified by the depth features. In order to balance the biases of two classes of saliency models, in this paper, different depth-based saliency maps are defined at low level and different depth-weighting saliency maps are defined at middle level. Note that with the additional depth information, many basic visual features such as texture, orientation and spatial distribution can be also extracted from the depth map for generating more feature maps. Therefore, saliency fusion becomes a more critical issue for saliency detection in RGBD/stereoscopic images than in 2D images.

Given the saliency map generated by any saliency model, rich beneficial cues can be leveraged for salient object segmentation, which aims to segment out salient objects from background. A number of salient object segmentation methods were proposed in the recent years. For example, in [19], the two-phase graph cut is used to segment salient objects based on the saliency map generated by kernel density estimation model. In [20], an improved iterative version of GrabCut [47] with the initialization from saliency map, namely SaliencyCut, is proposed to segment salient objects. In [48], segmentation seeds are derived from the saliency map and the Markov random field is applied to salient object segmentation by integrating image features including color, luminance and edge orientation. In [49], a SVM is used to select the salient regions, which are then clustered into the salient objects using the region merging method. Recently, a few tentative graph cut based works try to segment salient objects in RGBD/stereoscopic images with the use of depth information. In [43], a trimap obtained by thresholding the depth saliency map is fed to the GrabCut framework for salient object segmentation. In [50], saliency map, depth map and saliency weighted histogram are incorporated into the graph cut framework [51] to segment salient objects in one cut.

In this paper, based on the multiscale region segmentation result of input image, we measure saliency on different features at low-level, mid-level and high-level, by taking into account of the primary depth and appearance contrasts, different feature weighted factors and location priors, respectively. The key of our saliency model lies in the saliency fusion stage, in which a random forest regressor is trained to perform the nonlinear saliency fusion by automatically discovering the discriminative features. Then with the samples generated using the saliency map, we formulate salient object segmentation as a pixel-level binary labeling problem, which is solved under the framework of SVM. The main contributions of this paper can be briefly summarized as follows:

1) We propose a multiscale discriminative saliency fusion (MDSF) model for saliency detection in RGBD images and stereoscopic images. Different from the previous works on saliency fusion [3], [12], [31], [32], the proposed MDSF model exploits the discriminative saliency fusion as the core step to effectively integrate hundreds of regional saliency measures in a nonlinear manner.
2) We propose a salient object segmentation method based on bootstrap learning. Unlike most of the existing salient object segmentation methods for RGBD/stereoscopic images [43], [50], which mainly depend on the graph cut framework, our method is bootstrapped with samples based on saliency map and learns multiple kernel SVMs for salient object segmentation.
3) We evaluate the effectiveness of various categories of features at low-level, mid-level and high-level for saliency detection in RGBD/stereoscopic images via random forest. We can conclude that the depth based features are generally more discriminative than the color based features for saliency detection in RGBD/stereoscopic images.

Note that the proposed MDSF model (the first contribution) is an extension of our previous work in [52]. The main extensions are that multiscale region segmentation is adopted to improve the saliency detection performance, and in order to avoid bias on one depth source, the training set used in our MDSF model is extended to include both RGBD images and

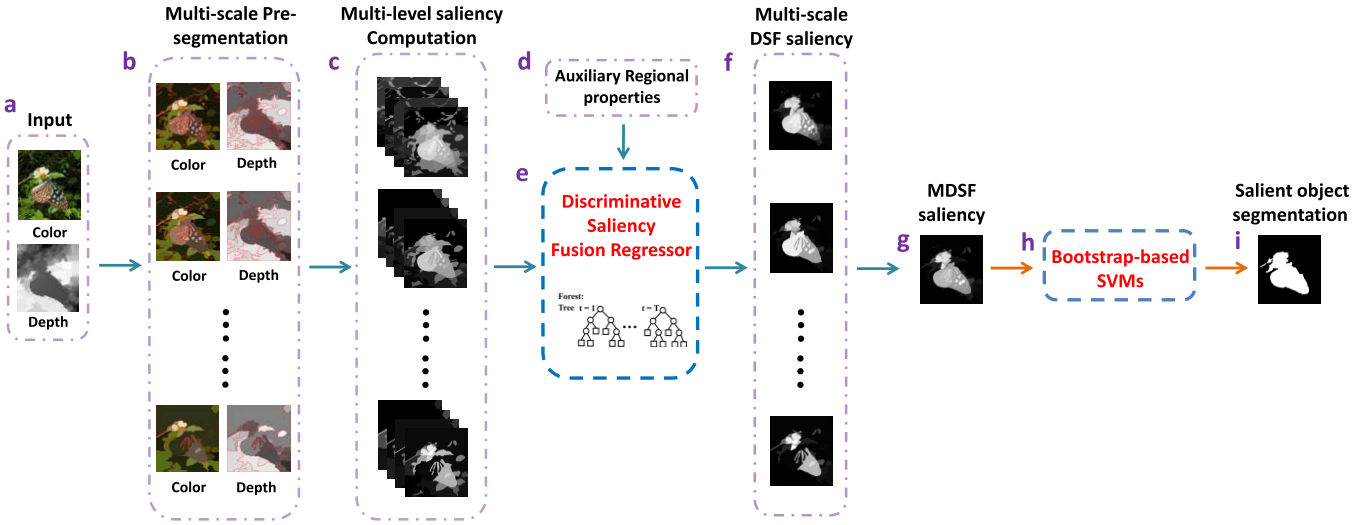

Fig. 1. Illustration of the proposed MDSF model for saliency detection and salient object segmentation. (a) Input color image and depth map; (b) multiscale region segmentation; (c) low-level, mid-level and high-level saliency computation at each scale; (d) auxiliary regional properties; (e) discriminative saliency fusion via DSF regressor at each scale; (f) DSF saliency maps at different scales; (g) MDSF saliency map by combining all the DSF saliency maps at different scales; (h) saliency-based SVMs via bootstrap for segmentation; (i) salient object segmentation mask.

stereoscopic images from two datasets. The proposed MDSF model achieves an obvious improvement on saliency detection performance compared to [52]. Besides, more meaningful experiments are conducted to analyze some key parameters of the proposed MDSF model and the effect of different features on saliency detection performance.

The rest of this paper is organized as follows. The details of the proposed MDSF model and salient object segmentation method are described in Sections II and III, respectively. Experimental evaluation and discussion are presented in Section IV, and conclusions are given in Section V.

II. Proposed Saliency Model

In this section, we propose a multiscale discriminative saliency fusion (MDSF) model for saliency detection in RGBD/stereoscopic images. First, the multiscale region segmentation is performed on the input color image and depth map. For each scale, saliency maps on various features at low-level, mid-level and high-level, respectively, are calculated. Then a random forest regressor is learned to perform the discriminative saliency fusion and generate the DSF saliency map at each scale. Finally, multiscale combination of different DSF saliency maps is exploited to generate the final MSDF saliency map. The whole process of MDSF model is illustrated in Fig. 1(a)–(g).

Note that the three levels of features are mainly defined based on the feature-integration theory [30]. The low-level features consist of some separable features such as color, texture and depth to simulate the early and parallel process of feature registration of visual attention mechanism. The mid-level features are defined based on the combination of some low-level features to simply simulate the feature integration process. Besides, similarly as the definition in [46], the high-level features involve some prior knowledge, i.e., center bias and border bias, which are irrelevant to the low-level separable features.

A. Multiscale Region Segmentation

In order to effectively measure saliency with well-defined boundaries, we first decompose the input color image as shown in Fig. 1(a) into homogeneous regions. Specifically, we use the *gPb-owt-ucm* [53] method, which exploits the globalized probability of boundary based contour detector and the oriented watershed transform, to generate the real-valued ultrametric contour map (UCM) for the color image. However, without knowing the size of salient object, it is nontrivial to determine a suitable scale for saliency detection. Therefore, we resort to the multiscale segmentation for robust saliency detection by tuning the maximum number of segmented regions, τ_m $(m = 1, 2, \cdots M)$ to obtain M segmentation results, S_m $(m = 1, 2, \cdots M)$. For clarity, we elaborate on the saliency detection process at a certain scale without the superscript m for the scale.

By thresholding the UCM based on the maximum number of segmented regions, a set of closed boundaries are retained to form a boundary map, which can be converted into a set of regions, R_i $(i = 1, \cdots, n)$, after removing too small regions. Then the mean color and mean depth of all pixels in each region are used to represent the region-level color and depth as shown in Fig. 1(b), in which the red lines represent the region boundaries.

B. Low-Level Saliency

From the observations on a variety of RGBD/stereoscopic images, salient object regions usually locate at a different depth level and show noticeable appearance contrasts with background regions. Thus the commonly used low-level center-surround contrast can still work as a fundamental principle of saliency detection. To obtain a group of saliency maps, on the basis of each region, we take into account multiple low-level regional features as follows: the average color of each channel; the color histograms in the RGB, HSV and Lab color spaces; the average depth value and depth histogram; texture

TABLE I
MULTI-LEVEL REGIONAL SALIENCY

Regional features		Dim	Feature difference (52)	Dim	Low-level saliency (104)	Mid-level saliency (166)	High-level saliency (52)	Total Dim (322)
Color	Average RGB value	3	$D^1_{i,j} \sim D^3_{i,j}$	3	$LS^1_{G/B} \sim LS^3_{G/B}$	$MS^1_{G/B,DP,1} \sim MS^3_{G/B,DP,3}$, $MS^1_{G/B,DG,4} \sim MS^3_{G/B,DG,6}$	$HS^1 \sim HS^3$	84
	RGB histogram	256	$D^4_{i,j}$	1	$LS^4_{G/B}$	$MS^4_{G/B,DP,7}$, $MS^4_{G/B,DG,8}$	HS^4	
	Average HSV value	3	$D^5_{i,j} \sim D^7_{i,j}$	3	$LS^5_{G/B} \sim LS^7_{G/B}$	$MS^5_{G/B,DP,9} \sim MS^7_{G/B,DP,11}$, $MS^5_{G/B,DG,12} \sim MS^7_{G/B,DG,14}$	$HS^5 \sim HS^7$	
	HSV histogram	256	$D^8_{i,j}$	1	$LS^8_{G/B}$	$MS^8_{G/B,DP,15}$, $MS^8_{G/B,DG,16}$	HS^8	
	Average Lab value	3	$D^9_{i,j} \sim D^{11}_{i,j}$	3	$LS^9_{G/B} \sim LS^{11}_{G/B}$	$MS^9_{G/B,DP,17} \sim MS^{11}_{G/B,DP,19}$, $MS^9_{G/B,DG,20} \sim MS^{11}_{G/B,DG,22}$	$HS^9 \sim HS^{11}$	
	Lab histogram	256	$D^{12}_{i,j}$	1	$LS^{12}_{G/B}$	$MS^{12}_{G/B,DP,23}$, $MS^{12}_{G/B,DG,24}$	HS^{12}	
Depth	Average depth value	1	$D^{13}_{i,j}$	1	$LS^{13}_{G/B}$	$MS^{13}_{G/B,CG,25}$	HS^{13}	10
	Depth histogram	256	$D^{14}_{i,j}$	1	$LS^{14}_{G/B}$	$MS^{14}_{G/B,CG,26}$	HS^{14}	
Texture	Absolute LM filter response	15×**2**	$D^{15}_{i,j} \sim D^{44}_{i,j}$	15×**2**	$LS^{15}_{G/B} \sim LS^{44}_{G/B}$	$MS^{15}_{G/B,DP,27} \sim MS^{29}_{G/B,DP,41}$, $MS^{15}_{G/B,DG,42} \sim MS^{29}_{G/B,DG,56}$ *and* $MS^{30}_{G/B,CG,57} \sim MS^{44}_{G/B,CG,71}$	$HS^{15} \sim HS^{44}$	216
	Max LM response histogram	15×**2**	$D^{45}_{i,j} \sim D^{46}_{i,j}$	1×**2**	$LS^{45}_{G/B} \sim LS^{46}_{G/B}$	$MS^{45}_{G/B,DP,72}$, $MS^{45}_{G/B,DG,73}$ *and* $MS^{46}_{G/B,CG,74}$	$HS^{45} \sim HS^{46}$	
	LBP histogram	59×**2**	$D^{47}_{i,j} \sim D^{48}_{i,j}$	1×**2**	$LS^{47}_{G/B} \sim LS^{48}_{G/B}$	$MS^{47}_{G/B,DP,75}$, $MS^{47}_{G/B,DG,76}$ *and* $MS^{48}_{G/B,CG,77}$	$HS^{47} \sim HS^{48}$	
	HOG feature	32×**2**	$D^{49}_{i,j} \sim D^{50}_{i,j}$	1×**2**	$LS^{49}_{G/B} \sim LS^{50}_{G/B}$	$MS^{49}_{G/B,DP,78}$, $MS^{49}_{G/B,DG,80}$ *and* $MS^{50}_{G/B,CG,81}$	$HS^{49} \sim HS^{50}$	
GD	Average GD value	1×**2**	$D^{51}_{i,j} \sim D^{52}_{i,j}$	1×**2**	$LS^{51}_{G/B} \sim LS^{52}_{G/B}$	$MS^{51}_{G/B,DP,81}$, $MS^{51}_{G/B,DG,82}$ *and* $MS^{52}_{G/B,CG,83}$	$HS^{51} \sim HS^{52}$	12

*Note that each $LS^k_{G/B}$ includes LS^k_G and LS^k_B. $MS^k_{G/B}$ includes MS^k_G and MS^k_B. The upperscript of each notation in low-level, mid-level and high-level saliency corresponds to the upperscript of feature difference in the same row. The "Total Dim" means the total number of saliency measures on four classes of features.

features including the absolute responses and the histograms of 15 LM filters [54], HOG [55] histograms and LBP [56] histograms; geodesic distance (GD) [6]. The details of these features are given in Table I. Note that the texture and geodesic distance features are extracted on both color image and depth map, and thus the dimensions of corresponding features are with a multiplication factor of two, "**×2**" in Table I.

With the above features, the low-level saliency (LS) of each region R_i based on the k^{th} feature contrast is evaluated with respect to all the other regions in the whole image and the image border, respectively, as follows:

$$LS^k_{G/B}(R_i) = \sum_{R_j \in G/B} w_{i,j} \cdot D^k_{i,j}, \tag{1}$$

where G denotes the set of all the other regions in the whole image, and B denotes the set of regions with the distance to the nearest image border less than 20 pixels. $D^k_{i,j}$ is the chi-square distance for the features with the form of histogram or the Euclidean distance for the features with other forms, on the k^{th} feature between R_i and R_j, i.e., the difference between f^k_i and f^k_j. The weight $w_{i,j}$ takes into account the factors of region size and spatial distance and is defined as follows:

$$w_{i,j} = \left|R_j\right| \cdot \exp\left(\frac{-\left\|c_i - c_j\right\|}{\alpha \cdot DL}\right), \tag{2}$$

where $\left|R_j\right|$ denotes the number of pixels in R_j, DL denotes the diagonal length of image, c_i denotes the spatial center position of R_i, and the coefficient α is set to a moderate value, 0.3, to control the influence of spatial distance between regions. Eqs. (1)-(2) indicate that a region R_j with a large size and closer to R_i contributes more to the low-level feature contrast of R_i. As shown in Table I, the total dimension of feature difference is 52. Since the low-level saliency is evaluated with respect to the whole image and the image borders, respectively, we obtain for each region R_i a 104D (52×2) low-level saliency vector, $\mathbf{v}^{LS}_i = \left[LS^1_G(R_i), ..., LS^{52}_G(R_i), LS^1_B(R_i), ..., LS^{52}_B(R_i)\right]$.

C. Mid-Level Saliency

The influence of depth cues in modeling saliency was studied in [39]–[41]. Based on fixations on 2D and 3D images, they concluded that human visual attention is correlated with the depth level of object, i.e., the fixated salient object regions are usually located at the closer depth ranges. However, based on the above conclusion alone, some background regions close to the camera are often mistaken for salient regions. On the other hand, although the center-surround principle on depth is also widely used for saliency measurement, the depth feature alone is not a reliable cue when the object regions show the weak depth contrast with background regions.

Therefore in this subsection, the depth value is considered as a weighting term to integrate with the low-level color-induced saliency. The depth weighted mid-level saliency (with the subscript *DP*) for each region R_i is defined as follows:

$$MS^k_{G/B,DP}(R_i) = \exp(-d_i) \cdot LS^k_{G/B}, \quad \forall f^k \in \Omega_C, \quad (3)$$

where d_i is the mean depth value of R_i, and Ω_C is the set of features (31 dimensions) involving color information mentioned in Section II-B. The exponential term $\exp(\cdot)$ indicates that the regions close to the camera tend to receive more visual attention.

Besides, the geodesic distance [6] is a simple yet effective feature indicating salient regions directly. Especially for depth geodesic distance, it is often more effective for measuring saliency. Therefore the depth/color geodesic distance weighted mid-level saliency is defined as follows:

$$MS^k_{G/B,DG}(R_i) = Geo_d(R_i) \cdot LS^k_{G/B}, \quad \forall f^k \in \Omega_C, \quad (4)$$

$$MS^k_{G/B,CG}(R_i) = Geo_c(R_i) \cdot LS^k_{G/B}, \quad \forall f^k \in \Omega_D, \quad (5)$$

where $Geo_d(R_i)$ and $Geo_c(R_i)$ is the depth and color geodesic distance, respectively. Ω_D is the set of features (21 dimensions) involving depth information mentioned in Section II-B.

Such a cross weighting of depth and color features in Eqs. (3)–(5) enables these mid-level saliency measures to comprehensively utilize both depth and color information. As shown in Table I, in order to count the number of mid-level saliency measures for each feature clearly, $MS^k_{G/B,DP}$, $MS^k_{G/B,DG}$ and $MS^k_{G/B,CG}$ are further added with a subscript of number from 1 to 83 ($31 + 31 + 21 = 83$). The three types of weighted saliency measures for each region R_i constitute a 166D (83×2) mid-level saliency vector, $\mathbf{v}_i^{MS} = \Big[MS^1_{G,DP,1}(R_i), ..., MS^{52}_{G,CG,83}(R_i), MS^1_{B,DP,1}(R_i), ..., MS^{52}_{B,CG,83}(R_i)\Big]$.

D. High-Level Saliency

In addition to the feature contrast, some high-level location priors are also important in identifying salient regions especially for some complicated images with the cluttered background and low contrast between salient object and background regions.

In this subsection, the high-level saliency (HS) arises from two basic observations. In most images, background regions generally have a higher ratio of connectivity with image borders than salient objects, and salient objects are more likely close to the image center. Based on the above two location biases, the location-based object prior (OP) for each region R_i is defined as follows:

$$OP(R_i) = \left[1 - \left(\frac{NB_i}{NB_{\max}}\right)^{\beta}\right] \cdot \exp\left(\frac{-SDC_i}{DL/2}\right), \quad (6)$$

where SDC_i denotes the Euclidean spatial distance from the center position of R_i to the image center position. NB_i is the number of image border pixels contained in R_i, and among all regions touching image borders, $NB_{\max}$ is the maximum number of image border pixels contained in a region.

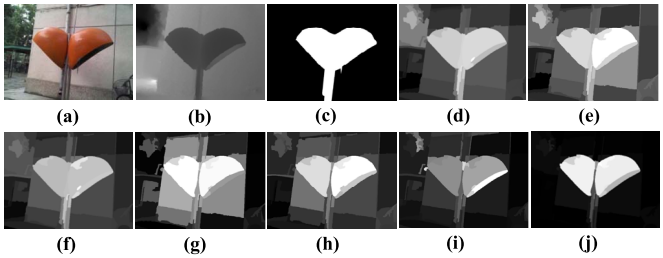

Fig. 2. Example of discriminative saliency fusion. (a) Color image; (b) depth map; (c) ground truth; from (d) to (h): the five most important regional saliency maps; (i) result of linear fusion; (j) result of the proposed discriminative saliency fusion.

The coefficient β is set to 0.25 for a moderate attenuation effect on location priors of those regions touching image borders.

To obtain the HS measures for each region R_i based on the k^{th} feature contrast, we exploit the k^{th} feature difference between R_i and all the other regions to assign similar location based saliency measures to regions with similar values on the k^{th} feature as follows:

$$HS^k(R_i) = OP(R_i) \cdot \frac{\sum_{j=1, j\neq i}^{n} OP(R_j) \cdot \left(1 - ND^k_{i,j}/ND^k_{\max}\right)}{\sum_{j=1, j\neq i}^{n} \left(1 - ND^k_{i,j}/ND^k_{\max}\right)}, \quad (7)$$

where $ND^k_{\max}$ is the maximum of $D^k_{i,j}$ between all the region pairs. In summary, we obtain for each region R_i a 52D high-level saliency vector $\mathbf{v}_i^{HS} = \big[HS^1(R_i), ..., HS^{52}(R_i)\big]$. The details of high-level saliency measures for different features are given in Table I.

For a certain scale, we obtain for each region a 322D saliency vector (104, 166 and 52 for low-level, mid-level and high-level saliency measures, respectively), and thus generate for the input image a total of 322 regional saliency maps as shown in Fig. 1(c).

E. Multiscale Discriminative Saliency Fusion

As shown in previous subsections, with the additional depth information, visual features can be extracted from depth map and color image in parallel for generating various saliency measures, which make saliency fusion become a more critical issue for saliency detection in RGBD/stereoscopic images than in conventional 2D images. Since a total of 322 regional saliency measures are calculated for each region, the traditional linear fusion is not able to effectively combine so many saliency measures on different features. For example, several most important regional saliency maps, which are selected based on the experiments of saliency contribution analysis in Section IV-C, are shown in Fig. 2(d)–(h). The linear fusion result shown in Fig. 2(i) fails to consistently highlight salient regions. Therefore, in this subsection, we aim to integrate various saliency measures in a nonlinear manner by automatically discovering the most discriminative ones using the random forest regressor. Besides, we consider some

TABLE II
VARIANCES FOR AUXILIARY REGIONAL PROPERTY

Variances of regional features		Total dim (108)
Color	RGB values	3
	HSV values	3
	Lab values	3
Depth values		1
Texture	LM filter response	15×**2**
	LBP histogram	1×**2**
	HOG feature	32×**2**
GD values		1×**2**

regional properties to better combine saliency measures. For each region R_i, we obtain a 118D auxiliary property vector $\mathbf{v}_i^{RP}$ consisting of a feature variances vector $\mathbf{v}_i^{fv}$ (108 dimensions), which is detailed in Table II, and the regional geometric feature vector (10 dimensions) $\mathbf{v}_i^{gf}$ in [7]. Then we concatenate regional saliency measures at three levels and auxiliary regional properties together to obtain the master regional saliency measure RS_i for each region R_i as follows:

$$RS_i = F\left(\mathbf{v}_i^{LS}, \mathbf{v}_i^{MS}, \mathbf{v}_i^{HS}, \mathbf{v}_i^{RP}\right) \tag{8}$$

where F is the DSF regressor based on random forest to perform the discriminative saliency fusion. F is an ensemble of T decision trees. For each decision tree, the training samples are randomly selected from the whole training set, and the feature set for node spiltting is randomly drawn from all the 322 regional saliency measures and auxiliary regional properties. The final prediction of the forest is the average of predictions over all the decision trees.

In order to obtain a strong generalization capability of the DSF regressor, the training set is critical for the learning process. In this paper, RGBD images captured by Kinect in RGBD-1000 dataset [46] and stereoscopic images with disparity maps in NJUDS-2000 dataset [43] are both selected into the training set to avoid overfitting on either type of depth sources. And the multiscale segmentation is performed on each training image to generate and expand training samples. As described in Section II-A, we set a group of parameters to control the maximum number of regions at each scale, and obtain a set of segmentation results $\{S_t\}_{t=1}^{M_t}$ with different number of regions for training. For a certain segmentation result S_t, we select those confident regions $\left\{R_{t,1}, R_{t,2}, \cdots, R_{t,i}, \cdots, R_{i,Q_t}\right\}$, in which the number of object/background pixels in each region exceeds 80 percent of the total number of pixels in the region. Then we set for each selected region its saliency score to 1/0, and obtain the binary-valued saliency score vector $\mathbf{A}_t = \left\{a_{t.1}, a_{t,2}, \cdots, a_{t,i} \cdots, a_{t,Q_t}\right\}$. As aforementioned, each region $R_{t,i}$ can be represented by using a 440D (322+118) saliency vector $\mathbf{x}_{t,i} = \left(\mathbf{v}_{t,i}^{LS}, \mathbf{v}_{t,i}^{MS}, \mathbf{v}_{t,i}^{HS}, \mathbf{v}_{t,i}^{RP}\right)$. The DSF regressor F based on random forest is learned by using the training data $\mathbf{X}_t = \left\{\mathbf{x}_{t,1}, \mathbf{x}_{t,2}, \cdots, \mathbf{x}_{t,i}, \ldots, \mathbf{x}_{t,Q_t}\right\}$ and the saliency score vector $\mathbf{A}_t$. During the process of learning the DSF regressor, the out-of-bag (OOB) error and the sum of squared residuals can be used to evaluate the contributions of each saliency measure to the final saliency measure. The details of saliency contribution analysis will be given in Section IV-C.

The DSF saliency map is generated by assigning RS_i to each pixel in the region R_i. We can observe from Fig. 2(j) that the DSF saliency map highlights the complete salient object with well-defined boundaries and consistently suppresses background regions compared to the result of linear fusion in Fig. 2(i).

A total of M DSF saliency maps, $DSF_m (m = 1, \ldots, M)$, are generated for all scales as shown in Fig. 1(f), and we combine them together to finally obtain the multiscale DSF based saliency map as follows:

$$MDSF = \frac{1}{M}\sum\nolimits_{m=1}^{M} DSF_m. \tag{9}$$

The linear fusion is adequate here for combining DSF saliency maps due to the relatively fewer DSF saliency maps ($M = 5$) and the similarity among them. The MDSF saliency map by combining these DSF saliency maps in Fig. 1(f) is shown in Fig. 1(g), which can better highlight the complete salient object and suppress background regions.

III. SALIENT OBJECT SEGMENTATION VIA BOOTSTRAP LEARNING

Given the MDSF saliency map generated in Section II, rich beneficial cues can be leveraged for salient object segmentation. A simple thresholding operation with the optimum tradeoff between precision and recall seems enough to segment salient objects with acceptable quality for some saliency maps. However, due to the inaccuracy and irregularity of saliency maps, some segmentation results would be poor by using the simple thresholding operation.

Generally, object segmentation can be formulated as a pixel-level binary labeling problem, which can be solved under the framework of SVM. Thus, for segmentation reliability on a wide range of saliency maps, we propose an effective salient object segmentation method via bootstrap learning. The proposed method is bootstrapped with samples based on the MDSF saliency map, which is generated for the input image, thereby alleviating the time-consuming offline training process.

A. Generating Training Samples

We sample N times based on an adaptive sampling strategy to collect accurate samples for training the SVM model. First, we calculate an adaptive threshold θ via the Otsu's method [57]. For the $q^{th}\,(q = 1, \cdots, N)$ sampling process, we generate the trimap as follows:

$$L_p = \begin{cases} 1, & MDSF\,(p) \geq \theta + \dfrac{255 - T}{N}\,(q-1) \\ 0, & MDSF\,(p) \leq \theta - \dfrac{T}{N}\,(q-1) \\ U, & otherwise, \end{cases} \tag{10}$$

where L_p is the label of pixel p, and U indicates that the pixel p is unlabeled. We randomly select Q pixels from the set of object pixels, $O = \{p \,|L_p = 1\}$, and the set of background pixels, $B = \{p \,|L_p = 0\}$, as the positive samples and negative samples, respectively. For a balance between the efficiency and performance, we set N to 4 and Q to 200 in our implementation.

For each pixel sample, various features are extracted as follows (along with the number of feature dimension in the parenthesis): average color of each channel in the RGB, HSV and Lab color spaces (9); average depth value (1); texture features including the absolute responses ($15 \times \mathbf{2}$) and max response ($1 \times \mathbf{2}$) of 15 LM filters, HOG feature ($32 \times \mathbf{2}$) and LBP feature ($1 \times \mathbf{2}$); geodesic distance ($1 \times \mathbf{2}$); the saliency value (1) and the coordinates (2) of each pixel. A multiplication factor of two, "$\times\mathbf{2}$", indicates that the features are extracted from color image and depth map simultaneously.

B. Learning the SVM Based on MKB

As described above, each sample can be represented by a 113D feature vector denoted as $\mathbf{v}_p$ and the segmentation label L_p. We adopt a simplified version of multiple kernel boosting (MKB) method as used in [11] to learn a label classifier, which maps the feature vector of each pixel to a segmentation label. Four types of kernels (linear, RBF, polynomial and sigmoid) are adopted and thus four weak SVMs are learned using the training samples. The final strong classifier is the weighted combination of the four weak SVMs as follows:

$$F\left(\mathbf{v}_p\right) = \sum_{h=1}^{4} \lambda_h z_h\left(\mathbf{v}_p\right), \tag{11}$$

where $z_h\left(\mathbf{v}_p\right) = \mu_h^T k_h\left(\mathbf{v}_p\right) + b_h$ denotes a SVM classifier. Let $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_Q\}$ denote the training samples, the kernel $k_h\left(\mathbf{v}_p\right)$ is represented as $\left[k_h\left(\mathbf{v}_p, \mathbf{v}_1\right), \ldots, k_h\left(\mathbf{v}_p, \mathbf{v}_Q\right)\right]^T$. The parameters μ_h, b_h and λ_h are learned by using the AdaBoost optimization process based on classification error for each SVM in an iterative way [58].

C. Segmentation Using the Trained Label Classifier

The boosted label classifier is applied to the test samples, i.e., all pixels in the input image, to obtain the pixel-wise salient object segmentation result. In order to obtain the well-defined boundaries, we map the pixel-wise segmentation result into M multiscale region segmentation results generated in Section II-A and obtain M pixel-wise salient object segmentation results, OS_m $(m = 1, 2, \cdots M)$. Therefore each pixel has M predicted labels, and the final label is determined by using the majority rule.

An example of the proposed salient object segmentation process is shown in Fig. 3. As can be seen from Fig. 3, along with the increasing number of sampling process, the quality of training data is refined and the quality of salient object segmentation result is also improved. For the input image in Fig. 1(a), the corresponding segmentation result is shown in Fig. 1(i), which accurately extracts the complete salient object.

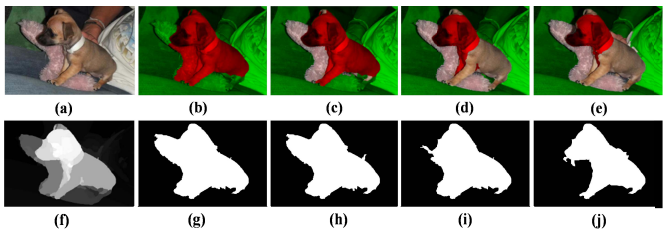

Fig. 3. Illustration of adaptive sampling and salient object segmentation. (a) Color image; (b-e) pixel samples obtained from the 1st to 4th sampling process; (f) saliency map; (g–j) salient object segmentation result with the increasing sampling number from 1 to 4. Object pixels are red, background pixels are green, and unknown regions are left unchanged.

IV. Experimental Results

A. Datasets and Experimental Settings

We performed experiments on two public datasets that are designed for depth-aware salient object detection and segmentation. The first dataset is RGBD-1000 [46], which includes 1000 color images and the corresponding depth maps captured by using Kinect directly, along with manually labeled ground truths. A variety of common objects in a series of indoor and outdoor scenes are captured in the dataset, and the depth information ranging from 0.5 to 10 meters for each image is repaired to obtain the smoothed depth map. The second dataset is NJUDS-2000 [43], which includes 2000 pairs of stereoscopic images with diverse objects and more complex and challenging scenes, along with manually labeled ground truths. The images are collected from Internet, 3D movies and photographs taken by a Fuji W3 stereo camera. Each disparity map, which is estimated from a pair of stereoscopic images using the optical flow method in [37], is used to represent the depth map.

Due to the different sources of depth information in the two datasets, before feature extraction with them, all the original depth maps and original disparity maps are uniformly normalized into the same range of [0, 255]. All the normalized disparity maps are then inverted, so that a small/large value of each pixel in them indicates a short/long distance from the camera, as the same as the normalized depth maps. Such a normalized depth/disparity map is used as the input depth map for the proposed MDSF model, and some examples are shown in the second column of Fig. 7. Then we randomly selected 500 images from the RGBD-1000 dataset and 500 images from the NJUDS-2000 dataset as the training set of our DSF regressor to avoid overfitting on either depth source. The remaining 500 images in the RGBD-1000 dataset and the remaining 1500 left-view images in the NJUDS-2000 dataset are used as the two testing sets.

We generated the saliency maps for all test images using the proposed MDSF saliency model, and compared the saliency detection performance with state-of-the-art saliency models including four depth-aware saliency models, i.e., SD [46], ACSD [43], CSD [44] and CDL [45], as well as three high-performing 2D saliency models including SO [10], ST [9] and DRFI [7]. For a fair comparison, we retrained a new model for DRFI using the same training set as our model. Additionally, we also compared with our previous work [52],

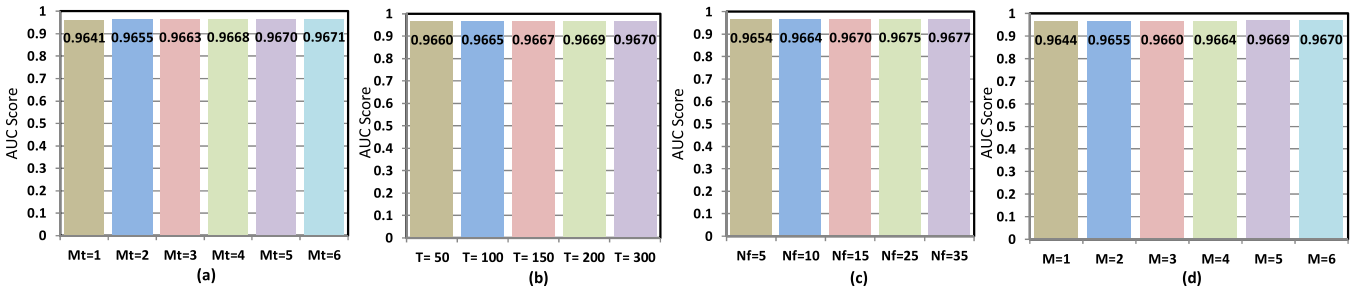

Fig. 4. AUC scores with different parameter settings during learning the DSF regressor. From left to right: comparison of saliency maps with different (a) number of segmentations to generate training samples, (b) number of decision trees, (c) number of features for node splitting in each tree and (d) number of scales to generate saliency maps.

i.e., DSFs model, which uses a single scale and only images from the RGBD-1000 dataset for training. Then, we performed salient object segmentation via bootstrap learning, and compared the segmentation performance with two state-of-the-art depth-aware salient object segmentation approaches, i.e., Ju's method [43] and Fan's method [50].

For evaluation of saliency detection performance, we adopt the commonly used objective measures, i.e., precision and recall, which are calculated by comparing the saliency map to the ground truth. Precision-recall (PR) curve is plotted by connecting the precision-recall values at all thresholds. Another important criterion is the AUC score (Area Under ROC Curve), which is generated based on true positive rates and false positive rates. For each test image, the binary ground truth is denoted by G, and the binary salient object mask, which is generated by binarizing the saliency map using an integer threshold from 0 to 255, is denoted as SM. In both G and SM, each object pixel is labeled as "1" and each background pixel is labeled as "0", the precision and recall are defined as follows:

$$Precision = \frac{\sum_{(x,y)} SM\,(x,y) \cdot G\,(x,y)}{\sum_{(x,y)} SM\,(x,y)}, \tag{12}$$

$$Recall = \frac{\sum_{(x,y)} SM\,(x,y) \cdot G\,(x,y)}{\sum_{(x,y)} G\,(x,y)}. \tag{13}$$

Besides, for evaluation of salient object segmentation performance, we use precision, recall and F-measure, which is defined as follows:

$$F_\gamma = \frac{(1+\gamma) \cdot precision \cdot recall}{\gamma \cdot precision + recall}, \tag{14}$$

where the coefficient γ is set to 0.5 to favor precision more than recall.

B. Analysis of Parameter Settings

In this subsection, we further analyze the performance of our MDSF saliency model against the settings of several key parameters during both training and testing phases. Five-fold cross-validation on the training set is run to select the parameters. We analyze the three parameters during the training process, i.e., M_t, the number of segmentation scales to generate training samples for each image, T, the number of decision trees, and N_f, the number of features sampled for splitting at each node when training the DSF regressor using random forest. Besides, we also analyze the key parameter for testing images, M, the number of scales. The average AUC scores resulting from cross-validation under different parameter settings are plotted in Fig. 4 for evaluation.

Regarding the number of segmentation scales to generate training samples, M_t, we gradually increase the number of training scales from the finest segmentation with 300 segmented regions to the coarsest segmentation with 50 segmented regions in our implementation. It can be seen from Fig. 4(a) that the performance of our MDSF saliency model with a larger value of M_t is higher. Since a larger value of M_t also results in a larger amount of training data, which benefits our random forest based regressor, we set $M_t = 5$, a moderate value, to obtain about 0.4 million samples for training.

We can observe from Fig. 4(b) more trees result in the higher performance, because the variances among the decision trees are smaller. In order to avoid time-consuming training process with a too large value of T, we set $T = 200$ in our implementation. Besides, the number of predictors sampled for splitting at each node, N_f, affects not only the discriminative saliency measures selected for each node but also the variances among decision trees. As shown in Fig. 4(c), we set $N_f = 25$ due that with an even larger value of N_f the performance gain is negligible but the training time obviously increases. As for the number of scales, M, for testing images, we use the same process from the finest segmentation to the coarsest segmentation, to increase the number of scales. As shown in Fig. 4(d), the AUC score steadily increases along with the increase of M, which also introduces a higher computational burden. Therefore we set $M = 5$ to balance effectiveness and computational efficiency.

The four parameters used in our MDSF saliency model, i.e., M_t, T and N_f during the training phase as well as M during the testing phase, are set based on the above analysis. Nonetheless, it can be seen from Fig. 4 that the performance of our MDSF saliency model is robust to different parameter settings.

C. Analysis of Saliency Contributions

In our MDSF saliency model, four classes of features from color image and depth map are extracted at low, middle and high levels, to generate totally 322 saliency maps. They are fused in our DSF regressor with the addition of auxiliary regional properties (AP). In this subsection, we try to

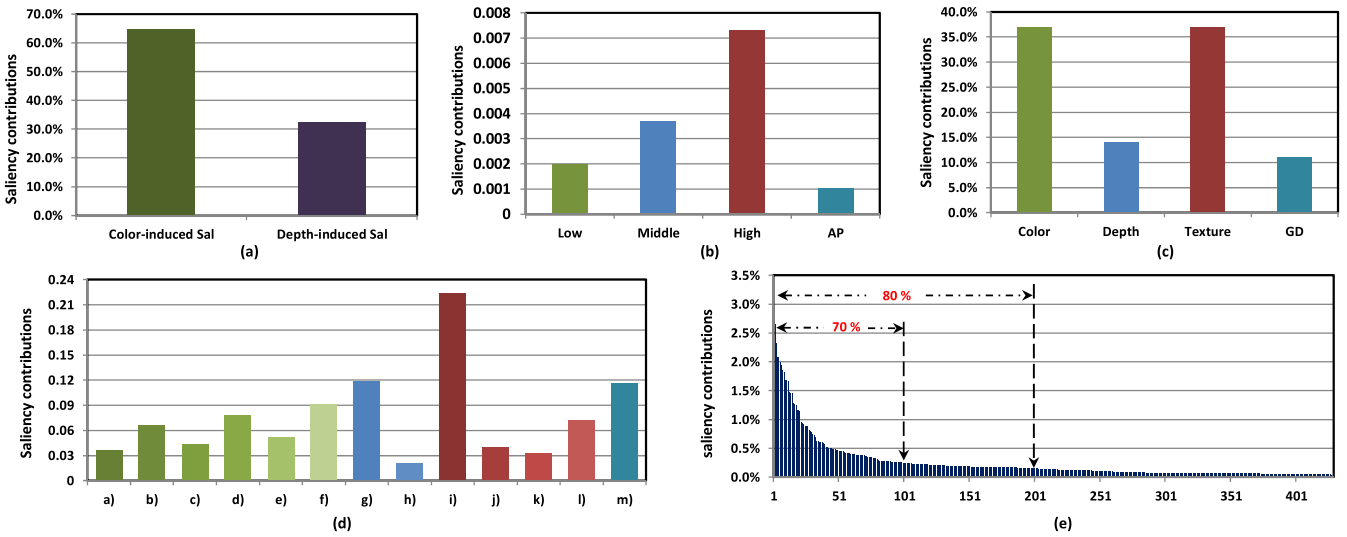

Fig. 5. Saliency contribution from different saliency measures in terms of OOB error. From left to right: saliency contributions of (a) color-induced saliency and depth-induced saliency, (b) low-level, mid-level, high-level saliency measures and auxiliary region properties, (c) color, depth, texture and geodesic distance features, (d) regional features from a) to m), which correspond to the regional features from top to bottom in the second column of TABLE I, and (e) rank of saliency contributions of the 440D saliency vector.

investigate how different saliency measures and regional properties contribute to the overall saliency detection performance. From different views, Fig. 5 shows the saliency contributions, which are estimated by using the OOB (out of bag) errors during the training of random forest regressor.

First, to investigate the importance of color information and depth information, we define saliency measures involving color, color texture and color geodesic distance as color-induced saliency measures, and those involving depth, depth texture and depth geodesic distance as depth-induced saliency measures. As shown in Fig. 5(a), the color-induced saliency measures (236 dimensions) and the depth-induced saliency measures (86 dimensions) contribute about two thirds and one third, respectively, of saliency detection performance. Considering the dimension of saliency measures, the depth-induced saliency measures are more discriminative.

Fig. 5(b) shows the contribution per saliency measure at different levels. As can be seen, the contribution per saliency measure at high level is higher than that at low level and middle level. The reason is that the two commonly used location priors at high level, e.g., border bias and center bias, are robust to most images to some extent, and the saliency detection performance is further promoted by using the feature similarity between regions. Besides, the weighting factors of depth and geodesic distance make the mid-level saliency measures more effective than low-level saliency measures. Last but not least, auxiliary region properties also play an important role on the overall saliency detection performance.

Saliency contributions from the four classes of features as classified in Table I are demonstrated in Fig. 5(c). As can be seen, color features (including color histograms and mean color in different color spaces) and texture features (including LM, LBP and HOG features on depth map and color image) are the two significant contributors to saliency detection performance. Note that although depth features and geodesic distance (GD) features are used to generate only 24 saliency measures out of 322, they contribute around a quarter of the whole energy, which clearly shows their effectiveness and discrimination for saliency detection.

Fig. 5(d) presents the performance contribution of regional features from a) to m), which correspond to the regional features from top to bottom in the second column of Table I. As can be seen in Fig. 5(d), each color feature from a) to f) among the three color spaces has its contributions. With a relatively high dimension, the absolute LM filter response, i.e., the feature i) in Fig. 5(d), has the most importance. Besides, the depth histogram, i.e., the feature h) in Fig. 5(d), is less discriminative than other histogram features, and this indicates that the depth values within each region are usually homogeneous.

In Fig. 5(e), the importance of all saliency measures and auxiliary regional properties are ranked in the descending order. Obviously, the majority of contribution derives from some discriminative saliency measures and regional properties. The most effective 100 and 200 saliency measures and regional properties occupy around 70% and 80%, respectively, of the total contribution. According to Fig. 5(e), we can select the most important saliency measures. As can be seen in Fig. 2, even the five most important saliency measures provide far less accurate information of salient objects, our MDSF model can achieve the higher saliency detection performance compared to the linear fusion method.

D. Evaluation of Saliency Models and Discussion

The objective evaluation of saliency detection performance is presented in Fig. 6. The PR curves and ROC curves are calculated on the two test sets, RGBD-1000 and NJUDS-2000, respectively. As shown in Fig. 6, thanks to the additional depth information, the depth-aware saliency models achieve the overall better performance than the three 2D models. Our MDSF model consistently outperforms all the other models on the two datasets with large margins in terms of PR curve

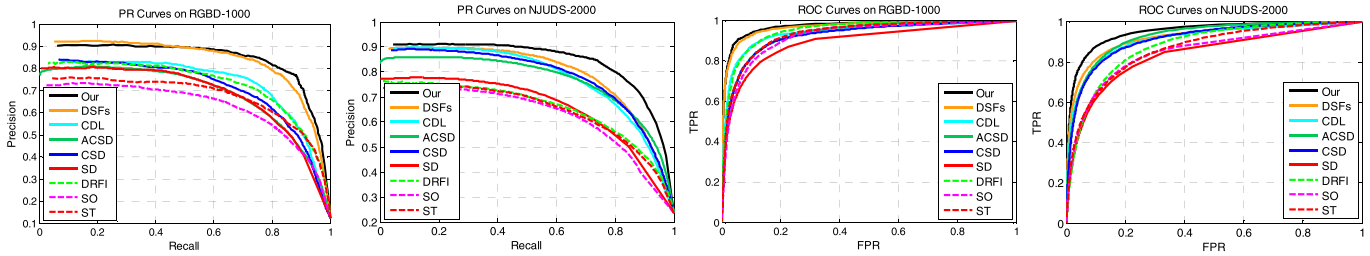

Fig. 6. (better viewed in color) Objective evaluation of different saliency models in terms of PR curves and ROC curves.

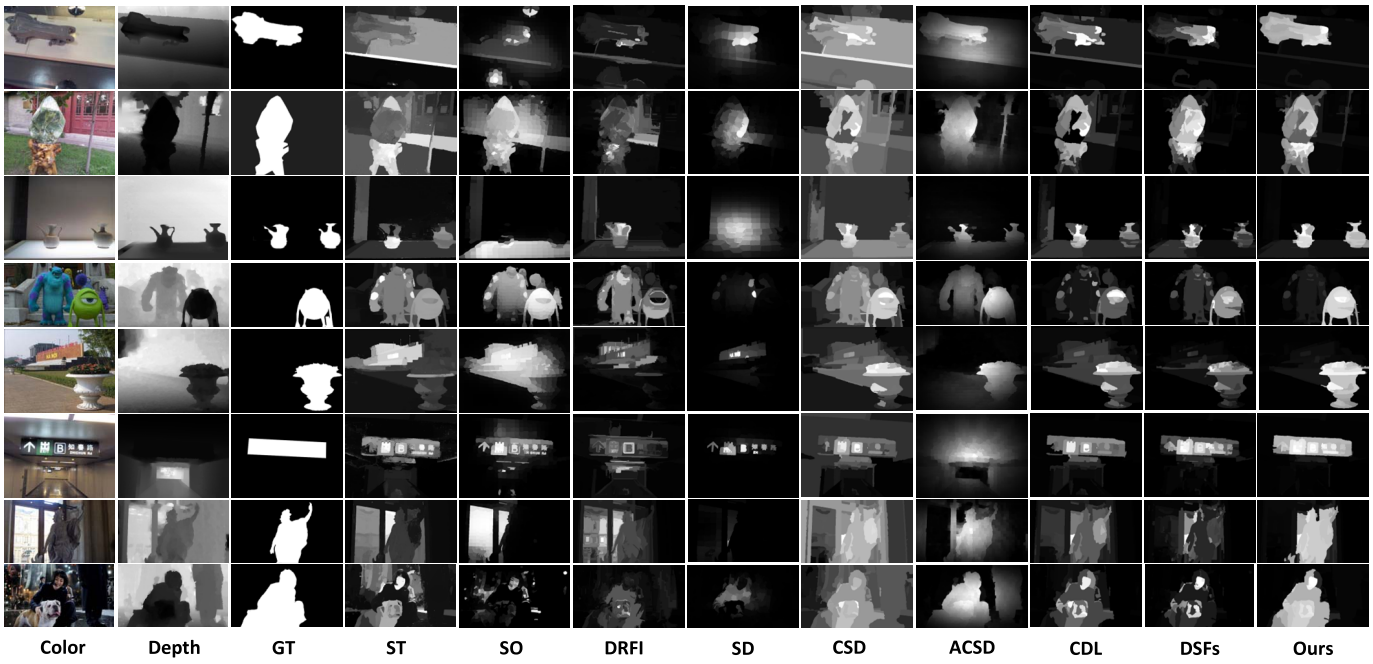

Fig. 7. Saliency maps generated by using different saliency models.

and ROC curve. This indicates that the quality of our MDSF saliency maps is generally better than the other classes of saliency maps for salient object segmentation.

For a subjective comparison, saliency maps of several example images are shown in Fig. 7. Compared to those 2D saliency models and other depth-aware saliency models, we can see from Fig. 7 that our model can generally better highlight salient objects with well-defined boundaries and suppress background regions effectively. Especially for some complicated images such as the bottom two examples, the low contrast between object and background as well as the cluttered background cause difficulties for 2D saliency models. However, in the depth maps, it is clear that the depth levels of salient regions are different from other regions. Compared to the 2D models, most depth-aware saliency models achieve the better saliency detection performance on the two example images. With the effective utilization of depth and color information, our MDSF model extracts various features of depth and color at different levels and effectively finds the most discriminative ones for generating the better saliency map. Therefore our MDSF model can highlight the complete salient objects with well-defined boundaries better than other depth-aware saliency models.

We further evaluate how the proposed discriminative saliency fusion and multiscale region segmentation affect the saliency detection performance. Two other versions of our saliency model are denoted as 1) DSF: the saliency map is generated via discriminative saliency fusion at a moderate scale with $\tau_m = 200$ using Eq. (8); and 2) LSF: the saliency map is generated by fusing all the saliency measures of three levels at a moderate scale with $\tau_m = 200$ in a linear way. The AUC values of the two versions and our MDSF model are shown in Fig. 8. It is apparent that DSF achieves a much better saliency detection performance than the linear version i.e., LSF, and this demonstrates the power of the discriminative

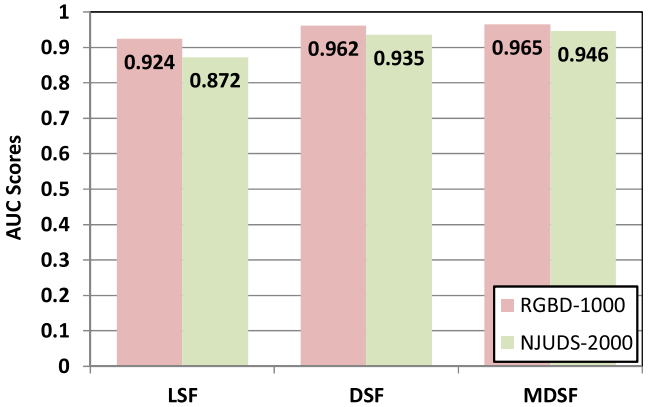

Fig. 8. Comparison with different versions of our saliency model.

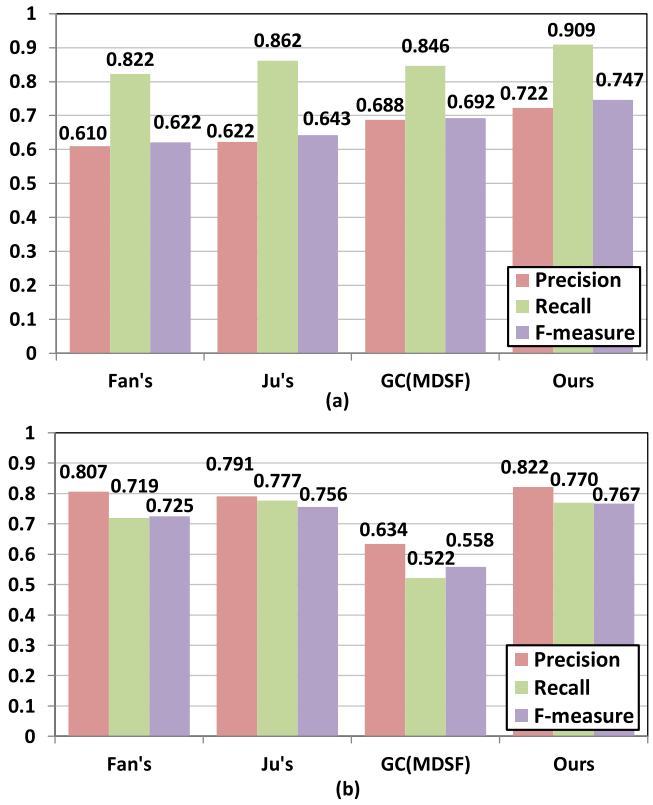

Fig. 9. Objective comparison of segmentation results on (a) RGBD-1000 and (b) NJUDS-2000.

saliency fusion method. Besides, the performance gain of MDSF over DSF shows that the introduction of multiscale region segmentation can further boost the saliency detection performance.

E. Evaluation of Salient Object Segmentation and Discussion

In order to evaluate the performance of salient object segmentation, we performed experiments on the images in the two test sets of RGBD-1000 and NJUDS-2000, respectively, using the proposed salient object segmentation method via bootstrap learning based on our MDSF saliency maps. Two state-of-the-art depth-aware salient object segmentation methods, i.e, Ju's method [43] and Fan's method [50], are compared with our method.

We objectively evaluate the quality of salient object segmentation results using the measures of precision, recall, and F-measure as defined in Eqs. (12)–(14). Fig. 9 shows the three measures achieved using different segmentation methods. As the overall performance comparison, our method achieves the highest F-measure on both datasets compared to the other two methods, and this demonstrates the better segmentation performance of our method. The advantage of our method is more significant on the RGBD-1000 dataset, since our method consistently outperforms the other two methods on precision, recall and F-measure.

The segmentation results for some test images are shown in Fig. 10. As shown in Fig. 10(a)–(d), for some salient objects with obvious contrast in either color image or depth map, all methods can segment salient objects, while our

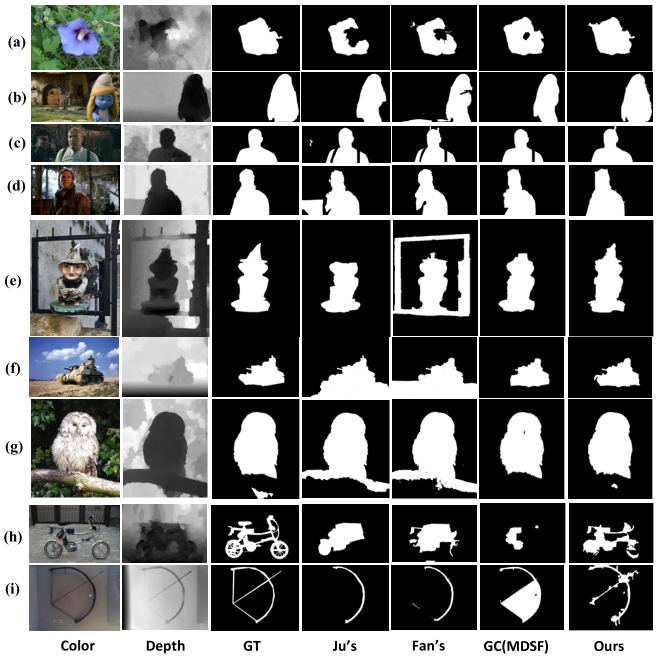

Fig. 10. Salient object segmentation results using different methods.

method can segment more complete objects due to the better saliency maps and full use of various depth features. For some complicated images with various scenes as shown in Fig. 10(e)–(g), it is difficult for Ju's method and Fan's method, which are based on graph cut framework using only color and depth features, to obtain high-quality salient object segmentation results. In Fig. 10(e)–(g), some background regions are quite similar with salient objects in both color image and depth map. Ju's method and Fan's method cannot accurately distinguish boundaries between salient object and background. In contrast, our method which takes into account multiple features including color, depth, texture and geodesic distance is more effective for such complicated scenes. For more complicated images with complex structures such as those in Fig. 10(h)–(i), our method as well as the other methods cannot accurately segment the precise structures. But thanks to the proposed adaptive sampling strategy, which is able to collect relatively accurate samples for training our SVM classifiers, the proposed method can segment the more complete objects.

Besides, the segmentation results using a popular graph cut based segmentation method (i.e., GrabCut [47]) with our MDSF saliency maps are denoted as GC(MDSF) and further compared with our bootstrap learning based method. Here, MDSF saliency maps are used for initializing object segmentation in place of user interaction, and the parameter settings follow that in [47]. The objective evaluation and segmentation results of GC(MDSF) are presented in Fig. 9 and Fig. 10, respectively. A comparison between our results and GC(MDSF) verifies the better performance of our bootstrap learning based segmentation method.

V. Conclusions

In this paper, we propose a novel depth-aware saliency model using multiscale discriminative saliency

fusion (MDSF). Saliency measures on four classes of features at three levels are calculated. A random forest regressor is learned to perform the discriminative saliency fusion and generate the DSF saliency map at each scale, and DSF saliency maps across multiple scales are combined to generate the final MDSF saliency map. Furthermore, we propose a bootstrap learning based salient object segmentation method, which is bootstrapped with samples based on the MDSF saliency map and learns multiple kernel SVMs. Both subjective and objective evaluations demonstrate that the proposed MDSF model and salient object segmentation method achieve the better saliency detection performance and the better segmentation performance, respectively, on both RGBD and stereoscopic image datasets. In our future work, we will investigate how the saliency detection performance varies with different learning based methods, and utilize more features for measuring saliency as well as reduce the redundancy of features.

REFERENCES

[1] P. Le Callet and E. Niebur, "Visual attention and applications in multimedia technologies," *Proc. IEEE*, vol. 101, no. 9, pp. 2058–2067, Sep. 2013.

[2] 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.

[3] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proc. IEEE CVPR*, Jun. 2007, pp. 1–8.

[4] Z. Liu, Y. Xue, H. Yan, and Z. Zhang, "Efficient saliency detection based on Gaussian models," *IET Image Process.*, vol. 5, no. 2, pp. 122–131, Mar. 2011.

[5] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. IEEE CVPR*, Jun. 2011, pp. 409–416.

[6] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. ECCV*, Sep. 2012, pp. 29–42.

[7] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE CVPR*, Jun. 2013, pp. 2083–2090.

[8] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE CVPR*, Jun. 2013, pp. 3166–3173.

[9] 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.

[10] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE CVPR*, Jun. 2014, pp. 2814–2821.

[11] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *Proc. IEEE CVPR*, Jun. 2015, pp. 1884–1892.

[12] P. Siva, C. Russell, T. Xiang, and L. Agapito, "Looking beyond the image: Unsupervised learning for object saliency and detection," in *Proc. IEEE CVPR*, Jun. 2013, pp. 3238–3245.

[13] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, "SUN: Top-down saliency using natural statistics," *Vis. Cognit.*, vol. 17, nos. 6–7, pp. 979–1003, 2009.

[14] J. Yang and M.-H. Yang, "Top-down visual saliency via joint CRF and dictionary learning," in *Proc. IEEE CVPR*, Jun. 2012, pp. 2296–2303.

[15] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. IEEE CVPR*, Jun. 2012, pp. 478–485.

[16] J. Wang, M. P. Da Silva, P. Le Callet, and V. Ricordel, "Computational model of stereoscopic 3D visual saliency," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2151–2165, Jun. 2013.

[17] J. Wang, D. M. Chandler, and P. Le Callet, "Quantifying the relationship between visual salience and visual importance," *Proc. SPIE*, vol. 7527, p. 75270K, Feb. 2010.

[18] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. ECCV*, Sep. 2010, pp. 366–379.

[19] Z. Liu, R. Shi, L. Shen, Y. Xue, K. N. Ngan, and Z. Zhang, "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.

[20] 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.

[21] Y. Luo, J. Yuan, P. Xue, and Q. Tian, "Saliency density maximization for efficient visual objects discovery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 12, pp. 1822–1834, Dec. 2011.

[22] R. Shi, Z. Liu, H. Du, X. Zhang, and L. Shen, "Region diversity maximization for salient object detection," *IEEE Signal Process. Lett.*, vol. 19, no. 4, pp. 215–218, Apr. 2012.

[23] J. Sun and H. Ling, "Scale and object aware image retargeting for thumbnail browsing," in *Proc. IEEE ICCV*, Nov. 2011, pp. 1511–1518.

[24] H. Du, Z. Liu, J. Jiang, and L. Shen, "Stretchability-aware block scaling for image retargeting," *J. Vis. Commun. Image Represent.*, vol. 24, no. 4, pp. 499–508, 2013.

[25] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[26] L. Shen, Z. Liu, and Z. Zhang, "A novel H.264 rate control algorithm with consideration of visual attention," *J. Multimedia Tools Appl.*, vol. 63, no. 3, pp. 709–727, 2013.

[27] S. Frintrop, A. Nuchter, H. Surmann, and J. Hertzberg, "Saliency-based object recognition in 3D data," in *Proc. IEEE ICIROS*, Sep. 2004, pp. 2167–2172.

[28] O. Le Meur and Z. Liu, "Saccadic model of eye movements for free-viewing condition," *Vis. Res.*, vol. 116, pp. 152–164, Nov. 2015.

[29] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *Proc. ECCV*, Oct. 2012, pp. 414–429.

[30] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[31] Q. Zhao and C. Koch, "Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost," *J. Vis.*, vol. 12, no. 6, pp. 1–15, 2012.

[32] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *J. Vis.*, vol. 13, no. 4, pp. 1–20, Mar. 2013.

[33] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: A data-driven approach," in *Proc. IEEE CVPR*, Jun. 2013, pp. 1131–1138.

[34] O. Le Meur and Z. Liu, "Saliency aggregation: Does unity make strength?" in *Proc. ACCV*, Nov. 2014, pp. 18–32.

[35] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. ECCV*, Sep. 2014, pp. 345–360.

[36] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2005, pp. 1161–1168.

[37] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *Proc. IEEE CVPR*, Jun. 2010, pp. 2432–2439.

[38] N. Ouerhani and H. Hugli, "Computing visual attention from scene depth," in *Proc. IEEE ICPR*, Sep. 2000, pp. 375–378.

[39] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *Proc. ECCV*, Oct. 2012, pp. 101–115.

[40] J. Gautier and O. Le Meur, "A time-dependent saliency model combining center and depth biases for 2D and 3D viewing conditions," *Cognit. Comput.*, vol. 4, no. 2, pp. 141–156, Jun. 2012.

[41] K. Desingh, K. M. Krishna, D. Rajan, and C. V. Jawahar, "Depth really matters: Improving visual salient region detection with depth," in *Proc. BMVC*, Sep. 2013, pp. 1–11.

[42] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE CVPR*, Jun. 2012, pp. 454–461.

[43] R. Ju, Y. Liu, T. Ren, L. Ge, and G. Wu, "Depth-aware salient object detection using anisotropic center-surround difference," *Signal Process., Image Commun.*, vol. 38, no. 10, pp. 115–126, Oct. 2015.

[44] X. Fan, Z. Liu, and G. Sun, "Salient region detection for stereoscopic images," in *Proc. IEEE DSP*, Aug. 2014, pp. 454–458.

[45] H. Song, Z. Liu, H. Du, G. Sun, and C. Bai, "Saliency detection for RGBD images," in *Proc. ICIMCS*, Aug. 2015, pp. 240–243.

[46] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. ECCV*, Sep. 2014, pp. 92–109.

[47] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

[48] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, Jan. 2006.
[49] B. C. Ko and J.-Y. Nam, "Object-of-interest image segmentation based on human attention and semantic region clustering," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 23, no. 10, pp. 2462–2470, Oct. 2006.
[50] X. Fan, Z. Liu, and L. Ye, "Salient object segmentation from stereoscopic Images," in *Proc. Graph-Based Representations Pattern Recognit.*, May 2015, pp. 272–281.
[51] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
[52] H. Song, Z. Liu, H. Du, and G. Sun, "Depth-aware saliency detection using discriminative saliency fusion," in *Proc. IEEE ICASSAP*, Mar. 2016, pp. 1626–1630.
[53] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.
[54] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vis.*, vol. 43, no. 1, pp. 29–44, Feb. 2001.
[55] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE CVPR*, Jun. 2005, pp. 886–893.
[56] M. Heikkilä, M. Pietikäinen, and C. Schmid, "Description of interest regions with local binary patterns," *Pattern Recognit.*, vol. 42, no. 3, pp. 425–436, 2009.
[57] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.
[58] F. Yang, H. Lu, and Y.-W. Chen, "Human tracking by multiple kernel boosting with locality affinity constraints," in *Proc. ACCV*, Nov. 2011, pp. 39–50.

Hangke Song received the B.E. degree from Hangzhou Dianzi University, Hangzhou, China, in 2014, and the M.E. degree from Shanghai University, Shanghai, China, in 2017. His research interests include saliency detection and salient object segmentation.

Zhi Liu (M'07–SM'15) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, and the Ph.D. degree from Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China, in 1999, 2002, and 2005, respectively. From Aug. 2012 to Aug. 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support by the EU FP7 Marie Curie Actions. He is currently a Professor with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. He has authored over 130 refereed technical papers in international journals and conferences. His research interests include image/video processing, machine learning, computer vision, and multimedia communication. He was a TPC member in the VCIP 2016, ICME 2014, WIAMIS 2013, IWVP 2011, PCM 2010, and ISPACS 2010. He co-organized special sessions on visual attention, saliency models, and applications at WIAMIS 2013 and ICME 2014. He is an Area Editor of *Signal Processing: Image Communication* and served as a Guest Editor for the special issue on Recent Advances in Saliency Models, Applications and Evaluations in *Signal Processing: Image Communication*.

Huan Du received the B.E. degree from China Center Normal University, Wuhan, China, and the M.E. degree from Shanghai University, Shanghai, China, in 2010 and 2013, respectively, where she is currently pursuing the Ph.D. degree. Since 2013, she has been with the Third Research Institute of Ministry of Public Security, Shanghai, China. Her research interests include saliency model, image retrieval, and face verification.

Guangling Sun received the B.S. degree in electronic engineering from Northeast Forestry University, China, in 1996, and the M.E. and Ph.D. degrees in computer application technology from the Harbin Institute of Technology, China, in 1998 and 2003, respectively. She was with the University of Maryland at College Park, College Park, as a visiting scholar from Dec. 2013 to Dec. 2014. Since 2006, she has been with the Faculty of the School of Communication and Information Engineering, Shanghai University, where she is currently an Associate Professor. Her research interests include saliency detection, face recognition, and image/video processing.

Olivier Le Meur received the Ph.D. degree from the University of Nantes, Nantes, France, in 2005. He was with the Media and Broadcasting Industry from 1999 to 2009. In 2003, he joined the Research Center of Thomson-Technicolor, Rennes, France, where he supervised a research project concerning the modeling of human visual attention. He has been an Associate Professor of image processing with the University of Rennes 1 since 2009. In the SIROCCO Team of IRISA/INRIA-Rennes, his current research interests include human visual attention, computational modeling of visual attention, and saliency-based applications, such as video compression, objective assessment of video quality, and retargeting.

Tongwei Ren (M'11) received the B.S., M.E., and Ph.D. degrees from Nanjing University, Nanjing, China, in 2004, 2006, and 2010, respectively. He was an Assistant Researcher with the Department of Computing, The Hong Kong Polytechnic University, China in 2008, and a Visiting Scholar with the School of Computing, National University of Singapore, Singapore from 2016 to 2017. He is currently an Associate Professor with the Software Institute, Nanjing University, China. He has authored over 40 refereed technical papers in international journals and conferences. His research interests include category-independent object analysis and stereo image processing.