Salient Object Detection via Multiple Instance Joint Re-Learning

Guangxiao Ma, Chenglizhao Chen, Shuai Li, Chong Peng, Aimin Hao, and Hong Qin

Abstract—**In recent years deep neural networks have been widely applied to visual saliency detection tasks with remarkable detection performance improvements. As for the salient object detection in single image, the automatically computed convolutional features frequently demonstrate high discriminative power to distinguish salient foregrounds from its non-salient surroundings in most cases. Yet, the obstinate feature conflicts still persist, which naturally gives rise to the learning ambiguity, arriving at massive failure detections. To solve such problem, we propose to jointly re-learn common consistency of inter-image saliency and then use it to boost the detection performance. Its core rationale is to utilize the easy-to-detect cases to re-boost much harder ones. Compared with the conventional methods, which focus on their problem domain within the single image scope, our method attempts to utilize those beyond-scope information to facilitate the current salient object detection. To validate our new approach, we have conducted a comprehensive quantitative comparisons between our approach and 13 state-of-the-art methods over 5 publicly available benchmarks, and all the results suggest the advantage of our approach in terms of accuracy, reliability, and versatility.**

Index Terms—**Salient Object Detection Inter-image Correspondence, Multiple Instance Learning, Joint Re-Learning.**

I. Introduction

THE problem of salient object detection is to locate the most eye-attracting object in a given scene. The strongest attractors of the human vision system are stimuli that pop-out from their near surroundings, usually referred as saliency. As one of the most frequently used pre-processing tools, the detected salient object can be fed into various down-stream applications, including adaptive compression of images [1], video surveillance [2]–[4], video saliency [5]–[8], person re-identification [9], medical image analysis [10], image classification [11], object tracking [12], video summarization [13], and video expression [14]. To reveal the image saliency, many conventional methods focus on the designation of the contrast computation over hand-crafted low-level cues, and such implementations are usually time-consuming with noticeable problems, i.e., the contrast itself is in terms of difference between visual cues.

After entering the deep learning era [15], [16], the performance of salient object detection has been significantly improved by using the high discriminative convolutional features. However, due to the varying nature of image scene, those similar regions which belong to different images may be assigned with completely different saliency value if their near surroundings are changed, and see demonstrations in Fig. 1.

In fact, the aforementioned problem is mainly induced by the feature conflicts, which evade the feature margin between the salient foregrounds and its non-salient near surroundings and finally lead to the learning ambiguity. Although various deep architectures have been proposed in recent years [17]–[19], the core rationale of these methods still follows the conventional common thread [20], i.e., by using the deep convolutional features to automatically support contrast computation within a multi-scale/multi-level manner [21], [22].

In terms of the co-saliency community [23]–[26], the problem of learning ambiguity can be slightly alleviated by applying the "inter-image" feature clustering [27] before saliency revealing. However, because these methods conduct their clustering procedure globally to group those sub-regions with similar non-local appearance, the performance improvement toward those hard-to-detect cases (i.e., usually with distinct feature pattern) is still limited. From the perspective of salient object detection over video data [28]–[34], the learning ambiguity problem can be much alleviated due to the newly available "inter-frame" motion clues, which consistently outperform the solely spatial information based on image salient object detection methods. Also noticing the phenomenon that massive easy-to-detect cases might exhibit strong "non-local" similarity to those hard ones, the obstinate learning ambiguity may be alleviated by learning common consistency of its inter-image saliency.

Manuscript received December 28, 2018; revised April 19, 2019 and June 4, 2019; accepted July 5, 2019. Date of publication July 22, 2019; date of current version January 24, 2020. This work was supported in part by National Key R&D Program of China under Grant 2017YFF0106407, in part by the National Natural Science Foundation of China under Grants 61802215 and 61806106, in part by the Natural Science Foundation of Shandong Province under Grants ZR2019BF011 and ZR2019QF009, and in part by the National Science Foundation of USA under Grants IIS-1715985, IIS-0949467, IIS-1047715, and IIS-1049448. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Jian Zhang. *(Guangxiao Ma, Chenglizhao Chen, and Shuai Li contributed equally to this work.) (Corresponding authors: Chenglizhao Chen; Chong Peng.)*

G. Ma is with the Qingdao Research Institute, Beihang University, Qingdao 266100, China (e-mail: mgx@buaa.edu.cn).

C. Chen is with the Qingdao Research Institute, Beihang University, Qingdao 266100, China, and also with the Qingdao University, Qingdao 266071, China (e-mail: cclz123@163.com).

S. Li is with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: lishuai@buaa.edu.cn).

C. Peng is with the Qingdao University, Qingdao 266071, China (e-mail: pchong1991@163.com).

A. Hao is with the Qingdao Research Institute, Beihang University, Qingdao 266100, China, and also with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: ham@buaa.edu.cn).

H. Qin is with the Stony Brook University, Stony Brook, NY 11794 USA (e-mail: qin@cs.stonybrook.edu).

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

Digital Object Identifier 10.1109/TMM.2019.2929943

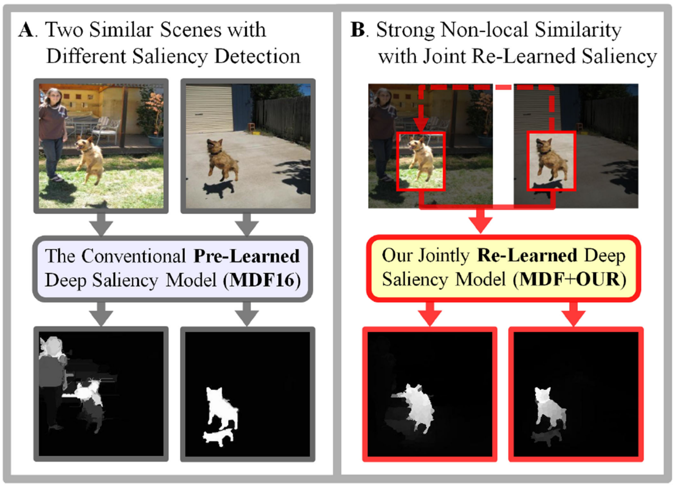
A. Two Similar Scenes with Different Saliency Detection
The Conventional Pre-Learned Deep Saliency Model (MDF16)
B. Strong Non-local Similarity with Joint Re-Learned Saliency
Our Jointly Re-Learned Deep Saliency Model (MDF+OUR)

Fig. 1. The motivation of our method. (A) demonstrates the conventional deep methods which easily output completely different saliency results even for those images with similar non-local regions. (B) shows the results of our joint re-learning rationale, which utilizes those easy-to-detect cases to facilitate those hard ones.

Inspired by all of the above, we propose to jointly re-learn the common consistency of inter-image saliency and then use it to promote the detection performance, and its underneath rationale is illustrated in Fig. 1(B). To achieve this, we interactively consider both the inter-image object-level similarity and the non-local correlation to build the inter-image correspondences. Thus, based on the saliency outputs of the current *off-the-shelf* saliency model, we can focus our saliency model on those hard-to-detect cases to improve the detection performance while using those easy cases with inter-image correspondences to alleviate the learning ambiguity. Specifically, we summarize the salient contributions of this paper as follows:

- We propose to construct the inter-image non-local correspondences, which is indispensable for our subsequent saliency re-learning to reduce the learning ambiguity.
- Our method can simultaneously learn the "most informative" inter-image consistency while suppressing the remaining conflicted inconsistency within the semi-supervised manner.
- We newly design a novel saliency re-learning framework to jointly reveal the common consistency of those hard-to-detect cases while converging towards the saliency-critical learning objective.
- The proposed re-learning scheme is universally effective to boost the detection performance of any existing state-of-the-art saliency model.

II. Related Works

A. The Conventional Methods Using Handcrafted Features

Image saliency detection has been extensively studied over the past decade. Early salient object detection methods [20], [35] follow the bottom-up methodology to assemble low-level saliency clues via elaborately designed contrast computations over the handcrafted features. Following the key rationale that the salient regions in the visual field would first pop out from their surroundings, the conventional saliency revealing schemes are mainly based on the multi-scale and multi-level contrast computation [31], [36], [37].

B. The Deep Learning Network Based Methods

After entering the deep learning era, the automatically computed deep convolutional features have exhibited huge advantages over the conventional handcrafted features. Li *et al.* [21] propose to utilize the convolutional neural network (CNN) based contextual deep features to represent the multi-scale contrast degree to perform superpixel-wise saliency prediction. Lee *et al.* [38] propose to pre-compute the low-level spatial saliency features and then feed it to the fully connected layers enhancing the non-local coherency to improve the detection performance. Liu *et al.* [39] build a two stage network, which respectively performs the coarse salient object location and the saliency detail refinement to achieve multi-level saliency detection. Although large performance improvements have been made, these methods are time-consuming due to the usage of fully connected layers, not to mention that the spatial topology info vanishing can induce performance bottle-neck.

Recently, significant performance improvements have been made in the full convolutional network (FCN) based methods due to the end-to-end flexible/efficient architecture, which can simultaneously conserve the spatial topology info while conducting the pixel-wise saliency predictions efficiently, e.g., Luo *et al.* [40] directly use the FCN providing non-local deep features for the high-quality saliency detection. Meanwhile, the most recent methods follow the rationale that the saliency common consistency of the FCN' inter-layer outputs as the fine-level information can effectively boost the detection performance (i.e., the convolutional steps cause salient object boundary blurring) of those coarse-level layers. Li *et al.* [41] propose to obtain the multi-level contextual info by stacking sequential FCN blocks with varying convolutional scope. Hou *et al.* [22] reveal the multi-scale and multi-level saliency consistency by introducing the short connections from the FCN' inner side layers to the current problem domain. Lu *et al.* [42] propose to utilize a multi-scale context-aware feature extraction module, which contains multiple dilated convolutions, to interactively robust the saliency detection. Wang *et al.* [43] propose to capture hierarchical deep saliency information via using a skip-layer network structure to predict human attention from multiple convolutional layers with various reception fields.

As Wang *et al.* [44] point out, although the usage of multi-scale and multi-level info can indeed boost the detection performance, massive failure detections still exist, e.g., the hollow effects for those large size salient object, the fragmental detections toward those salient object with complex appearance, and the false-alarm detections for those scenarios with distinct non-salient backgrounds. And we believe all these failure detections are mainly induced by the feature conflicts between the salient foregrounds and the non-salient backgrounds, which may easily lead to the learning ambiguity causing the above mentioned defects.

C. The Inter-image Co-Saliency Based Methods

Co-saliency detection aims at finding out multiple salient regions, which usually exhibit strong inter-image common consistency over the handcrafted feature spanned sub feature space, from a group of related images. Almost all the current state-of-the-art co-saliency methods [23]–[26], [45], [46] are following the common bottom-up thread, and the key rationale of these methods is to conduct feature clustering before applying the contrast based saliency revealing. And all these methods follow the assumption that similar regions should exhibit similar saliency detections. By assigning equally saliency value to those pre-grouped sub-regions [47], the overall saliency detection can be potentially boosted.

Liu *et al.* [23] propose to simultaneously conduct multi-level hyper feature clustering (e.g., object prior, regional similarity, boundary connectivity) over the hierarchical segmentations. Then, for each clustered region, the intra image saliency clues are fused as the final saliency estimation. Ge *et al.* [24] propose to utilize a two-round saliency propagation to boost the inter-image co-saliency, i.e., it utilizes the image-level propagation to coarsely locate the co-salient regions, and then the regional-level propagation is conduced to refine previously estimated saliency degree. Different from [23], [24] which mainly focus on the designation of the inter-image clustering procedure, Li *et al.* [25] focus their work on the designation of the fusion scheme, which attempts to utilize both averaging and multiplication to fuse multiple ranking scheme to reveal saliency clues. Further, Ye *et al.* [45] utilize the pre-revealed co-saliency as the binary labels to conduct weakly supervised learning. Then, the saliency predictions from multiple weak classifiers are integrated to boost overall performance. Li *et al.* [48] propose to utilize the pixel-wise correspondences between the input image and an example image to transfer saliency annotations from an existing example onto an input image. Similarly, Ren *et al.* [46] propose to directly transfer saliency values from the retrieved sub-group images using the newly constructed inter-image correspondences. Specially, Zhang *et al.* [26] propose to utilize the CNN based high-level representations to guide the inter-image patch-wise clustering. Meanwhile, by simultaneously seeking the intra-image contrast and the inter-clustering consistency, those salient foregrounds can be easily highlighted while compressing those non-salient backgrounds.

Although many performance improvements have been made by the above mentioned co-saliency methods, there still exists one obstinate limitation which leads the co-saliency methods to reach the performance bottle-neck, i.e., the real saliency degree of those clustered inter-image regions may be totally different (see demonstrations in Fig. 3), which easily lead to the learning ambiguity and then cause massive failure detections. Different from all these co-saliency methods which only focus on their clustering from the appearance similarity perspective, our method further constraints the inter-image alignments to bias toward the saliency consistency.

III. Re-learning Preliminaries

Since we propose to utilize the inter-image common consistency to solve the learning ambiguity problem, we need to acquire accurate inter-image correspondences first. In fact, similar images should have large probability to contain non-local regions with strong correlation toward the saliency assignment. Given an input image (**I**), we utilize the common GIST+HOG descriptor to retrieve a sub group images (**IS**, with maximum image number being 5) which share similar scene topology to the given target image. And only those inter-image correspondences between **I** and **IS** are needed to be built.

A. Bi-Level Inter-Image Alignment

We demonstrate our bi-level inter-image alignment in Fig. 2 (marked with yellow color). Based on the SIFT-Flow [49] providing inter-image pixel-wise correspondences, our alignment procedure interactively considers both the superpixel based mid-level correlation and the object proposal based object-level similarity and see demonstrations in Fig. 4.

We utilize the SLIC [50] to conduct mid-level superpixel decomposition (with total 500 superpixels). And the EdgeBoxes [51] method is adopted to formulate the object-level rectangle proposals (with total 1000 potential proposals). Meanwhile, the pixel-wise inter-image feature distance can be obtained via SIFT-Flow method [49], which can be represented as $||f(i), f(j)||_2$, where f denotes the SIFT-Flow feature representation, i denotes the i-th pixel in the given input image **I**, and j denotes the j-th pixel belonging to the retrieved sub-set images **IS**. Therefore, we can formulate the inter-image mid-level binary alignments $\mathbf{QM} = \{0, 1\}^{\{\phi\times\phi\}}$ and the inter-image object-level binary alignments $\mathbf{QP} = \{0, 1\}^{\{\theta\times\theta\}}$ as Eq. (1), where ϕ and θ respectively denote the aligned superpixel number and the aligned object proposal number.

$$\begin{aligned} \mathop{\arg\min}_{\mathbf{QM},\mathbf{QP}} & \sum_{i=1}^{\phi_{\mathbf{I}}} \sum_{j=1}^{\phi_{\mathbf{IS}}} \mathbf{QM}(i,j) \cdot d(S_i, S_j) \\ & + \sum_{k=1}^{\theta_{\mathbf{I}}} \sum_{l=1}^{\theta_{\mathbf{IS}}} \mathbf{QP}(k,l) \cdot g(P_k, P_l), \\ \text{s.t.,} \quad & \mathbf{QM} \times 1^{\{\phi\times 1\}} = 1^{\{1\times\phi\}},\ \mathbf{QP} \times 1^{\{\theta\times 1\}} = 1^{\{1\times\theta\}}, \end{aligned} \tag{1}$$

Here S_i denotes the i-th superpixel, and P_k denotes the k-th object proposal. In fact, the first term Eq. (1) measures the mid-level correlation between the given input image **I** and one of its retrieved images **IS**, and the feature distance d can be detailed as Eq. (2).

$$d(S_i, S_j) = \frac{\sum_{u\in S_i} \sum_{v\in S_j} \mathbf{QS}(u,v) \cdot ||f(u), f(v)||_2}{\sum_{u\in S_i} \sum_{v\in S_j} \mathbf{QS}(u,v)}, \tag{2}$$

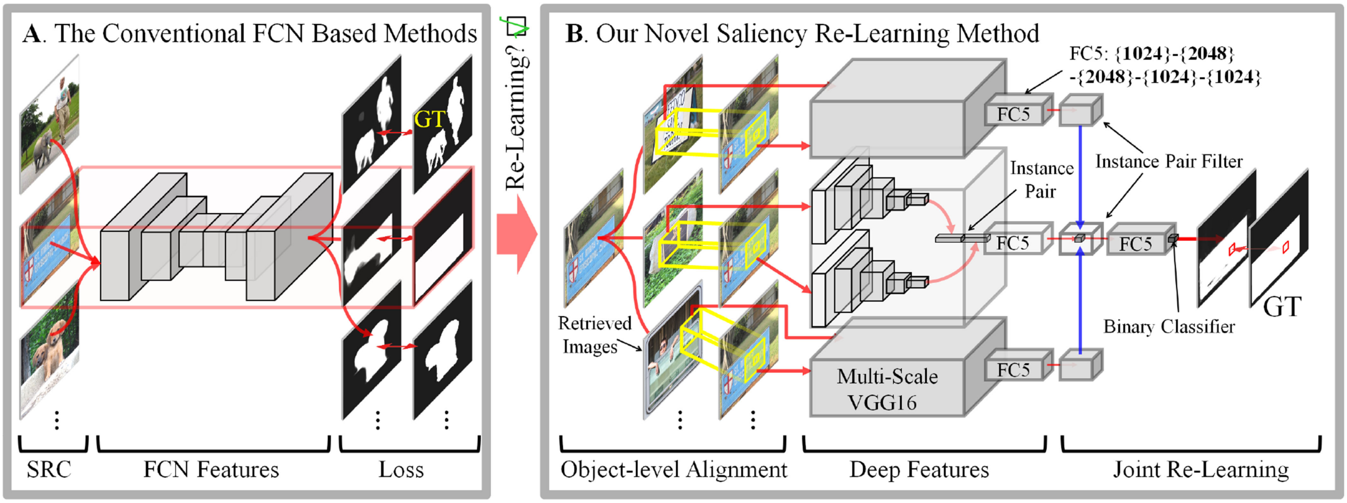

Fig. 2. Method Overview. (A) The current commonly-used approaches are frequently using the FCN method to conduct end-to-end training for image saliency detection. (B) Our method is to reveal the common consistency between the current image and other similar images which contain similar non-local regions with the current salient foregrounds.

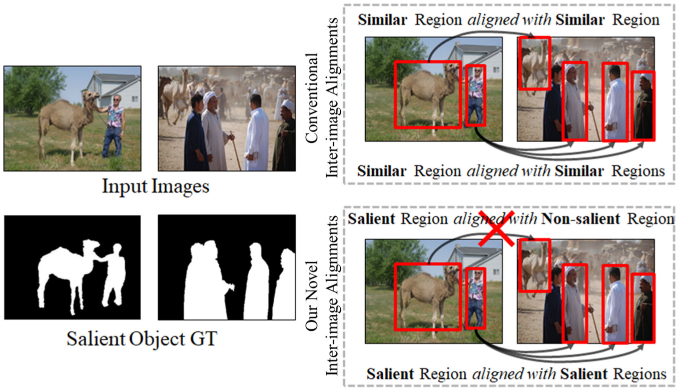
Input Images
Salient Object GT
Conventional Inter-image Alignments
Similar Region aligned with Similar Region
Similar Region aligned with Similar Regions
Our Novel Inter-image Alignments
Salient Region aligned with Non-salient Region
Salient Region aligned with Salient Regions

Fig. 3. Comparison of our inter-image alignments and the co-saliency methods frequently adopted alignments. The conventional methods only focus on the appearance similarity, while our method further bias toward the inter-image saliency consistency. Actually, our method can effectively avoid to align salient regions with non-salient regions, even those regions are similar in appearance.

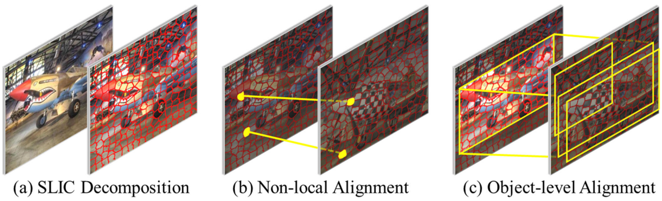
(a) SLIC Decomposition (b) Non-local Alignment (c) Object-level Alignment

Fig. 4. The demonstration of our bi-level inter-image alignment.

where $\mathbf{QS} = \{0,1\}^{\{W\times H\}}$ denotes the SIFT-Flow guiding pixel-wise correspondences, and W and H respectively represent the image width and height.

Meanwhile, the right term Eq. (1) denotes the object-level similarity, and its behind rationale is that those inter-image aligned suerpixels should exhibit strong object-level similarity, which can be formulated as Eq. (3).

$$g(P_k, P_l) = \sum_{u\in P_k}\sum_{v\in P_l} d(S_u, S_v). \tag{3}$$

To this end, we can obtain the inter-image object-level alignment info $\mathbf{Q}$ by pursuing the optimization over Eq. (1), which can be effectively solved via the one-fixed-the-other-solved iterations. Since both parts of Eq. (1) are mutually constraint with each other, for each iteration steps, we respectively perform the top-K (i.e., we initially set K $=$ 10) best alignments alternatively. Meanwhile, to guarantee the convergency, we shrink the slack variable K←K $-$ 1 after each iteration.

Also, in order to strike the trade-off between performance and computation cost of Eq. (1), we simultaneously consider both the objectness score and the overlapping rate to initially select V object-level rectangle proposals for each image from the Edgeboxes provided potential proposals. Therefore, our initially selected object-level proposals can simultaneously exhibit high objectness score while averagely distributed over the entire image. Meanwhile, the total number of the selected object-level proposals (we empirically assign $V = 100$) should also be as few as possible while maintaining informative to constraint accurate superpixel-level alignment, which can be detailed as following:

$$\{P^*\} = \mathop{\arg\max}_{i,j\in[1,V]} \sum_{i=1}^{V} score_i + \sum_{i\neq j} ||c_i, c_j||_2, \tag{4}$$

where $score_i$ denotes the objectness score of the i-th object-level proposal, and the right term controls the overlapping rate of the selected object proposal (P^*), c_j denotes the center location of the j-th proposal.

Compared to the solely SIFT-Flow correspondence guided inter-image non-local alignment method [52], our method can improve the alignment average correct rate from 69% to 89% for all the adopted benchmark dataset.

B. Multi-Scale Deep Feature Computation

By using the alignment info **QP** (Eq. (1)), we can easily align each object proposal in the given image (we name it as "target proposal") to multiple object proposals from the retrieved sub-group images. Thus, the training dataset can be formulated as object-level instance pairs. As we mentioned previously, our deep model focuses on the inter-image common saliency consistency learning rather than the conventional appearance similarity. So the desired deep features should be able to represent the non-local contrast info of the given instance pair. Therefore, for each instance pair, we utilize VGG16 to automatically compute the high-dimensional deep feature to represent the "target superpixel" which is located at the center location of the given target proposal. Meanwhile, we also utilize the VGG16 deep feature of the near surrounded regions of the target superpixel to represent the non-local contrast. Therefore, we can get a 16384-dimensional features (4096×4) to represent the inter-image object-level contrast info. And the deep feature of the i-th instance pairs can be formulated as : $\mathrm{CF_i} = \{\xi_\mathrm{i}, \xi_\mathrm{j}\}$, where ξ_i presents the two scale convolutional feature (4096×2) of the i-th superpixel in the given target image, and ξ_j also denotes the two scale convolutional feature of the j-th superpixel in the corresponding retrieved sub-group image.

IV. Our Novel Saliency Re-Learning Model

Given a pre-learned saliency model, our method proposes to re-learn those "hard-to-detect instances" whose correct saliency detection failed to be performed by the given saliency model. The behind rationale of our method is that those hard-to-detect instances may be correctively re-learned if its aligned inter-image object-level info is already correctly classified by the given saliency model. To achieve this, our re-learning model should simultaneously acquire the following three attributes:

First, in order to improve the detection performance while avoiding the over-fitting problem, our re-learning model should focus on those hard-to-detect instances;

Second, those hard-to-detect instances aligned info should be potentially able to alleviate the conventional feature conflicts, thus only those Instance Paris that may "potentially bring performance improvements" will be feeded to our re-learning model;

Third, for those hard-to-detect instances, our re-learned saliency model should be able to conduct high-quality image salient object detection.

To full-fill the above mentioned first two aspects, we propose to utilize semi-supervised Instance Pair Filter (Fig. 2) to identify those valuable Instance Pairs. Thus, those filtered Instance Pairs are all with the formulation as: $\{\mathrm{CFhard} + \mathrm{CFeasy}\}$. Finally, our saliency model will be jointly re-learned based on those filtered Instance Pairs, and the overall architecture of our method can be found in Fig. 2.

A. Bi-level Pseudo GT for Image Pool

Since the inter-image alignment performance is positively correlated to the image pool size, we propose to train our Instance Pair Filter within the semi-supervised manner to ensure the scalability of the adopted image pool. Thus, given an existing pre-trained deep saliency model from which our method attempt to re-learn, we propose to utilize a bi-level CRF (Conditional Random Fields) to formulate pseudo-GT of the adopted image pool, and the CRF kernel function between pixel i and pixel j can be formulated as Eq. (5).

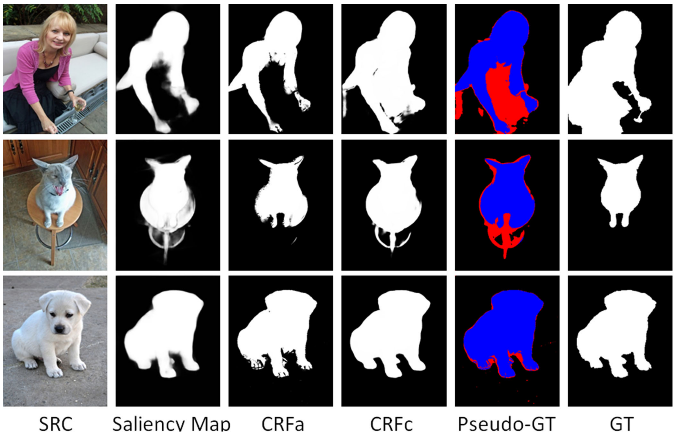

Fig. 5. The demonstration of our bi-level CRF saliency assumptions, where the BLUE represents the pseudo salient regions, the BLACK represents the pseudo non-salient regions, and those filtered regions are marked by RED color.

Based on the saliency predictions of the given deep saliency model, we respectively conduct two CRF binary assumptions with two sets of pre-defined CRF parameters, i.e., the proximity parameter θ_α, the similarity degree parameter θ_β, and the smoothness degree parameter θ_γ.

$$k(i,j) = \underbrace{\omega^{(1)} exp\left(-\frac{|\{x,y\}_i - \{x,y\}_j|^2}{2\theta_\alpha^2} - \frac{|\mathbf{I}_i - \mathbf{I}_j|^2}{2\theta_\beta^2}\right)}_{appearance} + \underbrace{\omega^{(2)} exp\left(-\frac{|\{x,y\}_i - \{x,y\}_j|^2}{2\theta_\gamma^2}\right)}_{smoothness} \tag{5}$$

where the $\{x,y\}_i$ denotes the location of the i-th pixel, $\mathbf{I}_i$ denotes the corresponding color vector, ω is the weighting parameter which is identical to the paper [53]. To formulate pseudo GT for our adopted image pool, we attempt to assign CRF with aggressive $\theta_{\alpha,\beta,\gamma} = \{5, 1, 2\}$ (CRFa) to focus on the binary assumption's accuracy. Meanwhile, we use another CRF with conservative $\theta_{\alpha,\beta,\gamma} = \{200, 50, 20\}$ (CRFc) to bias toward the binary assumption's integrity. Hence, based on the above computed two-level binary saliency assumptions (i.e., CRFa and CRFc), the corresponding pseudo GT can be formulated as followed:

1) We regard the intersections of the obtained two-level binary CRF assumptions (i.e., $(\mathrm{CRFc} \cap \mathrm{CRFa})_+$ as those most trustworthy salient foreground regions and see BLUE regions in Fig. 5.

TABLE I
LABELING SCHEME FOR INSTANCE PAIR FILTER TRAINING

Conditions	Labels
$\mathrm{IOU}(\mathrm{pGT}^{\mathrm{I}}_{\mathrm{i,j}}, \mathrm{GT}^{\mathrm{I}}_{\mathrm{i,j}}) \leq \mathrm{T}_1$ and $\mathrm{IOU}(\mathrm{GT}^{\mathrm{I}}_{\mathrm{i,j}}, \mathrm{pGT}^{\mathrm{IS}}_{\mathrm{i,j}}) > \mathrm{T}_2$	1-Positive
Otherwise	0-Negative

2) We regard the residual between the entire image and the union of two-level binary CRF assumptions as those most trustworthy non-salient background regions and see BLACK regions in Fig. 5.

Also, because of the large uncertainty degree, we filter out those remaining regions and see RED regions in Fig. 5. By using our bi-level pseudo GT, the accuracy lower bound of our inter-image alignment achieves about 81%.

B. Semi-Supervised Instance Pair Filter

We propose to perform weakly supervised learning to train our Instance Pair Filter, i.e., FC5 (see details in Fig. 2) full connected layers followed with one binary classifier, and then we use it to identify those "valuable" instance pairs. Here we formulate the training label of those instance pairs in Table I where $\mathrm{IOU}(\cdot)$ denotes the Intersection Overunion Ratio, GT and pGT respectively denote the real ground truth and the previously obtained pseudo ground truth, superscript I and IS respectively denote the given target image and its retrieved and aligned sub-group images, the subscript i and j respectively indicate the image index and the superpixel index, and $\mathrm{T}_1, \mathrm{T}_2$ are two predefined hard threshold to control the trustworthy degree toward our labeling scheme, which we empirically assign $\mathrm{T}_1 = 15\%, \mathrm{T}_2 = 90\%$ to balance the trade-off between training sample number and the corresponding trustworthy degree. In Table I, the condition $\mathrm{IOU}(\mathrm{pGT}^{\mathrm{I}}_{\mathrm{i,j}}, \mathrm{GT}^{\mathrm{I}}_{\mathrm{i,j}}) \leq \mathrm{T}_1$ can well locate those hard-to-detect cases, and the condition $\mathrm{IOU}(\mathrm{GT}^{\mathrm{I}}_{\mathrm{i,j}}, \mathrm{pGT}^{\mathrm{IS}}_{\mathrm{i,j}}) > \mathrm{T}_2$ ensures those aligned instance, exhibiting identical saliency degree. By using the labeling scheme in Table I, all Instance Pairs have already been assigned with training label (1/0), and we randomly select 30 k positive instance pairs and 50 k negative instance pairs to construct our training dataset to train our Instance Pair Filter. The qualitative demonstrations toward the advantage of our Instance Pair Filter can be found in Fig. 6.

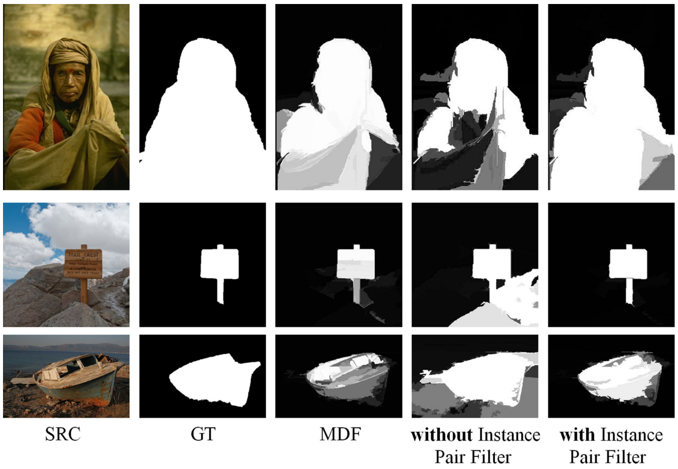
SRC
GT
MDF
without Instance Pair Filter
with Instance Pair Filter

Fig. 6. Demonstrations of the advantage of our Instance Pair Filter.

C. Joint Saliency Re-Learning

Since those filtered instance pairs can well handle the feature conflicts and alleviate the learning ambiguity, we propose to train multiple binary classifiers for the initial saliency estimations. And then, we utilize the joint learning scheme to achieve high-quality final saliency predictions.

We formulate our saliency training dataset (containing N filtered instance pairs) as $(\{\mathrm{CFhard}_{\mathrm{n}}, \ \mathrm{CFeasy}_{\mathrm{m}}\}, \mathrm{Z}_{\mathrm{n}})$, where $n = 1, ..., N, \ 1 \leq m \leq M, \ Z$ denotes the binary saliency ground truth of the n-th superpixel, and M denotes the maximum number of aligned inter-image info, which we empirically set it to 5. Thus, by taking the trust degree as descent ranking order, we can parallel learn m binary saliency classifiers, and L_{m} denotes the corresponding loss. Meanwhile, we utilize the aforementioned FC5 full connected layers followed with an additional binary classifier to assemble these binary saliency classifiers to respect the joint saliency loss $\mathrm{L}_{\mathrm{joint}}$, which can be formulated as follows:

$$\mathrm{L}_{\mathrm{joint}} = \sigma\left(\mathrm{Z}, \mathrm{h}\left(\sum_{\mathrm{m}=1}^{\mathrm{M}} \delta_{\mathrm{m}} \cdot \mathrm{g}_{\mathrm{m}} \cdot \mathrm{w}_{\mathrm{m}}\right)\right), \tag{6}$$

where $\delta_m \in \{0, 1\}$ denotes the output of our Instance Pair Filter, $\sigma(\cdot, \cdot)$ measures the image-level class-based entropy loss [22], $h(\cdot)$ denotes the sigmoid function, g_m denotes the similarity between the target instance and its m-th aligned info (Eq. (3)), and w_m denotes the output of the last FC5 layer when training the m-th binary saliency classifier. Thus, we can formulate the assembled final loss (i.e., totally M+1 classifiers) as Eq. (7) to achieve high-quality saliency prediction.

$$\mathrm{L}_{\mathrm{final}} = \mathrm{L}_{\mathrm{joint}} + \sum_{\mathrm{m}=1}^{\mathrm{M}} \alpha_{\mathrm{m}} \cdot \mathrm{L}_{\mathrm{m}}, \tag{7}$$

where α_m (Eq. (8)) denotes the assemble weight which is inversely correlating to the m-th classifier' error e_m,

$$\alpha_m = \frac{1}{2} \log \frac{1 - e_m}{e_m} \tag{8}$$

To this end, the re-learned final saliency can be predicted via performing the spatial-weighting scheme [20] over the outputs of the re-learned saliency model, and we show the pictorial demonstrations in Fig. 7 and Fig. 8.

D. Differences Between [52] and Our Method

Here we summarize the main differences between [52] and our method in the following 3 aspects:

1) the inter-image correspondences constructed by [52] are solely based on the SIFT-Flow providing pixel -level/patch-level similarly, which may produce incorrect

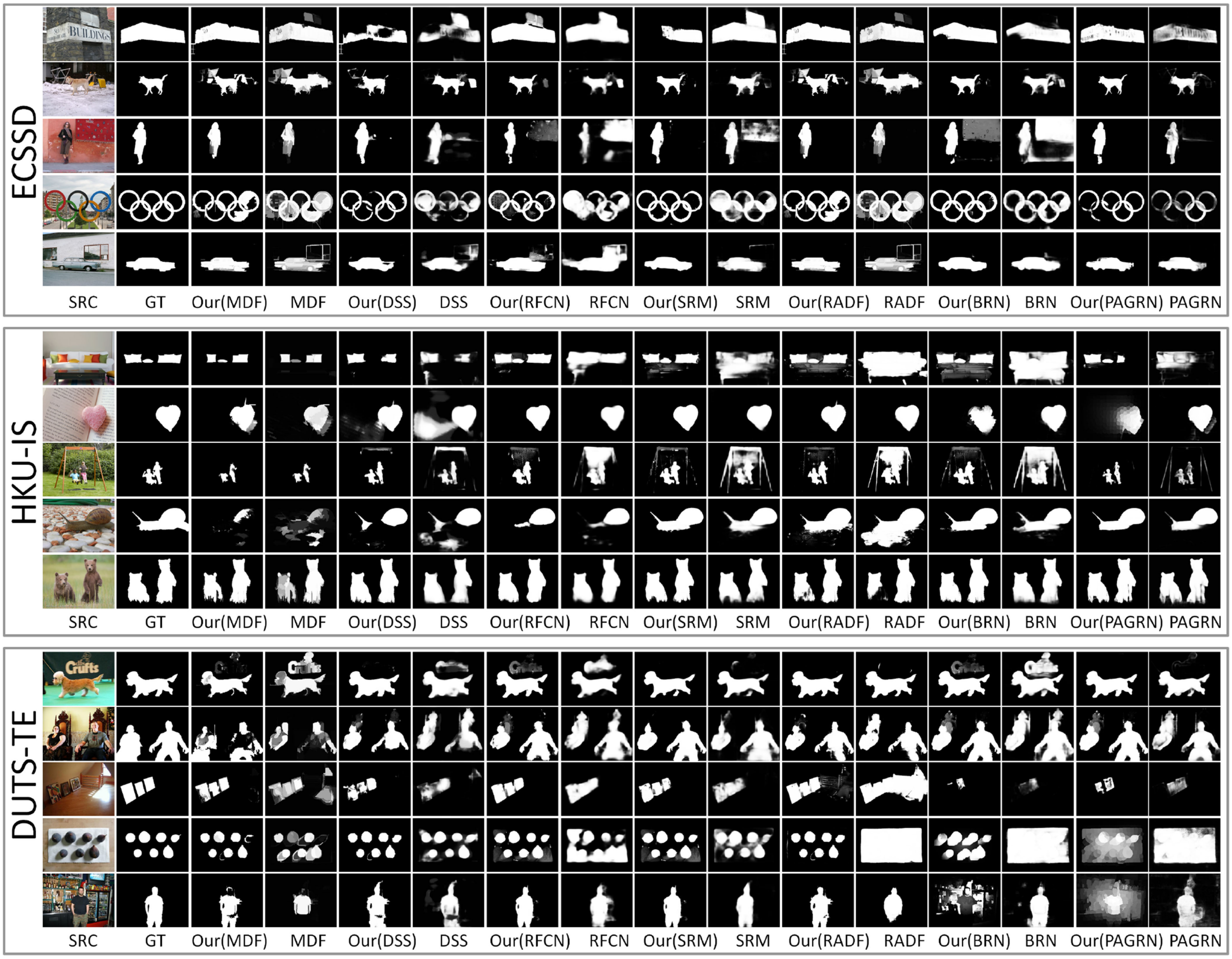

Fig. 7. Visual comparison of saliency maps (Part I). Note that GT stands for Ground truth. Apparently, our method can produces more accurate saliency maps than others.

alignments toward those large inter-image displacement. As for our method, we interactively consider both the mid-level correlation and the object-level similarity to alleviate the displacement induced problem, improving the inter-image correspondences accuracy.

2) our method utilizes the newly designed instance filter to further exclude those inter-image correspondences which may not be really helpful for the saliency re-learning (e.g., those incorrect correspondences), while the [52] is less considerate of such perspective.
3) the core rationale of [52] is to directly formulate the saliency degree via utilizing the inter-image correspondences to transfer saliency labels/anotations from the retrieved images. Thus, [52] needs to know the saliency labels/anotations beforehand, yet such saliency ground truth is not needed in our "semi-supervised" saliency re-learning framework. Also, by increasing the image pool size, we can effectively avoid the situation that the similar images can not be retrieved.

V. Experiments and Results

We have conducted massive quantitative experiments to validate the effectiveness of our method. We also have compared our method with 13 state-of-the-art methods over 5 public available dataset to demonstrate the advantages of our method.

A. Evaluation Datasets

ECSSD [35]. This dataset consists of 1,000 semantically meaningful natural images coupling with complex backgrounds, which includes many semantically meaningful but structurally complex images for evaluation. The images are acquired from the internet and 5 helpers were asked to produce the ground truth masks.

HKU-IS [17]. This dataset consists of 4,447 images, and most of these images simultaneously contain multiple salient objects.

PASCAL-S [64]. This dataset consists of 850 images with multiple salient objects. This dataset was built using the

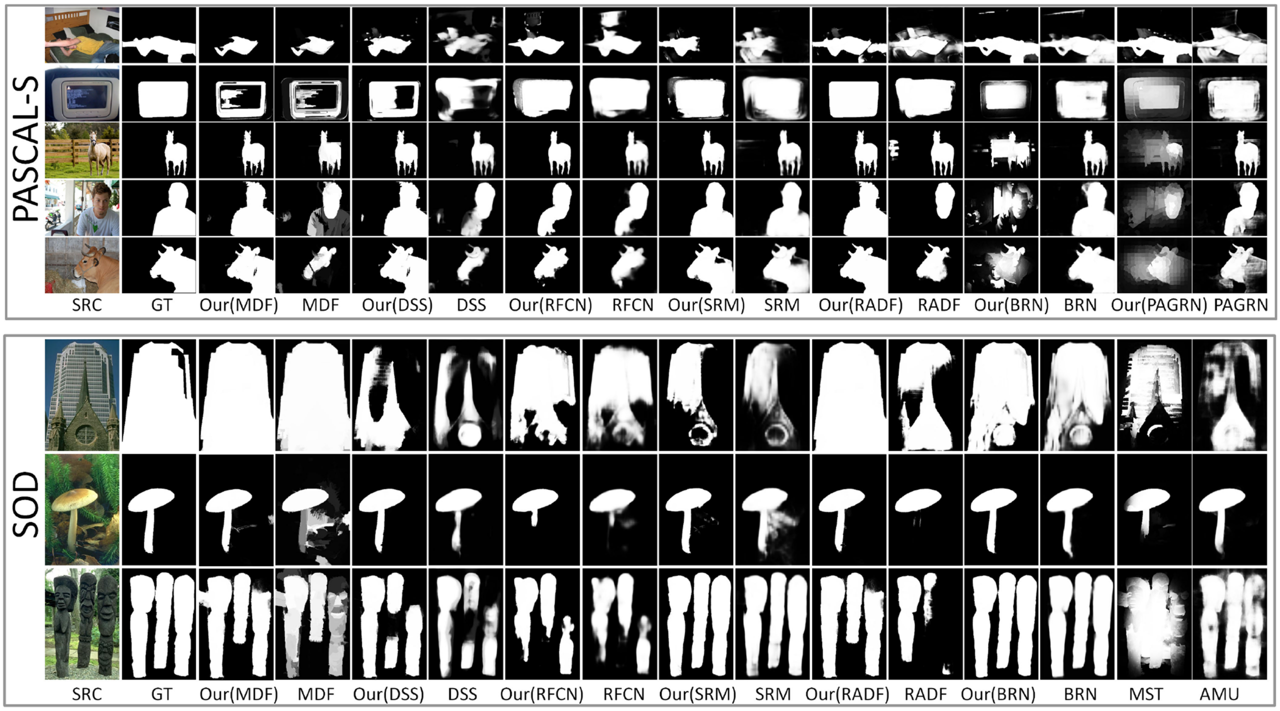

Fig. 8. Visual comparison of saliency maps (Part II). Note that GT stands for Ground truth. Apparently, our method can produces more accurate saliency maps than others.

validation set of the PASCAL VOC 2010 segmentation challenge. The ground truth saliency masks were labeled by 12 subjects.

SOD [65]. This dataset consists of 300 challenging images, and it was originally designed for image segmentation. This dataset is very challenging since many images contain multiple salient objects either with low contrast or overlapping with the image boundary.

DUTS-TE [63]. DUTS dataset is currently the largest saliency detection benchmark, and contains training images (DUTS-TR) and test images (DUTS-TE) with high quality pixel-wise annotations. Both the training and test sets contain very challenging scenarios for saliency detection. This dataset consists of 5,019 images, which has moderate complex and diversified contents.

B. Training Details

Since our bi-level inter-image alignment requires a large number of images to ensure its effectiveness, an additional training dataset (randomly selected 5 K images from DUTS-TR [66]) is adopted to facilitate the re-learning procedure of the original method, i.e., training set of Our(DSS) is composed of the original adopted MSRA-B and the newly adopted DUTS-TR-5 K. We trained the proposed network for 80 K iterations, using Stochastic Gradient Descent (SGD) with a moment 0.9, weight decay 0.005, and learning rate 0.001.

C. Evaluation Metrics

We evaluate our model using four widely adopted metrics as suggested by Borji *et al.* [67] including the precision-recall (PR) curve, the F-measure, Mean Absolute Error (MAE) and Area Under Curve (AUC). Given a predicted saliency map, we perform binary segmentation with hard threshold T, resulting in a pair of precision and recall values when the binary mask is compared against the ground truth. If it is deemed as successful detection, the final precision-recall curves are obtained by varying T from 0 to 255. As the recall rate is inversely proportional to the precision, the tendency of the trade-off between precision and recall can truly indicate the overall video saliency detection performance. And the F-measure is an important performance indicator when the precision rate conflicts the recall rate, which can be computed by

$$\mathrm{F-measure} = \frac{(1+\beta^2) \times \mathrm{Pre} \times \mathrm{Rec}}{\beta^2 \times \mathrm{Pre} + \mathrm{Rec}}, \tag{9}$$

where the Pre and the Rec denote the corresponding precision rate and recall rate respectively. According to the suggestion proposed by Achantay *et al.* [68] , we set $\beta^2 = 0.3$ to weigh precision more than recall. We report the maximum F-measure score among all pairs of precision and recall values. Although commonly used, PR curves have limited value because they fail to consider true negative pixels. For a more balanced comparison, we adopt the MAE as another evaluation criterion. MAE measures the numerical distance between the ground truth and the estimated saliency map, and is more meaningful in evaluating the applicability of a saliency model in a task such as object segmentation. It is defined as the average pixelwise absolute difference between the binary ground truth (GT) and the saliency map (SAL). The MAE evaluates the saliency detection accuracy

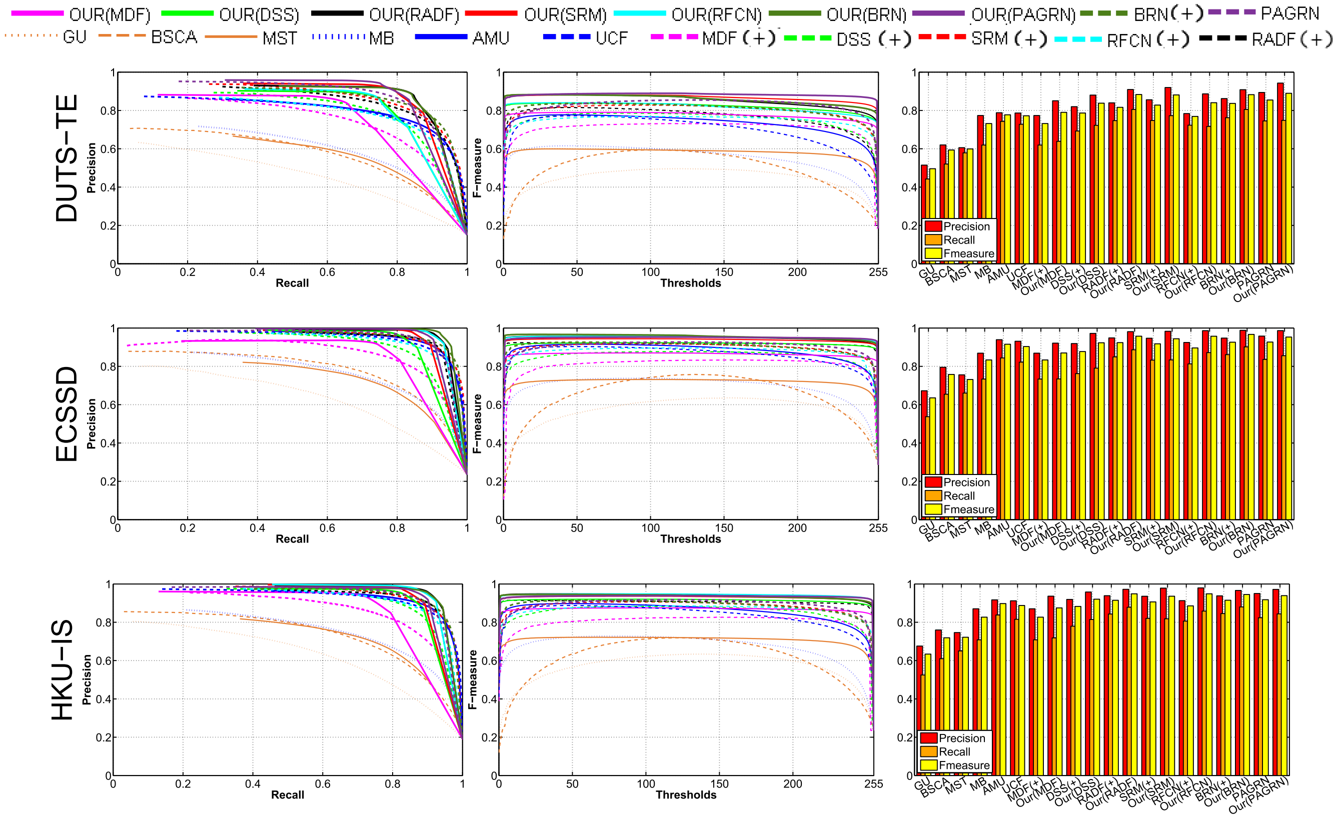

Fig. 9. Quantitative comparisons of the proposed approach and 13 baseline methods including GU13 [54], MB15 [55], BSCA15 [56], MDF16 [21], MST16 [57], DSS17 [22], AMU17 [15], SRM17 [58], UCF17 [59], RADF18 [16] RFCN18 [60], BRN18 [61] and PAGRN18 [62] over **DUTS-TE** [63], **ECSSD** [35], **HKU-IS** [17] dataset. The first two columns respectively demonstrate the PR curves and the F-measure curves under different thresholds, and the last column demonstrate the averaged precision rate and averaged recall rate corresponding to the maximum F-measure. The mark (+) denotes that an additional DUST-TR-5K dataset is added to the training dataset.

by Eq. (10).

$$\mathrm{MAE} = \frac{1}{\mathrm{W} \times \mathrm{H}} \sum_{x=1}^{W} \sum_{y=1}^{H} \left| \mathrm{SAL}(\mathrm{x}, \mathrm{y}) - \mathrm{GT}(\mathrm{x}, \mathrm{y}) \right|, \qquad (10)$$

where W and H represent width and height of the given image, $\mathrm{SAL}(\mathrm{x}, \mathrm{y})$ denotes the saliency value of the pixel with coordinates (x, y), and GT denotes the corresponding binary ground truth. The ROC curve can be conveniently generated according to the true positive rates and false positive rates obtained during the calculation of the PR curve. The AUC is the area under the ROC curve. A perfect model will score an AUC of 1, while random guessing will score an AUC around 0.5.

$$\mathrm{AUC} = \frac{\sum_{\mathrm{i} \in \mathrm{Pos}} \mathrm{Rank_i} - \frac{\mathrm{M} \times (\mathrm{M}+1)}{2}}{\mathrm{M} \times \mathrm{N}}, \qquad (11)$$

where $\mathrm{Rank_i}$ represents the serial order of the i-th element, M is the number of positive samples and N is the number of negative samples.

D. Comparison Results

We have compared our method with 13 state-of-the-art methods, including GU13 [54], MB15 [55], BSCA15 [56], MDF16 [21], MST16 [57], DSS17 [22], AMU17 [15], SRM17 [58], UCF17 [59] , RADF18 [16] RFCN18 [60], BRN18 [61] and PAGRN18 [62]. For fair comparison, the saliency map/executable code of the compared methods are all provided by authors with parameters/implementations unchanged.

To validate the effectiveness of our method, we evaluate the performance of our re-learning scheme by respectively combining 7 most recent state-of-the-art methods, i.e., Our-MDF16, Our-DSS17, Our-SRM17, Our-RADF18, Our-RFCN18, Our-BRN18 and Our-PAGRN18. And both the qualitative and the quantitative results suggest the significant performance improvements, and the details can be found in Fig. 7, Fig. 8, Fig. 9 and Fig. 10.

As we can see from the PR curves and F-measure curves in Fig. 9 and Fig. 10, almost the top-ranked curves are all our re-learned method, i.e., all those re-learned deep model $\mathrm{Our}(\cdot)$ are marked with solid lines, while the original baseline methods are marked with dash lines. Also it can be found in the last column of Fig. 9 and Fig. 10 that our method can consistently achieves almost 4%$\sim$8% F-max score improvements. Meanwhile, as for PR curves with small dynamic thresholds (i.e., with large Rec rate), the Pre rates of our method are also substantially larger than the corresponding baseline methods.

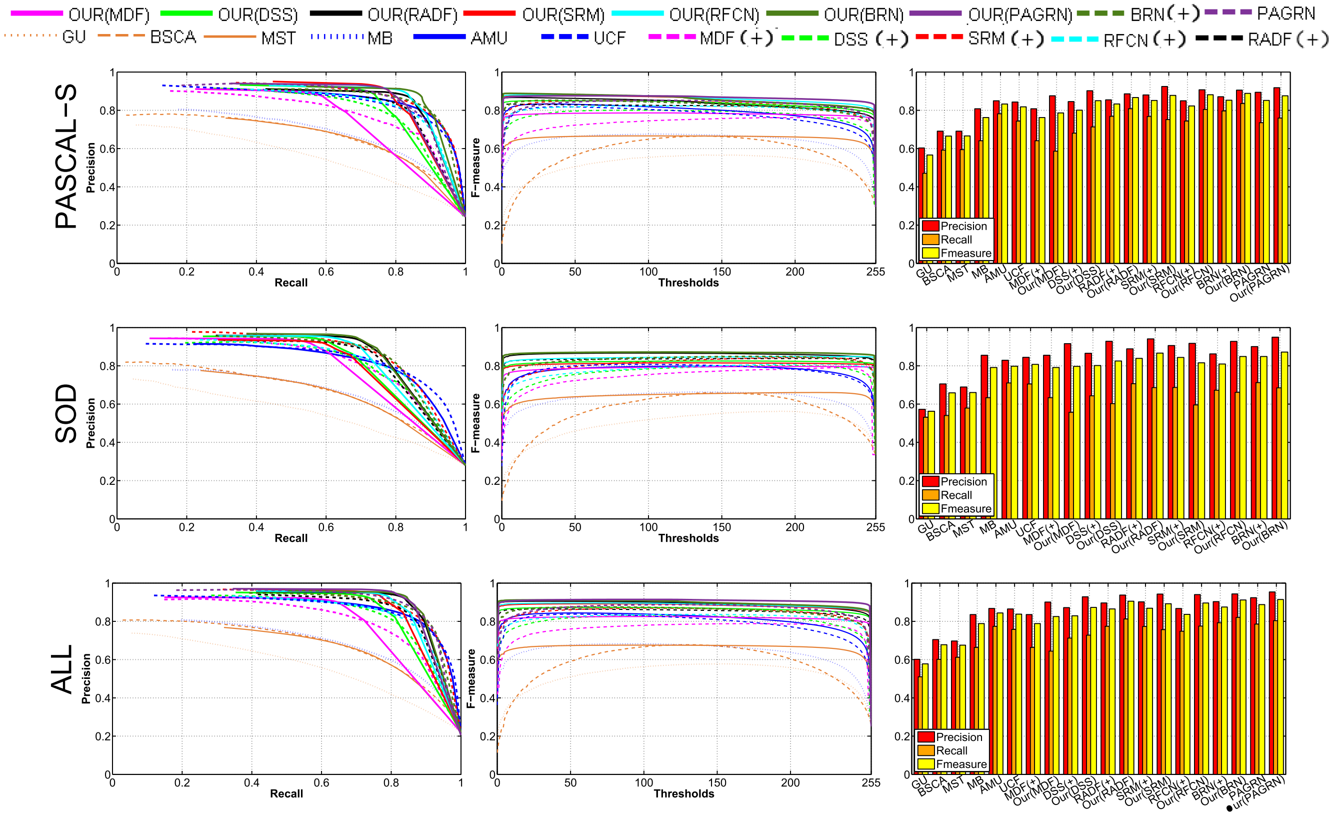

Fig. 10. Quantitative comparisons of the proposed approach and 13 baseline methods including GU13 [54], MB15 [55], BSCA15 [56], MDF16 [21], MST16 [57], DSS17 [22], AMU17 [15], SRM17 [58], UCF17 [59] , RADF18 [16] RFCN18 [60], BRN18 [61] and PAGRN18 [62] over **PASCAL-S** [64] and **SOD** [65] dataset. It should be noted that the result of PAGRN18 over the SOD dataset is not available. The mark (+) denotes that an additional DUST-TR-5K dataset is added to the training dataset.

We also compare our approach with the existing methods in terms of F-meature and MAE scores. The quantitative results are shown in Table II and III. Our approach achieves the best score (F_β and MAE) on easy datasets, such as ECCSD [36]. Besides, we also observe that the proposed approach behaves even better on more difficult datasets, such as HKUIS [21], DUTS [66], and PASCAL-S [64], which contain a large number of images with multiple salient objects.

As we can see in Fig. 7 and Fig. 8, our method can effectively correct those low contrast foregrounds and backgrounds induced by false-alarm detections. As we can see in the last row of Fig. 10, the overall performance of the adopted baseline methods can be ranked as $\text{PAGRN18} > \text{BRN18} > \text{RADF18} > \text{RFCN18} > \text{SRM17} > \text{DSS17} > \text{MDF16}$. After using our re-learning scheme to boost the detection performance, the overall performance can be significantly improved while maintaining identical ranking order, i.e., $\text{Our(PAGRN18)} > \text{Our(BRN18)} > \text{Our(RADF18)} > \text{Our(RFCN18)} > \text{Our (SRM17)} > \text{Our(DSS17)} > \text{Our(MDF16)}$. Meanwhile, as the performance of our re-learning scheme is dependent on the quality of the inter-image alignments, the performance improvement degree may be limited if the given target image exhibits large difference toward the retrieved sub-group images, and this problem can be effectively solved via expanding the image pool size.

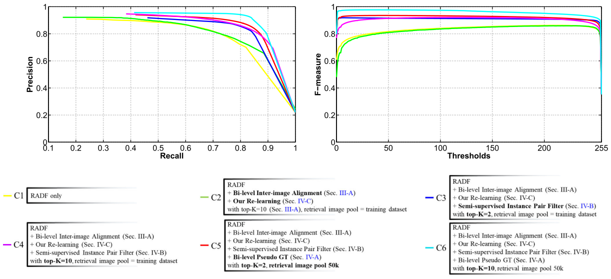

Fig. 11. Component evaluation results, including the averaged PR curve and the averaged F-measure curve over the **HKU-IS** [17] dataset, **ECSSD** [35] and **SOD** [65] dataset.

E. Component Evaluation

To validate the effectiveness of our method, we perform the component evaluation via using averaged PR curve and averaged F-measure curve over the **HKU-IS** [17] dataset, **ECSSD** [35] and **SOD** [65] dataset. Here we regard the RADF18 as the baseline method.

As we can see the detailed results in Fig. 11, where the conventional RADF (C1) exhibits the worst performance. Also we can notice that solely using our re-learning framework over the RADF can not guarantee to achieve an improved performance (C2), because the constructed inter-image correspondences may

TABLE II

Comparison of Quantitative Results Including Maximum F-measure (Larger is Better), MAE (Smaller is Better) and AUC (Larger is Better), and **the Mark (+) Denotes that an Additional DUST-TR-5K Dataset is Added to the Training Dataset**. The Top Three Results Are Highlighted in RED, GREEN, and BLUE, Respectively. The Result of PAGRN18 Over SOD Dataset Is not Available, so the Result of SOD-Our(PAGRN) Is Absent

Method(Year)	DUTS-TE [63]			ECSSD [35]			HKUIS [17]			PASCAL-S [64]			SOD [65]		
	F_β	MAE	AUC	F_β	MAE	AUC	F_β	MAE	AUC	F_β	MAE	AUC	F_β	MAE	AUC
Our(RADF18)	.886	.058	.954	.957	.034	.971	.948	.028	.970	.867	.081	.933	.867	.108	.879
Our(RFCN18)	.859	.063	.953	.957	.047	.965	.948	.046	.974	.881	.086	.938	.849	.128	.895
Our(BRN18)	.882	.037	.949	.966	.035	.975	.945	.032	.969	.888	.058	.949	.872	.093	.869
Our(SRM17)	.861	.032	.969	.944	.033	.945	.836	.027	.988	.878	.053	.951	.815	.110	.899
Our(DSS17)	.838	.064	.941	.932	.065	.956	.920	.052	.942	.850	.090	.917	.825	.136	.834
Our(MDF16)	.777	.082	.922	.870	.072	.930	.875	.094	.932	.786	.119	.821	.797	.135	.859
Our(PAGRN18)	.889	.038	.914	.953	.042	.950	.938	.030	.949	.876	.079	.911	/	/	/
RADF18(+) [16]	.841	.061	.940	.941	.039	.969	.944	.033	.966	.845	.099	.917	.846	.113	.866
RFCN18(+) [60]	.798	.070	.949	.912	.060	.966	.901	.048	.971	.833	.095	.932	.822	.141	.889
BRN18(+) [61]	.843	.051	.966	.936	.044	.983	.922	.041	.981	.856	.068	.953	.851	.105	.902
SRM17(+) [58]	.829	.048	.967	.928	.041	.988	.917	.041	.984	.851	.080	.960	.846	.107	.919
DSS17(+) [22]	.831	.068	.941	.921	.072	.955	.912	.058	.968	.813	.097	.915	.818	.141	.891
MDF16(+) [21]	.731	.092	.918	.830	.113	.939	.841	.111	.946	.771	.135	.909	.803	.152	.881
PAGRN18 [62]	.856	.056	.954	.926	.061	.968	.917	.048	.970	.851	.092	.935	/	/	/

TABLE III

Continued Table : Comparison of Quantitative Results Including Maximum F-measure (Larger is Better) , MAE (Smaller is Better) and AUC (Larger is Better). The Top Three Results Are Highlighted in red, green, and blue, Respectively

Method(Year)	DUTS-TE [63]			ECSSD [35]			HKUIS [17]			PASCAL-S [64]			SOD [65]		
	F_β	MAE	AUC	F_β	MAE	AUC	F_β	MAE	AUC	F_β	MAE	AUC	F_β	MAE	AUC
BSCA15 [56]	0.593	0.199	0.875	0.758	0.183	0.922	0.719	0.174	0.910	0.666	0.224	0.871	0.659	0.266	0.840
MB15 [55]	0.616	0.176	0.898	0.739	0.171	0.919	0.728	0.150	0.929	0.676	0.198	0.884	0.664	0.233	0.853
MST16 [57]	0.599	0.152	0.857	0.731	0.149	0.884	0.721	0.128	0.893	0.666	0.186	0.847	0.660	0.219	0.798
AMU17 [15]	0.777	0.085	0.960	0.915	0.059	0.982	0.897	0.053	0.983	0.833	0.095	0.956	0.799	0.146	0.913
UCF17 [59]	0.772	0.112	0.961	0.903	0.078	0.982	0.888	0.073	0.984	0.821	0.120	0.957	0.808	0.164	0.930
GU13 [54]	0.496	0.225	0.772	0.636	0.223	0.808	0.637	0.201	0.827	0.567	0.258	0.765	0.563	0.280	0.736

not be really helpful to alleviate the learning ambiguity problem. On the contrary, those incorrect inter-image correspondences may further lead the learning ambiguity even worse. Thus, we propose to utilize the “semi-supervised instance pair filter” to alleviate the above mentioned limitations toward the inter-image regional alignments, achieving a significant performance improvement, and see the quantitative curves C3 and C4.

Also, since the inter-image alignment performance is positively correlated to the retrieval image pool size, the overall detection performance can be further improved by using the pseudo GT to increase the available retrieval pool size, and see the quantitative curves C4 and C5. Meanwhile, it should be noted that the top-K choice in Section III-A is also important for the overall detection performance (see the differences respectively between {C3, C4} and {C5, C6}), and we assign it to 10 as the optimal choice.

F. Limitations

Our method tends to be time-consuming in general. In fact, excluding the time consumption of the pre-learned deep saliency mode (i.e., RADF, DSS, MDF), our method needs an additional 0.9 s. In more detail, on a desktop computer with i7-6700 k 4.00 GHz CPU, GTX 1080 GPU, 32 GB RAM, the image retrieval takes about 0.01 s (with CPU parallel computation), the SIFT-Flow guided inter-image alignment takes about 0.4 s (with CPU parallel computation), the deep feature computation takes about 0.4 s, and the saliency prediction takes about 0.1 s. Since the most time-consuming steps of our method are mainly induced by the CPU related computations, we will attempt to full implement these steps on GPU (i.e., the CUDA acceleration) in our near future work.

VI. Conclusion

In this paper, we have proposed a novel saliency relearning method to improve the detection performance of those pre-learned deep saliency model. Based on any pre-learned deep saliency model, our objective is to utilize those easy-to-detect training instance to facilitate those hard ones. We establish the inter-image correspondences by interactively considering both the object-level similarity and the non-local superpixel-wise correlation, thus the conventional feature conflicts can be effectively resolved. Meanwhile, we have newly designed a novel deep network to jointly learn common consistency of inter-image saliency for the high-quality image saliency detection.The quantitative comparisons between our new method and the state-of-the-art methods have confirmed the effectiveness of our method.

References

[1] S. Li, M. Xu, Y. Ren, and Z. Wang, "Closed-form optimization on saliency-guided image compression for HEVC-MSP," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 155–170, Jan. 2018.

[2] C. Chen, S. Li, H. Qin, and A. Hao, "Robust salient motion detection in non-stationary videos via novel integrated strategies of spatio-temporal coherency clues and low-rank analysis," *Pattern Recognit.*, vol. 52, pp. 410–432, 2016.

[3] C. Chen, Y. Li, S. Li, H. Qin, and A. Hao, "A novel bottom-up saliency detection method for video with dynamic background," *IEEE Signal Process. Lett.*, vol. 25, no. 2, pp. 154–158, Feb. 2018.

[4] C. Peng, C. Chen, Z. Kang, J. Li, and Q. Cheng, "RES-PCA: A scalable approach to recovering low-rank matrices," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 7317–7325.

[5] X. Zhou, Z. Liu, K. Li, and G. Sun, "Video saliency detection via bagging-based prediction and spatiotemporal propagation," *J. Visual Commun. Image Representation*, vol. 51, pp. 131–143, 2018.

[6] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2527–2542, Dec. 2017.

[7] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan. 2018.

[8] C. Chen, G. Wang, and C. Peng, "Structure-aware adaptive diffusion for video saliency detection," *IEEE Access*, vol. 7, pp. 79 770–79 782, 2019.

[9] S. Bi, G. Li, and Y. Yu, "Person re-identification using multiple experts with random subspaces," *J. Image Graph.*, vol. 2, no. 2, 2014.

[10] W. Song *et al.*, "Multi-task cascade convolution neural networks for automatic thyroid nodule detection and recognition," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1215–1224, May 2019.

[11] 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.

[12] C. Chen, S. Li, H. Qin, and A. Hao, "Real-time and robust object tracking in video via low-rank coherency analysis in feature space," *Pattern Recognit.*, vol. 48, no. 9, pp. 2885–2905, 2015.

[13] G. Evangelopoulos *et al.*, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, Nov. 2013.

[14] F. Zhou, S. Bing Kang, and M. F. Cohen, "Time-mapping using space-time saliency," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 3358–3365.

[15] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 202–211.

[16] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 6943–6950.

[17] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 478–487.

[18] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 1711–1720.

[19] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Trans. Pattern Anal. Mach. Intell*, to be published.

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

[21] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep cnn features," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016.

[22] Q. Hou *et al.*, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.

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

[24] C. Ge, K. Fu, F. Liu, L. Bai, and J. Yang, "Co-saliency detection via inter and intra saliency propagation," *Image Commun.*, vol. 44, pp. 69–83, 2016.

[25] Y. Li, K. Fu, Z. Liu, and J. Yang, "Efficient saliency-model-guided visual co-saliency detection," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 588–592, May 2015.

[26] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 2994–3002.

[27] C. Peng, Z. Kang, S. Cai, and Q. Cheng, "Integrate and conquer: Double-sided two-dimensional k-means via integrating of projection and manifold construction," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 5, pp. 57:1–57:25, 2018.

[28] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, Sep. 2014.

[29] 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.

[30] W. Wang, J. Shen, and S. Ling, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, Jan. 2018.

[31] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3156–3170, Jul. 2017.

[32] G. Li, Y. Xie, T. Wei, K. Wang, and L. Lin, "Flow guided recurrent neural encoder for video salient object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 3243–3252.

[33] C. Chen, S. Li, H. Qin, Z. Pan, and G. Yang, "Bi-level feature learning for video saliency detection," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3324–3336, Dec. 2018.

[34] X. Zhou, Z. Liu, C. Gong, and W. Liu, "Improving video saliency detection via localized estimation and spatiotemporal refinement," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 2993–3007, Nov. 2018.

[35] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 3166–3173.

[36] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 1155–1162.

[37] C. Chen, S. Li, H. Qin, and A. Hao, "Structure-sensitive saliency detection via multilevel rank analysis in intrinsic feature space," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2303–2316, Aug. 2015.

[38] L. Gayoung, T. Y.-Wing, and K. Junmo, "Deep saliency with encoded low level distance map and high level features," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 660–668.

[39] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 678–686.

[40] Z. Luo *et al.*, "Non-local deep features for salient object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6609–6617.

[41] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 384–393.

[42] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 1741–1750.

[43] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[44] W. Wang, Q. Lai, H. Fu, J. Shen, and H. Ling, "Salient object detection in the deep learning era: An in-depth survey," 2019, *arXiv:1904.09146*.

[45] L. Ye, Z. Liu, X. Zhou, L. Shen, and J. Zhang, "Saliency detection via similar image retrieval," *IEEE Signal Process. Lett.*, vol. 23, no. 6, pp. 838–842, Jun. 2016.

[46] J. Ren, Z. Liu, X. Zhou, G. Sun, and C. Bai, "Saliency integration driven by similar images," *J. Vis. Commun. Image Representation*, vol. 50, pp. 227–236, 2018.

[47] 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.

[48] X. Li, F. Yang, L. Chen, and H. Cai, “Saliency transfer: An example-based method for salient object detection,” in *Proc. Int. Joint Conf. Artif. Intell.*, no. 7, 2016, pp. 3411–3417.

[49] C. Liu, J. Yuen, and A. Torralba, “SIFT flow: Dense correspondence across scenes and its applications,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

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

[51] C. L. Zitnick and P. Dollár, “Edge boxes: Locating object proposals from edges,” in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 391–405.

[52] W. Wang, J. Shen, L. Shao, and F. Porikli, “Correspondence driven saliency transfer,” *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5025–5034, Nov. 2016.

[53] P. Krähenbühl and V. Koltun, “Efficient inference in fully connected CRFS with gaussian edge potentials,” in *Proc. Advances Neural Inf. Process. Syst.*, 2011, pp. 109–117.

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

[55] J. Zhang *et al.*, “Minimum barrier salient object detection at 80 fps,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 1404–1412.

[56] Y. Qin, H. Lu, Y. Xu, and H. Wang, “Saliency detection via cellular automata,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 110–119.

[57] W. Tu, S. He, Q. Yang, and S. Chien, “Real-time salient object detection with a minimum spanning tree,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2334–2342.

[58] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, “A stagewise refinement model for detecting salient objects in images,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 4039–4048.

[59] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, “Learning uncertain convolutional features for accurate saliency detection,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 212–221.

[60] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, “Salient object detection with recurrent fully convolutional networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1734–1746, Jul. 2019.

[61] T. Wang *et al.*, “Detect globally, refine locally: A novel approach to saliency detection,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 3127–3135.

[62] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, “Progressive attention guided recurrent network for salient object detection,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 714–722.

[63] R. Zhao, W. Ouyang, H. Li, and X. Wang, “Saliency detection by multi-context deep learning,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, IEEE, 2015, pp. 1265–1274.

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

[65] V. Movahedi and J. H. Elder, “Design and perceptual validation of performance measures for salient object segmentation,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit. Workshops*, 2010, pp. 49–56.

[66] L. Wang *et al.*, “Learning to detect salient objects with image-level supervision,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 136–145.

[67] 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.

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

Authors’ photographs and biographies not available at the time of publication.