This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2381471, IEEE Transactions on Circuits and Systems for Video Technology

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2381471, IEEE Transactions on Circuits and Systems for Video Technology

Background Prior Based Salient Object Detection via Deep Reconstruction Residual

Junwei Han, Dingwen Zhang, Xintao Hu, Lei Guo, Jinchang Ren, and Feng Wu

***Abstract*—Detection of salient objects from images is gaining increasing research interest in recent years as it can substantially facilitate a wide range of content-based multimedia applications. Based on the assumption that foreground salient regions are distinctive within a certain context, most conventional approaches rely on a number of hand designed features and their distinctiveness measured using local or global contrast. Although these approaches have shown effective in dealing with simple images, their limited capability may cause difficulties when dealing with more complicated images. This paper proposes a novel framework for saliency detection by first modeling the background and then separating salient objects from the background. We develop stacked denoising autoencoders with deep learning architectures to model the background where latent patterns are explored and more powerful representations of data are learnt in an unsupervised and bottom up manner. Afterwards, we formulate the separation of salient objects from the background as a problem of measuring reconstruction residuals of deep autoencoders. Comprehensive evaluations on three benchmark datasets and comparisons with 9 state-of-the-art algorithms demonstrate the superiority of the proposed work.**

***Index Terms*—salient object detection, stacked denoising autoencoder, background prior, deep reconstruction residual.**

I. Introduction

SALIENT object detection aiming to discover the most important and informative parts in an image is gaining intensive research attention recently as it can serve as a base for a large number of multimedia applications such as image resizing, image montage, action analysis and visual recognition [1-4]. Based on the underlying hypothesis that the salient stimulus is distinct from its contextual stimuli, most existing saliency detection models need to solve two key problems: i) extract effective features to represent the image and, ii) develop an optimal mechanism to measure the distinctiveness over the extracted features.

The performance of saliency detection models heavily relies on the features (data representations) being used. In the last 15 years, a variety of features have been proposed for the task of image saliency detection. The earliest saliency computation model by Itti *et al*. [5] proposed three biological plausible features including color, intensity, and orientation. In Judd *et al.* [6], besides Itti's three features, several new features were introduced to characterize image content, which include the local energy of the steerable pyramid filters, subband pyramids based features, 3D color histogram, and horizon line detector. As visual attention could be directed by specific objects, some detectors of face, car, and person were treated as features for detecting saliency [6, 7]. All these feature representations are hand-designed and require significant amounts of domain knowledge. However, hand-designed features in general suffer poor generalization capability for different images, especially due to the lack of thorough understanding of the biological mechanisms and principles of human visual attention as well as weak human intuition involved. A few recent approaches tried to learn better representations from natural scenes for saliency detection by using independent component analysis (ICA) [8], sparse coding [9, 10], and low-rank matrix recovery [11]. Nevertheless, due to the shallow-structured architectures used these methods still have limited representational power and are insufficient to capture high-level information and latent patterns of complex image data. To overcome such drawbacks, in this paper, we investigate the feasibility of learning more powerful representation directly from the raw image data itself in an unsupervised way for the task of saliency detection.

The saliency or distinctiveness is typically measured by image contrast computation over features, where various contrast measures have been presented. Depending on the extent of context in which the contrast is calculated, these approaches can be classified into local-contrast based methods and global-contrast based methods. Local-contrast based methods estimate the saliency of an image pixel or an image patch by calculating the contrast against its local neighborhood, and some representative local methods include the center-surround difference [5, 6, 12, 13], incremental coding length [10], and self-resemblance [14]. Global-contrast based methods characterize the saliency of an image region as the uniqueness in the entire image. Previous literatures have proposed a variety of approaches to model the global contrast from different perspectives. To be specific, in [15] and [16] the global contrast is derived in the frequency domain with the hypothesis that salient regions are normally less frequent. Han *et al.* [9] and Zhang *et al.* [8] utilized the Gaussian models to

Manuscript received on April 14, 2014. This work was partially supported by the National Science Foundation of China under Grant 61103061, 91120005, and 61473231.

Junwei Han, Dingwen Zhang, Xintao Hu, and Lei Guo are with School of Automation, Northwestern Polytechnical University, Xi'an China. (phone and fax: 86-29-88431318; e-mail: junweihan2010@gmail.com).

Jinchang Ren is with the Department of Electronic and Electrical Engineering, University of Strathclyde, UK.

Feng Wu is with School of Information Science, University of Science and Technology of China.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2381471, IEEE Transactions on Circuits and Systems for Video Technology

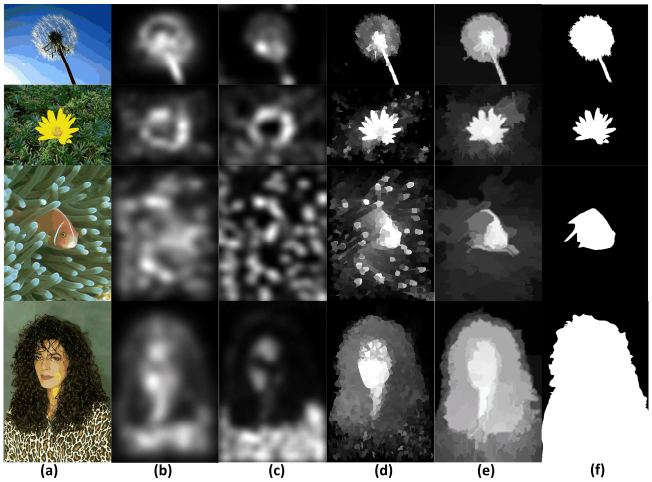

Fig. 1. Some examples of saliency detection. (a) Input images. (b) Results from one local contrast method [5]. (c) Results from one global contrast method [15]. (d) Results from the background prior based method [18]. (e) Results from the proposed method. (f) Ground truth salient object masks.

calculate the global contrast. Cheng *et al.* [17] proposed to model the global contrast on the region level where each region's contrast is generated by a weighted summation of the differences between itself and all other regions. Shen *et al.* [11] represented a whole image as a low-rank matrix with sparse noises where sparse noises denote the salient regions.

In spite of extensive efforts, local and global contrast based approaches still suffer from some drawbacks. First, these approaches normally can only highlight object boundaries but fail to detect the whole target region uniformly as shown in the examples given in Fig. 1. This problem may be alleviated in some global-contrast based methods while the results yielded are still unsatisfactory. Second, although the salient objects often present high contrast, the inverse might unnecessarily be true [11]. In many complex images (as shown in the third example of Fig. 1), the background contains small-scale high-contrast patterns which may lead to previous contrast-based methods fail in such cases.

Essentially, the true aim of salient object detection is to find objects that are distinctive from the image background. It needs to calculate the contrast between the objects and the image background and then select those with high contrast as the salient objects. However, the local and global contrast-based methods do not identify which regions form the image background. They blindly assume the neighboring regions or the entire image to be the background and then calculate the contrast between each location and the assumed background. As their assumed background may not be the real one, the determined contrast also becomes incorrect, which in turn reduces the performance of saliency detection. To overcome these problems, a few emerging methods [18, 19] using background priors were proposed based on the idea of modeling the property of background first and thereby separating salient objects from the background. Specially, Wei *et al.* [18] exploited the boundary and connectivity priors about the background in natural images and detected saliency based on the geodesic distance. Considering that the salient object may be partially cropped on the boundary, this work adopts an existing saliency detection method [33] to compute the saliency of boundary patches and generates weights for the virtual background nodes. However, in some challenging images where the work [33] could not calculate the saliency of boundary patches precisely, the method of [18] is difficult to obtain satisfactory results. Yang *et al.* [19] modeled saliency detection as a manifold ranking problem and proposed a two-stage scheme for graph labelling. They represent the image as a close-loop graph with superpixels as nodes. In saliency detection, they first use the nodes on the image boundary as background seeds to rank other nodes in the graph. Then, in the second stage, they select the salient nodes from the detection results of the first stage and use them to refine the saliency of other nodes in the graph. On the assumption that the image boundary is mostly background, these methods result in a background template. As a result, the contrast between salient object and background can be precisely obtained. By incorporating background priors into traditional contrast-based methods, they show improved results in saliency detection.

However, existing background prior based methods still have certain limitations. Typically, there are four scenarios where performing background prior based saliency detection as summarized below.

1) The entire image boundary is a large and smoothly connected region (see the first row of Fig. 1);

2) The regions defined within the image boundary look different whereas they may share certain latent pattern (see the second row of Fig. 1);

3) The background is complex (for example, containing small-scale high-contrast patterns) and regions of image boundary are different as shown in the third row of Fig. 1;

4) Salient objects significantly touch the image boundary and parts of them are wrongly considered as background as shown in the fourth row of Fig. 1.

As can be seen in Fig. 1, existing background prior based approaches [18] are effective for the first scenario and moderately effective for the second scenario. However, unsatisfactory results are produced in dealing with the last two scenarios. In this paper, we propose a novel background prior based saliency detection framework using stacked denoising autoencoder (SDAE) with deep learning architectures. In the proposed work, SDAE is used to model image background. Rather than adopting hand-designed features as used in previous works [18, 19], the deep-structured SDAE is employed to learn more powerful representation directly from the raw image data in an unsupervised way, which also enables to capture the latent pattern of the input data hierarchically. It thus helps to deal with the second scenario (shown in the second row of Fig. 1) where the background regions share latent patterns. Then, the measure of contrast between salient objects and the background is formulated as the reconstruction residuals in the deep-structured SDAE. Different from the previous works [18, 19] which mainly focused on the way to calculate the similarity or distinctiveness between a certain image patch and the image boundary, the proposed work pays more attention to modeling the background regions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2381471, IEEE Transactions on Circuits and Systems for Video Technology

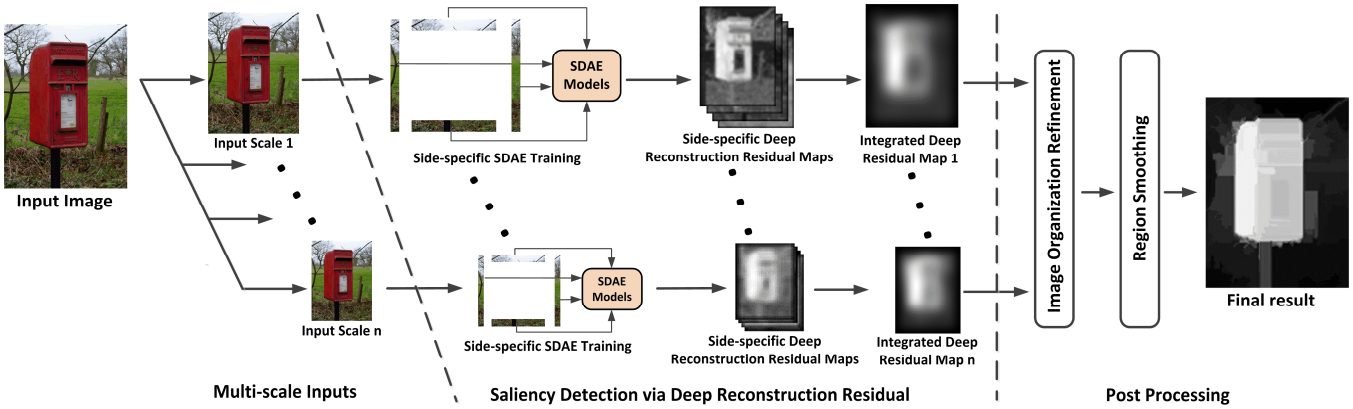

Fig. 2. The workflow of the proposed framework.

Specifically, the sparsity is considered when training SDAE models, which is helpful to suppress the saliency of the background regions. Therefore, it is robust in handling the third scenario (shown in the third row of Fig. 1) where the most challenging task is to avoid mis-highlighting the small-scale high-contrast background regions in the saliency maps. In addition, the learning process of SDAE with the usage of stochastic corruption criteria is helpful to train a deep model for better robustness and feature representation. Thus, the trained robust SDAE shows promising performance in these scenarios.

Fig. 2 illustrates the workflow of the proposed framework. First, we down sample the original image to multiple scales to generate the multi-scale inputs. Afterwards, we explore the background prior via SDAE and detect salient regions by deep reconstruction residuals which can reflect the distinctness between the background and salient regions. Finally, post processes are applied to integrate the salient object detection results for each scale of input and generate the final saliency map by image organization refinement and region smoothing.

The rest of the paper is organized as follows. Section II introduces the proposed approach in details. Section III presents experimental results with quantitative evaluation in comparison with a group of state-of-the-art approaches. Finally, several concluding remarks are drawn in Section IV.

II. The Proposed Approach

In this section, we discuss the proposed method for salient object detection in details. It includes three subsections, which in turn introduce SDAE, the proposed salient detection framework, and two useful post-processing steps, respectively.

A. Stacked Denoising Autoencoder (SDAE)

Autoencoders are simple learning neural networks which aim to transform inputs into outputs with the least possible amount of distortion for learning latent patterns of the given data. While conceptually simple, they play an important role in machine learning and feature representation. More recently, autoencoders have taken center stage again in the "deep architecture" approaches [20-23], where autoencoders are stacked and pre-trained in an unsupervised fashion. These deep architectures have been shown to lead to state-of-the-art results on a number of classification and regression problems [24].

As a form of neural network, the classical autoencoder [24] is an unsupervised learning algorithm that applies back-propagation and sets the target values of the network outputs to be equal to the inputs. Specifically, it includes an encoding process and a decoding process. The encoding process uses an encoding function $f(\boldsymbol{x}_i, \theta_f)$ to take a nonlinear mapping from the visible input vector $\boldsymbol{x}_i$ to a hidden representation vector $\boldsymbol{y}_i$ by using an affine transformation with a projection matrix $\mathbf{W}$ and a bias $\boldsymbol{b}$. Normally, the sigmoid function $sigm(\eta) = 1/(1+exp(-\eta))$ is used as the deterministic mapping as follows:

$$\boldsymbol{y}_i = f(\boldsymbol{x}_i, \theta_f) = sigm(\mathbf{W}\boldsymbol{x}_i + \boldsymbol{b}) \quad (1)$$

A decoding function $g(\boldsymbol{y}_i, \theta_g)$ is adopted to map the hidden representation $\boldsymbol{y}_i$ back to a reconstruction representation $\boldsymbol{z}_i$ through a similar transformation:

$$\boldsymbol{z}_i = g(\boldsymbol{y}_i, \theta_g) = sigm(\mathbf{W}'\boldsymbol{y}_i + \boldsymbol{b}') \quad (2)$$

After the decoding process, the obtained reconstruction is taken as a prediction of input $\boldsymbol{x}_i$. The training of an autoencoder is to optimize the parameters $\theta_f = \{\mathbf{W}, \boldsymbol{b}\}$ and $\theta_g = \{\mathbf{W}', \boldsymbol{b}'\}$ by minimizing the mean-squared reconstruction error between the training data and their reconstructed data via:

$$\arg\min_{\theta_f, \theta_g} L(\mathbf{X}, \mathbf{Z}) \quad (3)$$

$$L(\mathbf{X}, \mathbf{Z}) = \frac{1}{2}\sum_{i=1}^{m} || \boldsymbol{x}_i - \boldsymbol{z}_i ||_2^2 \quad (4)$$

where $\mathbf{X}=\{\boldsymbol{x}_i\}$, $\mathbf{Z}=\{\boldsymbol{z}_i\}$, $i \in [1, m]$ denote all the training and reconstructed data, respectively.

Stacked autoencoder (SAE) is a deep learning architecture of the classical autoencoders, which is built by stacking additional unsupervised feature learning layers, and can be trained using greedy methods for each additional layer. Specifically, once the first layer is trained, the hidden representation of the first layer can be treated as the input of the second layer. As a result, any number of the *K* layers in this deep architecture can be trained effectively. This deep architecture allows SAE to learn more complex mapping from the input to hidden representations and capture the latent patterns which reflects the most homogametic property shared among the training data.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2381471, IEEE Transactions on Circuits and Systems for Video Technology

The stacked denoising autoencoder (SDAE) [25] is an extension of the SAE. It builds a deep architecture by stacking multiple layers of the denoising autoencoder (DAE) which reconstructs the input into a corrupted and partially destroyed version. By introducing stochastic corruption to the training samples, SDAE can avoid over-fitting and achieve better learnt features, where non-trivial features are robust to input noise and useful for the further tasks. For a two-layered SDAE, it is done by first corrupting the initial input $\boldsymbol{x}_i \in \mathbf{X}$ into $\tilde{\boldsymbol{x}}_i$ by using a stochastic mapping $\tilde{\boldsymbol{x}}_i = qD(\tilde{\boldsymbol{x}}_i \mid \boldsymbol{x}_i)$. According to [24, 25], $\tilde{\boldsymbol{x}}_i = qD(\tilde{\boldsymbol{x}}_i \mid \boldsymbol{x}_i)$ is implemented by randomly selecting a fraction (10% in this paper) of the input data and forcing them to be zero. In the bottom layer, corrupted input $\tilde{\boldsymbol{x}}_i$ is then mapped to a hidden representation $\boldsymbol{y}_i^{(1)} = f^{(1)}(\tilde{\boldsymbol{x}}_i, \theta_f^{(1)})$ from which we reconstruct a $\boldsymbol{z}_i^{(1)} = g^{(1)}(\boldsymbol{y}_i^{(1)}, \theta_g^{(1)})$.

Once the bottom layer is trained, the hidden representation of the bottom layer $\boldsymbol{y}_i^{(1)}$ is henceforth used as the input of the second layer $\boldsymbol{x}_i^{(2)}$ to train a new denoising autoencoder as follows:

$$\tilde{\boldsymbol{x}}_i^{(2)} = qD(\tilde{\boldsymbol{x}}_i^{(2)} \mid \boldsymbol{x}_i^{(2)}) \tag{5}$$

$$\boldsymbol{y}_i^{(2)} = f^{(2)}(\tilde{\boldsymbol{x}}_i^{(2)}, \theta_f^{(2)}) \tag{6}$$

$$\boldsymbol{z}_i^{(2)} = g^{(2)}(\tilde{\boldsymbol{y}}_i^{(2)}, \theta_g^{(2)}) \tag{7}$$

Note that SDAE still minimizes the reconstruction loss between a clean input $\mathbf{X}$ and its reconstruction representation $\mathbf{Z}$. It thus forces the learning of a far more clever mapping than the identity, e.g. extracting useful features for denoising [25]. Motivated by the physiological evidence that describing patterns with less active neurons minimizes the probability of destructive cross talk, a regularization term that penalizes a deviation of the expected activation of the hidden units (representation vector) from a fixed (low) level ρ is applied to constrain the sparsity to the target activation function [26]. By taking a single layer autoencoder for example, the target activation function with sparsity constraint can be written as:

$$\arg\min_{\theta_f, \theta_g} L_{sparsity}(\mathbf{X}, \mathbf{Z}, \rho, \hat{\rho}_j) \tag{8}$$

$$L_{sparsity}(\mathbf{X}, \mathbf{Z}, \rho, \hat{\rho}_j) = L(\mathbf{X}, \mathbf{Z}) + \beta \sum_{j=1}^{N} \mathrm{KL}(\rho \| \hat{\rho}_j) \tag{9}$$

$$\mathrm{KL}(\rho \| \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j} \tag{10}$$

where β is the weight of the sparsity penalty, N is the number of features in the weight matrix, ρ is the target average activation of the hidden units, and $\hat{\rho}_j = \sum_{i=1}^{m} [\boldsymbol{y}_j]_i \big/ m$ is the average activation of the jth hidden unit $\boldsymbol{y}_j$ over the m training data. The Kullback-Leibler divergence $\mathrm{KL}(\cdot)$ provides the sparsity constraint. As in sparse coding, a non-redundant over-complete feature set is learned when ρ is small. Here we set ρ=0.05 as suggested in [26]. Usually, training a DAE is straightforward, where the back-propagation algorithm can be used to compute the gradient of the objective function [26, 27], and the same target activation function can be used in all the layers when training SDAE. As the labels of the input data are not needed in the training process above, the layer-wise training step is actually unsupervised.

B. Saliency Detection via Deep Reconstruction Residual

As we mentioned in Section I, local and global contrast-based methods lack the ability to precisely compute the contrast between foreground objects and the background. Inspired by the success of [18], this paper develops the framework along the pipeline of modeling the background and thereby separating salient objects from the background. We follow the basic rule of photographic composition and assume that the image boundary is mostly background. Then, the contrast between salient object and the background can be more precisely obtained. Specifically, we separately define four boundaries for each image as shown in side-specific SDAE training of Fig. 2. The height of two horizontal boundaries is then percent of the image height and their width is the image width. Similarly, the width of two vertical boundaries is then percent of the image width and their height is the image height. To valid the assumption that the image boundary is mostly background, we compute the percentage of foreground pixels (labeled in the ground truth) within the defined image boundaries in two widely used databases (the SOD database [40] and the SED dataset [50]). The statistic result shows that, for most images, only less than 10% of pixels in the image boundary are foreground pixels, which demonstrates that our assumption is reasonable. For the small number of foreground patches, the learning process of SDAE could decrease their influence by minimizing the objective function with the reconstruction error term when modeling the background.

As shown in Fig. 2, the proposed framework mainly consists of three components: multi-scale inputs generation, salient region detection via deep reconstruction residual, and post processing. According to [28, 29], scale is an important factor for identifying objects of different sizes. Similar to [28], we use five scales as $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}\}$ of the original image size to generate multi-scale inputs. It is more sensitive to small objects at the large scale whereas it is more likely to highlight the inner regions of large objects at the small scale.

Afterwards, we model the background using SDAEs described in last subsection and then detect saliency by deep reconstruction residuals for each scale. Specifically, we construct four deep residual maps based on four boundaries (Side-specific deep reconstruction residual maps shown in Fig. 2) and integrate them for the final map, which is referred to as the separation/combination (SC) approach [19]. Specifically, each image boundary is divided into patches of 6×6 pixels with an overlapping of 2 pixels in each direction. Afterwards, we establish the SDAE model with a visible (input) layer with $6\times6\times3$=108 visible units and two hidden layers. According to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2381471, IEEE Transactions on Circuits and Systems for Video Technology

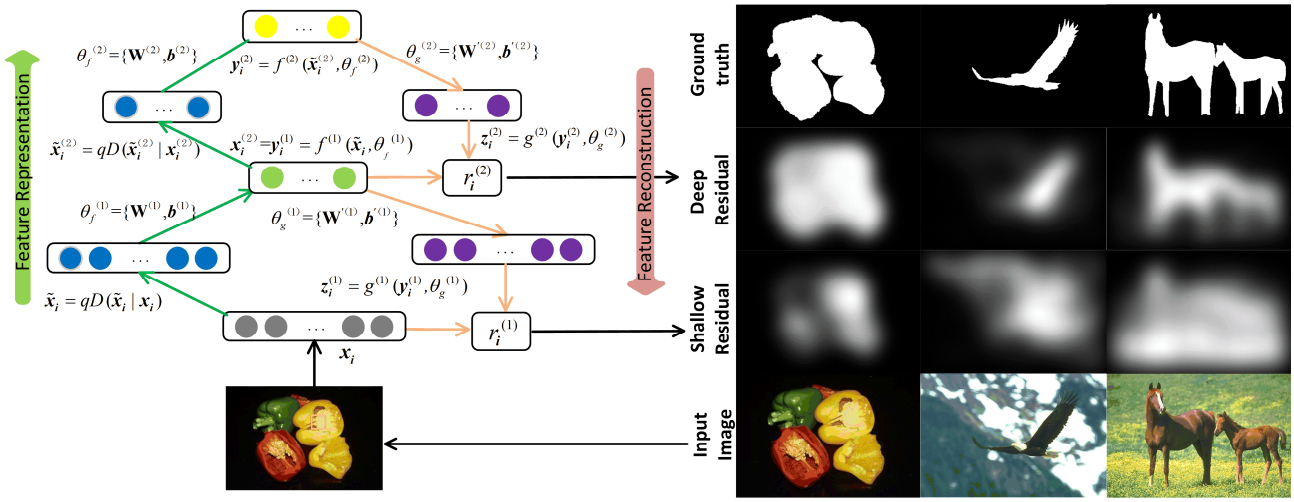

Fig. 3. The process to generate deep reconstruction residual map.

[30], setting the same size for all layers can generally achieve good results. As the number of units in the visible layer is 108, we set each hidden layer to have 100 hidden units, which is approximately equal to the number of units in the visible layer. As pointed out in [31, 32], data preprocessing plays an important role in many deep learning algorithms. In our approach, we perform Zero-phase Component Analysis (ZCA) whitening suggested in [32] to make the input data less redundant. ZCA whitening is implemented by using PCA to make the input vectors to be uncorrelated and then enabling them to have covariance equal to the identity matrix. In the unsupervised training phase, we gradually train the SDAE model layer by layer to learn the feature representation and extract latent patterns for image boundaries by optimizing the objective function in (8-10) for each layer.

Next, we calculate the deep reconstruction residuals for each patch in the image as shown in Fig. 3. Specifically, for each input patch $\boldsymbol{x}_i$, its feature representation $\boldsymbol{y}_i^{(1)}$, $\boldsymbol{y}_i^{(2)}$ and reconstruction vector $\boldsymbol{z}_i^{(1)}$, $\boldsymbol{z}_i^{(2)}$ are obtained by using (5-7) with the trained projection matrixes and biases $\theta_f^{(1)}=\{\mathbf{W}^{(1)},\boldsymbol{b}^{(1)}\}$, $\theta_f^{(2)}=\{\mathbf{W}^{(2)},\boldsymbol{b}^{(2)}\}$, $\theta_g^{(1)}=\{\mathbf{W}'^{(1)},\boldsymbol{b}'^{(1)}\}$, and $\theta_g^{(2)}=\{\mathbf{W}'^{(2)},\boldsymbol{b}'^{(2)}\}$. The deep reconstruction residuals are defined as:

$$r_i^d = r_i^{(2)} = \frac{1}{2}\| \boldsymbol{x}_i^{(2)} - \boldsymbol{z}_i^{(2)} \|_2^2 \tag{11}$$

Here we use the deep reconstruction residual rather than the shallow reconstruction residual, which is generated by only using one layer denoising autoencoder, to measure the saliency. This is because the feature representation in the deep layer captures more intrinsic and latent patterns of the image boundary, which generally leads to more promising saliency detection results in line with human perception (see Fig. 3). Similar to [29], we first assign the patch-level deep residual to each pixel within the corresponding patch and then sum the deep residuals of each pixel assigned from multiple overlapped patches to generate the pixel-level deep residual map.

After normalization, the deep reconstruction residual map R_{top}, R_{bottom}, R_{left}, and R_{right} are obtained based on the SDAE models for the top, bottom, left and right image boundary subsets, respectively. Finally, the four residual maps are linearly combined to generate the saliency map S_R.

$$S_R = (R_{top} + R_{bottom} + R_{left} + R_{right}) / 4 \tag{12}$$

C. Post Processing

As discussed above, we compute saliency map S_R at five different image scales to account for scale changes in salient objects. To integrate salient regions in different scales, we use the average value of the five single scale saliency maps to generate the multi-scale integrated saliency map $\overline{S}_R$. Then this map is normalized to the range of [0, 1]. To further refine the results, two post-processing steps are adopted on the basis of the image organization priors and the region property as presented in details below.

1) Image organization refinement

After obtaining the integrated saliency map $\overline{S}_R$, we observe that although most integrated saliency maps can separate salient regions from the background, there are still some cases where the highlighted regions contain several undesired background regions (such as the branches shown in the first row of Fig. 4) or omit a bit of real foreground regions (such as some parts of the coral shown in the second row of Fig. 4). According to the visual organization rules in [33], these cases can be refined by considering the visual contextual effect. As a result, we propose to use an image organization refinement approach with two components to tackle this problem.

In the first component, as suggested by [34], which states that the salient pixels tend to group together, as they typically correspond to real objects in the scene, we propose to use a self-adaptive threshold $t = mean(\overline{S}_R)$ to obtain the salient cluster firstly. Then the center of the salient cluster G_{cm} is computed, and its location is placed using a Gaussian with σ_c=10000 as suggested in [34] to modify the salient values

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2381471, IEEE Transactions on Circuits and Systems for Video Technology

according to their distance to the salient cluster center.

In the second component, to deal with the case where highlighted regions omit a bit of real foreground, we follow [35] to include the immediate context by weighting the saliency value of each pixel based on their distance to the high salient pixel locations. This is because the context of the dominant objects is as essential as the objects themselves [35]. To encode immediate context information, high salient pixel locations $\Phi = \overline{S}_R > t$ are found and the saliency value at all pixel locations are weighted by their distance to Φ.

Finally, the whole image organization refinement is implemented by integrating these two components via:

$$S_{OR}(p) = \overline{S}_R(p) \times \sum_{y_c \in N_{64}(p,\Phi)} \exp\left(-\frac{(l(p)-l(G_c))^2}{\sigma_l}\right) + \overline{S}_R(p) \times \exp\left(-\frac{(l(p)-l(G_{cm}))^2}{\sigma_c}\right) \quad (13)$$

where $G_c \in N_{64}(p,\Phi)$ is the 64 nearest neighbor of pixel p in Φ, $l(\cdot)$ is the normalized image coordinate of pixels, and S_{OR} is the saliency map after the organization refinement. G_{cm} indicates the center of the salient cluster.

2) Region smoothing

In order to highlight the entire salient object uniformly and recover more edge information, inspired by [35], we refine the saliency of each pixel using the region information. Specifically, a graph based segmentation algorithm [36] is used to decompose the image into a number of small regions and the final saliency of each region is calculated by the average saliency value of all the pixels within it. Examples of region smoothing results are shown in the fifth column of Fig. 4.

III. EXPERIMENTS

To evaluate the performance of the proposed salient object detection framework, we compared it with 9 state-of-the-art approaches, which have been published within last 3 years and in top journals or conferences. These approaches include SVO [37], RC [17], CBS [38], CNTX [33], GS-G [18], GS-S [18], BLSM [39], PD [34] and GBMR [19]. For the work of [18], we used both the grid patch based geodesic saliency (GS-G) and the superpixel based geodesic saliency (GS-S) in our comparison. To obtain the performance of these 9 methods, we adopted either the author-provided implementations or author-provided saliency maps.

Evaluations were constructed on three publicly available benchmark datasets including the ASD dataset [13, 16], the SOD dataset [40] and, the SED dataset [50]. The ASD dataset consists of 1,000 images with manually labeled ground truth. To our best knowledge, this dataset is one of the largest test sets for salient object detection whose ground truth is in the form of manually labeled accurate object contours instead of rough bounding boxes. The SOD dataset consists of 300 images, which generally contain complex background and multiple salient objects with vague appearance. Some images contain foreground objects with very similar color to the background,

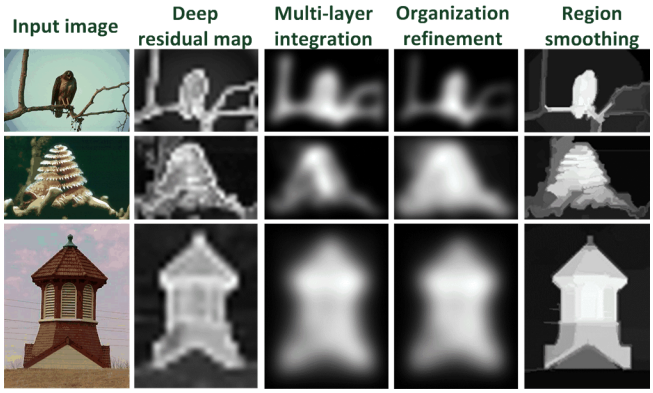

Fig. 4. Experimental results of some examples after each step in the proposed method.

which makes it difficult to be precisely separated. For many images, even the ground truth annotated by multiple subjects shows inconsistency. Consequently, it has been regarded as the most challenging dataset for salient object detection by a recent survey paper [41]. Another challenging dataset is the SED dataset, which contains 100 images with one salient object and complex background and another 100 images with two salient objects. The SED dataset also provides accurate human-labelled ground truth for each image. These three benchmark datasets have been widely utilized by a variety of saliency detection approaches for performance evaluation.

Fig. 5 illustrates a number of saliency maps yielded by using the proposed method and the 9 state-of-the-art algorithms. The subjective evaluations by comparing with the ground truth suggest that the proposed method can yield saliency maps correctly and robustly in all three datasets. It can be observed that, compared with PD, GBMR, GS-S, GS-G, BLSM, and CNTX, the proposed method can highlight salient region more uniformly. Compared with SVO and RC, our approach can achieve higher distinctness between foreground regions and background regions. In the next subsections, further comprehensive experiments are designed for both parameter analysis and quantitative performance assessment.

A. Evaluation Metrics

By following previous works of [9, 12, 15, 16, 34, 41-43], four metrics are adopted in our experiments to quantitatively measure the performance of saliency map, which include the receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), precision recall (PR) curve, and the average precision (AP). ROC and AUC are generated by classifying the pixels in a saliency map into salience or non-salience by varying the quantization threshold within the range [0, 255]. The resulting false positive rate versus true positive rate at each threshold value forms the ROC curve. Similarly, PR and AP are generated using the precision rate and the true positive rate (or the recall rate). The precision PRE, true positive rate TPR and false positive rate FPR values are respectively defined by

$$PRE = \frac{|SF \cap GF|}{|SF|} \quad TPR = \frac{|SF \cap GF|}{|GF|} \quad FPR = \frac{|SF \cap GB|}{|GB|} \quad (14)$$

where SF, GF and GB denote the set of segmented

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2381471, IEEE Transactions on Circuits and Systems for Video Technology

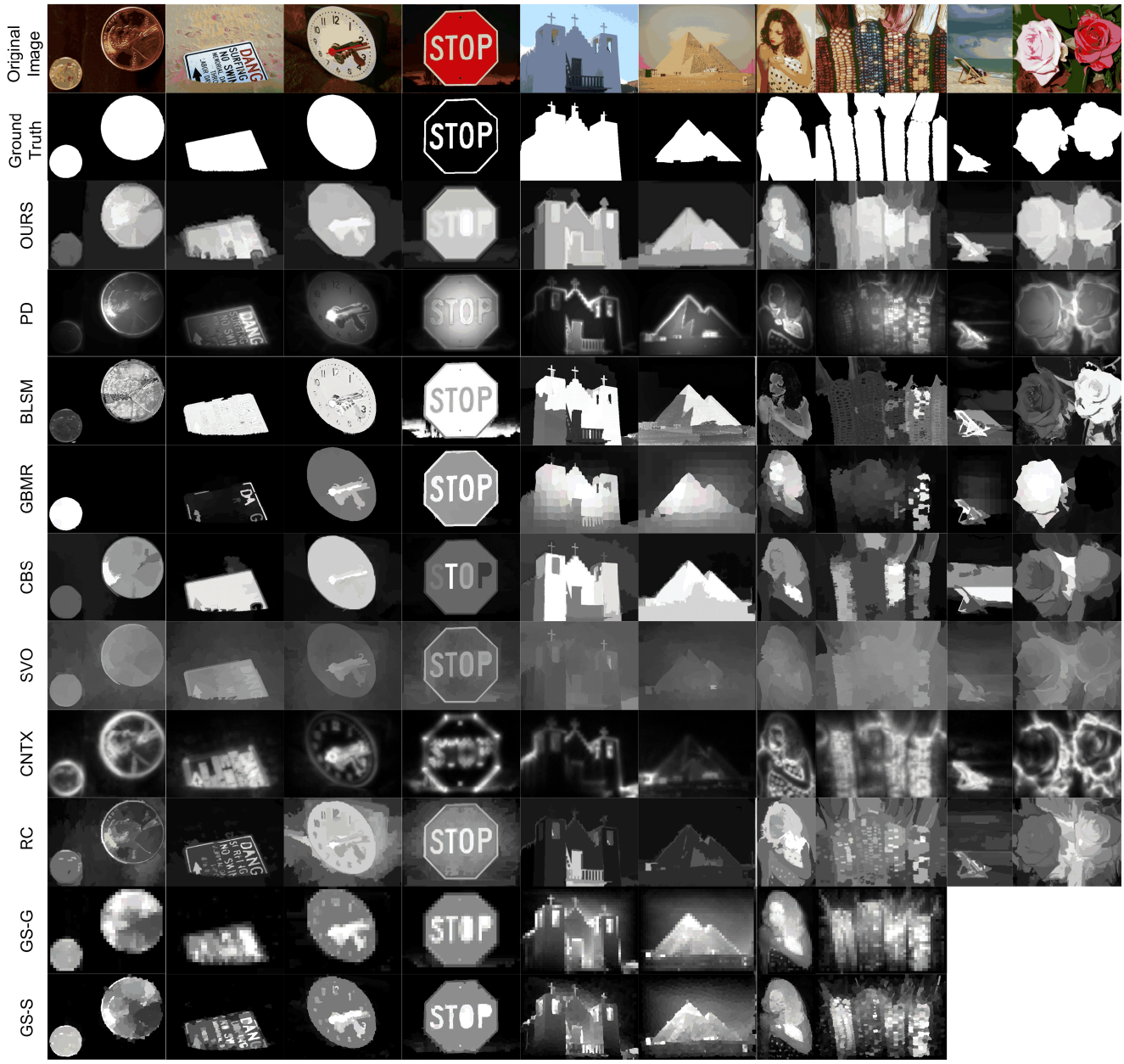

Fig. 5. A number of comparison results of ours, 9 state-of-the-art approaches, and the ground truth. From the left to the right, the first four examples are from the ASD dataset, the middle four examples are from the SOD dataset, and the last two examples are from the SED dataset.

foreground pixels after a binary segmentation using a certain threshold, the set of ground truth foreground pixels and the set of ground truth background pixels, respectively.

To further evaluate the performance of the proposed method for salient object segmentation, we report the performance in segmenting the saliency map using a self-adaptive threshold. Observing the Gaussian-like distributions of the saliency value in the proposed saliency maps, an adaptive threshold $T = \mu + \sigma$ as suggested in [44] is used to segment the saliency maps. Here, μ and σ are the mean saliency value and the standard deviation of the saliency map, respectively. For each segmented foreground binary map SF_T under the adaptive threshold T, we follow [51] to evaluate it by using the weighted F-measure. $E = |G - SF_T|$ denotes the absolute error of detection, where G is the column-stack representation of the binary ground truth. In order to take into consideration both the dependency between pixels and the location of the errors, a weighting function is applied to the errors as $E^w = min(E, E\,\mathbf{A}) \cdot \mathbf{B}$. As defined in [51], the matrix $\mathbf{A}$ captures the dependency between foreground pixels based on the Euclidean distance and the matrix $\mathbf{B}$ assigns importance weights to false detections according to their distance from the foreground. Then, the weighted true positive TP^w, the weighted false positive FP^w and the weighted false negative FN^w can be calculated by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2381471, IEEE Transactions on Circuits and Systems for Video Technology

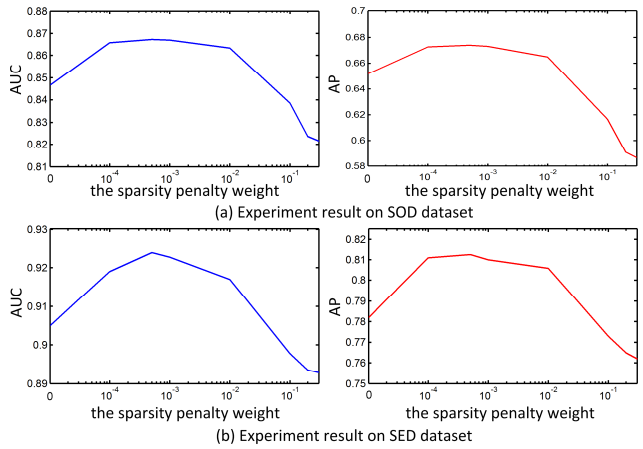

Fig. 6. AUCs and APs with different sparsity penalty weights.

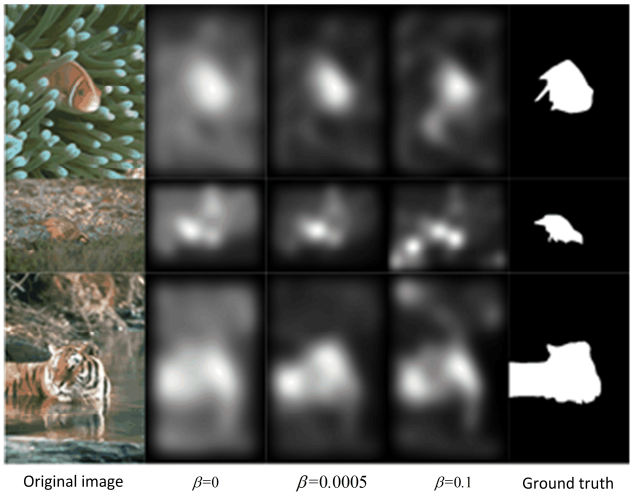

Fig. 7. Some experimental results obtained under different parameter setting.

$$TP^w = (1 - E^w) \cdot G \tag{15}$$

$$FP^w = E^w \cdot (1 - G) \tag{16}$$

$$FN^w = E^w \cdot G \tag{17}$$

Finally, the weighted precision PRE^w, the weighted recall REC^w, and the weighted F-measure F_α^w for the segmented foreground binary map can be obtained by:

$$PRE^w = \frac{TP^w}{TP^w + FP^w} \quad REC^w = \frac{TP^w}{TP^w + FN^w} \tag{18}$$

$$F_\alpha^w = (1+\alpha^2)\frac{PRE^w \cdot REC^w}{\alpha^2 \cdot PRE^w + REC^w} \tag{19}$$

where α is set to 1 as suggested in [45-47] to balance the determined precision and recall measures.

B. *Parameters Analysis and Model Evaluation*

In this section, we analyze the effect of a few key parameters in the proposed model on performance. Here we conducted the evaluation on the SOD and SED datasets. In SDAE, the weight of the sparsity penalty β in (9) is a parameter to balance squares error term and the sparsity penalty term. Essentially, the tradeoff parameter β has a notable influence on the saliency detection performance. In this paper, we empirically generated the saliency map using the proposed approach by varying β between 0 and 0.3. Fig. 6 illustrates AUCs and APs with different values of β. As can be seen, the proposed algorithm is reasonably sensitive to β and SDAE can work well under a range of parameter settings from 0.0001 to 0.001. In all subsequent experiments, β was fixed at 0.0005. Some examples of the experimental results obtained under different β are also given in Fig. 7. From the second and the third column of Fig. 7, we can see that for the images with clustered background, the sparsity is an essential element for suppressing the saliency of the background regions. However, if the sparsity constraint is set too big, it normally leads to less stable and discontinuous detection results (as shown in the forth column of Fig. 7). Similar phenomenon is also discovered in [48, 49].

TABLE I
DEMONSTRATION OF THE EFFECTIVENESS OF THE KL DIVERGENCE

		100 hidden nodes with the KL divergence term (β=0.0005)	100 hidden nodes without the KL divergence term	50 hidden nodes without the KL divergence term
SOD	AUC	0.8673	0.8489	0.8262
	AP	0.6738	0.6515	0.6001
SED	AUC	0.9240	0.9042	0.8895
	AP	0.8128	0.7818	0.7538

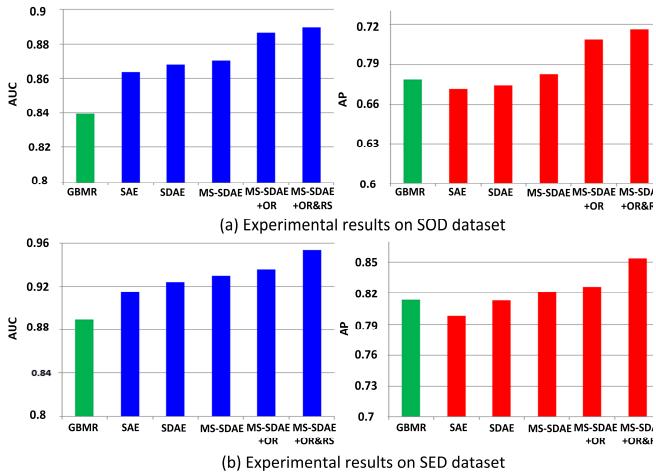

Fig. 8. AUCs and APs of different models.

To demonstrate the effectiveness of the KL divergence used in the sparsity constraint, we also compared our SDAE model with the SDAE models without the KL divergence term and using less hidden nodes. Experimental results are shown in Table I. As can be seen from the results, the use of KL divergence can improve the model performance. It should be mentioned that in above experiments (results shown in Fig. 6 and Table I), we only used the SDAE model without multi-scale inputs and post processing steps to evaluate the effect of the sparsity constraint clearly.

Besides sparsity, the denoising criterion, multi-scale inputs, and post processing (i.e. the image organization refinement (OR) and region smoothing (RS)) are three other critical factors in the proposed framework. In order to show the effect of these factors on performance, we compared the proposed approach

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2381471, IEEE Transactions on Circuits and Systems for Video Technology

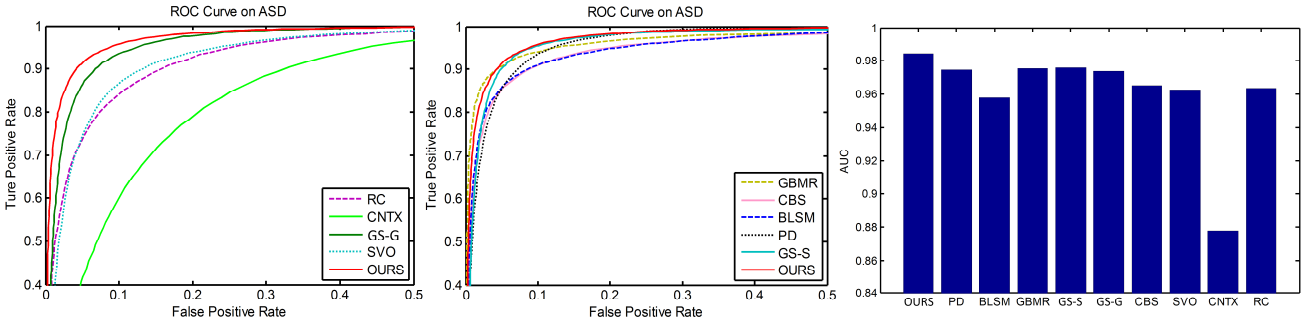

Fig. 9. The ROC curves and AUC scores for the proposed method and 9 state-of-the-art methods on the ASD dataset.

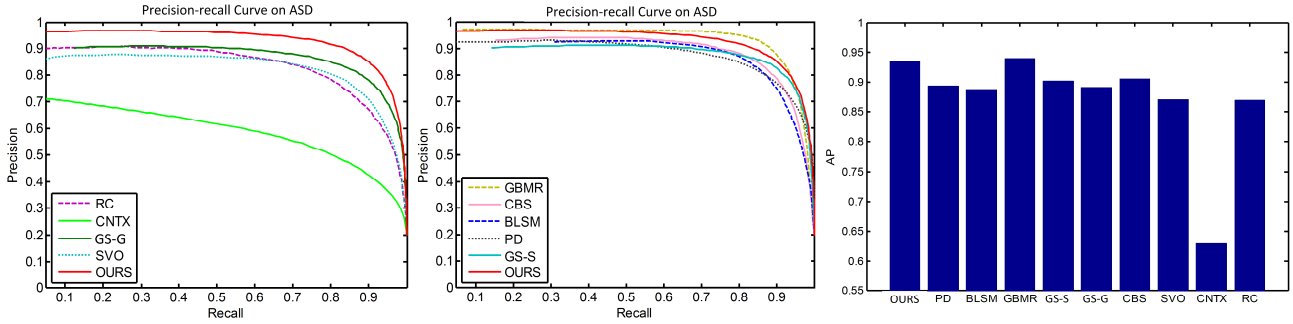

Fig. 10. The PR curves and AP scores for the proposed method and 9 state-of-the-art methods on the ASD dataset.

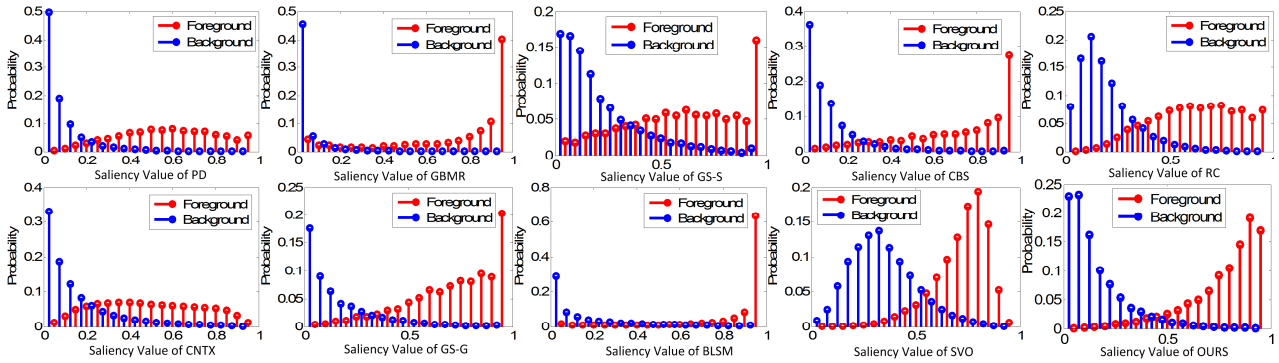

Fig. 12. Saliency value distribution of all foreground and background pixels in the ASD dataset.

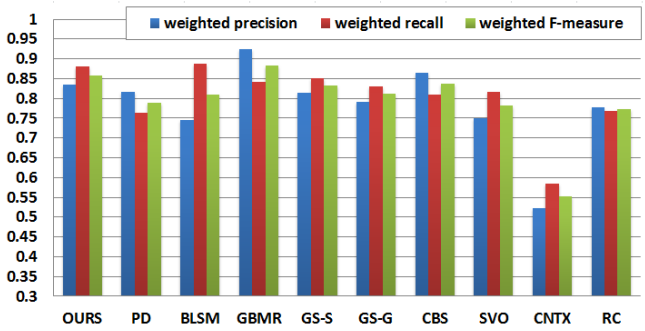

Fig. 11. The results of weighted F-measure for the proposed method and 9 state-of-the-art methods on the ASD dataset.

(indicated by MS-SDAE+OR&RS in Fig. 8) with the method using SAE which does not corrupt the input data and ignores the scale problem (indicated by SAE in Fig. 8), the method using SDAE with single scale input which ignores the scale problem (indicated by SDAE in Fig. 8), and the method without using two post-processing steps (indicated by MS-SDAE in Fig. 8), and the method with using only the first post-processing step (indicated by MS-SDAE+OR in Fig. 8). As GBMR [18] has the excellent performance on salient object detection and is one of the state-of-the-art methods, we treat it as the baseline method in this group of experiments. As can be seen in the results given in Fig. 8, with the denoising criterion, the obtained SDAE can model the background more robustly and achieve higher scores compared with the SAE. In addition, the development of SDAE on multi-scale inputs can further improve the performance. It is worth mentioning that key techniques involved in our two post-processing steps have often been adopted in other state-of-the-art saliency detection approaches such as [33, 35, 18] although they may not be explicitly called as post-processing steps. Based on the comparison results, two points are observed. First, post processing used in our proposed approach can improve the performance of saliency detection in terms of AUC and AP scores. Second, even without using any post-processing steps, the proposed deep learning based method (MS-SDAE) still outperforms the baseline method of GBMR in both the SOD and the SED datasets.

C. *Evaluations on the ASD Dataset*

We conducted quantitative comparisons on the ASD dataset

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2381471, IEEE Transactions on Circuits and Systems for Video Technology

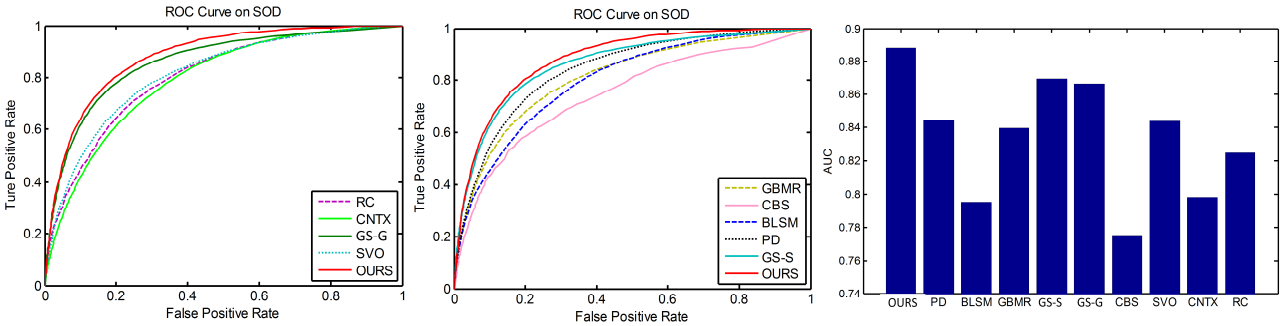

Fig.13. The ROC curves and AUC scores for the proposed method and 9 state-of-the-art methods on the SOD dataset.

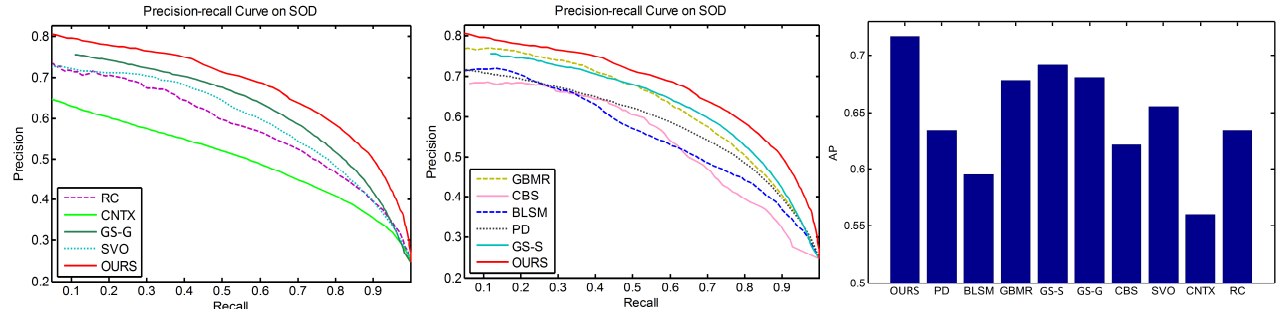

Fig. 14. The PR curves and AP scores for the proposed method and 9 state-of-the-art methods on the SOD dataset.

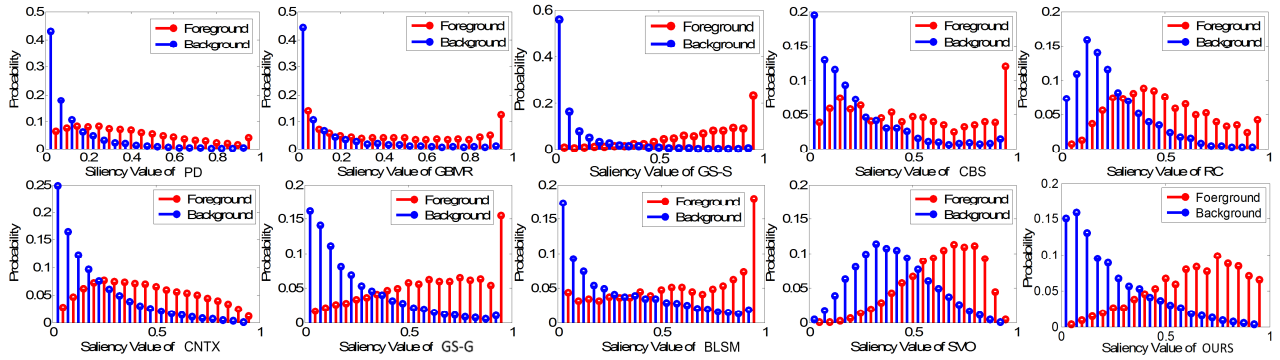

Fig. 16. Saliency value distribution of all foreground and background pixels in the SOD dataset.

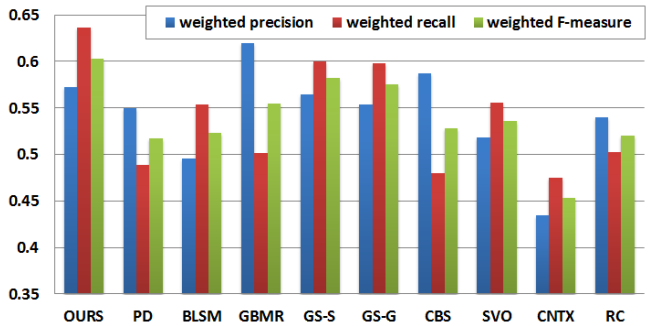

Fig. 15. The results of weighted F-measure for the proposed method and 9 state-of-the-art methods on the SOD dataset.

using ROC, PR, AUC, AP, and weighted F-measure as the performance metric. Fig. 9 shows the comparison results of the ROC curve and AUC score. From the ROC results, we can see that the proposed method achieves the highest true positive rate when the false positive rate is between about 0.05 and 1. As a result, the proposed method outperforms other 9 algorithms in terms of ROC and AUC. According to the PR curves in Fig. 10, the proposed method performs the best at a recall rate among [0, 0.6] and [0.9, 1] whereas its precision is lower than GBMR when the recall rate is among [0.6, 0.9], which leads to the slightly lower AP value than GBMR. However, the performance of the proposed method is still better than other 8 algorithms in terms of PR and AP. Fig. 11 shows the comparisons of the weighted F-measure when segmenting salient objects. It is seen that the proposed method can achieve the second best performance for salient object segmentation.

To further analyze how the detection result of the proposed method differs from other methods, we followed [18] to show in Fig. 12 the saliency value distributions of all foreground and background pixels in the ASD dataset for each method by using the ground truth masks. The statistics results can reflect the distributions of the true salient pixels and true background pixels on the calculated saliency value. Ideally, it should be a clear boundary or a quite small overlap between the foreground distribution and background distribution. As shown in Fig. 12, the proposed method can achieve a satisfactory performance, while the approximately ideal distributions of GBMR explains its highest AP score in the PR curve.

D. Evaluations on the SOD Dataset

We also conducted the comparisons on the more challenging

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2381471, IEEE Transactions on Circuits and Systems for Video Technology

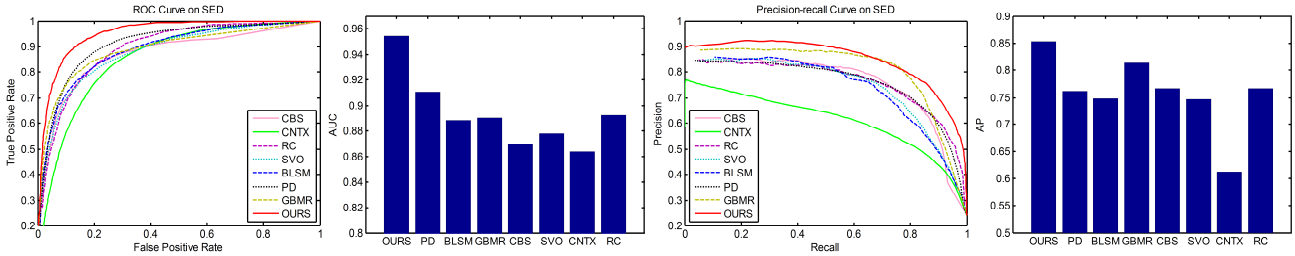

Fig. 17. The ROC curves, AUC scores, PR curves, and AP scores for the proposed method and 7 state-of-the-art methods on the SED dataset.

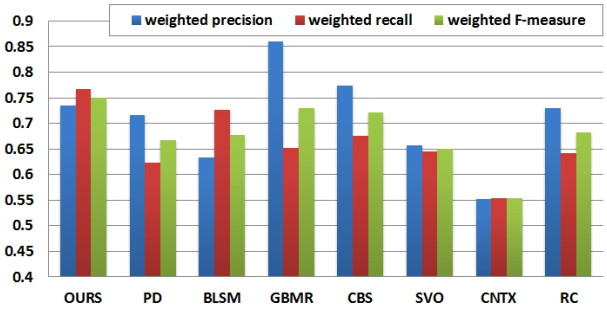

Fig. 18. The results of weighted F-measure for the proposed method and 7 state-of-the-art methods on the SED dataset.

SOD database. All the comparison results, including ROC, AUC, PR, AP, and weighted F-measure, are shown in Figs. 13-15. As can be seen, the performance of the proposed method outperforms other 9 state-of-the-art algorithms in terms of all metrics. Specifically, the proposed method has achieved higher true positive rate on the whole ROC curve, and the precision values on the whole PR curve of the proposed method are almost higher than all other state-of-the-art approaches as well. From Fig. 15, it is observed that the proposed approach can achieve the highest weighted F-measure score. In addition, it also shows that the weighted recall values of most of the state-of-the-art are less than 0.6 whereas the proposed approach can achieve the highest weighted recall value that is around 0.64, which indicates the proposed method tends to highlight the entire salient objects. It is also worth noting that because a large number of images in the SOD dataset contain complicated content and multiple salient objects, many excellent approaches, such as GS-G, GS-S, and GBMR, cannot work effectively in this dataset though they can achieve promising performance on the ASD dataset. On the contrary, the proposed method has the capability to yield consistently satisfactory results on both of the datasets, especially on the more challenging SOD dataset.

For the foreground distribution and background distribution, similar observations can be found in comparison of results obtained from different approaches. As shown in Fig. 16, the distributions on the SOD dataset tend to worse obviously. Only the GS (including GS-S and GS-G) and the proposed method have the relatively clear separation between the two distributions. This may explain why the proposed method can achieve the highest AUC and AP score on the SOD dataset.

E. Evaluations on the SED dataset

The proposed approach was also tested on the SED database, another challenging dataset. As GS-S and GS-G have not provided their codes and their results on this dataset, we are unable to compare with these two algorithms. To this end, we compared the proposed approach with the remaining 7 state-of-the-art methods in this group of experiments. All the comparison results, including ROC, AUC, PR, AP, and the weighted F-measure, are shown in Figs. 17 and 18. As can be seen, the performance of the proposed method outperforms other state-of-the-art algorithms in terms of all metrics. More encouragingly, compared with other state-of-the-art algorithms, the proposed method has achieved the higher true positive rate in the whole ROC curve, and the higher precision values along almost the whole PR curve as well. From Fig. 18, it is observed that the proposed approach can also achieve the highest weighted F-measure value compared with other 7 state-of-the-art methods. Similar to the SOD dataset, SED dataset also contains a large number of images with complicated content and multiple salient objects. The experimental results show that the proposed algorithm has more powerful capability to handle these tough scenarios.

TABLE II
COMPARISON OF AVERAGE EXECUTION TIME (SECONDS PER IMAGE)

Method	OURS	PD	BLSM	GBMR
Time(s)	0.82	4.15	98.37	0.54
Method	CBS	SVO	CNTX	RC
Time(s)	2.68	72.11	28.72	0.046

F. Running time

Table II lists the average execution time in processing an image of size 400×300 by using different approaches. All experiments were run on a 24-core Lenovo Server with Intel Xeon CPU of 2.8 GHz and 64 GB RAM. For the implementation of the proposed method, we used the parallel computing toolbox of MATLAB and executed the code on the NVIDIA GPU named GeForce GTX Titan Black. For other state-of-the-art approaches, we used the source codes provided by their authors. We did not compare with GS-G and GS-S because the corresponding codes have not been released by the authors. As can be seen, the proposed algorithm has the moderate computational complexity. In the proposed algorithm, the training of background model spent most of the time. We can speed up this training algorithm by improved weight quantization schemes, optimization algorithms or initialization strategies, which will be one of our future research issues.

IV. CONCLUSION

In this paper, we have proposed a bottom-up salient object detection framework based on the background prior. The novelty that can discriminate the proposed work from existing

approaches is twofold. First, instead of using traditional hand-designed features, the proposed algorithm adopted SDAE with deep structures to learn more powerful representations for saliency computation. Second, the proposed work casted separation of salient objects from the background as a problem of calculating reconstruction residual of SDAE. Comprehensive experiments on three publicly available benchmarks have demonstrated the effectiveness of the proposed work. To the best of our knowledge, this work might be among the earliest efforts to explore the feasibility of deep learning for salient object detection.

For the further work, we tend to extend the proposed work in the following directions. First, we improve the proposed work by combining a number of top-down cues. Second, the proposed method can be extended to saliency detection in dynamic videos and many other applications such as image retrieval, image categorization, and image collection visualization.

REFERENCES

[1] J. Sun, and H. Ling, "Scale and object aware image retargeting for thumbnail browsing," in *Proc. IEEE Int. Conf. Comput. Vision*, 2011. pp. 1511-1518.

[2] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2Photo: internet image montage," *ACM Trans. Graph*., Vol. 28. No. 5. pp. 124, Dec. 2009.

[3] X. Zhen, L. Shao and X. Li, "Action Recognition by Spatio-Temporal Oriented Energies," *Inform. Science*, vol. 281, pp. 295-309, Oct. 2014.

[4] F. Zhu and L. Shao, "Weakly-Supervised Cross-Domain Dictionary Learning for Visual Recognition," *Int. J. Comput. Vis.*, vol. 109, no. 1-2, pp. 42-59, Aug. 2014.

[5] 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.

[6] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vision*, 2009. pp. 2106-2113.

[7] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 438-445.

[8] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vision,* vol. 8, no. 7, pp. 32, Dec. 2008.

[9] J. Han, S. He, X. Qian, D. Wang, L. Guo, and T. Liu, "An object-oriented visual saliency detection framework based on sparse coding representations," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 23, no. 12, pp.2009 -2021, Dec. 2013.

[10] X. Hou, and L. Zhang, "Dynamic visual attention: searching for coding length increments," in *Proc. Conf. Adv. Neural Inform. Process. Syst.*, 2008. pp. 7.

[11] X. Shen, and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 853-860.

[12] B. Han, H. Zhu, and Y. Ding, "Bottom-up saliency based on weighted sparse coding residual," in *Proc. ACM Int. Conf. Multimedia*, 2011. pp. 1117-1120.

[13] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 33, no. 2, pp. 353-367, Feb. 2011.

[14] H. Seo, and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vision,* vol. 9, no. 12, Nov. 2009.

[15] X. Hou, and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2007, pp. 1-8.

[16] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 1597-1604.

[17] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2011, pp. 409-416.

[18] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 29-42.

[19] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency Detection via Graph-Based Manifold Ranking," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 3166-3173.

[20] Y. Bengio, and Y. LeCun, "Scaling learning algorithms towards AI," *Large-Scale Kernel Machines,* L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, eds., MIT Press, 2007.

[21] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?," *J. Mach. Learn. Res*., vol. 11, pp. 625-660, March 2010.

[22] L. Shao, D. Wu and X. Li. Learning Deep and Wide: A Spectral Method for Learning Deep Networks. *IEEE Trans. Neural Netw. Learn. Syst.*, Online published, 2014.

[23] G. Hinton, and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science,* vol. 313, no. 5786, pp. 504-507, July 2006.

[24] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, pp. 1798–1828, Aug. 2013.

[25] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.,* vol. 9999, pp. 3371-3408, Jan. 2010.

[26] A. Ng, "Sparse autoencoder," *CS294A Lecture notes*, Stanford University, pp. 72, 2011.

[27] H. Shin, M. Orton, D. Collins, S. Doran, and M. Leach, "Stacked Autoencoders for Unsupervised Feature Learning and Multiple Organ Detection in a Pilot Study Using 4D Patient Data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1930-1943, Aug. 2013.

[28] W. Kim, and C. Kim, "Spatiotemporal Saliency Detection Using Textural Contrast and Its Applications," *IEEE Trans. Circuits Syst. Video Technol.,*vo. 24, no.4, pp. 646 - 659, April 2013.

[29] X. Qian, J. Han, G. Cheng, and L. Guo, "Optimal contrast based saliency detection," *Pattern Recognit. Lett.*, vol. 34, no. 11, pp. 1270-1278, Aug. 2013.

[30] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," *Neural Networks: Tricks of the Trade*, K.-R. Mu¨ller, G. Montavon, and G.B. Orr, eds., Springer 2013.

[31] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *Proc. IEEE Int. Conf. Machine Learning*, 2013.

[32] N. Wang, J. Melchior, and L. Wiskott, "An analysis of Gaussian-binary restricted Boltzmann machines for natural images," in *Proc. ESANN*, 2012. pp. 287-292.

[33] 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.

[34] R. Margolin, A. Tal, and L. Zelnik-Manor, "What Makes a Patch Distinct?," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 1139-1146.

[35] P. Siva, C. Russell, T. Xiang, and L. Agapito, "Looking Beyond the Image: Unsupervised Learning for Object Saliency and Detection," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 3238-3245.

[36] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.,* vol. 59, no. 2, pp. 167-181, Sept. 2004.

[37] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2011, pp. 914-921.

[38] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior," in *Proc. Brit. Machine Vision Conf.*, 2011. pp. 7.

[39] Y. Xie, H. Lu, and M. Yang, "Bayesian saliency via low and mid level cues," *IEEE Trans. Image Process.,* vol. 22, no. 5, pp. 1689-1698, May 2013.

[40] V. Movahedi, and J. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *Proc. IEEE Comput. Soc. Workshop Perceptual Org. Comput. Vis*, 2010, pp. 49-56.

[41] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," *in Proc. Eur. Conf. Comput. Vision*, 2012, pp. 414-429.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2014.2381471, IEEE Transactions on Circuits and Systems for Video Technology

[42] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Conf. Adv. Neural Inform. Process. Syst.*, 2006. pp. 545-552.
[43] X. Hou, and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. Conf. Adv. Neural Inform. Process. Syst.*, 2008. pp. 681-688.
[44] Y. Jia, and M. Han, "Category-Independent Object-level Saliency Detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013.
[45] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Trans. Syst. Man Cybern. Part B-Cybern.,* vol. 43, no. 2, pp. 660-672, April 2013.
[46] D. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 26, no. 5, pp. 530-549, May 2004.
[47] L. Sun, and T. Shibata, "Unsupervised Object Extraction by Contour Delineation and Texture Discrimination Based on Oriented Edge Features," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 24, no. 5, pp. 780-788, May 2014.
[48] C. Lang, G. Liu, J. Yu, and S. Yan, "Saliency detection by multitask sparsity pursuit," *IEEE Trans. Image Process.,* vol. 21, no. 3, pp. 1327-1338, March 2012.
[49] Y. Li, Y. Zhou, L. Xu, X. Yang, and J. Yang, "Incremental sparse saliency detection," in *Proc. IEEE Int. Conf. Image Process.,* 2009, pp. 3093-3096.
[50] S. Alpert, M. Galun, R. Basri, A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, 2007, pp. 1-8.
[51] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to Evaluate Foreground Maps?" in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.,* 2014.

Junwei Han received his Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2003. He is currently a professor with Northwestern Polytechnical University. His research interests include computer vision and multimedia processing.

Dingwen Zhang received his M.S. degree from the Northwestern Polytechnical University, China, in 2011. He is currently pursuing the Ph.D. degree at Northwestern Polytechnical University. His research interests include computer vision and multimedia processing.

Xintao Hu received his M.S. and Ph.D degrees from the Northwestern Polytechnical University, China, in 2005 and 2011, respectively. He is currently a postdoc with School of Automation at NWPU. His research interests include computational brain imaging and its application in computer vision.

Lei Guo received his Ph.D. degree from Xidian University, Xi'an, China, in 1994. He is currently a professor in Northwestern Polytechnical University, China. His research interests include computer vision, pattern recognition, and medical image processing.

Jinchang Ren received his B. E., MEng, DEng degrees all from Northwestern Polytechnical University, Xi'an, China. He was also awarded a PhD in Electronic Imaging and Media Communication from Bradford University, U.K. Currently he is with University of Strathclyde, U.K. His research interests focus on visual computing and multimedia processing, especially on semantic content extraction for video analysis and understanding and more recently hyperspectral imaging.

Feng Wu (M'99-SM'06-F'13) received the B.S. degree in Electrical Engineering from XIDIAN University in 1992. He received the M.S. and Ph.D. degrees in Computer Science from Harbin Institute of Technology in 1996 and 1999, respectively. Now he is a professor in University of Science and Technology of China. Before that, he was principle researcher and research manager with Microsoft Research Asia. His research interests include image and video compression, media communication, and media analysis and synthesis. He has authored or co-authored over 200 high quality papers (including several dozens of IEEE Transaction papers) and top conference papers on MOBICOM, SIGIR, CVPR and ACM MM. He has 77 granted US patents. His 15 techniques have been adopted into international video coding standards. As a co-author, he got the best paper award in IEEE T-CSVT 2009, PCM 2008 and SPIE VCIP 2007. Wu has been a Fellow of IEEE. He serves as an associate editor in IEEE Transactions on Circuits and System for Video Technology, IEEE Transactions on Multimedia and several other International journals. He got IEEE Circuits and Systems Society 2012 Best Associate Editor Award. He also serves as TPC chair in MMSP 2011, VCIP 2010 and PCM 2009, and Special sessions chair in ICME 2010 and ISCAS 2013.