Cosaliency Detection Based on Intrasaliency Prior Transfer and Deep Intersaliency Mining

Dingwen Zhang, Junwei Han, Jungong Han, and Ling Shao, *Senior Member, IEEE*

***Abstract*— As an interesting and emerging topic, cosaliency detection aims at simultaneously extracting common salient objects in multiple related images. It differs from the conventional saliency detection paradigm in which saliency detection for each image is determined one by one independently without taking advantage of the homogeneity in the data pool of multiple related images. In this paper, we propose a novel cosaliency detection approach using deep learning models. Two new concepts, called intrasaliency prior transfer and deep intersaliency mining, are introduced and explored in the proposed work. For the intrasaliency prior transfer, we build a stacked denoising autoencoder (SDAE) to learn the saliency prior knowledge from auxiliary annotated data sets and then transfer the learned knowledge to estimate the intrasaliency for each image in cosaliency data sets. For the deep intersaliency mining, we formulate it by using the deep reconstruction residual obtained in the highest hidden layer of a self-trained SDAE. The obtained deep intersaliency can extract more intrinsic and general hidden patterns to discover the homogeneity of cosalient objects in terms of some higher level concepts. Finally, the cosaliency maps are generated by weighted integration of the proposed intrasaliency prior, deep intersaliency, and traditional shallow intersaliency. Comprehensive experiments over diverse publicly available benchmark data sets demonstrate consistent performance gains of the proposed method over the state-of-the-art cosaliency detection methods.**

***Index Terms*— Cosaliency detection, deep learning, prior transfer, stacked denoising autoencoder (SDAE).**

I. Introduction

SALIENCY detection has been an extensively studied topic in the past few decades. It enables a computer vision system to select a subset of interesting regions in each input image for further processing and analysis [1]–[3], [46]–[48]. More recently, the growing popularity of photosharing websites, such as Flickr and Facebook, has taught us that people love taking photographs, and there is a rich collection of related pictures sharing the common foreground regions of the same object or event [4]. When detecting these cosalient foregrounds, the direct use of conventional saliency detection methods that process each of these images individually may lead to unsatisfactory performance (see the second row of Fig. 1). This, thus, triggers a new and interesting research area named cosaliency detection with the goal of discovering the consistent salient patterns in multiple related images and, finally, extracting the common salient foreground regions in the image group (see the third row of Fig. 1). Different from cosegmentation [4]–[6] that considers not only common salient foreground regions but also similar nonsalient background areas in images, cosaliency detection focuses on exploring the most important information, i.e., the common foreground regions, among the image group with a reduced computational demand by implying priorities based on human visual attention. Cosaliency detection can serve as a more promising preprocessing step for many high-level visual information understanding tasks, such as video foreground extraction [7], image retrieval [8], object detection [9], [52], and image matching [10].

Fig. 1. Illustration of the difference between conventional saliency detection and cosaliency detection. First row: input images. Second row: saliency detection results obtained by region-based contrast [3]. Third row: cosaliency detection results obtained by the proposed approach.

As shown in [11], cosalient image regions usually have two properties: 1) they should be prominent or noticeable regions with respect to the background in each image and 2) high homogeneity should be observed for such regions across multiple related images. To explore the first property, some earlier cosaliency models proposed in [12] and [13] directly combine several existing saliency detection methods for predicting the salient regions within each single image. For obtaining better performance, Fu *et al.* [14] and Liu *et al.* [15] proposed novel algorithms for intrasaliency prediction by

Manuscript received November 10, 2014; revised October 10, 2015; accepted October 22, 2015. Date of publication November 11, 2015; date of current version May 16, 2016. This work was supported in part by the National Science Foundation of China under Grant 61473231 and Grant 61522207 and in part by the Doctorate Foundation through Northwestern Polytechnical University. *(Corresponding author: Junwei Han.)*

D. Zhang and J. Han are with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: zhangdingwen2006yyy@gmail.com; junweihan2010@gmail.com).

J. Han is with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne NE1 8ST, U.K. (e-mail: jungong.han@northumbria.ac.uk).

L. Shao is with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne NE1 8ST, U.K., and also with the College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: ling.shao@ieee.org).

modifying the existing unsupervised saliency detection models. To explore the second property, most previous approaches discover the homogeneity of cosalient regions within each image pair [12], [16]–[19]. To extend beyond pairwise relations, Fu *et al.* [14] employed CIE Lab color and Gabor filter to represent each pixel, and extracted contrast cue, spatial cue, and correspondence cue from the image group for generating the cluster level cosaliency maps. Liu *et al.* [15] proposed to derive the global similarity measures of image regions over the image set based on the quantized color features.

As can be seen, corresponding to the above two properties, the key problems in cosaliency detection lie in two aspects: 1) predicting the saliency of image regions within each single image, i.e., the intrasaliency, robustly and 2) developing an optimal mechanism to explore the homogeneity of cosalient objects, i.e., the intersaliency, among multiple related images. For the first problem, most existing approaches only directly apply or manually modify the previous unsupervised saliency detection algorithms for a single image to cosaliency detection. However, they cannot yield promising results as unsupervised saliency detection algorithms tend to lack robustness and be influenced by the complex backgrounds. In addition, the recent progress of saliency detection in a single image has acquired more prior knowledge on saliency. Knowledge transfer from single saliency detection will be certainly beneficial to the intrasaliency in cosaliency detection. For the second problem, the existing approaches mainly focus on exploring the homogeneity based on the low-level features, such as color, texture, or corner descriptors. In this paper, we call it shallow intersaliency, because they only formulate the homogeneity of the low-level visual stimulus, while the homogeneity in deeper insights into higher level concepts could not be captured. In addition, low-level features are easily influenced by the variation in luminance, shape, or viewpoint, leading to unsatisfactory performance of cosaliency detection.

In order to tackle these problems and further improve the performance of cosaliency detection, we adopt deep learning models in this paper for better solving the problems in both the generation of the robust intrasaliency prior and the discovery of the intersaliency patterns. Instead of using humans as a transfer machine, where researchers learn the knowledge of how to formulate saliency from the conventional unsupervised saliency detection approaches and, then, manually modify these approaches for predicting intrasaliency, inspired by the studies in [20] and [21], we propose an alternative framework to design a real transfer machine that can learn the saliency prior knowledge from the auxiliary annotated data sets automatically and, then, transfer the learned knowledge to predict the intrasaliency for each image in cosaliency data sets. As we know, saliency is an abstract concept that relates to the contrast between the certain image regions and the image backgrounds, as well as the content within the image regions. This relationship holds true regardless of the object category. Thus, according to [22], this kind of an abstract concept is more likely to be suitable for transfer learning. In addition, the training data in cosaliency data sets appears to be limited (about 17 images per group). When the labeled training data are scarce, transfer learning of the relevant knowledge from the auxiliary data sets would yield a significant performance improvement [23], [24]. In order to capture saliency prior from the data in the source domain and transfer it to predict the intrasaliency for the data in the target domain, we design a novel framework by adopting the stacked denoising autoencoder (SDAE). As SDAE has been demonstrated to be a powerful deep model that can learn more abstract representations based on its hierarchical architecture and take advantage of the out-of-distribution data for knowledge transfer [22], [25], the proposed transfer learning framework would be an effective way to predict the intrasaliency.

Deep learning has shown outstanding performance on mining deep and hidden patterns for building powerful representations in many challenging tasks, such as visual classification and object localization [26], [27]. In this paper, we attempt to leverage deep learning for the discovery of higher level homogeneity among cosalient regions. Specifically, we present the concept of deep intersaliency, which is formulated using the deep reconstruction residual obtained in the highest hidden layer of a self-trained SDAE. As the SDAE is trained on the image regions with higher intrasaliency priors among the multiple related images, it can extract more intrinsic and general hidden patterns to discover the homogeneity of cosalient objects in terms of some higher level concepts. Consequently, the obtained deep intersaliency could alleviate the influence of variance in luminance, shape, and view point, and should become a novel and useful cue when generating the final cosaliency map.

The flowchart of the proposed approach is shown in Fig. 2. First, the input images are decomposed hierarchically into fine-level superpixels and coarse-level segments. Then, the saliency prior in this paper is formulated based on the contrast prior and the object prior. We train the contrast model and the objectness model in the auxiliary data sets, and transfer them to generate the contrast prior map and the object prior map for each image in the cosaliency data sets, respectively. The intrasaliency prior is obtained by combining the contrast prior and the object prior. Afterward, we simultaneously explore the homogeneity among the multiple related images based on low-level feature matching and high-level pattern mining to establish the shallow intersaliency and the deep intersaliency, respectively. Finally, the cosaliency maps are generated by weighted integration of the proposed intrasaliency prior, shallow intersaliency, and deep intersaliency.

We notice that some early works [42] have applied deep models to solve problems in saliency detection. However, most of those algorithms are proposed for the task of eye fixation prediction rather than the task in this paper, i.e., cosaliency detection. More specifically, the deep model proposed in [42] is used for extracting low- and mid-level features and computing local contrast. However, the deep learning model proposed in this paper is used for the intrasaliency prior transfer and the deep intersaliency pattern mining.

In summary, the major contributions of this paper are threefold.

1) In this paper, we make the earliest effort to cast the intrasaliency prediction in cosaliency detection as

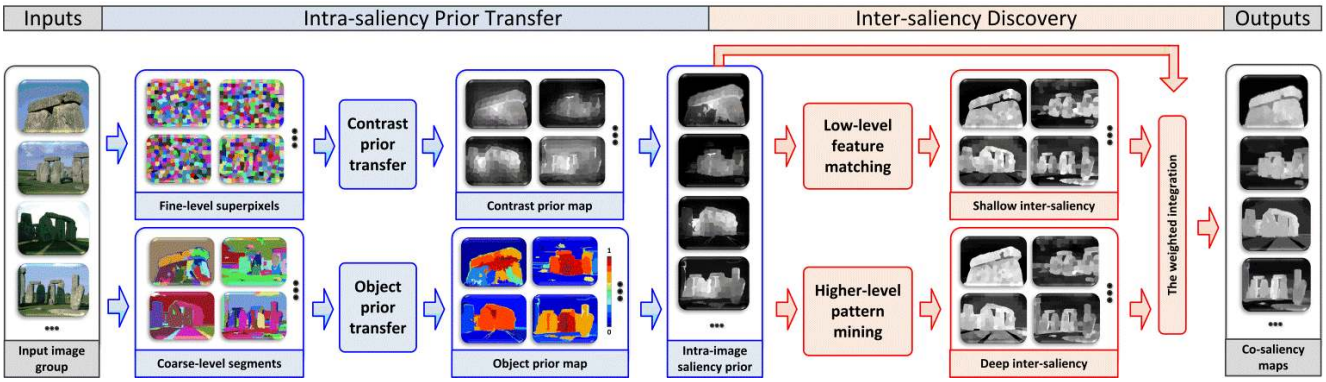


Fig. 2. Flowchart of the proposed cosaliency detection approach.

a problem of prior knowledge transfer, which could take advantage of the auxiliary fully annotated data sets and generate robust intrasaliency.

2) Besides exploring the shallow intersaliency, we also propose to mine the deep intersaliency for discovering higher level homogeneity of the cosalient objects in the image group. The generated deep intersaliency map is demonstrated in our experiments to be another critical factor in cosaliency detection.
3) SDAEs are used in this paper for better solving the problems both in the generation of the robust intrasaliency prior and in mining deep intersaliency patterns, which is the earliest effort to introduce deep learning to cosaliency detection.

The rest of this paper is organized as follows. Section II reviews the related works. Section III describes the proposed approach in detail. Section IV presents the experimental results with a quantitative evaluation in comparison with a number of the state-of-the-art approaches. Finally, the conclusions are drawn in Section V.

II. Related Works

Most early approaches for cosaliency detection explore the joint information provided by the image pair to find cosalient regions [12], [16]–[19]. However, these methods only seek to detect the cosaliency of two images at a time, not accounting for the discovery of the global coherent information that may exist when there are more than two images. This results in a direct limitation for cosalient pattern exploration when extending beyond pairwise relations. To tackle this problem, some recent works [11], [13]–[15], [28], [49] have been proposed to simulate the attention mechanisms for cosaliency detection in a group of images. Based on their assumption and formulation, these methods can be subdivided into three categories.

The first category is based on the assumption that the salient areas detected by the single image-based saliency detection approaches always contain parts of the foreground object, and the cosalient regions can be decided by selecting the areas frequently occurring among the multiple related images from the detected salient areas. The most representative work for this class was proposed in [13], where the cosaliency was formulated by a simple hard constraint of the distinctness (i.e., saliency in an individual image) and the repeatedness (i.e., the consistence measured in an image group) as Cosaliency = Distinctness × Repeatedness. This algorithm gives better performance than the conventional single image-based saliency detection methods in the task of cosaliency detection. However, it still appears to be ineffective due to its idealized assumption.

To mitigate this limitation, the second category of cosaliency detection approaches [11], [12], [14], [15] relieves the hard constraint to the soft constraint, which usually considers the intrasaliency, intersaliency, and other useful factors as independent information cues, and generates the final cosaliency map through the weighted integration of these cues. Specifically, Li *et al.* [11] proposed to generate an intraimage saliency map and an interimage saliency map based on multiscale segmentation and pairwise similarity ranking, respectively. Then, the cosaliency map was modeled as a linear combination of the two saliency maps. Fu *et al.* [14] extracted contrast cue, spatial cue, and corresponding cue through clustering and weighted integration of these information cues based on the probability formulation. Liu *et al.* [15] proposed a hierarchical segmentation-based cosaliency model, where the regional contrasts, global similarity, and object prior are calculated based on segmentations of multiple levels. The final cosaliency map was generated by effectively fusing the intrasaliency map and the object prior map.

Cao *et al.* [19], [28] proposed another category of algorithms for cosaliency detection, which focus on finding ways to integrate the existing saliency and cosaliency cues more reasonably. Rather than engaging to discover homogeneous information from the collection of multiple related images for representing cosalient objects, these methods mainly exploit the relationship of the obtained maps of multiple existing saliency and cosaliency approaches to obtain the self-adaptive weights for generating the final cosaliency map. Based on the most recent achievements in saliency detection and cosaliency detection, these methods produce a relatively satisfactory performance. However, the large time costs for preparing the existing saliency and cosaliency maps before the fusion process become their major limitations.

III. Proposed Approach

In this section, we first introduce the basic idea of the SDAE algorithm. Then, the overall procedure of the proposed

algorithm is briefly introduced. Afterward, two major components of the proposed framework, i.e., the robust intraimage saliency prior transfer and the intersaliency pattern mining, are described in detail. The generation of the final cosaliency map is introduced in Section III-E.

A. Stacked Denoising Autoencoder

SDAE is one kind of state-of-the-art deep learning models, which seeks to exploit the unknown structure in the input distribution at multiple layers to make the learned higher level representations more abstract and informative [22]. Compared with the convolutional neural network, SDAE can lean informative patterns from the input data in an unsupervised manner, which is what we need in Section III-D for mining deep intersaliency patterns. In addition, compared with the other unsupervised deep learning models, e.g., the deep Boltzmann machine (DBM), SDAE has fourfold advantages in this paper. First, SDAE is a better way to extract stable and deterministic numerical feature vectors, since it can directly learn the parametric mappings from input data to their representations [26]. However, although DBM can learn latent random variables to describe a posterior distribution over the observed data, the learnt posterior distribution is not yet the simple usable feature vectors in some cases [26]. Second, SDAE is demonstrated in [22] and [25] to have the capability to handle domain adaption. Thus, it is more suitable to use SDAE to transfer the prior knowledge for cosaliency detection, as described in Section III-C. Third, SDAE is a reconstruction-based model, and the generated reconstruction residual is what we need to formulate the deep intersaliency in Section III-D. However, we cannot obtain such a term from DBM. Finally, SDAE is simpler to train and explain, provides an efficient inference, and yields the results comparable or better than the RBM-based models in series of experiments [25]. All the above-mentioned advantages motivate us to use SDAE instead of other deep models in this paper.

As a basic building block in SDAE, an AE consists of an encoding process and a decoding process. With the aim to transform the input vector into output reconstructions with the least possible amount of distortion, it would learn useful representations and latent patterns of the given data. Specifically, the encoding process uses an encoding function $f(\boldsymbol{x_i}, \theta_f)$ to map from the input vector $\boldsymbol{x_i}$ to a hidden representation vector $\boldsymbol{y_i}$, where θ_f indicates the encoding parameters including an encoding projection matrix $\mathbf{W}^{(1)}$ and an encoding bias $\boldsymbol{b}^{(1)}$. Normally, the sigmoid function $\mathrm{sigm}(\eta) = 1/(1 + \exp(-\eta))$ is used in the encoding function

$$\boldsymbol{y_i} = f(\boldsymbol{x_i}, \theta_f) = \mathrm{sigm}(\mathbf{W}^{(1)}\boldsymbol{x_i} + \boldsymbol{b}^{(1)}). \quad (1)$$

Then, with the decoding parameters $\theta_g = \{\mathbf{W}^{(2)}, \boldsymbol{b}^{(2)}\}$, a decoding function $g(\boldsymbol{y_i}, \theta_g)$ is utilized to map the hidden representation $\boldsymbol{y_i}$ back to a reconstruction representation $\boldsymbol{z_i}$ through

$$\boldsymbol{z_i} = g(\boldsymbol{y_i}, \theta_g) = \mathrm{sigm}(\mathbf{W}^{(2)}\boldsymbol{y_i} + \boldsymbol{b}^{(2)}). \quad (2)$$

After encoding and decoding, the obtained reconstruction representation $\boldsymbol{z_i}$ can be taken as a prediction of input $\boldsymbol{x_i}$, which is based on the patterns encoded in the network. To learn appropriate parameters in θ_f and θ_g, the training process of such a network is to minimize the cost function with two important terms. The first one is the reconstruction error constraint, which is a basic constraint to reflect the difference between the original input data and the reconstruction output of the network. The second one is called sparsity constraint, which penalizes the deviation of the expected activation of the hidden units (in representation vector) from a fixed (low) level. With these two constraint terms, the cost function is written as

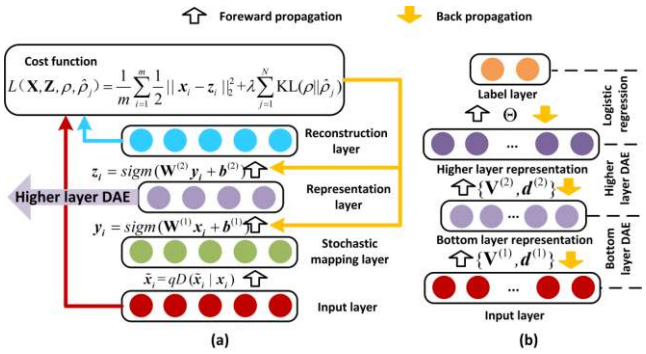


Fig. 3. Illustration of the architecture of DAE and SDAE. (a) DAE acted as one unit for building the SDAE. (b) SDAE built by two DAE layers and a logistic regression layer.

$$L(\mathbf{X}, \mathbf{Z}, \rho, \hat{\rho_j}) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}||\boldsymbol{x_i} - \boldsymbol{z_i}||_2^2 + \lambda\sum_{j=1}^{n}\mathrm{KL}(\rho||\hat{\rho}_j) \quad (3)$$

$$\mathrm{KL}(\rho||\hat{\rho}_j) = \rho\log\frac{\rho}{\hat{\rho_j}} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho_j}} \quad (4)$$

where m denotes the number of all the training and reconstructed data, respectively. λ is the weight of the sparsity constraint term, n is the dimension of the hidden representation vector, ρ is the target average activation of the hidden units, and $\hat{\rho_j} = \frac{1}{m}\sum_i[\boldsymbol{y}_1]_i/m$ is the average activation of the jth hidden unit $\boldsymbol{y}_j$ over the m training data. $\mathrm{KL}(\cdot)$ indicates the Kullback–Leibler divergence for providing the sparsity constraint. Like in sparse coding, a nonredundant overcomplete feature set is learned when ρ is small.

For further improving the effectiveness of AEs, Vincent *et al.* [29] propose to build DAEs by reconstructing the input data into a corrupted and partially destroyed version. In DAE [see Fig. 3(a)], the stochastic mapping function $\dot{\boldsymbol{x}}_i = qD(\dot{\boldsymbol{x}}_i|\boldsymbol{x_i})$ is first added to the original input data by randomly forcing 30% of them to be zero, while the objective function is still to minimize the reconstruction loss between a clean input $\boldsymbol{x_i}$ and its reconstruction output $\boldsymbol{z_i}$. Thus, it forces the learning of far more clever mapping than the identity [29]. Usually, training a DAE is straightforward using the gradient descent optimization algorithm to update the parameters $\mathbf{W}^{(l)} = \{W_{ij}^{(l)}\}$ and $\boldsymbol{b}^{(l)} = \{b_i^{(l)}\}$ in iterations, where $l = \{1, 2\}$ indicates the representation layer and the reconstruction layer. Specifically, all these parameters are randomly initialized, and then they are updated with the

updating rules

$$\Delta W_{ij}^{(l)} = -\varepsilon \frac{\partial L(\mathbf{X}, \mathbf{Z}, \rho, \hat{\rho_j})}{\partial W_{ij}^{(l)}} \quad (5)$$

$$\Delta b_i^{(l)} = -\varepsilon \frac{\partial L(\mathbf{X}, \mathbf{Z}, \rho, \hat{\rho_j})}{\partial b_i^{(l)}} \quad (6)$$

where ε is the learning rate. The partial derivatives in (5) and (6) are calculated by the backpropagation algorithm [30].

Based on the observation that the layerwise stacking of feature extraction often yields better representations [26], SDAE is built by stacking additional DAE layers to form the deep architecture [29] [see Fig. 3(b)]. Just as other deep neural networks, training SDAE could be done in two phases: 1) layerwise self-learning and 2) fine-tuning. Given a set of training data, the layerwise self-learning allows the usage of DAE as independent blocks for training the whole deep network. The key concept in this phase is to train one layer DAE at a time. As shown in Fig. 3(b), the bottom layer DAE is first trained with the original input data to obtain its encoding parameters. Then, the obtained hidden representations are used as the input data for training the higher layer DAE. As the labels of the input data are not needed in this process, the layerwise self-learning becomes to a task-free process focusing on learning hierarchical generative representations in an unsupervised manner. After the layerwise self-learning, a logistic regression layer can be added on the top of DAEs, as shown in Fig. 3(b), enabling the established deep architecture to capture more discriminative information under the supervision of the specific task.

Suppose, we have a training set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ with its label set $\{4_1, 4_2, \ldots, 4_m\}$. For each input data $\boldsymbol{x}_{i\in[1,m]}$, its higher (second) layer representation, as shown in Fig. 3(b), is denoted by $H_{\mathbf{V},\boldsymbol{d}}(x_i)$, where $\mathbf{V}$ and $\boldsymbol{d}$ indicate the parameters in the bottom two-layer neural network. These parameters include the weight matrix $\mathbf{V}^{(1)}$ and offset vector $\boldsymbol{d}^{(1)}$ between the input layer and the bottom representation layer, and $\mathbf{V}^{(2)}$ and $\boldsymbol{d}^{(2)}$ between the bottom representation layer and the higher representation layer. In the logistic regression layer, the hypothesis function is

$$h_{\mathbb{C}}(H_{\mathbf{V},\boldsymbol{d}}(\boldsymbol{x}_i)) = \frac{1}{1 + \exp(-\mathbb{C}^T H_{\mathbf{V},\boldsymbol{d}}(\boldsymbol{x}_i))} \quad (7)$$

where $\mathbb{C}$ is the parameter learned in logistic regression by minimizing the cost function

$$J = -\frac{1}{m}\sum_{i=1}^{m} 4_i \log h_{\mathbb{C}}(H_{\mathbf{V},\boldsymbol{d}}(\boldsymbol{x}_i)) + (1 - 4_i)\log(1 - h_{\mathbb{C}}(H_{\mathbf{V},\boldsymbol{d}}(\boldsymbol{x}_i))) \quad (8)$$

In this training phase, the parameters $\mathbf{V}$ and $\boldsymbol{d}$ are initialized by the layerwise self-learning, while $\mathbb{C}$ is initialized by random values. Then, all these parameters are optimized under the supervised information in the top logistic regression layer, which is implemented by using the gradient descent algorithm with backpropagation to minimize the cost function in (8).

Algorithm 1: Overall Procedure of Our Algorithm

Input: A group of images;
Output: Co-saliency maps of these images;
1: Generate fine-level superpixels $\{Sup_p\}$ for each image and extract the feature vectors $\{x_p\}$;
2: Train boundary-specific contrast SDAE models via **Algorithm** 2 and use the learnt SDAE models to calculate the final contrast prior CP_p via **Eq.** 9 and **Eq.** 10;
3: Generate course-level segments $\{Seg_q\}$ for each image;
4: Use the objectness model learnt in [40] to calculate the object prior for each segment via **Eq.** 11 and **Eq.** 12;
5: Use the pixel-wise mean of the contrast prior and object prior to generate the intra-saliency prior S^{in};
6: Calculate the shallow inter-saliency S_p^{sh} by using **Eq.** 13, **Eq.** 14, and **Eq.** 15;
7: Train a SDAE model via **Algorithm** 3 and use it to calculate the deep inter-saliency S_p^{dp} using **Eq.** 16 and **Eq.** 17;
8: Use the obtained S^{in}, S_p^{sh}, and S_p^{dp} to generate the final co-saliency maps via **Eq.** 18.

The notations in this Algorithm are defined in Section III.

B. Overall Algorithm

By using the SDAE model introduced above, we can transfer contrast prior knowledge (in Section III-C) and explore deep intersaliency (in Section III-D) in the proposed cosaliency detection framework. The overall algorithm flow of the proposed algorithm is shown in Algorithm 1.

In contrast prior transfer, the core problem is how to learn and transfer the prior knowledge of image contrast, which is a relationship between superpixels in the image foreground and background. To solve this problem, we use the generated sample pairs in an auxiliary data set, i.e., the accurate-segmented saliency detection (ASD) data set, as the input data to train SDAEs through greedy layerwise pretraining and supervised fine-tuning, as shown in Algorithm 2. By inputting the sample pairs from the cosaliency data sets into the trained SDAEs, we can obtain the outputted boundary specific contrast prior values for each superpixel that will be fused to generate the final contrast prior, as described in Section III-C.

In deep intersaliency pattern mining, the problem is how to capture the homogeneity of the cosalient objects in terms of some higher level concepts. To solve this problem, we use the selected superpixels with higher intrasaliency as the input data to train an SDAE via greedy layerwise unsupervised learning, which is shown in Algorithm 3. Then, the obtained deep model is used to output the deep reconstruction residuals for each input superpixel to formulate the deep intersaliency, as described in Section III-D.

C. Intrasaliency Prior Transfer

Contrast and objectness are two critical concepts for visual attention modeling [31]. More importantly, these two concepts are the most general knowledge about how much certain regions are visually different from the background and likely

Algorithm 2: Train SDAE Models for Transferring Contrast Prior

Input: Superpixels and their features in an auxiliary dataset;

Output: The learnt boundary-specific contrast SDAE models;

1: For Boundary = [top, left, bottom, right]
2: Collect the boundary-specific CB sample pairs and their labels;
3: Use the boundary-specific CB sample pairs as input data to layer-wise train the boundary-specific SDAE in an unsupervised manner;
4: Use the labels of the input data to fin-turning the boundary-specific SDAE model by using back-propagation.
5: End for

Algorithm 3: Train the SDAE Model to Formulate Deep Intersaliency

Input: The features and intra-saliency prior values of superpixels in each image of an image group;
Output: The learnt SDAE model;
1: Use the adaptive threshold in each image to select superpixels with higher intra-saliency prior;
2: Collect all the selected superpixels in the image group to form the training data;
3: Train the SDAE model in a completely unsupervised layer-wise manner.

to be parts of the salient objects. Regardless of the specific object category, these concepts would have less constraint on the choice of the auxiliary data set and be easy to transfer from the auxiliary data to the target data [20]. Inspired by this insight, we propose to transfer the saliency priors from the auxiliary annotated data sets for better solving the problems in generating a robust intrasaliency map.

1) Contrast Prior Transfer: Image contrast is one of the most widely used information for saliency detection in a single image [32], [33], because the contrast operator simulates the human receptive fields [14]. As a result, image regions that are distinct from the background would capture more human visual attention and become the salient regions in the image. By following the basic rule of photographic composition, we assume most image boundaries belong to the background area and formulate saliency based on the contrast between each image region and the image boundaries. As suggested by [34], image boundaries are separated into four sides, i.e., the top boundary, left boundary, bottom boundary and right boundary, and the final contrast prior would be obtained by combining the four side-specific contrast priors.

In this paper, we choose the ASD data set [35] as the auxiliary data set for learning and transferring the contrast model. Since the ASD data set is one of the largest benchmark data set for saliency detection containing 1000 images and the ground truth is manually labeled, we can use it to learn the contrast model to formulate the mechanism of human visual attention and, then, transfer the learned model to calculate the contrast prior for each image in the cosaliency data sets. Specifically, for each image, we first apply the simple linear iterative clustering algorithm [36] to decompose it into K_{fin} fine-level superpixels $\{\text{Sup}_p\}$, $p \in [1, K_{\text{fin}}]$. Then, we extract low-level visual features of 53 dimensions for each pixel as suggested in [33], including a 5-D color feature (three RGB color values as well as the hue and the saturation components), 12-D steerable pyramid filter responses, and 36-D Gabor filter responses. For each Sup_p, we use the mean features of the pixels within this superpixel as its feature vector $\boldsymbol{x}_p$.

In the learning process, we train four individual contrast models to formulate the image contrast specific to the top boundary, left boundary, bottom boundary, and right boundary, respectively. Because superpixels in different image boundaries are often dissimilar, we use them separately for better performance [34]. For each image boundary, we first collect the center-boundary (CB) sample pairs (where center indicates a superpixel not in the image boundary) to generate the pairwise inputs as well as their labels, which are determined by the ground truth mask within the center superpixels. Then, a four-layer SDAE is trained based on the generated inputs and labels to formulate the side-specific contrast. Taking the top image boundary as an example [see Fig. 4(a)], the superpixels within the top boundary (in purple) and a center superpixel (in yellow) are collected to form a CB sample pair. Afterward, all the superpixels in the CB sample pair are represented by the extracted low-level features. In order to establish the relationship between the center superpixel and boundary superpixels, all the image superpixel features in one CB sample pair should be concatenated into a single feature vector for representing the CB pair. Since the number of boundary superpixels is far more than that of the center superpixel, we average the feature vectors of boundary superpixels into one vector to address the imbalanced data dimension problem, and then concatenate it with the feature vector of the center superpixel. Therefore, the dimension of input vectors of SDAE should be twice of that of each superpixel representation. For training SDAE, we first use layerwise self-learning to determine the parameters among the input layer and two hidden layers, which helps to reduce the risk of falling into a poor local optimum of the whole network. Then, the supervised fine-tuning is applied with the label layer and the cost function in (8) to optimize the parameters ($\mathbf{V}$, $\boldsymbol{d}$, and ©) of the deep network. Thus, it could learn more complex mapping relations between the CB pair inputs and the corresponding saliency of the center superpixels.

After the learning process, the obtained SDAE models can capture the mutual patterns among CB pairs and infer their contrast hierarchically. Since the abstract concepts learned by SDAE could share a statistical strength across different but related types of examples coming from other domains than the task domain [25], it is convenient to transfer the trained SDAE models to calculate the contrast prior for the images in the cosaliency data sets without additional steps for domain adaption. Specifically, for each image in the cosaliency data set [see Fig. 4(b)], we first sample each center image superpixel [the yellow superpixel in the top-left image of Fig. 4(b)]

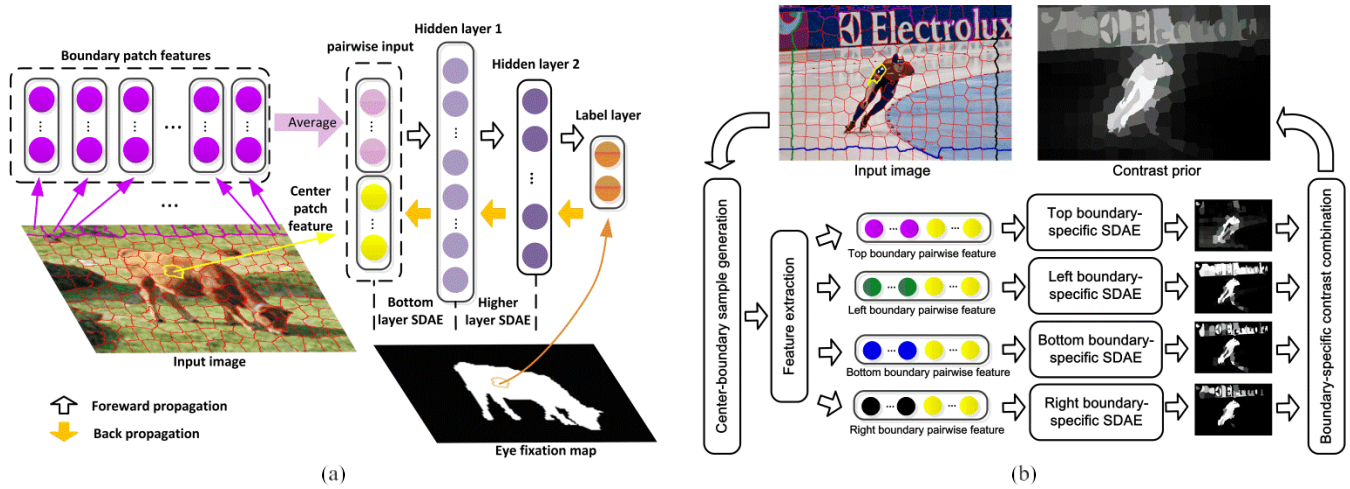


Fig. 4. Illustration of learning and transferring the contrast prior. (a) Learning process of the top-side-specific SDAE model. (b) Transferring contrast model for generating contrast prior. Yellow nodes: feature representation of the yellow center superpixel. Violet, green, blue, and black nodes: feature representations of the top, left, bottom, and right image boundaries, respectively.

with the boundary superpixels in each side to generate the sample pairs. Then, the low-level features are extracted to form the side-specific pairwise features. Putting these pairwise features into the corresponding contrast model, we can obtain the boundary specific contrast prior, i.e., $\mathrm{cp}_p^{\mathrm{top}}$, $\mathrm{cp}_p^{\mathrm{left}}$, $\mathrm{cp}_p^{\mathrm{bot}}$, and $\mathrm{cp}_p^{\mathrm{rig}}$. By taking into account the spatial consistency, the final contrast prior of each input image is obtained by

$$\mathrm{CP}_p = \frac{\sum_{\mathrm{Sup}_\tau \in N(\mathrm{Sup}_p)} \overline{\mathrm{cp}}_\tau \cdot \exp(-D(\boldsymbol{x}_p, \boldsymbol{x}_\tau))}{\sum_{\mathrm{Sup}_\tau \in N(\mathrm{Sup}_p)} \exp(-D(\boldsymbol{x}_p, \boldsymbol{x}_\tau))} \tag{9}$$

$$\overline{\mathrm{cp}}_p = \left(\mathrm{cp}_p^{\mathrm{top}} + \mathrm{cp}_p^{\mathrm{left}} + \mathrm{cp}_p^{\mathrm{bot}} + \mathrm{cp}_p^{\mathrm{rig}}\right)/4 \tag{10}$$

where $N(\mathrm{Sup}_p)$ denotes the neighborhood of Sup_p and $D(\boldsymbol{x}_p, \boldsymbol{x}_\tau)$ indicates the Euclidean distance between the two feature vectors.

2) Object Prior Transfer: The object prior in this paper is a generic measurement over various classes, which is different from the category specific detectors, such as faces or cars. It indicates how likely it is for an image window to contain an object of any class rather than background, such as sky and lawn. In contrast to object detectors extensively trained from a large number of category specific training samples, our approach is relatively less expensive and easy to obtain, but it is effective to salient object detection.

According to [15], the object prior is more suitably evaluated on the coarse segmentation. Thus, we apply a graph-based segmentation algorithm [37] to decompose an image into K_{coa} coarse-level segments $\{\mathrm{Seg}_q\}$, $q \in [1, K_{\mathrm{coa}}]$. Inspired by the studies in [21], [38], and [39], the objectness [40] is used in this paper, which is trained on PASCAL VOC07 data set to distinguish windows containing an object with a well-defined boundary from amorphous background windows based on several low-level image cues. It is then transferred to the cosaliency data sets to evaluate whether an image window contains an object or not. For each image, we can obtain a set of image windows W_k with their corresponding objectness probabilities ψ_k, where $k \in [1, 1000]$ as suggested in [40]. Afterward, all of these windows are integrated to form the objectness map OB by the pixelwise mean of their objectness probabilities

$$\mathrm{OB}_{\mathrm{pix}} = \frac{1}{|\gamma_{\mathrm{pix}}|} \sum_{W_k \in \gamma_{\mathrm{pix}}} \psi_k \tag{11}$$

where the subscript pix denotes a pixel in the objectness map and γ_{pix} indicates the collection of the windows that contain the certain pixel.

Inspired by the work in [15], we also use the appearance characteristics of real world backgrounds in images to improve the object prior map, which assumes that the background regions are usually large and homogeneous, and have a higher ratio of connectivity with image boundaries than salient objects. Consequently, the proposed object prior for each segment can be formulated by

$$\mathrm{OP}_q = \exp\left(-\gamma \frac{|\mathrm{Seg}_q \cap \mathrm{Bou}|}{\mathrm{per}_q} + \sum_{\mathrm{pix} \in \mathrm{Seg}_q} \frac{\mathrm{OB}_{\mathrm{pix}}}{|\mathrm{pix} \in \mathrm{Seg}_q|}\right) \tag{12}$$

where Bou denotes the image boundary, per_q indicates the perimeter of Seg_q, and $|\cdot|$ refers to the number of elements. γ is a decay factor set to be 2 as suggested in [15]. Finally, the intrasaliency prior S^{in} is obtained by the pixelwise mean of the contrast prior and object prior.

D. Intersaliency Pattern Mining

Mining the intersaliency patterns from the data pool of the multiple related images is another important component in our proposed cosaliency detection framework. Based on the intrasaliency prior, both the shallow and deep intersaliency patterns are explored in this paper to extract the common patterns of the cosalient objects among the image group (see Fig. 5).

The shallow intersaliency is explored based on the observation that the cosalient regions should be the visual similar regions sharing consistent color or texture and having

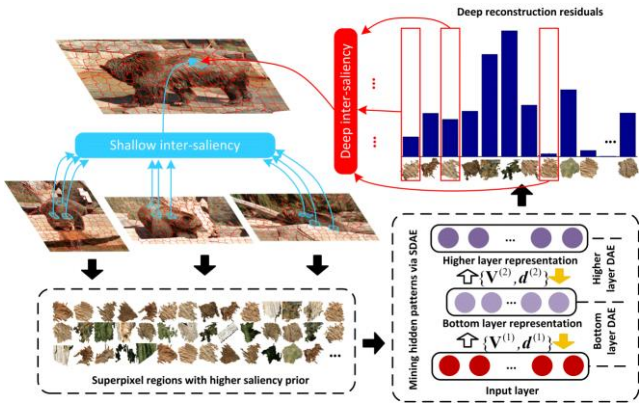


Fig. 5. Illustration of the mining of intersaliency based on shallow and deep cosalient cues.

Fig. 6. Some examples of the shallow intersaliency and the deep intersaliency.

As can be seen, the deep intersaliency alleviates the influence of variations in luminance, shape, and viewpoint to highlight the cosalient objects more uniformly.

higher intrasaliency prior. In other words, the cosalient regions normally have a higher global saliency and visual similarity. Specifically, for each Sup_p in the image group with M images, its K_{sim} most similar regions in each of the other images are searched based on the Euclidean distance of their features to form the collection $\{\text{Sup}_t\}$, $t \in [1, (M-1)K_{\text{sim}}]$. Thus, we can calculate the global saliency for each superpixel by

$$S_p^{\text{gl}} = \sum_{t=1}^{(M-1)K_{\text{sim}}} S_t^{\text{in}} \tag{13}$$

where S_t^{in} indicates the intrasaliency prior of Sup_t and S_p^{gl} denotes the global saliency of Sup_p. In order to further encourage the salient regions frequently appearing in multiple images and suppress the uncommon regions that only occur in a small number of images, we calculate the global similarity for each superpixel as follows:

$$O_p = \sum_{t=1}^{(M-1)K_{\text{sim}}} D(x_p, x_t) \tag{14}$$

where a small O_p indicates a large similarity among the image group and vice versa. By considering these two terms, the proposed shallow intersaliency is defined as

$$S_p^{\text{sh}} = S_p^{\text{gl}} \cdot \exp(-O_p). \tag{15}$$

Note that it is difficult to uniformly highlight the cosalient objects by mining of the shallow intersaliency based on the low-level features due to the influence of variations in luminance, shape, and viewpoint (see the second row of Fig. 6). To this end, we also propose to mine the deep intersaliency among the image groups. Unlike the shallow intersaliency, the deep intersaliency can capture the homogeneity of the cosalient objects in terms of some higher level concepts. This is important in cosaliency detection, whereas it is unexplored in the previous works. In this paper, the deep intersaliency is formulated by using the deep reconstruction residual obtained in a three-layer self-trained SDAE. Specifically, we first use an adaptive threshold, i.e., twice of the mean intrasaliency prior value, in each image to select superpixels with a higher priority. Then, all of these superpixels obtained in the image group are collected to form a data pool, which is then used by an SDAE model for the deep intersaliency pattern mining. With the help of the unsupervised self-training, the SDAE can abstract the generative and representative patterns layer to layer, and encode them into its weight matrices $\{\mathbf{V}^{(1)}, \mathbf{V}^{(2)}\}$. When using these learned patterns to represent input superpixels, the ones homogeneous with the cosalient regions are well represented with small reconstruction residuals and vice versa. Since the DAE trained in higher layer can capture more intrinsic and latent patterns of the cosalient regions [26], we propose to utilize the deep reconstruction residuals to formulate the deep intersaliency as

$$S_p^{\text{dp}} = \exp(-O_p) \sum_{t=1}^{(M-1)K_{\text{sim}}} \frac{1}{\text{DR}_t} \tag{16}$$

$$\text{DR}_t = \frac{1}{2} \left\| x_t^{(2)} - z_t^{(2)} \right\|_2^2 \tag{17}$$

where DR_t indicates the deep reconstruction residual of Sup_t, and $x_t^{(2)}$ and $z_t^{(2)}$ indicate the input vector and reconstruction vector of the higher (second) layer DAE in the SDAE model, respectively.

E. Cosaliency Map Generation

Until now, three critical information cues, i.e., the intrasaliency prior, the shallow intersaliency, and the deep intersaliency, have been introduced for cosaliency detection. Since each of these information cues only partially reflects one aspect of characteristics of the cosalient regions, we utilize a weighted linear combination in this paper to calculate cosaliency for each superpixel by

$$S_p^{\text{co}} = \beta \cdot \alpha S_p^{\text{dp}} + (1-\alpha) S_p^{\text{sh}} + (1-\beta) S_p^{\text{in}} \tag{18}$$

where S_p^{co} denotes the cosaliency value of Sup_p, α and β are two free parameters with values between 0 and 1. The final cosaliency map is generated by extracting the mean cosaliency values within the coarse segments of each image. As can be seen, α and β are two important parameters for the fusion process. α reflects the significance of mining intrinsic and deep structures for exploring the common patterns among multiple images, while β indicates the importance of exploring the common patterns among multiple images for the task of cosaliency detection. The final cosaliency is positively correlated with all the three information cues.

IV. Experiment

In this section, we evaluate the proposed approach both on image pair cosaliency detection and multiple images cosaliency detection. Qualitative and quantitative analyses of

TABLE I
HYPERPARAMETERS IN THE SDAE MODELS

	SDAE models used in contrast prior transfer			SDAE model used to mine the deep inter-saliency	
	Hidden layer 1	Hidden layer 2	Label layer	Bottom representation layer	Higher representation layer
Number of units	300	150	1	150	100
Target mean activation ρ	0.01	0.005	-	0.01	0.01

the experimental results are presented, which include the comparisons with some state-of-the-art methods on a variety of benchmark data sets.

A. Experimental Settings

1) Data Sets: Basically, we evaluate the proposed algorithm on two public benchmark data sets: the Image Pair data set [12] and iCoseg data set [4]. The Image Pair data set [12] contains 105 image pairs (i.e., 210 images) with manually labeled ground truth data. It is the earliest benchmark data set built for evaluating the performance of cosaliency detection, in which each image pair contains one or more similar objects with different backgrounds. The iCoseg data set [4] may be the largest publicly available data set so far that can be used for cosaliency detection. It consists of 38 image groups of totally 643 images along with pixel ground truth hand annotations. Since most images in the iCoseg data set contain complex background and multiple cosalient objects, and it is difficult to discover the useful information among multiple images, the iCoseg data set is considered as a more challenging data set for cosaliency detection.

2) Evaluation Metrics: To evaluate the performance of the proposed method, we adopted four widely used criteria that include the receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), the precision recall (PR) curve, and the average precision (AP). Like in [14], [15], and [34], ROC and AUC are generated by thresholding pixels in a saliency map into binary cosalient object masks with a series of fixed integers from 0 to 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). Specifically, the precision PRE, true positive rate TPR, and false positive rate FPR values are, respectively, defined as

$$\text{PRE} = \frac{|\text{SF} \cap \text{GF}|}{|\text{SF}|} \quad \text{TPR} = \frac{|\text{SF} \cap \text{GF}|}{|\text{GF}|} \quad \text{FPR} = \frac{|\text{SF} \cap \text{GB}|}{|\text{GB}|} \tag{19}$$

where SF, GF, and GB denote the set of segmented 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.

3) Implementation Details and Parameter Analysis: It is known that there are many hyperparameters involved in such deep neural networks, affecting the performance of the model. More specifically, we used a publicly available library in http://cn.mathworks.com/MATLABcentral/fileexchange/38310-deep-learning-toolbox, where the SDAE models are first initialized randomly and then trained with several hyperparameters, e.g., the target mean activation ρ, the weight of the sparsity penalty, the learning rate for the backpropagation optimization, and the number of units at each hidden layer. Before training, we follow [29], [43], and [44] to build a three-layer network for unsupervised layerwise learning and add another label layer for supervised fine-tuning (when necessary). Then, according to [26], we set the target mean activation ρ and the number of units empirically, as shown in Table I. For the other hyperparameters, we use a coordinate ascentlike method [41], [45] to optimize them for each layer. In addition, we show the relationship between the number of training samples and the performance of the contrast prior transfer in Fig. 7. As can be seen, the performance of such transfer process reasonably relies on the number of training samples and using all the images in the ASD data sets is able to generate the best transfer performance.

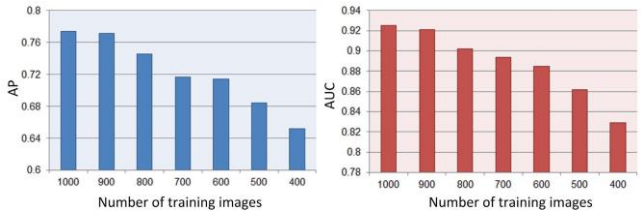


Fig. 7. Illustration of relationship between the number of training images and the performance of the transferred contrast priors.

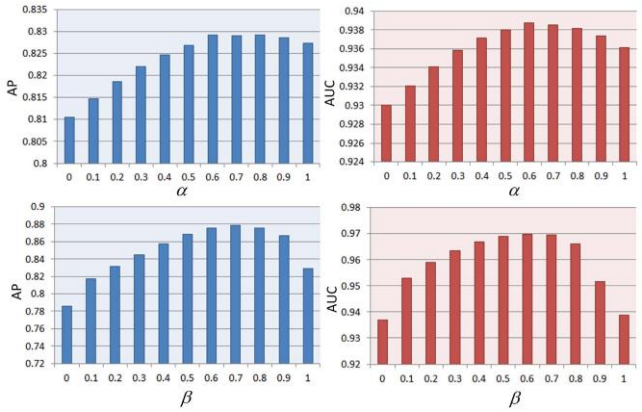


Fig. 8. Illustration of the mining of intersaliency based on the shallow and deep cosalient cues.

Besides the hyperparameters in the SDAE models, the parameter K_{sim} in the intersaliency pattern mining is empirically set to 3. In the experiments, we observe that the cosaliency detection results are reasonably sensitive to the parameters in (18). Thus, we set α and β to be 0.6 and 0.7, respectively, for the best performance. The detailed experiment and discussion

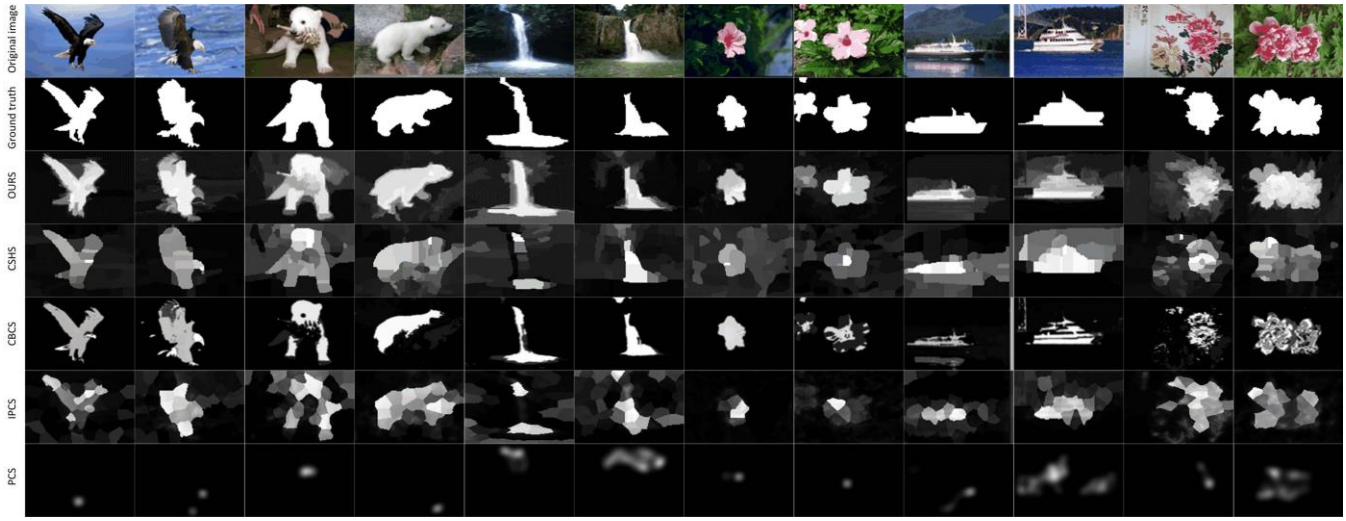


Fig. 9. Qualitative comparison of cosaliency maps on the Image Pair data set.

of these two parameters can be found in the next paragraph. For a hierarchical image segmentation, we generate fine-level superpixels and coarse-level segments by setting the number of superpixels in each image to be 200 and the pixels within each segment to be larger than 200, respectively. A unified set of parameters was utilized in all experiments.

In order to discuss the main parameters in (18) and investigate the contributions of the three information cues, i.e., the intrasaliency prior, the shallow intersaliency, and the deep intersaliency, on the overall performance based on the AUC curve, AP curve, AUC score, and AP score, we conduct an experiment on the iCoseg data set. The reason is that it contains more images that can be used for more comprehensive analysis. Specifically, we first set $\beta = 1$ to investigate the contributions of the shallow intersaliency and the deep intersaliency by varying α from 0 to 1. As shown in the top two histograms in Fig. 8, the performance of the deep intersaliency ($\alpha = 1$) is better than the shallow intersaliency ($\alpha = 0$). In addition, it also shows that the best performance for the intersaliency pattern mining can be achieved when α is $\sim$0.6. This implies that the deep patterns are more important in mining of the intersaliency. Afterward, we fix α to be 0.6 and vary β (0-1) to investigate the contributions of the intrasaliency prior and the intersaliency mined among the related images. From the bottom two histograms in Fig. 7, we can observe that the obtained intersaliency ($\beta = 1$) achieves better performance than the intrasaliency ($\beta = 0$) does, especially when looking at the AP score. In addition, it also can be found that the best fusion performance is reached when $\beta = 0.7$, indicating that mining intersaliency patterns plays a more important role in cosaliency detection.

B. Evaluation on the Image Pair Data Set

In this experiment, we first compared our cosaliency detection algorithm with a number of state-of-the-art cosalient detection algorithms, i.e., IPCS [12], CBCS [14], CSHS [15], and PCS [17]. Fig. 9 shows some comparison results of six pairs of images from the Image Pair data set, where the common objects exhibit distinct diversities in a color or shape property. The subjective evaluations by comparing with the ground truth reveal that the proposed method can yield cosaliency maps more correctly and robustly in these image pairs.

To provide quantitative comparison, we plotted the ROC and PRC for each approach and calculated the corresponding AUC and AP scores. As shown in Fig. 10, compared with the state-of-the-art cosaliency detection algorithms (i.e., IPCS, CBCS, CSHS, and PCS), the proposed approach can consistently achieve the highest true positive rates on the whole ROC curve and the highest precisions on the whole PR curve.

To demonstrate the effectiveness of the proposed saliency prior transfer method, we compared the proposed saliency prior transfer method with the intrasaliency detection method CBCS-S [14], the two state-of-the-art unsupervised single image saliency detection algorithms HS [32] and LR [33], and another outstanding supervised single image saliency detection method DRFI [46]. The experimental results shown in Fig. 10 demonstrate that transferring a contrast prior and an object prior from the auxiliary data sets is a promising way to formulate intraimage saliency, which outperforms both the intraimage saliency detection methods proposed in the state-of-the-art cosaliency detection and the recent single image saliency detection algorithms. The AUC and AP scores for each method are listed in Table II, from which we can observe that the proposed approach achieves the best performance with respect to both the AUC score and the AP score.

C. Evaluation on the iCoseg Data Set

We further evaluate the proposed algorithm on the iCoseg data set in which each image group may contain much more (17 on average) related images. Since IPCS [12] and PCS [17] are not valid on more than two images, we only compared the proposed approach with the two state-of-the-art cosaliency detection methods, i.e., CBCS [14] and CSHS [15], in this data set. Some experimental results are shown in Fig. 11, which contains five image groups, i.e., the Cheetah group, the Elephants group, the Gymnastics group, the Stonehenge

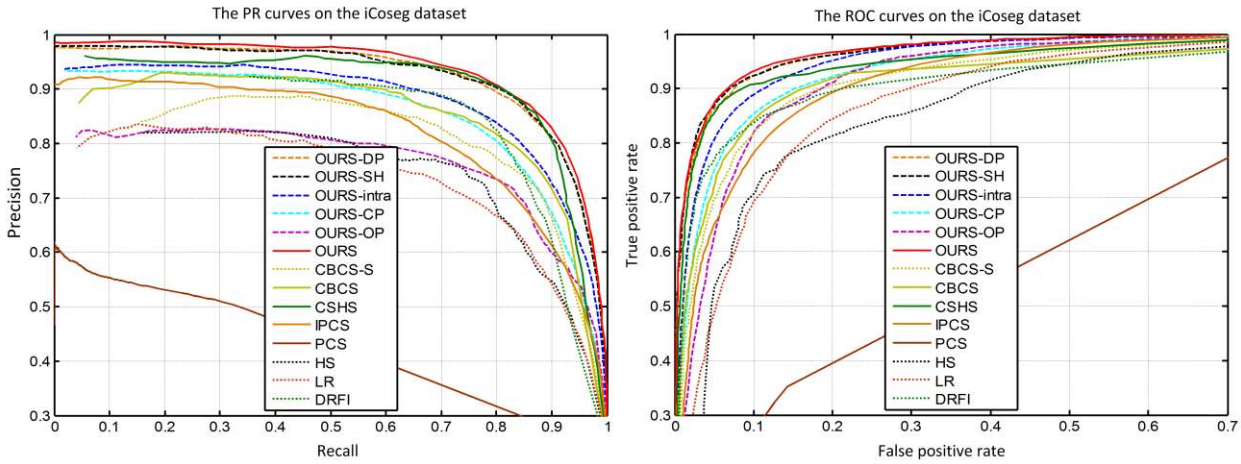


Fig. 10. ROC curves and PR curves for the proposed approach and other state-of-the-art algorithms (including the cosaliency methods and the single image methods) on the Image Pair data set. Solid lines: methods for cosaliency detection. Dashed lines: approaches used for intraimage saliency detection. OURS-intra corresponds to the performance of the proposed intraimage saliency prior. OURS-CP and OURS-OP are the curves of the proposed contrast prior and object prior, respectively. OURS-DP and OURS-SH are the curves of the proposed deep intersaliency and shallow intersaliency, respectively. CBCS-S is the intraimage saliency detection approach proposed in [14].

TABLE II

COMPARISON OF AUC AND AP SCORES BETWEEN THE PROPOSED APPROACH AND THE OTHER STATE-OF-THE-ART METHODS ON THE IMAGE PAIR DATA SET

	LR	HS	DRFI	PCS	IPCS	CSHS	CBCS	CBCS-S	OURS -OP	OURS -CP	OURS -intra	OURS -SH	OURS -DP	OURS
AP	0.733	0.758	0.857	0.427	0.819	0.903	0.852	0.818	0.766	0.853	0.878	0.922	0.923	**0.933**
AUC	0.891	0.874	0.921	0.608	0.925	0.954	0.926	0.928	0.931	0.944	0.959	0.969	0.971	**0.973**

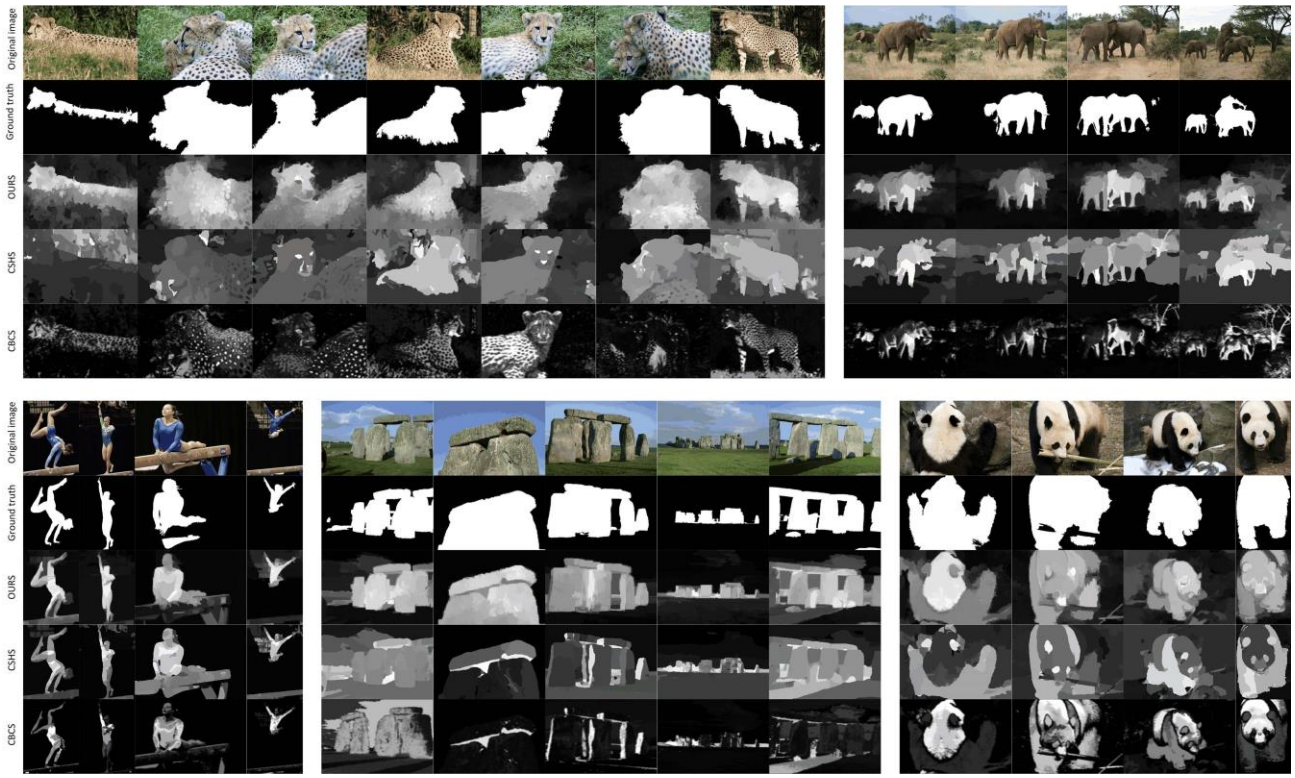


Fig. 11. Qualitative comparisons of cosaliency maps on the iCoseg data set.

group, and the Panda group. As can be seen, the proposed approach can obtain robust performance in the sense that it suppresses the cluttered and complex background regions (see the top two groups in Fig. 11), and meanwhile, uniformly highlights the cosalient objects with different viewpoints and shapes (see the bottom three groups in Fig. 11).

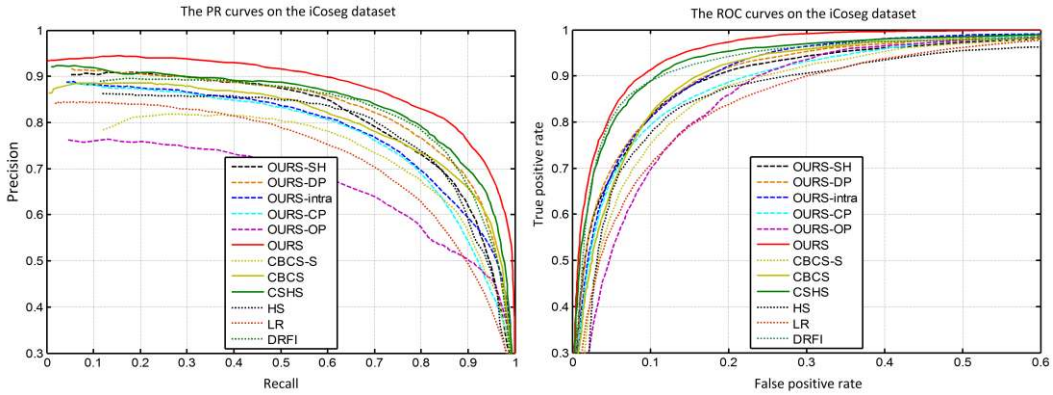


Fig. 12. ROC curves and PR curves for the proposed approach and other state-of-the-art algorithms (including the cosaliency methods and the single image methods) on the iCoseg data set. Solid lines: methods for cosaliency detection. Dashed lines: approaches used for intraimage saliency detection. OURS-intra corresponds to the performance of the proposed intraimage saliency prior. OURS-CP and OURS-OP are the curves of the proposed contrast prior and object prior, respectively. OURS-DP and OURS-SH are the curves of the proposed deep intersaliency and shallow intersaliency, respectively. CBCS-S is the intraimage saliency detection approach proposed in [14].

TABLE III

COMPARISON OF AUC AND AP SCORES BETWEEN THE PROPOSED APPROACH AND THE OTHER STATE-OF-THE-ART METHODS ON THE iCoseg DATA SET

	LR	HS	DRFI	CSHS	CBCS	CBCS-S	OURS-OP	OURS-CP	OURS-intra	OURS-SH	OURS-DP	OURS
AP	0.727	0.798	0.828	0.839	0.801	0.747	0.671	0.773	0.786	0.811	0.827	**0.878**
AUC	0.898	0.899	0.949	0.955	0.932	0.911	0.904	0.925	0.937	0.930	0.936	**0.969**

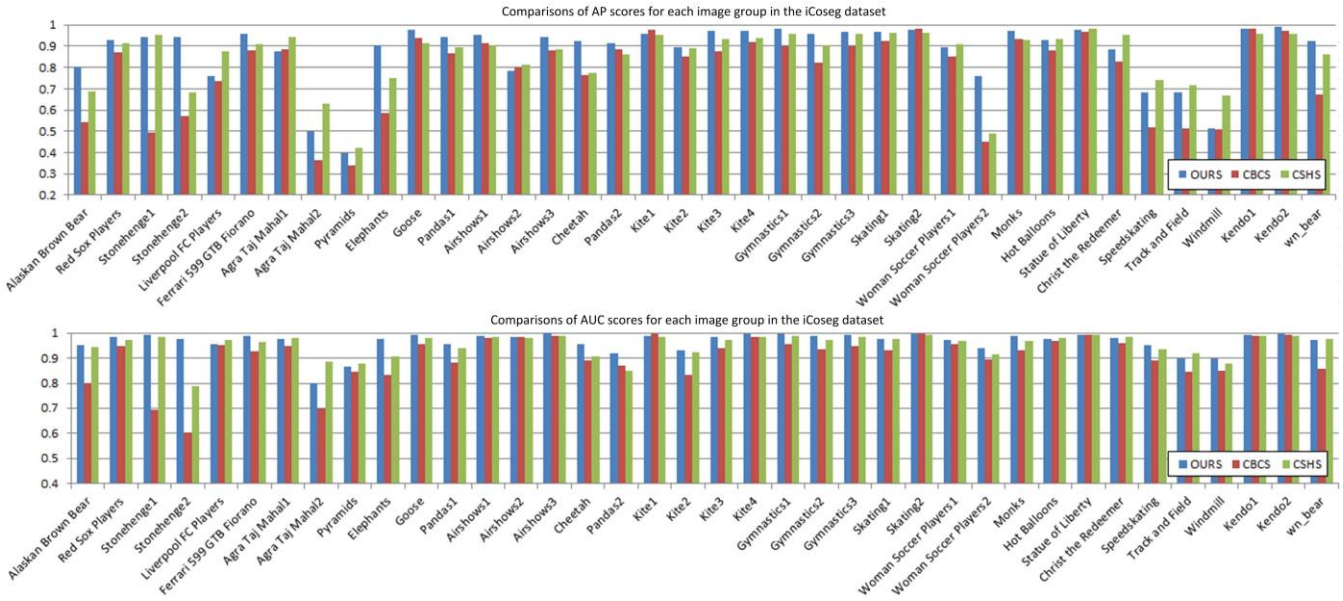


Fig. 13. Comparison of AUC and AP scores between the proposed approach and the other state-of-the-art cosaliency detection methods for each image group in the iCoseg data set.

Similar to what we did in the Image Pair data set, we also compared the proposed saliency prior transfer method with the intrasaliency detection method CBCS-S [14], and another three state-of-the-art single image saliency detection algorithms HS [32], LR [33], and DRFT [46] in the iCoseg data set. The ROC curves and PR curves of these approaches were drawn in Fig. 12, and the corresponding AUC scores and AP scores were listed in Table III. From Fig. 12 and Table III, it shows that the proposed saliency prior transfer method still obtains satisfactory performance, which is better than those two unsupervised single saliency models LR and HS, but worse than the supervised single saliency method DRFI. Due to our analysis, the reason for the promising performance of DRFI mainly lies in some additional considered factors, e.g., discriminative regional description and learning-based multilevel saliency fusion. This finding suggests a potential utility in transferring more useful knowledge for cosaliency detection in our future work. More importantly, like in the Image Pair data set, the cosaliency detection results of the proposed method could also outperform all other state-of-the-art

TABLE IV
AVERAGE RUNTIME (S) PER IMAGE

	IPCS	CBCS	CSHS	OURS*
Image Pair dataset	468.12	0.78	15.73	11.3
iCoseg dataset	N/A	1.63	103.36	19.6

* *The reported time does not include the time to train the boundary-specific contrast SDAE models off-line.*

algorithms and achieve the highest true positive rates on the whole ROC curve as well as the highest precisions on the whole PR curve consistently.

To perform further verification, we compared the AUC and AP scores between the proposed approach and the other state-of-the-art cosaliency detection methods for each image group in the iCoseg data set in Fig. 13. As can be seen, the proposed approach is superior to the other state-of-the-art algorithms in 25 image groups among the overall 38 image groups. For some image groups, e.g., Stonehenge2, Elephants, and Woman Soccer Players2, the proposed approach improves the performance of the existing cosaliency detection algorithms to a large extent.

D. Computational Cost and Runtime

Given an image group with M images, the time complexity of the proposed algorithm for generating cosaliency maps for these images is $O(M\tau \log\tau) + O(M^2)$, where τ indicates the number of pixels in each image. For intuitional comparison, Table IV lists the average execution time for each image by using different approaches. The experiment was run on a PC with Intel i3-2130 3.4-GHz CPU and 8-GB RAM. The code was implemented in MATLAB without optimization. For IPCS [12] and CBCS [14], we run the source codes provided by the authors on the same environment. Since the authors of CSHS [15] did not release their source code, we directly reported its runtime listed in their paper, which was run on the PC with a similar configuration to ours. As can be seen, the proposed algorithm achieves the best performance with the moderate computational complexity.

V. CONCLUSION

In this paper, we have proposed a novel cosaliency detection framework, which is one of the earliest efforts to investigate the feasibility of using deep learning in cosaliency detection. For better solving the problems in generating a robust intrasaliency map, this paper made the earliest effort to transfer useful knowledge from the auxiliary annotated data sets. Rather than just exploring the shallow intersaliency, we also proposed to mine the deep intersaliency by discovering the intrinsic and coherent structures of the cosalient objects. Comprehensive experiments on two publicly available benchmarks have demonstrated the effectiveness of the proposed work.

For the further work, we tend to extend the proposed work in the following directions. First, we will improve the proposed work by using more principled integration framework to fuse the obtained information cues. Second, we will embed the cosaliency detection process into weakly supervised learning framework [50], [51] for helping the object selecting with weakly labeled images. Third, the proposed method can also be extended and applied to a wide range of video processing tasks, such as video foreground extraction, video categorization, and video memorability computation [53].

REFERENCES

[1] 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.

[2] J. Han, D. Wang, L. Shao, X. Qian, G. Cheng, and J. Han, "Image visual attention computation and application via the learning of object attributes," *Mach. Vis. Appl.*, vol. 25, no. 7, pp. 1671–1683, Oct. 2013.

[3] 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.

[4] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "iCoseg: Interactive co-segmentation with intelligent scribble guidance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 3169–3176.

[5] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 169–176.

[6] L. Mukherjee, V. Singh, and J. Peng, "Scale invariant cosegmentation for image groups," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2011, pp. 1881–1888.

[7] H. Fu, D. Xu, B. Zhang, and S. Lin, "Object-based multiple foreground video co-segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2013, pp. 3166–3173.

[8] Y. Yang, B. Geng, Y. Cai, A. Hanjalic, and X.-S. Hua, "Object retrieval using visual query context," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1295–1307, Dec. 2011.

[9] X. Guo, D. Liu, B. Jou, M. Zhu, A. Cai, and S.-F. Chang, "Robust object co-detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 3206–3213.

[10] A. Toshev, J. Shi, and K. Daniilidis, "Image matching via saliency region correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.

[11] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1896–1909, Dec. 2013.

[12] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3375, Dec. 2011.

[13] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2011, pp. 2129–2136.

[14] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.

[15] Z. Liu, W. Zou, L. Li, L. Shen, and O. Le Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 88–92, Jan. 2014.

[16] D. E. Jacobs, D. B. Goldman, and E. Shechtman, "Cosaliency: Where people look when comparing images," in *Proc. Annu. ACM Symp. User Interface Softw. Technol.*, New York, NY, USA, Oct. 2010, pp. 219–228.

[17] H.-T. Chen, "Preattentive co-saliency detection," in *Proc. IEEE Int. Conf. Image Process.*, Hong Kong, Sep. 2010, pp. 1117–1120.

[18] Z. Tan, L. Wan, W. Feng, and C.-M. Pun, "Image co-saliency detection by propagating superpixel affinities," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 2114–2118.

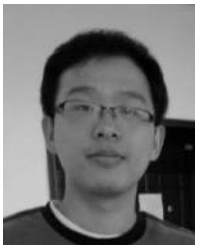
[19] X. Cao, Z. Tao, B. Zhang, H. Fu, and X. Li, "Saliency map fusion based on rank-one constraint," in *Proc. IEEE Int. Conf. Multimedia Expo*, San Jose, CA, USA, Jul. 2013, pp. 1–6.

[20] Z. Shi, P. Siva, and T. Xiang, "Transfer learning by ranking for weakly supervised object annotation," in *Proc. Brit. Mach. Vis. Conf.*, Guildford, U.K., Sep. 2012, pp. 1–5.

[21] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *Int. J. Comput. Vis.*, vol. 100, no. 3, pp. 275–293, Dec. 2012.

[22] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proc. Unsupervised Transf. Learn. Challenge Workshop*, Bellevue, WA, USA, Jul. 2011, pp. 17–36.

[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 580–587.
[24] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2015.
[25] Y. Bengio *et al.*, "Deep learners benefit more from out-of-distribution examples," in *Proc. Int. Conf. Artif. Intell. Statist.*, Fort Lauderdale, FL, USA, Apr. 2011, pp. 164–172.
[26] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
[27] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, Zürich, Switzerland, Sep. 2014, pp. 346–361.
[28] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4175–4186, Sep. 2014.
[29] 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. 11, pp. 3371–3408, Jan. 2010.
[30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Representations by Back-Propagating Errors*. Cambridge, MA, USA: MIT Press, 1988.
[31] 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.
[32] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 1155–1162.
[33] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 853–860.
[34] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 3166–3173.
[35] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1597–1604.
[36] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[37] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep. 2004.
[38] A. Prest, C. Schmid, and V. Ferrari, "Weakly supervised learning of interactions between humans and objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 601–614, Mar. 2012.
[39] D. Kuettel and V. Ferrari, "Figure-ground segmentation by transferring window masks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 558–565.
[40] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.
[41] H.-C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. 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.
[42] Y. Lin, S. Kong, D. Wang, and Y. Zhuang, "Saliency detection within a deep convolutional architecture," in *Proc. Workshops AAAI Conf. Artif. Intell.*, Québec City, QC, Canada, Jul. 2014.
[43] Y. Kang, K.-T. Lee, J. Eun, S. Park, and S. Choi, "Stacked denoising autoencoders for face pose normalization," in *Proc. Neural Inf. Process.*, Stateline, NV, USA, Dec. 2013, pp. 241–248.
[44] A. Sankaran, P. Pandey, M. Vatsa, and R. Singh, "On latent fingerprint minutiae extraction using stacked denoising sparse autoencoders," in *Proc. IEEE Int. Joint Conf. Biometrics*, Clearwater, FL, USA, Oct. 2014, pp. 1–7.
[45] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade* (Lecture Notes in Computer Science), K.-R. Müller, G. Montavon, and G. B. Orr, Eds. Berlin, Germany: Springer-Verlag, 1998.
[46] 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 Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 2083–2090.
[47] J. Han, D. Zhang, S. Wen, L. Guo, T. Liu, and X. Li, "Two-stage learning to predict human eye fixations via stacked denoising autoencoders," *IEEE Trans. Cybern.*, to be published.
[48] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1309–1321, Aug. 2015.
[49] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 2994–3002.
[50] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.
[51] D. Zhang, J. Han, G. Cheng, Z. Liu, S. Bu, and L. Guo, "Weakly supervised learning for target detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 701–705, Apr. 2015.
[52] J. Han *et al.*, "Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding," *ISPRS J. Photogramm. Remote Sens.*, vol. 89, pp. 37–48, Mar. 2014.
[53] J. Han, C. Chen, L. Shao, X. Hu, J. Han, and T. Liu, "Learning computational models of video memorability from fMRI brain imaging," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1692–1703, Aug. 2015.



Dingwen Zhang received the B.E. degree from Northwestern Polytechnical University, Xi'an, China, in 2012, where he is currently pursuing the Ph.D. degree.

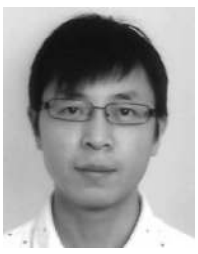
His current research interests include computer vision and multimedia processing, in particular, on saliency detection, co-saliency detection, and weakly supervised learning.

Junwei Han is currently a Full Professor with Northwestern Polytechnical University, Xi'an, China. His current research interests include computer vision, multimedia processing, and brain imaging analysis.

Prof. Han is an Associate Editor of the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, *Neurocomputing*, and *Multidimensional Systems and Signal Processing*.

Jungong Han is currently a Senior Lecturer with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, U.K. His current research interests include multimedia content identification, multisensor data fusion, computer vision, and multimedia security.

Dr. Han is an Associate Editor of *Neurocomputing* (Elsevier), and an Editorial Board Member of *Multimedia Tools and Applications* (Springer).



Ling Shao (M'09–SM'10) is currently a Professor with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, U.K., and a Guest Professor with the Nanjing University of Information Science and Technology, Nanjing, China. His current research interests include computer vision, image/video processing, and machine learning.

Prof. Shao is a fellow of the British Computer Society and the Institution of Engineering and Technology. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CYBERNETICS, and several other journals.