# Correspondence Driven Saliency Transfer

Wenguan Wang, Jianbing Shen, *Senior Member, IEEE*, Ling Shao, *Senior Member, IEEE*,
and Fatih Porikli, *Fellow, IEEE*

*Abstract*—In this paper, we show that large annotated data sets have great potential to provide strong priors for saliency estimation rather than merely serving for benchmark evaluations. To this end, we present a novel image saliency detection method called saliency transfer. Given an input image, we first retrieve a support set of best matches from the large database of saliency annotated images. Then, we assign the transitional saliency scores by warping the support set annotations onto the input image according to computed dense correspondences. To incorporate context, we employ two complementary correspondence strategies: a global matching scheme based on scene-level analysis and a local matching scheme based on patch-level inference. We then introduce two refinement measures to further refine the saliency maps and apply the random-walk-with-restart by exploring the global saliency structure to estimate the affinity between foreground and background assignments. Extensive experimental results on four publicly available benchmark data sets demonstrate that the proposed saliency algorithm consistently outperforms the current state-of-the-art methods.

*Index Terms*—Image saliency, salient object detection, saliency transfer, correspondence, random-walk-with-restart.

## I. INTRODUCTION

SALIENCY detection is an important research problem in both neuroscience and computer vision. According to the studies of psychology and cognitive science, the human vision system is remarkably effective in localizing the most visually important regions in a scene. In order to simulate such attentional and selective capability of human perception, early saliency detection algorithms aimed at predicting scene locations where a human observer may fixate, which are mostly based on cognitive theories (*e.g.*, feature integration theory (FIT) [1]) and biologically inspired visual attention models (*e.g.*, Koch and Ullman [2] and Itti *et al.* [5]). In recent years, intensive research has been carried out for salient object detection to accurately extract the most informative and noticeable regions or objects. This new trend is driven by object based vision applications, such as object detection [6], content-aware image resizing [3], image segmentation [4], [40], and other applications [38], [39], [41], [43]. In this work, we focus on the salient object detection, and the algorithm outputs a gray saliency image, where a brighter pixel stands for a higher saliency value.

A large number of salient object detection methods have been proposed in the past few years. From the perspective of information processing, those saliency algorithms can be broadly categorized as either top-down or bottom-up approaches. Top-down approaches [7]–[10] are goal-directed and usually adopt supervised learning with a specific class. Most of the saliency detection methods are based on bottom-up visual attention mechanisms [11]–[15], [17], [18], [21], which are independent of the knowledge of the content in the image and utilize various low level features, such as intensity, color and orientation. Those bottom-up saliency models are generally based on different mathematical formulations of center-surround contrast or treat the image boundary as the background. Albeit previous saliency models have achieved success in their own aspects, a few commonly noticeable and critically influencing issues still exist. Firstly, traditional stimuli-driven saliency models are often constructed by simple bottom-up and low-level heuristics and lack of adaptability to capture image content for describing complex scenarios and object structures. Secondly, for top-down saliency approaches, the salient object classes are usually limited and constrained to the training images, which restricts its applicability seriously. Thirdly, existing saliency models, no matter top-down or bottom-up, ignore the contextual information in saliency detection. Therefore, it is unclear how current models perform on complex, cluttered scenes. An example is presented in Fig. 1. In the depicted scene, the state-of-the-art methods unsurprisingly fail since they omit key contextual information.

Here, we explore the value of the contextual information and introduce a correspondence-based saliency transfer approach that infers foreground regions from a support set of annotated images (see Fig. 1-f) that share similar context to the input image. The algorithm is essentially an example-driven mechanism, which is more generally valid than traditional heuristics methods. For an input image, our method first retrieves a support set of its most similar matches from a large database of images annotated with salient regions. The support images only share high-level scene characteristics, yet they provide the contextual information that we are after. Instead of estimating saliency only from the features within the query image, we transfer the annotations from the support images into the query

W. Wang and J. Shen are with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: wenguanwang@bit.edu.cn; shenjianbing@bit.edu.cn).

L. Shao is with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, NE1 8ST, U.K. (e-mail: ling.shao@ieee.org).

F. Porikli is with the Research School of Engineering, Australian National University, Canberra, ACT 0200, Australia, and also with the NICTA, NSW 2015, Australia (e-mail: fatih.porikli@anu.edu.au).
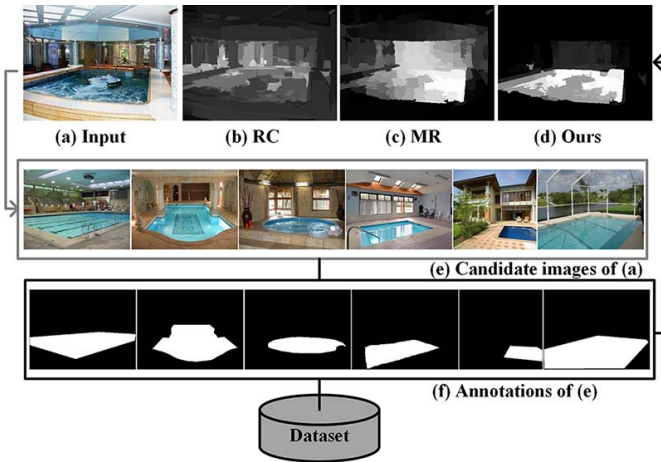
Fig. 1. We can ask how to identify correctly the salient region in complex scenario (a). The state-of-the-art methods, e.g., (b) the contrast prior based RC [11] and (c) the background prior based MR [15], face with ambiguity since they have no mechanism to incorporate additional contextual information. Our correspondence-based saliency transfer method (d) utilizes the saliency prior (f) from a set of support images (e) that share similar contextual scene information with the input image.

image according to their global and local correspondences. We employ the deformable spatial pyramid matching [22] that simultaneously regularizes match consistency at multiple spatial extents ranging from global-level of the entire image, to local patch-level, and to every single pixel in order to establish dense correspondences for each pair of query-support images. Then, we map the annotations of the support images onto a transitional saliency map according to their dense correspondences, and utilize two refinement measures to refine the saliency maps (Sec. III-C). Finally, we apply the random-walk-with-restart (RWR) segmentation (Sec. III-D) to obtain the final saliency map. Our source code will be available at.[1]

Compared to the existing approaches, the proposed method offers the following contributions:

- A novel saliency technique, called saliency transfer, is proposed for transferring the labels from existing annotated images to the input image through dense scene correspondences.
- Scene level and patch level matching strategies are proposed for selecting nearest-support images and transferring saliency.
- Two complementary saliency distance measurements and an RWR based approach are incorporated for inferring the saliency assignment.
- Saliency transfer is an example-driven mechanism relying on semantic correspondence, which is more generalizable compared with traditional heuristics models.

## II. RELATED WORK

Image saliency is a classic problem that has been extensively studied for decades. Instead of surveying the large volume of literature, which is impractical here, we mainly focus on recent bottom-up saliency methods and top-down models, and

[1]http://github.com/shenjianbing/saliencytransfer

analyze their properties and limitations. We refer the readers to [23] and [24] for more detailed reviews of the saliency models.

### A. Top-Down Saliency Detection

Saliency detection can be regarded as a specific task, which assumes a priori knowledge or constraints on scenes, and thus performs in a top-down manner [8]–[10], [25]. In [8], a multi-task rank learning was proposed for inferring multiple saliency models that apply to different scene clusters. Liu *et al.* [9] presented a conditional random field based supervised approach to detect a salient object in an image or sequential images. Borji *et al.* [10] proposed a Bayesian approach to model task-driven visual attention by utilizing the sequential nature of real-world tasks. Several sources of information, including global context of a scene, previous attended locations, and previous motor actions, are integrated over time to predict the next attended location. Li *et al.* [25] presented a top-down saliency approach to incorporate low-level features and the objectness measure via label propagation. Generally, such task-driven methods are useful especially for object recognition [7], but they require knowledge learning that increases the complexity of saliency detection in general.

### B. Bottom-Up Saliency Detection

Bottom-up saliency detection methods are largely independent of the knowledge of content in the image and can be broadly classified as either contrast prior based or boundary prior based approaches. As argued by the pioneering perceptual research studies [26], [27], *contrast* is one of the influential factors in low-level visual saliency. Since the salient regions in the visual field would first pop out through different low-level features from their surroundings, numerous bottom-up models [11]–[13], [28]–[30], [44] have been proposed to detect salient regions in images based on different mathematical principles. These saliency approaches built saliency models focusing on high contrast regions between candidate foreground objects and their surrounding backgrounds. More specifically, Cheng *et al.* [28] aimed at two saliency indicators: global appearance contrast and spatially compact distribution. Goferman *et al.* [12] built a content-aware saliency detection model with the consideration of the contrast from both local and global perspectives. Klein and Frintrop [13] presented a saliency detection framework based on the fusion of different feature channels and the local center-surround hypothesis. Such methods, however, may suffer from the internal attenuation problem that causes emphasizing mainly object boundaries rather than highlighting the entire object region, and are limited by the high complexity and large variety of object appearances in real scenarios.

The core of those contrast prior based mechanisms is performing saliency detection via exploring the notion of "what salient object is". More recently, alternative approaches attempted to tackle this problem from an opposite viewpoint by focusing on "what the background should look like". These methods treat image boundaries as background, further enhancing saliency computation. Wei *et al.* [31] exploited

*boundary prior* by noting that the image boundary regions are more likely to belong to the background. Similarly, many follow-up studies [14]–[16], [18], [21] were proposed to treat image boundaries as background samples. To improve the performance, these approaches also explore more robust boundary priors. For example, Jiang *et al.* [14] proposed a graph-based method that models boundary regions as the absorbing nodes in a Markov chain and computes the saliency according to the absorption time in a random walk propagation. In [15], the saliency of image regions was measured by their relevance to the image boundary via a manifold ranking scheme. The work in [18] constructed a robust boundary prior based on boundary connectivity. Qin *et al.* [21] used the clustered boundary seeds into a cellular automata. While these methods have demonstrated impressive results, they also encounter critical issues. Their performance may deteriorate when the object connects with an image boundary. Furthermore, when the background is close to the center of the image, extra efforts should be paid for this situation.

## III. SALIENCY TRANSFER

### A. Overview

In this work, we introduce a novel method to predict what is salient or interesting in a scene using a saliency transfer strategy. Our algorithm can be decomposed into three phases. In the first stage, we introduce a correspondence-based transitional saliency estimation. This method is based on an observation that saliency can be estimated by transferring the labels from semantically related images and patches. We introduce global as well as local matching strategies for transferring saliency, which are based on scene level and patch level, respectively. This stage produces rough and initial saliency estimation, which is detailed in Sec. III-B. After that, we introduce two refinement measures to further improve the initial saliency constructed in the first stage. A detailed description on separating the salient region from the background based on these two refinement measures is given in Sec. III-C. Finally, in Sec. III-D, we utilize an RWR based method to further modify the saliency map generated at the second stage.

### B. Correspondence-Based Transitional Saliency

This stage provides initial saliency estimates by making the best use of the available saliency annotations in a large reference dataset. We start our system by finding a support group of the input image, which consists of the nearest neighbors of the input image from the annotated dataset. We use a *scene retrieval* technique to find *M-best support images* that share similar scene configuration with the input. The distance between the query image and the support images is measured via the GIST descriptor, which can model the scene characteristic and is widely used in image retrieval [20].

After this, we establish the correspondences from the input image to each support image using a deformable scene matching scheme [22] by comparing dense, pixel-wise SIFT descriptors on a spatial pyramid that divides the image recursively into four rectangular grid cells until it reaches the pixel-level
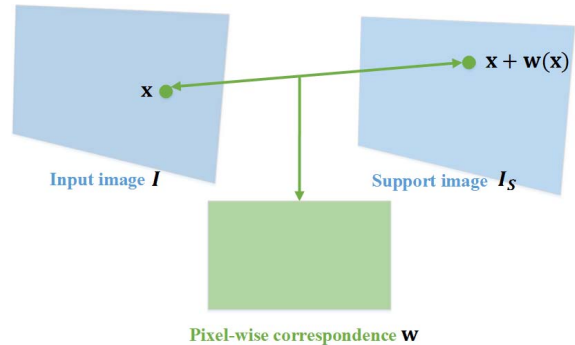


Fig. 2. Illustration of the dense correspondences $\mathbf{w}$ between the input image $I$ and a support image $I_S$ via scene matching.

resolution. The SIFT descriptor is an efficient representation for matching objects under different views. Suppose $I$ and $I_S$ denote the input image and a support image respectively, reliable pixel-wise correspondences $\mathbf{w}$ between $I$ and $I_S$ can be established. As shown in Fig. 2, pixel $\mathbf{x}$ of image $I$ is associated with pixel $\mathbf{x} + \mathbf{w}(\mathbf{x})$ of support image $I_S$. Based on such pixel-wise correspondences, we can warp support images and the annotations. The warped images are closer to the input image according to the correspondences than support image $I_S$, and the warped annotations are used for inferring the saliency of the input image.

We introduce the following two correspondence based matching strategies for excluding noisy images and incorrect assignments, and transferring saliency on both image and patch levels.

*1) Global Correspondence-Based Saliency:* For input image $I$, we build a SIFT feature map $f_I$ that contains the detected SIFT feature landmark points $\mathbf{x} = (x, y)$ and their descriptors $f_I(\mathbf{x})$. Then we establish a set of $M$-best support images $\{I_i, f_i, g_i\}_{i=1:M}$, where $I_i$ is the $i$-th nearest support image through GIST matching; $f_i$ is the SIFT feature map of the warped image of $I_i$; $g_i$ is the warped annotation of $I_i$. The warped image of $I_i$ and the warped annotation $g_i$ are obtained according to the correspondences [22] to the input image. Fig. 3 illustrates such correspondence-based warping.

We further select $N$ ($N < M$) closest support images as candidates for image $I$, and the matching accuracy is measured using the distance between the SIFT image $f_I$ and the SIFT image $f_i$ from the warped image of the $i$-th support image. This candidate set is used for transferring the available saliency annotation onto the input image. The $N$ closest candidates $\{I_{t_j}, f_{t_j}, g_{t_j}\}_{j=1:N, t_j \in \{1:M\}}$ are determined by:

$$\underset{\substack{t_{1:N} \\ t_j \in \{1:M\}}}{\operatorname{argmin}} \sum_{t_j} \sum_{\mathbf{x}} (f_I(\mathbf{x}) - f_{t_j}(\mathbf{x}))^2. \tag{1}$$

Via (1), $N$ closest candidate images are selected according to the difference between the warped image and the input image via their SIFT feature distances. Support images with inaccurate matching correspondences can be excluded as their warped images are largely different from the input image measured via accumulated pixel-wise SIFT feature distance. Fig. 4-a directly illustrates the global matching strategy. In our global matching process, the $N$ candidate images
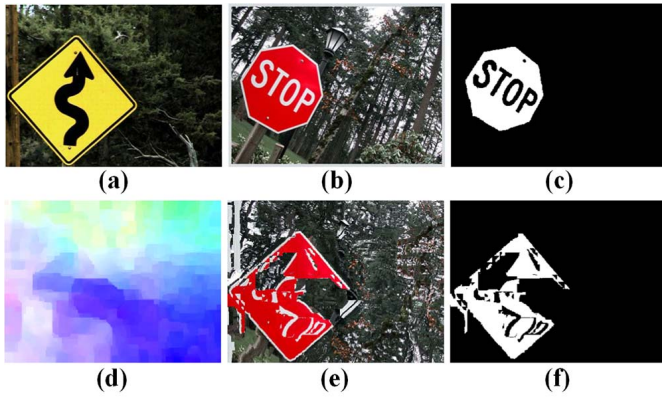
Fig. 3. Illustration of correspondence-based warping. We choose the N best matching images and patches according to the feature similarity between the warped image and the input image using SIFT descriptors.(a) Input image $I$. (b) A support image $I_i$ retrieved from the reference dataset via GIST matching. (c) The annotation of $I_i$. (d) Pixel-wise correspondences between $I$ and $I_i$ established via [22]. (e) Warped image of $I_i$ according to the correspondences in (d), which is similar to test image $I$. (f) Warped annotation $g_i$ of (c) according to the correspondences in (d).
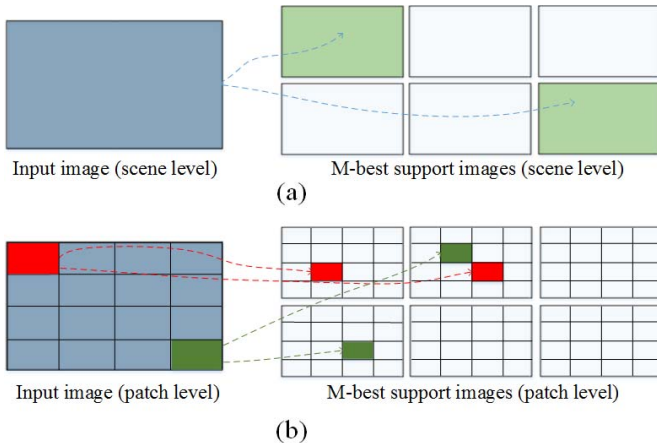


Fig. 4. Illustration of our global as well as local matching strategies. (a) Our global matching strategy considers the similarity between the input image and the warped images of $M$-best support images on the scene level. The green images indicate the $N$ candidate images. (b) Our local matching strategy is on the patch level. The red (green) patches of the warped images of support images are selected as the $N$ candidate patches for the red (green) one of the input image.

(green images in (a)) are selected from the $M$-best support images according to the SIFT feature similarities between their warped images and the input image.

Based on the candidate set $\{I_{t_j}, f_{t_j}, g_{t_j}\}_{j=1:N}$, we adopt a voting strategy for estimating saliency:

$$S_g = \frac{1}{N} \sum_j g_{t_j}, \qquad (2)$$

where $g_{t_j}$ is computed through warping the annotation of candidate image $I_{t_j}$ according to the correspondence using scene matching (see Fig. 3-f).

*2) Local Correspondence-Based Saliency:* In (1), we consider global matching of two images. This strategy, however, ignores patch-wise details in the matching process despite that only small part of the input image and the support images

might be consistent. To address this, we introduce a local correspondence method that improves matching accuracy via patch alignment instead of matching the whole scene.

For each $4 \times 4$ patch $\mathbf{p}$ in image $I$, we select $N$ most similar patches $\{I_{t'_j}(\mathbf{p}), f_{t'_j}(\mathbf{p}), g_{t'_j}(\mathbf{p})\}_{j=1:N, t'_j \in \{1:M\}}$ as its candidates, where $f_{j'}(\mathbf{p})$ and $g_{j'}(\mathbf{p})$ are extracted from patch $\mathbf{p}$ of the $j'$th warped support image and warped annotation. The candidate patch set for patch $\mathbf{p}$ is selected by:

$$\underset{\substack{t'_{1:N} \\ t'_j \in \{1:M\}}}{\operatorname{argmin}} \sum_{t'_j} \sum_{\mathbf{x} \in \mathbf{p}} (f_I(\mathbf{x}) - f_{t'_j}(\mathbf{x}))^2. \qquad (3)$$

Based on (3), a patch in input image $I$ is matched with its $N$ closest patches from different support images according to their SIFT feature similarity scores. As shown in Fig. 4-b, the $N$ green patches of support images are selected as the candidate patches for the green patch of the input image. Similarly, the red patches of support images correspond to the red one of the input image. The candidate patch sets of different patches from image $I$ are different since they may come from different support images. In contrast, the global matching based candidate set is same to all the pixels of $I$.

Via the candidate patch set $\{I_{t'_j}(\mathbf{p}), f_{t'_j}(\mathbf{p}), g_{t'_j}(\mathbf{p})\}_{j=1:N}$ for patch $\mathbf{p}$, a voting strategy is used for local saliency:

$$S_l(\mathbf{p}) = \frac{1}{N} \sum_{t'_j} g_{t'_j}(\mathbf{p}). \qquad (4)$$

More specifically, we resize the input image, the warped support images, and their annotations to a quarter of the original size, thus one pixel in the resized image corresponds to a $4 \times 4$ patch of the original image. We compute the SIFT feature of the resized image, and the SIFT feature of each pixel in the resized image is treated as the SIFT feature of the corresponding $4 \times 4$ patch of the original image.

Evidently, these two correspondence-based saliency cues are complementary, therefore we merge $S_g$ and $S_l$ into a transitional saliency $S_{Ini}$ via:

$$S_{Ini} = S_g \cdot S_l. \qquad (5)$$

Example results of the correspondence-based saliency estimation are given in Fig. 5.

### C. Saliency Refinement via Foreground and Background Cues

Our correspondence-based transitional saliency estimation can roughly infer the position of the foreground and the background. For precisely separating salient object from the background, we introduce two types of saliency distance measures based on foreground and background cues. The first one $d_1$ is based on a principle that a pixel which is spatially closer to salient pixels should have a higher saliency value. That is because the salient object regions are usually relatively compact in spatial distribution. The second saliency distance $d_2$ is based on the observation that a pixel which is more different with unsalient regions should gain higher saliency. Both $d_1$ and $d_2$ terms are explained next.
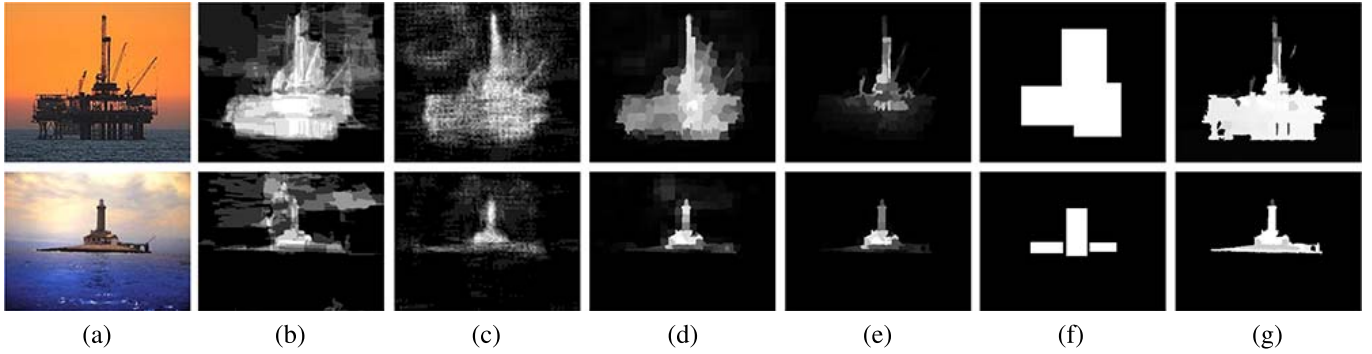
Fig. 5. Saliency maps obtained at each phase of our method. (a) Input image $I$. (b) Global correspondence-based saliency $S_g$ via (1). (b) Local correspondence-based saliency $S_l$ via (4). (d) Initial saliency $S_{Ini}$ via (5). (e) Refined saliency estimation $S_{Ref}$ via (11). (f) Object's rectangular area **W** indicated by the white area. We found the $K = 3$ is a suitable number to represent salient regions in our model. (g) Final saliency $S_{Fin}$ using RWR via (17).

We oversegment the image $I$ into superpixels via the SLIC algorithm [32] for computational efficiency. For region $r_i$, we express the saliency measurements via the above distances $d_1$ and $d_2$ as:

$$D(r_i) = d_1(r_i) \cdot d_2(r_i). \qquad (6)$$

Inspired by the first observation, we design a saliency bias, which enhances the saliency value of the regions near the saliency center while suppresses the saliency of the regions far away from the saliency center. Such a saliency bias is expressed by a two-dimensional Gaussian distribution $G(x, y | \mu_x, \mu_y, \sigma_x, \sigma_y)$. For a region $r_i$ in an image, the first distance $d_1$ is defined as:

$$d_1(r_i) = G(x_{r_i}, y_{r_i} | \mu_x, \mu_y, \sigma_x, \sigma_y), \qquad (7)$$

where $(x_{r_i}, y_{r_i})$ indicates the coordinate of the center of region $r_i$. The center $(\mu_x, \mu_y)$ of Gaussian distribution $G(x, y)$ is computed as:

$$(\mu_x, \mu_y) = \frac{\sum_{\mathbf{x}} \exp(\theta \cdot S_{Ini}(\mathbf{x})) \cdot \mathbf{x}}{\sum_{\mathbf{x}} \exp(\theta \cdot S_{Ini}(\mathbf{x}))}. \qquad (8)$$

In above equation, $(\mu_x, \mu_y)$ is computed as the geometric centroid of the pixels weighted by $\exp(\theta \cdot S_{Ini}(\mathbf{x}))$. When $\theta$ is large, the pixels with large saliency are emphasized. While $\theta$ is set as small as zero, (8) is reduced to the *center-bias* prior, which is based on the fact that the objects near the center of an image are more likely to be salient. We set $\theta = 1$ to allow a balance between compactness of salient regions and the center-bias prior. The horizontal variance $\sigma_x$ and the vertical variance $\sigma_y$ of Gaussian distribution $G(x, y)$ are computed as follows:

$$\sigma_x = \left( \frac{\sum_{\mathbf{x}} \exp(\theta \cdot S_{Ini}(\mathbf{x})) \cdot (x - \mu_x)^2}{\sum_{\mathbf{x}} \exp(\theta \cdot S_{Ini}(\mathbf{x}))} \right)^{\frac{1}{2}},$$

$$\sigma_y = \left( \frac{\sum_{\mathbf{x}} \exp(\theta \cdot S_{Ini}(\mathbf{x})) \cdot (y - \mu_y)^2}{\sum_{\mathbf{x}} \exp(\theta \cdot S_{Ini}(\mathbf{x}))} \right)^{\frac{1}{2}}. \qquad (9)$$

For the second principle, we measure the saliency of a region by its shortest distance to the boundary regions on geodesics. Geodesic distance is a powerful measurement for saliency detection [18], [31]. Additionally, our correspondence-based saliency offers an indication of a boundary region whether it belongs to the background or not.

We then construct an undirected weighted graph by connecting all adjacent superpixels $(r_i, r_j)$ and assigning their weights $w(r_i, r_j)$ as the Euclidean distance (normalized to [0, 1]) between their mean colors. Following [31], a virtual node $v$ is added to connect all boundary regions $[r_g]$ and we define $w(v, r_g) = S_{Ini}(r_g)$. The second type of saliency distance $d_2$ is defined as the geodesic distance between superpixel $r_i$ and virtual node $v$:

$$d_2(r_i) = \min_{C_{r_i,v}} \sum_{p,q} w(p, q), \qquad p, q \in C_{r_i,v}, \qquad (10)$$

where $C_{r_i,v}$ is a path connecting nodes $r_i$ and $v$. $d_2$ for region $r_i$ is computed as the accumulated distance along its shortest path to the virtual node.

As mentioned earlier, these two measurements $d_1$ and $d_2$ are complementary. The former explores saliency of a region via its spatial distance to the saliency center, while the latter exploits the saliency in an opposite view based on its geodesic distance to the background. Then, the refined saliency estimation is:

$$S_{Ref} = S_{Ini} \cdot D, \qquad (11)$$

where term $D$ is defined in (6) based on $d_1$ and $d_2$. Example refined saliency results are shown in Fig. 5-e.

### D. RWR-Based Final Saliency Derivation

While most saliency maps produced in Sec. III-C well identify the salient object and the background, there are still partial foregrounds of the saliency maps that are not uniformly highlighted, which can be seen in the two examples of Fig. 5. To alleviate this issue, we extract foreground and background samples from previous saliency results $S_{Ref}$ and apply RWR to generate final accurate saliency.

Our intuition here is straightforward. We aim to simultaneously use foreground and background samples into a graph based segmentation method to obtain spatially consistent results. RWR is a variant of the conventional random walk, and has been widely employed in several applications [19], [42], including data mining [33] and image segmentation [34]. Image $I$ is represented as an undirected, weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with superpixels as nodes $\mathcal{V}$. Edge $e_{ij}$ in the edge set $\mathcal{E}$ connects adjacent superpixels $r_i$ and $r_j$ in $\mathcal{V}$. Edges $\mathcal{E}$ are

weighted by an affinity matrix $\mathcal{W} = [w_{ij}]$, which is defined as a typical Gaussian weighting function [34]:

$$w_{ij} = \exp(-\frac{\|c_i - c_j\|^2}{\tau}), \tag{12}$$

where $c_i$ indicates the colors of region $r_i$ and $\tau$ is a scale parameter. RWR iteratively transmits to its neighborhood according to the transition probability which is proportional to the edge weight between them. The transition probability $p_{ij}$ between nodes $i, j$ is defined as: $p_{ij} = w_{ij}/\sum_k w_{kj}$. The transition matrix $\mathbf{P} = [p_{ij}]$ is computed by normalizing each column of affinity matrix $\mathcal{W}$.

After each walk, the random walker returns to the starting node with a *restart* probability $\epsilon$. Let $\mathbf{b}^l = [b_1^l, \dots, b_i^l, \dots]^T, l \in \{0, 1\}$ be the *restart* distribution for the random walker, the random walk process converges and the random walker finally has a *stationary distribution* $\boldsymbol{\pi}^l = [\pi_1^l, \dots, \pi_i^l, \dots]^T$. The $i$-th element $\pi_i^l$ is the probability that the random walker stays at node $i$ in the equilibrium condition. The stationary distribution $\boldsymbol{\pi}^l$ can be obtained via:

$$\boldsymbol{\pi}^l = \epsilon(\mathbf{I} - (1 - \epsilon)\mathbf{P})^{-1}\mathbf{b}^l. \tag{13}$$

A simple strategy that derives the foreground and background samples via thresholding the saliency map is not an excellent choice. That is merely because that RWR is sensitive to the restart distribution; a low-quality distribution often leads to unfavorable results. Therefore, we design a more intelligent strategy that extracts a reliable restart distribution from a saliency map. We first extract salient regions $R^s = [r_i^s]$ according to our saliency result $S_{Ref}$, where $S_{Ref}(r_i^s) > mean(S_{Ref})$. We employ the K-means algorithm to divide the salient regions into $K$ clusters according to the coordinates of their centers. We empirically set the number of boundary clusters $K = 3$ in this paper.

The salient superpixels belonging to cluster $k$ is represented as $R_k^s = [r_{i,k}^s]$, where $r_{i,k}^s \in R^s$ and $k = 1, 2, \dots, K$. For each cluster $k$, we build an object rectangle $W_k$, where the center $(x_{W_k}, y_{W_k})$ of $W_k$ is the center of cluster $k$. The width $w_k$ and height $h_k$ of $W_k$ are defined as:

$$w_k = min\Big(\frac{2}{|R_k^s|}\big(\sum_p(x_p - x_{W_k})^2\big)^{\frac{1}{2}}, \ 0.3 \times w_I\Big),$$

$$h_k = min\Big(\frac{2}{|R_k^s|}\big(\sum_p(y_p - y_{W_k})^2\big)^{\frac{1}{2}}, \ 0.3 \times h_I\Big). \tag{14}$$

where $(x_p, y_p)$ denotes the coordinate of the center of the salient region $p \in R_k^s$, and $w_I$ and $h_I$ are the width and height of image $I$. Based on (14), the salient object is represented by $K$ components, which can be observed in Fig. 5-f. This is beneficial for representing salient objects with complex structures on the one hand, while occupying a smaller portion of the background, on the other.

The height and width of rectangle $W_k$ is defined as twice as the mean $\ell_2$-normalized distance from the regions of $R_k^s$ to the center of cluster $k$. For a region $r_i$ outside the object rectangle area $\mathbf{W} = [W_k]$, we define its restart distribution $b_i^l$ as:

$$b_i^l = \begin{cases} 0 & \text{if } l = \text{foreground}; \\ 1 & \text{if } l = \text{background}. \end{cases} \tag{15}$$

For a region $r_i$ inside the object rectangle area $\mathbf{W} = [W_k]$, its restart distribution $b_i^l$ is defined as:

$$b_i^l = \begin{cases} S_{Ref}(r_i) & \text{if } l = \text{foreground}; \\ 0 & \text{if } l = \text{background}. \end{cases} \tag{16}$$

We bias the regions outside the salient window $\mathbf{W}$ with high probabilities for the background by (15) while give the regions inside the salient window $\mathbf{W}$ respectively conservative tendency for the foreground by (16). This is mainly due to the fact that our saliency estimation tends to be more accurate for the background than foreground. Additionally, we observe that the background areas usually are much larger than salient regions (on average, 4~6 times larger according to statistics from typical saliency datasets). According to the stationary distribution in (13), the final saliency for $r_i$ is computed as:

$$S_{Fin}(r_i) = \frac{\pi_i^{fore}}{\pi_i^{fore} + \pi_i^{back}}. \tag{17}$$

Some results of our RWR algorithm based saliency optimization are presented in Fig. 5-g.

## IV. EXPERIMENTAL RESULTS

Our saliency transfer method can identify the salient area in an image by transferring saliency from candidates that share a similar scene with the input. In this section, we provide exhaustive comparison results to demonstrate the effectiveness of our approach. We compare our method to 8 top performing saliency detection methods: geodesic saliency (GS12) [31], saliency filter (SF12) [30], hierarchical saliency (HS13) [35], saliency model via absorbing markov chain (MC13) [14], saliency model via graph-based manifold ranking (MR13) [15], saliency model via robust background detection (wCtr14) [18], saliency model via cellular automata (BSCA15) [21], and saliency model via bootstrap learning (BL15) [36].

*Parameter Settings:* In Sec. III-B, the algorithm retrieves $M$-best support images for the input and selects $N$ closest support images/patches as candidates for voting saliency. We empirically set $M = 50$ and $N = 10$ for all the experiments. In Sec. III-D, an RWR based saliency approach is introduced for generating more accurate saliency results. We set the restart probability $\epsilon = 4 \times 10^{-4}$ of RWR. In our experiments, all the parameters of our algorithm are fixed to unity.

*Datasets:* We mainly evaluate our method on four benchmark datasets: MSRA-5000 [9], ECCSD [35], DUT-OMRON [15] and PASCAL-S [37]. The MSRA-5000 dataset, containing 5000 natural images, is widely used for saliency detection and covers a large variety of image contents. The ECCSD dataset consists of 1000 images with multiple objects with complex structures. Some of the images
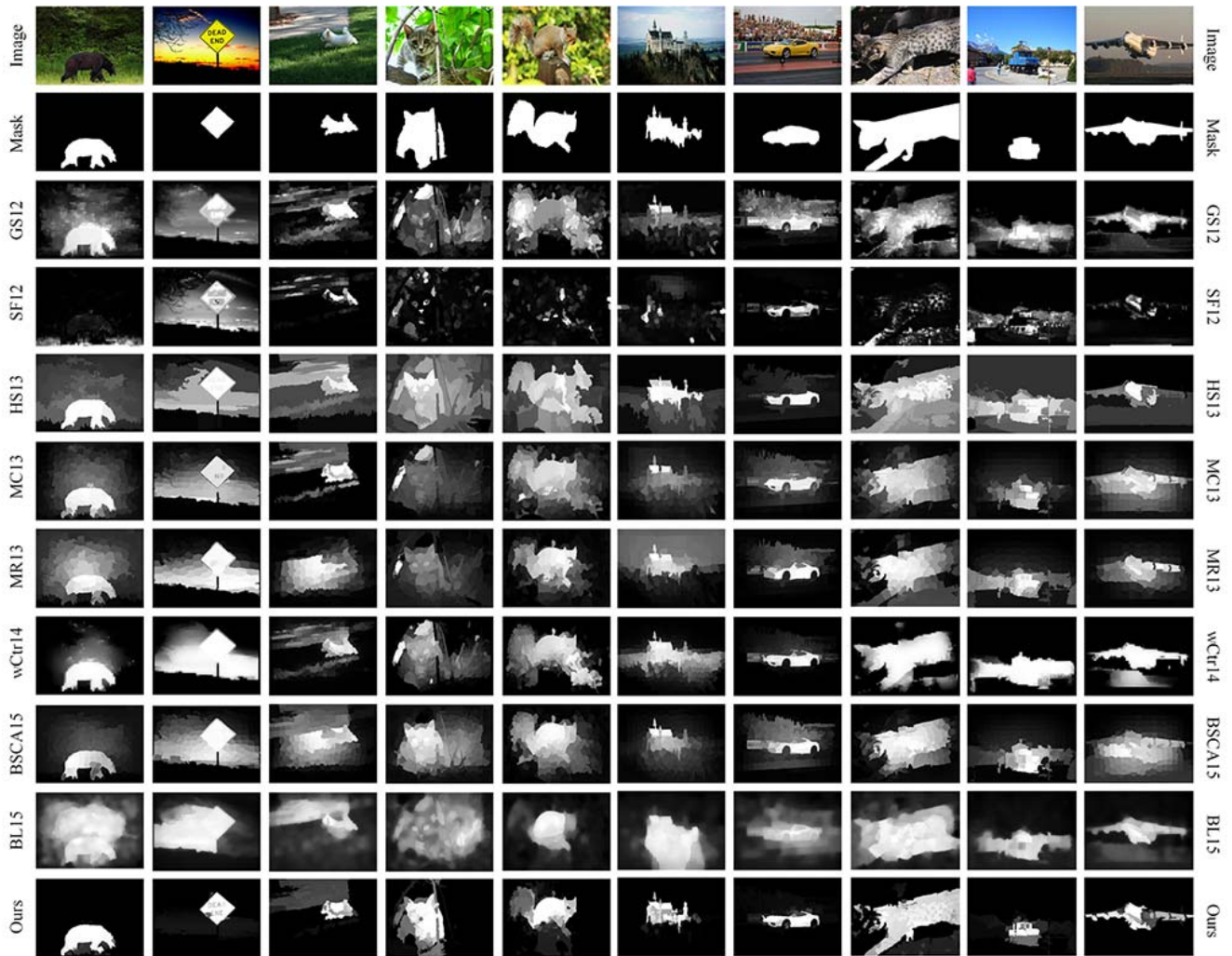
Fig. 6. Comparison of saliency maps with eight state-of-the-art methods. From top to bottom: Input images, ground-truth, saliency maps generated by GS12 [31], SF12 [30], HS13 [35], MC13 [14], MR13 [15], wCtr14 [18], BSCA15 [21], BL15 [36] and our method. Note that the proposed method generates more reasonable saliency maps compared with the state-of-the-art.

come from the challenging Berkeley-300 dataset. The DUT-OMRON dataset is another challenging saliency dataset and contains 5172 images with high background clutter. We further report our performance on the newly developed PASCAL-S dataset [37], which is one of the most challenging saliency benchmarks. It contains 850 natural images where in most cases multiple objects with various locations and scales, and/or highly cluttered backgrounds. Unlike the traditional benchmarks, PASCAL-S is believed to eliminate the dataset design bias. For all the datasets, pixel-wise groundtruth annotation for each image is available. In our experiments, unless stated otherwise, 40% of the images from each dataset are randomly selected for testing. The remaining images are used for transferring saliency.

### A. Performance Comparison

To evaluate the quality of the proposed approach, we provide in this section quantitative comparison for performance of the proposed method against eight top-performing alternatives: GS12 [31], SF12 [30], HS13 [35], MC13 [14], MR13 [15], wCtr14 [18], BSCA15 [21], BL15 [36] on

MSRA-5000 [9], ECCSD [35], DUT-OMRON [15] and PASCAL-S [37] datasets. For a fair comparison, all saliency maps generated using different saliency models are normalized into the same range of [0, 255] with the full resolution of original images.

*1) Qualitative Results:* To provide qualitative comparison of the different saliency outputs, we present results of saliency maps generated by our method and eight state-of-the-art methods in Fig. 6. The top row shows input images. The second row shows the ground truth detection results of salient objects.

We observe that the proposed algorithm captures foreground salient objects faithfully in most test cases. In particular, the proposed algorithm yields good performance on more challenging scenarios, even for objects on image boundaries and blurred backgrounds. This can be attributed to the use of contextual information based on saliency transfer. Thanks to our RWR based optimization, our method is able to detect salient objects accurately despite similar appearance to the background regions. The proposed saliency model can highlight salient object regions more completely with well-defined boundaries, and suppress background regions more effectively compared to previous saliency models.
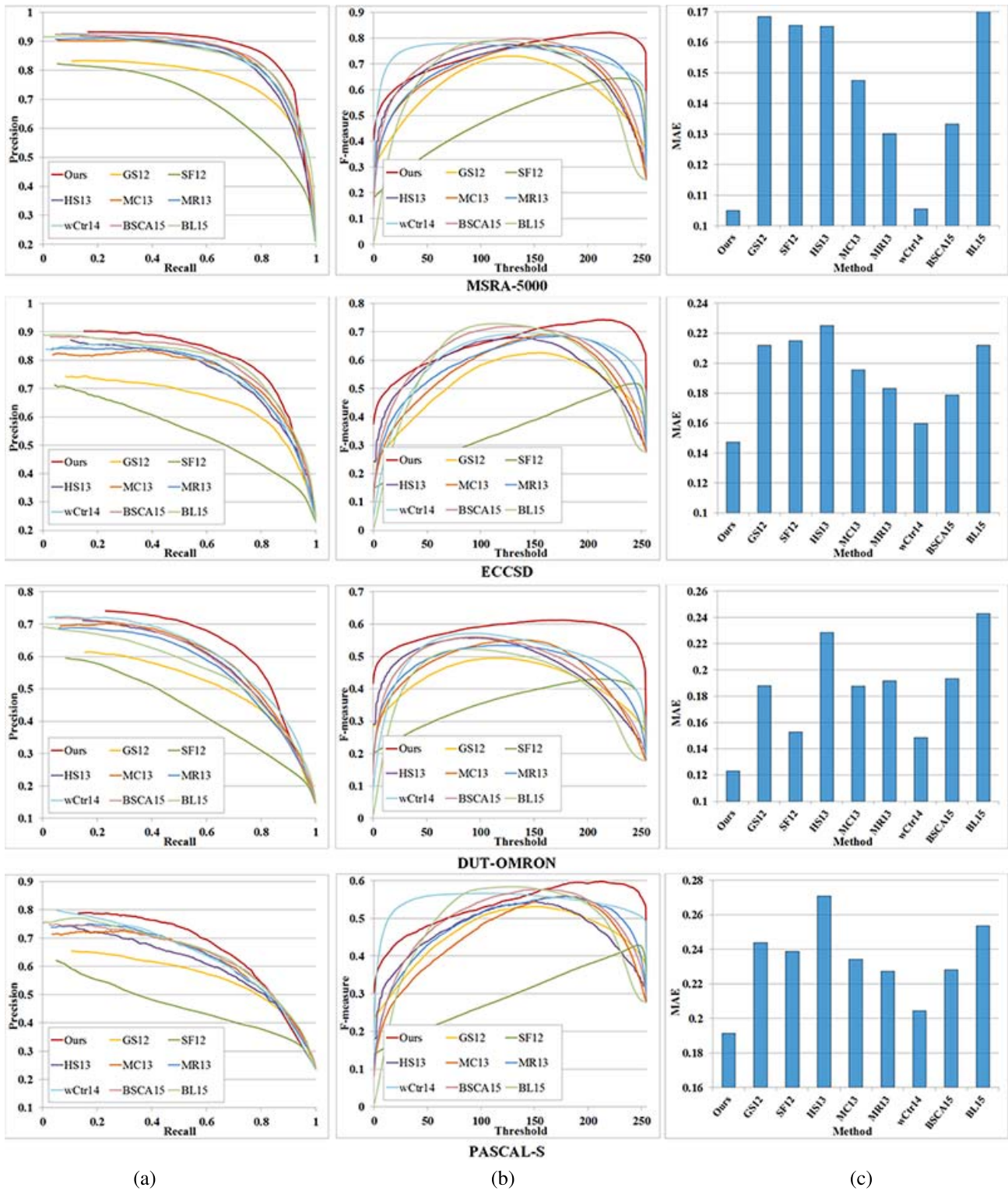
Fig. 7. Statistical comparison with 8 alternative saliency detection methods using MSRA-5000 [9], ECCSD [35], DUT-OMRON [15] and PASCAL-S [37] datasets: (a) PR curves, (b) F-measure, (c) MAE.

*2) Quantitative Results:* Three measures are employed for the quantitative evaluation: precision-recall (PR) curves, F-measure and MAE. We first use precision-recall (PR) curves for performance evaluation. Fig. 7-a shows the PR curves. Our saliency transfer method performs superior on all datasets. The minimum recall value in these curves can be regarded

as an indicator of robustness. As can be seen, minimum recall scores of GS12, SF12, HS13, MC13, wCtr14 and BSCA15 become very small, and the recall scores of MR13 and BL15 shrink to 0. This is because those saliency maps do not correspond well to the ground truth objects. To our advantage, the minimum recall score of our method is about 0.2, which is
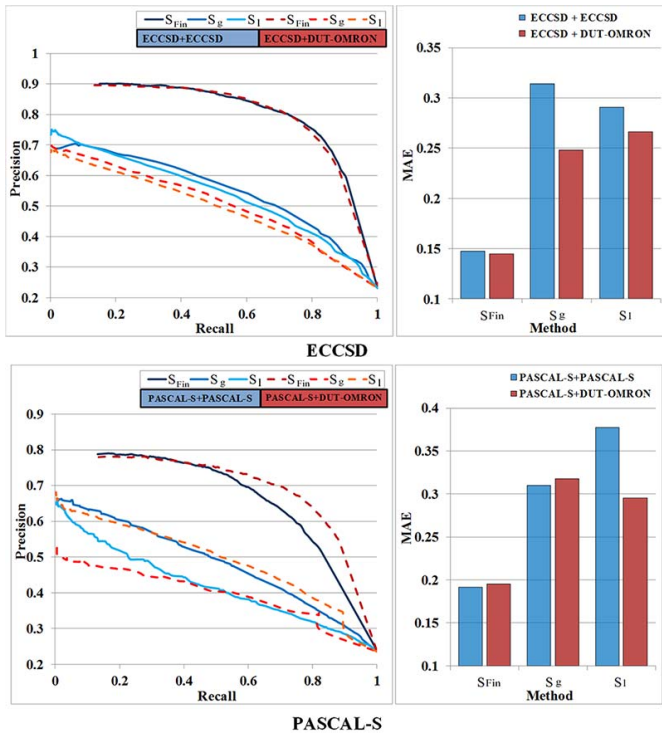
Fig. 8. Cross-dataset validation of the proposed method. **Top**: evaluation results on the ECCSD [35] database. **Bottom**: evaluation results on the PASCAL-S [37] dataset. Fairly close performance in PR curves and MAE with different settings consistently demonstrate effectiveness of our method.

higher than other methods. This demonstrates that our saliency maps align better with the correct objects. In addition, our saliency method achieves the best precision rates over other algorithms, which shows it is more precise and responsive to the actual salient information. The resulting F-measure scores on famous datasets are given in Fig. 7-b. Our method again gives the highest F-measure scores among all approaches, which indicates the effectiveness of the proposed method.

The MAE estimates the approximation degree between the saliency map and the ground truth map, and it is normalized into [0, 1]. The MAE provides a direct way of measuring how close a saliency map is to the ground truth. The MAE results are presented in Fig. 7-c. Our algorithm achieves the lowest MAE scores on the four corresponding datasets, which indicates that the resultant maps are closest to ground truth.

### B. Cross-Dataset Validation

In the above experiments, our approach searches and builds the support group of the input image from the same dataset. To test the generalization of our idea, cross-dataset validation is provided here. We use 40% of the images from the ECCSD [35] dataset, which are the images used in the previous experiment, as the test images. And the DUT-OMRON [15] dataset is used to establish the support group. In Fig. 8, PR curves and MAE in the ECCSD dataset using the DUT-OMRON dataset for transferring saliency are plotted. *ECCSD + ECCSD* indicates the saliency results using the ECCSD to build the support group, which are plotted in blue. *ECCSD + DUT-OMRON* corresponds to the saliency results

using the DUT-OMRON to transfer saliency, which are plotted in red. We can observe that the performance with different settings is fairly comparable. Interestingly, the performance of *ECCSD + DUT-OMRON* in MAE is even slightly better than that of *ECCSD + ECCSD*. A cross-dataset validation is also performed on the PASCAL-S [37] dataset using the same settings with the previoustest on the ECCSD dataset. A similar conclusion can also be drawn from this experiment on the PASCAL-S dataset. All the above observations are further evidence for the generalization and effectiveness of our correspondence-based saliency transfer.

### C. Runtime Analysis

We carry out time analysis on a personal computer equipped with Intel Core 2 Duo E8400 3-GHz CPU and 4GB RAM. The computational cost of our method consists of three parts. The first is our correspondence-based transitional saliency computation in Sec. III-B, including SIFT descriptor based scene matching [22], which typically requires 15s for image and patch level based saliency estimation with $N = 10$. Scene matching [22] occupies almost all the computation time, since the GIST descriptor can be pre-stored for an annotated dataset and scene retrieval takes little time. The second part is the saliency refinement stage in Sec. III-C, including SLIC superpixel segmentation [32], which takes 0.15s. The third is the final stage in Sec. III-D, which costs 0.1s.

## V. CONCLUSION

In this paper, we have presented a novel saliency transfer method to take the advantage of the existing large annotated datasets for identifying the primary and smooth connected salient object areas from an image. The proposed algorithm emphasized the value of contextual information through transferring the saliency from candidate images and patches to an input image using dense scene matching. Based on pixel-wise correspondences, we warp support images and the annotations. The warped image was used to transfer its warped annotations and infer the saliency of the input image. Aiming to select closest support images and exclude the images with unsatisfactory correspondences, we introduced two matching strategies that are based on scene level and patch level respectively. Based on the saliency transferred from those selected image or patch candidates, we refined the saliency estimation of a region according to its distance to the saliency center and the geodesic distance to the background. Furthermore, accurate saliency maps were finally generated via the RWR algorithm. Extensive experiment results on four benchmark datasets showed that the proposed method achieves superior performance compared with the state-of-the-art techniques.

## REFERENCES

[1] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 2, no. 1, pp. 97–136, Jan. 1980.

[2] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," in *Matters of Intelligence*, 1987, pp. 115–141.

[3] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, Sep. 2012.

[4] C. Jung and C. Kim, "A unified spectral-domain approach for saliency detection and its application to automatic object segmentation," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1272–1283, Mar. 2012.

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

[7] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.

[8] J. Li, Y. Tian, T. Huang, and W. Gao, "Multi-task rank learning for visual saliency estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 5, pp. 623–636, May 2011.

[9] T. Liu, J. Sun, N.-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.

[10] A. Borji, D. N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 470–477.

[11] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 409–416.

[12] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2010, pp. 2376–2383.

[13] D. A. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 2214–2219.

[14] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing Markov chain," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1665–1672.

[15] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2013, pp. 3166–3173.

[16] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3395–3402.

[17] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: A data-driven approach," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2013, pp. 1131–1138.

[18] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 2814–2821.

[19] J. Shen, Y. Du, W. Wang, and X. Li, "Lazy random walks for superpixel segmentation," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1451–1462, Apr. 2014.

[20] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, Jan. 2011.

[21] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 110–119.

[22] J. Kim, C. Liu, F. Sha, and K. Grauman, "Deformable spatial pyramid matching for fast dense correspondences," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2013, pp. 2307–2314.

[23] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[24] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[25] H. Li, H. Lu, Z. Lin, X. Shen, and B. Price, "Inner and inter label propagation: Salient object detection in the wild," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3176–3186, Oct. 2015.

[26] W. Einhäuser and P. König, "Does luminance-contrast contribute to a saliency map for overt visual attention?" *Eur. J. Neurosci.*, vol. 17, no. 5, pp. 1089–1097, Mar. 2003.

[27] P. Reinagel and A. M. Zador, "Natural scene statistics at the centre of gaze," in *Proc. Netw., Comput. Neural Syst.*, Jul. 1999, pp. 341–350.

[28] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2013, pp. 1529–1536.

[29] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 497–504.

[30] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 733–740.

[31] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2012, pp. 29–42.

[32] 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.

[33] H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," in *Proc. Int. Conf. Data Mining (ICDM)*, 2006.

[34] T. H. Kim, K. M. Lee, and S. U. Lee, "Generative image segmentation using random walks with restart," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 264–275.

[35] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2013, pp. 1155–1162.

[36] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1884–1892.

[37] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille, "The secrets of salient object segmentation," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 280–287.

[38] W. Wang, J. Shen, X. Li, and F. Porikli, "Robust video object cosegmentation," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3137–3148, Oct. 2015.

[39] 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.

[40] W. Wang and J. Shen, "Higher-order image co-segmentation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1011–1021, Jun. 2016.

[41] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, and D. Tao, "Stacked convolutional denoising auto-encoders for feature representation," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2536638.2016.

[42] X. Dong, J. Shen, L. Shao, and L. Van Gool, "Sub-Markov random walk for image segmentation," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 516–527, Feb. 2016.

[43] X. Li, Q. Guo, and X. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3329–3342, Jul. 2016.

[44] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4185–4196, Nov. 2015.