Image Co-Saliency Detection and Co-Segmentation via Progressive Joint Optimization

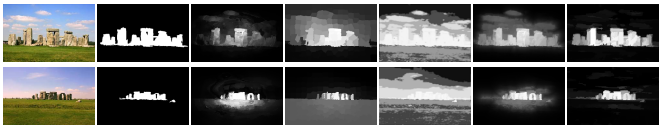
Chung-Chi Tsai, *Student Member, IEEE*, Weizhi Li, Kuang-Jui Hsu, Xiaoning Qian, *Senior Member, IEEE*, and Yen-Yu Lin, *Member, IEEE*

***Abstract*— We present a novel computational model for simultaneous image co-saliency detection and co-segmentation that concurrently explores the concepts of saliency and objectness in multiple images. It has been shown that the co-saliency detection via aggregating multiple saliency proposals by diverse visual cues can better highlight the salient objects; however, the optimal proposals are typically region-dependent and the fusion process often leads to blurred results. Co-segmentation can help preserve object boundaries, but it may suffer from complex scenes. To address these issues, we develop a unified method that addresses co-saliency detection and co-segmentation jointly via solving an energy minimization problem over a graph. Our method iteratively carries out the region-wise adaptive saliency map fusion and object segmentation to transfer useful information between the two complementary tasks. Through the optimization iterations, sharp saliency maps are gradually obtained to recover entire salient objects by referring to object segmentation, while these segmentations are progressively improved owing to the better saliency prior. We evaluate our method on four public benchmark data sets while comparing it to the state-of-the-art methods. Extensive experiments demonstrate that our method can provide consistently higher-quality results on both co-saliency detection and co-segmentation.**

***Index Terms*— Co-saliency detection, co-segmentation, locally adaptive proposal fusion, energy minimization, joint optimization.**

I. Introduction

IMAGE co-saliency detection and object co-segmentation are two fundamental and active research topics in computer vision and image analysis. They are highly relevant but different. Co-saliency detection is a weakly supervised extension of saliency detection to locate the eye-catching regions that are commonly present in multiple images. Compared to single-image saliency detection, co-saliency detection leverages not only intra-image but also inter-image evidence to better highlight regions of interest. As a key component of image analysis, it is essential to a broad set of applications, such as object detection [1], co-localization [2], and video compression [3]. In a different manner, object co-segmentation focuses on jointly extracting common objects from a group of images. It has been studied extensively, since it can borrow signal strengths across images to improve segmentation and it enhances action extraction [4] and image matching [5]. In this work, we investigate the strengths and weaknesses of co-saliency detection and co-segmentation. Motivated by the close relationship between the two tasks, we derive a new unified approach to solve them simultaneously. In this way, the complementary information can be transferred between both tasks to improve their performances.



(a) (b) (c) (d) (e) (f) (g)

Fig. 1. (a) A pair of images for co-saliency detection. (b) The ground truth. (c) ∼ (e) Three saliency proposals generated by DSR [6], MR [7], and SpC [8] respectively. (f) The detection results by the fusion-based method SACS [9]. (g) The detection results by our method.

We motivate our joint co-saliency detection and co-segmentation by first considering the requirements to achieve high-quality co-saliency detection. To capture complex image content, many modern co-saliency methods favor fusing multiple (co-)saliency proposals, each of which is generated from particular saliency evidence, via either *fixed-weight summation* [10]–[12], *fixed-weight multiplication* [8], [12] or *adaptive-weight summation* [9], [13]. Fig. 1(c) ∼ Fig. 1(e) show different saliency proposals generated by the method DSR [6], the method MR [7], and using the multi-image spatial cue (SpC) [8], respectively. None of them gives satisfactory results. The algorithm SACS [9] implements adaptive weighted summation of the three proposals, and significantly improves the detection results as shown in Fig. 1(f). Despite the effectiveness of proposal fusion, two major issues arise. First, the fusion-based methods mentioned above are of map-wise fashion; Namely, the fusion weights are assigned to the whole saliency proposals. However, the optimal saliency proposals often vary from image region to region, as mentioned in our prior work [14], [15]. Secondly, weighted combinations of different saliency proposals typically lead to blurred results,

Manuscript received October 6, 2017; revised June 19, 2018; accepted July 18, 2018. Date of publication July 31, 2018; date of current version September 19, 2018. This work was supported in part by the Ministry of Science and Technology (MOST) under Grant 105-2221-E-001-030-MY2 and Grant 107-2628-E-001-005-MY3, and in part by the National Science Foundation under Awards #1547557 and #1553281. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tolga Tasdizen. *(Corresponding authors: Xiaoning Qian; Yen-Yu Lin.)*

C.-C. Tsai is with the Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX 77843 USA, and also with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan (e-mail: chungchi@tamu.edu).

W. Li and X. Qian are with the Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: wayne0908@tamu.edu; xqian@ece.tamu.edu).

K.-J. Hsu and Y.-Y. Lin are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan (e-mail: kjhsu@citi.sinica.edu.tw; yylin@citi.sinica.edu.tw).

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

Digital Object Identifier 10.1109/TIP.2018.2861217

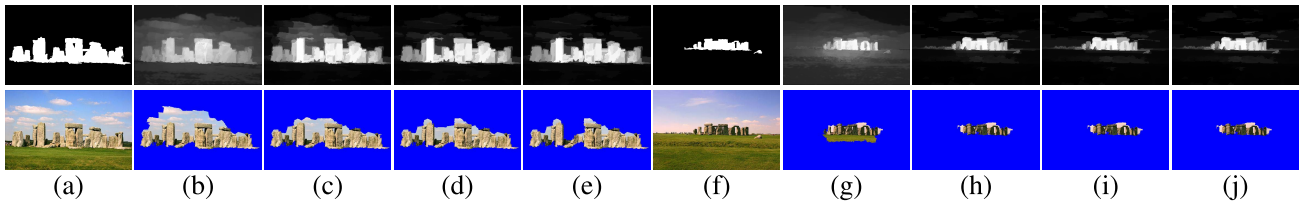


Fig. 2. Our approach enables the progressive improvement of co-saliency detection and co-segmentation. (a) & (f) Two images and the ground truth for co-saliency detection. (b) ∼ (e) The results of co-saliency detection (top row) and co-segmentation (bottom row) at the first four iterations for the image in (a). (g) ∼ (j) The results for the image in (b).

especially near the surrounding areas of objects. The evidence of *objectness* from co-segmentation can guide region-wise saliency proposal fusion and help recover sharp object boundaries [15]. Our approach can integrate co-segmentation into co-saliency detection, and achieves the superior results displayed in Fig. 1(g).

The second motivation of our method is that object co-segmentation often suffers from large intra-object variations or complex background, which may lead to over- or under-segmentation. Saliency detection identifies the focus in images by human visual processing. The detection results provide important evidence for figure-ground separation in image segmentation, which alleviate the ambiguity caused by large intra-object variations or complex background. Thus, (co-)saliency detection can serve as an intrinsic component of object (co-)segmentation to improve performance.

The mutual dependency between co-saliency detection and co-segmentation motivates a unified approach to accomplish the two tasks simultaneously with the complementary information transferred between them to help each other. Our method optimizes a coupled objective function over a graph structure that links the two tasks. Through alternating optimization, the concept of *objectness* attained via co-segmentation helps the region-wise proposal fusion to better highlight salient regions. Meanwhile, the improved co-saliency maps enhance co-segmentation with more favorable *saliency* priors. Fig. 2 shows an example of the progressive improvement of co-saliency detection and co-segmentation by our method. Given a pair of images in Figs. 2(a) and 2(f), our method carries out co-saliency detection and co-segmentation simultaneously. At the first iteration, the co-saliency detection results inherit the noise from different saliency proposals, while the co-segmentation masks contain some false positives. Through the optimization process, co-saliency maps of higher quality are attained with less false positives and sharper object boundaries. Meanwhile, gradually improved co-segmentation masks are obtained and used to guide saliency detection at the next iteration. At the end, both tasks help each other to stable high-quality solution after a few iterations as shown in Figs. 2(e) and 2(j).

II. Related Work

We review relevant topics to the development of our approach in this section, including saliency detection, co-saliency detection, and co-segmentation.

A. Saliency Detection

The literature of saliency detection is extensive. Methods for saliency detection can be roughly sorted into human *visual attention prediction* [16]–[20] and *salient object detection (SOD)* [6], [7], [21]–[38]. Methods for visual attention prediction usually generate a heat map consisting of blob-like regions indicating the eye-fixation likelihood. Inspired by human visual systems, Itti *et al*. [16] presented a pioneering saliency detection model based on local contrast computed from the center-surround differences across multiple scales. Borgi and Itti [18] fused complementary global rarity cues of a scene and local contrast evidence in both the RGB and L*a*b* color spaces to enhance the performance. Without using any image features or high-level priors, Hou and Zhang [17] defined the saliency through the residual on the Fourier domain of an input image; and Xia *et al*. [19] thought using spatial domain residual is more correlated to our visual attention. Visual fixation methods usually spotlight object boundaries because the design principles abide human visual systems to target on the place of rapid scene change first; thus it is not as suitable as salient object prediction to support a wide-range of multimedia applications by showing regions of interest.

Salient object detection (SOD) aims to spotlight entire salient objects, instead of merely their boundaries or discriminative parts in visual attention prediction. To separate the conspicuous foreground from the background, traditional methods highly rely on the contrast cues. For instance, Achanta *et al*. [21] approximated saliency based on the deviation between a low-pass filtered image and the average color of the whole image. Perazzi *et al*. [24] jointly considered the color contrast with surrounding pixels and the spatial compactness of saliency distribution. Besides pixel-level saliency models, several region-based models, e.g. [6], [7], [22], [23], [25]–[29], [31], were developed to reduce the computation load and ease the influence of image noise. In addition to low-level features, Shen and Wu [25] further took high-level knowledge, such as face locations and center priors, into account. Some approaches to saliency detection, such as [6], [7], and [26], concentrated on the derivation of correct background. Specifically, these approaches consider regions near image boundaries as background and predict a superpixel as salient or non-salient based on its difference from the background. Zhu *et al*. [29] further integrated global contrast with the improved background priors to achieve better performance. Moreover, methods based on graph-based clustering, e.g. [23], [28], [31], were proposed to better locate the potential objects. Stemming from the unsupervised nature, the performance of these methods based on either the learned or handcrafted features for single-image saliency detection is still limited.

Recent research efforts, e.g. [32]–[36], have been made to use *convolutional neural networks* (CNN) for saliency detection. Due to the availability of large-scale training data such as ImageNet, the features learned by CNN for object recognition

usually give better performance than conventional handcrafted features. As mentioned in [39], recent models [34], [36] based on *fully convolutional networks (FCN)* [40] can achieve superior saliency inference by incorporating the contextual image information with end-to-end learning. With the intrinsic interdependence between saliency detection and semantic image segmentation, some methods, e.g. [33], have formulated a multitask objective for joint feature learning for such two correlated tasks. However, CNN-based methods rely on labeled training with object masks or extra information sources for tuning the deep models, which are generally unavailable in saliency detection, and such heavy annotation cost makes these methods less practical. To alleviate the requirement of a large set of training masks, weakly supervised salient object detection becomes widely studied to infer the exact object locations given only training images with weak image-level labels [37], [38]. However, it is restricted to detecting saliency objects whose categories have been covered by training data.

B. Co-Saliency Detection

Co-saliency detection is another branch of weakly supervised extension of single-image saliency detection by exploring the visual cues shared across multiple images to identify salient objects better. Chang *et al*. [41] formulated co-saliency as a combination of intra-image saliency and inter-image repetitiveness. Fu *et al*. [8] proposed a clustering-based algorithm for co-saliency detection by considering intra-cluster evidence such as pixel distribution, contrast, and correspondences. Then, co-saliency is carried out via Bayesian inference of each pixel belonging to the clusters. To prevent detecting common background as salient foreground, Zhang *et al*. [42] incorporated object proposals from other image groups into the testing group to better distill the intra-image contrast and intra-group consistency to generate the co-saliency score in a Bayesian framework. Different from the existing methods that focus on RGB images and assume all images contain co-salient objects of a single category, Cong *et al*. [43] proposed a novel co-saliency detection model for RGBD images, whereas Yao *et al*. [44] used an efficient clustering-based principle to achieve multi-class co-saliency detection on cluttered datasets that contain an arbitrary number of object categories. Despite their effectiveness, co-saliency detection remains a challenging task in practice due to various unfavorable image variations, such as small objects or background clutters.

A research trend in saliency detection lies in fusing a set of saliency proposals, each of which is obtained based on particular image evidence. The fused saliency map is derived to leverage the most information with these proposals while excluding their individual biases. Li *et al*. [10] and Fu *et al*. [8] respectively proposed normalized summation and multiplication to combine saliency proposals; however, simple arithmetic operations are insufficient to effectively wipe out non-salient regions as well as keep the salient foregrounds. Hence, Cao *et al*. [9], [13] sought adaptive fusion weights based on a low-rank constraint on different salient foreground color content. Huang *et al*. [30] obtained multi-scale saliency proposals and fused them via the low-rank constraint to extract the shared intrinsic saliency information.

The aforementioned fusion-based methods carry out *image-wise* proposal fusion, while the optimal saliency proposals often vary from region to region. To address this issue, Tsai *et al*. [14] formulated adaptive region-wise fusion as an optimization problem where local consensus, spatial consistency and global correspondence are jointly taken into account. Huang *et al*. [45] adopted a hybrid strategy that adaptively selects a summation or multiplication fusion scheme for each superpixel. Despite the effectiveness, the common drawback for fusion-based approaches, e.g. [9], [14], [30], [45], is that the resultant saliency maps are typically blurred, especially near the object boundaries. Thus, post-processing is often required; but it is *ad-hoc* and may degenerate the performance.

Segmentation or boundary detection has been commonly integrated into co-saliency detection e.g. [11], [15], [46]–[48] to enhance the performance since foreground segments directly gives estimated objects in a secne. Li *et al*. [11] applied *GrabCut* [49] to multi-scale initialization windows, and utilize the commonly appeared segment-based object proposals for intra-image saliency estimation. Jerripothula *et al*. [48] utilized the segmentation masks to adaptively determine the penalty for superpixels in fusing saliency proposals. However, these methods derive image segmentation and saliency detection in separated steps. Hence, complementary information between image segmentation and saliency detection cannot be mutually transferred to enhance each other's performance.

C. Image Co-Segmentation

Image co-segmentation is closely related to co-saliency detection as it targets at segmenting the common but not necessarily salient parts across multiple images. Rother *et al*. [50] introduced the pioneering work of co-segmentation by minimizing the unnormalized foreground histogram dissimilarity in *Markov random field* (MRF). Hochbaum and Singh [51] used a sub-modular rewarding term to encourage similar pixels having same labels and efficiently solved it by *graph-cut*. Joulin *et al*. [52], [53] utilized discriminative clustering to separate the common foreground superpixels from the background.

Co-saliency detection can be adopted in the pre-processing step of co-segmentation, and replaces the interactive supervision process. It provides the prior knowledge of the common objects in multiple images, and can deal with the difficulties due to complex background and large intra-object variations. Chang *et al*. [41] introduced a co-saliency guided method for co-segmentation by taking into account foreground similarity and figure-ground dissimilarity. Yu *et al*. [54] used a *Gaussian mixture model* (GMM) to compute figure-ground statistics, and embedded co-saliency information in the unary term of MRF for co-segmentation.

Saliency information can also be used to improve the object appearance models and enhance co-segmentation. For instance, Fu *et al*. [55] used depth enhanced co-saliency maps for co-segmentation. Meng *et al*. [56] cast co-segmentation as the shortest path problem on a directed graph constructed by referring to object proposals, region similarities, and co-saliency information. Rubinstein *et al*. [57] built several energy terms by using saliency and correspondence information for co-segmentation. To reduce the interference from similar backgrounds in images, Han *et al*. [58] proposed an optimization framework where background knowledge derived from boundary superpixels is exploited for co-segmentation. However, treating prior knowledge generation separately from the segmentation process potentially impedes the effective and adaptive transfer of useful information across different tasks.

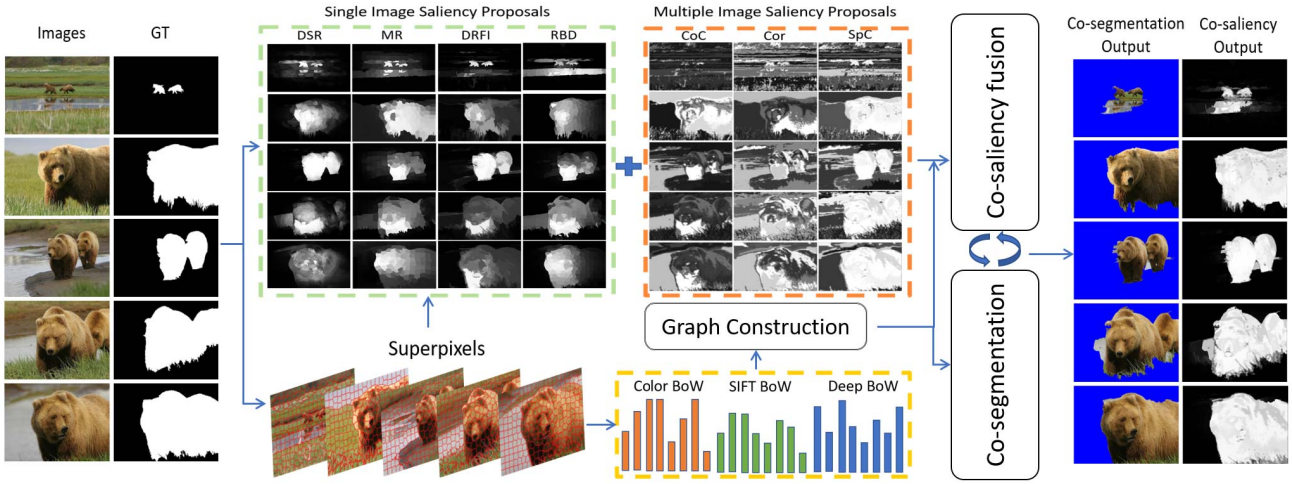


Fig. 3. The proposed framework for joint co-saliency detection and co-segmentation. Given images of a particular object category, we process the input images by compiling their superpixel representation, extracting features from the superpixels, and computing a set of saliency proposals. The proposed approach takes the processed data as input, and performs alternating co-saliency detection and co-segmentation until convergence.

Co-saliency detection and co-segmentation are highly relevant to each other. Their combination has been explored in existing methods. Nevertheless, these methods treat the two tasks as separated steps. Thus the combination is *unidirectional*. Namely, these methods either use co-segmentation to improve co-saliency detection, e.g. [11], [15], [46]–[48], or leverage co-saliency detection to help co-segmentation, e.g. [41], [54]–[58]. Our approach instead enables simultaneous co-saliency saliency and co-segmentation. It *bidirectionally* links the two tasks in the domain of superpixels whose pairwise relationships are modeled by a graph. The joint objective function on both tasks is designed on the graph. Through an alternating optimization process, both tasks are progressively improved via sharing information. As an unidirectional approach, our prior work [15] integrates prior knowledge attached via segmentation into region-wise proposal fusion for saliency detection. We will show in the experiments that this bidirectional method here consistently outperforms our prior work [15] for co-saliency detection. More importantly, this work further improves co-segmentation with the integration of co-saliency detection, and make extension to four datasets.

III. Proposed Method

We introduce our method in this section. First, the problem definition is given. Then, the steps of image processing, feature extraction, and graph construction are applied to the input images. Finally, the proposed objective function for joint co-saliency detection and co-segmentation as well as its optimization are specified.

A. Problem Definition

Considering a set of n images $\mathcal{I} = \{I_1, I_2, \cdots, I_n\}$, we apply several existing (co-)saliency detection algorithms, e.g. [6]–[8], [27], [29], to obtain M saliency proposals for each image. Each image I_j is decomposed into N_j *superpixels*, which serve as the domain of joint co-saliency detection and co-segmentation because they preserve intrinsic image structures and abstract unnecessary details. Total $N = \sum_j N_j$, $j \in \{1, 2, \cdots, n\}$ superpixels are yielded for the image set $\mathcal{I}$.

For co-saliency detection, our goal is to seek a plausible weight vector $\mathbf{y}_i = [y_{i,1} \;\; y_{i,2} \;\; \cdots \;\; y_{i,M}]^\top \in [0,1]^M$ for each superpixel $i \in \{1, 2, \cdots, N\}$ to accomplish the saliency detection by region-wise combining the M saliency proposals. For co-segmentation, we optimize the segmentation masks represented by superpixel figure-ground indicators $z_i \in \{0, 1\}$, $i \in \{1, 2, \cdots, N\}$.

Fig. 3 illustrates our framework where co-saliency detection and co-segmentation are carried out simultaneously. By iteratively transferring useful information to regularize each other, both tasks are progressively improved and converge rapidly to favorable results. For instance, in Fig. 3, most of our adopted saliency proposals, especially the multiple-image saliency proposals, are interfered by the common background regions across images, such as lake areas. Thanks to the collaborated and iterative refinement framework, potential adversary effect is minimized in the joint outputs.

B. Superpixel and Feature Extraction

In our implementation, each image I_j is decomposed into $N_j \approx 200$ superpixels by using the *SLIC* algorithm [59]. In addition to the color and SIFT [60] features, we also exploit the deep features produced by the CNN-S network [61] to describe semantic characteristics of objects. Combining the three complementary types of features typically results in more comprehensive description of the co-salient regions. To extract deep features, we up-sample and concatenate the feature maps in layers of the CNN-S network, conv_relu1 (96 channels), conv_relu2 (256 channels), conv_relu3 (512 channels), conv_relu4 (512 channels) and conv_relu5 (512 channels), to yield a 1888-dimensional *hypercolumn* representation for each pixel. Next, we use the *bag-of-words* (BoWs) model for superpixel representation. Specifically, for color features, the k-means clustering algorithm is applied to pixels in three color spaces, i.e. `RGB`, `L*a*b*`, and `YCbCr`, and generates 20 *visual words*. To ensure having stable result, we run k-means 20 times, and select the clustering with the minimal sum of the squared distances between data and their cluster centers. The color BoWs representation of the i-th superpixel $\mathbf{h}_i^c$ is then a 20-dimensional histogram.

The SIFT and deep BoWs representations, denoted by $\mathbf{h}_i^s$ and $\mathbf{h}_i^d$ respectively, are similarly set. Lastly, we concatenate them and yield a 60-dimensional feature representation for the i-th superpixel, $\mathbf{h}_i = [\mathbf{h}_i^c, \mathbf{h}_i^s, \mathbf{h}_i^d]$. The similarity between two superpixels i and $\hat{i}$ is defined by

$$s(i,\hat{i}) = \exp(-\frac{\chi^2(\mathbf{h}_i, \mathbf{h}_{\hat{i}})}{\sigma}), \tag{1}$$

where the constant σ is set to the average pair-wise distance between all superpixels under this feature representation.

C. Graph Construction

A graph $\mathcal{G} = (\mathcal{V} = \cup\, \mathcal{V}_j, \mathcal{E} = \cup\, \mathcal{E}_j)$ is constructed to encode the spatial relationships among superpixels. $\mathcal{V}_j$ corresponds to all the superpixels in I_j, thus $|\mathcal{V}| = N$. Edge set $\mathcal{E}_j$ represents the adjacency relationships between superpixels in $\mathcal{V}_j$. Namely, edge $e_{i\hat{i}} \in \mathcal{E}_j$ is added for linking v_i and $v_{\hat{i}}$ if superpixels i and $\hat{i}$ in I_j are spatially connected. We set the weight of edge $e_{i\hat{i}}$ as

$$A(i,\hat{i}) = s(i,\hat{i}) * b(i,\hat{i}), \tag{2}$$

where $b(i,\hat{i})$ is the counts of pairs of adjacent pixels across the boundary of superpixels i and $\hat{i}$. The design of the edge weights is crucial. Considering both the content and shared boundary lengths of superpixels can better describe the inherent structure of images, and boost the performance. With affinity matrix $A \in \mathbb{R}^{N\times N}$ in (2), the associated *graph Laplacian* $L \in \mathbb{R}^{N\times N}$ can be computed.

D. Objective Function

We seek plausible weights $Y = [\mathbf{y}_1\ \mathbf{y}_2\ \ldots\ \mathbf{y}_N] \in \mathbb{R}^{M\times N}$ for superpixel-wise saliency map fusion as well as figure-ground configuration $Z = [z_1\ z_2\ \ldots\ z_N] \in \{0,1\}^N$ for co-segmentation by minimizing the following objective function:

$$\begin{aligned} &J(Y,Z) \\ &= \|Y\|_2^2 + \alpha_1 \sum_{i:v_i\in\mathcal{V}} U(\mathbf{y}_i) + \alpha_2 \sum_{1\le j<\hat{j}\le n} D(\mathbf{z}_j, \mathbf{z}_{\hat{j}}) \\ &\quad + \alpha_3 \sum_{i:v_i\in\mathcal{V}} C(\mathbf{y}_i, z_i) + \alpha_4 \sum_{e_{i\hat{i}}\in\mathcal{E}} B_1(\mathbf{y}_i, \mathbf{y}_{\hat{i}}) \\ &\quad + \alpha_5 \sum_{e_{i\hat{i}}\in\mathcal{E}} B_2(z_i, z_{\hat{i}}) \\ &\quad \text{s.t. } \|\mathbf{y}_i\|_1 = 1, \mathbf{y}_i \ge \bar{\mathbf{0}}, z_i \in \{0,1\}, \quad \text{for } 1 \le i \le N, \end{aligned} \tag{3}$$

where $\bar{\mathbf{0}}$ is an all-zero vector, and α_1, α_2, α_3, α_4 and α_5 are five positive constants. $\mathbf{z}_j = \{z_i | i \in V_j\}$ denotes the figure-ground configuration of image I_j. $\mathbf{z}_{\hat{j}}$ is similarly defined. Real-valued $y_{i,m} \in [0,1]$ is the fusion weight of saliency proposal m on superpixel i. Binary variable z_i takes value 1 if superpixel i belongs to the foreground, and 0 otherwise. Y and Z are optimized jointly so that the useful information can be shared for transferring object-aware boundaries from co-segmentation to co-saliency as well as transferring saliency priors from co-saliency to co-segmentation. In (3), $U(\mathbf{y}_i)$ and $B_1(\mathbf{y}_i, \mathbf{y}_{\hat{i}})$ are the unary and pairwise terms for co-saliency detection, respectively. $D(\mathbf{z}_j, \mathbf{z}_{\hat{j}})$ and $B_2(z_i, z_{\hat{i}})$ are the discriminative and pairwise terms for co-segmentation, respectively. The coupling term $C(\mathbf{y}_i, z_i)$ is included to encourage the coherence between the co-saliency map and the figure-ground co-segmentation. Lastly, the term $||Y||_2^2$ is introduced for regularization. These terms are detailed as follows.

1) Unary Term $U(\mathbf{y}_i)$ for Co-Saliency Detection: We follow the co-saliency formula

$$\textit{Co-saliency} = \textit{Saliency} \times \textit{Repetitiveness},$$

to design this unary term. Thus, this term contains two parts that leverage the intra- and inter-image cues to infer the goodness of each saliency proposal in terms of *saliency* and *repetitiveness* on superpixel i, respectively. The two parts are respectively shown in the blue and yellow diagrams of Fig. 4.

For the intra-image cue, we intend to assign a higher weight to a saliency proposal that is consistent with others. It helps exclude individual biases. Inspired by [62], we employ a low-rank formulation to conduct this task. We further generalize it to *locally* estimate the quality of saliency proposals. For superpixel i, we find its $K(=50)$ spatially nearest superpixels. See the blue colored region on I_2 of Fig. 4 as an example. Let $\mathbf{x}_{i,m} \in \mathbb{R}^{256}$ be a histogram denoting the 256-bin distribution of saliency values on the saliency proposal m for the region covered by these K superpixels, i.e. the blue contours in the blue diagram of Fig. 4. By stacking the M vectors derived from all the saliency maps, $X_i = [\mathbf{x}_{i,1}\ \mathbf{x}_{i,2}\ \ldots\ \mathbf{x}_{i,M}] \in \mathbb{R}^{256\times M}$, we infer the consistency by seeking a low-rank representation of X_i. Specifically, *robust PCA* (RPCA) [63] is adopted to decompose X_i into a low-rank approximation R_i and a residual matrix E_i by solving

$$\min_{R_i,E_i} (||R_i||_* + \lambda||E_i||_1), \quad \text{s.t. } X_i = R_i + E_i, \tag{4}$$

where $||R_i||_*$ is the nuclear norm of R_i. λ is a constant and we set it to 0.05 in this work. After solving (4), we convert normalized errors $E_i = [\mathbf{e}_{i,1}\ \ldots\ \mathbf{e}_{i,M}]$ to *belief*:

$$b_{i,m} = \frac{\exp(-||\mathbf{e}_{i,m}||_2^2)}{\sum_{k=1}^{M}\exp(-||\mathbf{e}_{i,k}||_2^2)}, \quad \text{for } 1 \le m \le M. \tag{5}$$

For energy minimization, the associated penalty variable l_i computed from intra-image evidence for superpixel i using the saliency proposal m is then defined by

$$l_{i,m} = \frac{\exp(1-b_{i,m})}{\sum_{k=1}^{M}\exp(1-b_{i,k})}. \tag{6}$$

For the inter-image cue, we explore inter-image correspondences to evaluate the property of *repetitiveness*. Let $c_{i,j} \in [0,1]$ be the similarity, computed via (1), between superpixel i and its most similar superpixel $\hat{i}$ in image I_j, $j \in \{1,2,...,n\}$. See the bottom part in the yellow diagram of Fig. 4 for an example where the most similar superpixels in other images are pointed by black arrows. We take into account the similarities of all correspondences of superpixel i, and define the correspondence cue as

$$c_i = \frac{\texttt{mean}(\{c_{i,j} | 1 \le j \le n\})}{\texttt{var}(\{c_{i,j} | 1 \le j \le n\}) + 1}. \tag{7}$$

Large c_i means that superpixel i is consistently matched across images and the degree of *repetitiveness* is high. On the contrary, low c_i implies that superpixel i probably belongs to distinct background. To make this cue more robust, we normalize $\{c_i\}$ of all superpixels in an image as a probability indication of recurrent regions.

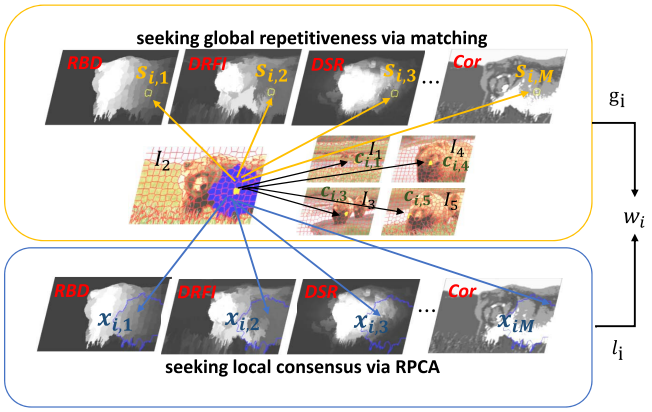


Fig. 4. Illustration of the unary term U term for co-saliency detection. See the text for the details.

Let $s_{i,m}$ denote the mean saliency value of saliency proposal m on superpixel i, the yellow circled region on the saliency proposals in the yellow diagram of Fig. 4. We prefer saliency map m if the value of $s_{i,m}$ is proportionate to that of c_i. We introduce a variable $g_{i,m}$ that penalizes the case where just one of c_i and $s_{i,m}$ is large, i.e.

$$g_{i,m} = \frac{\exp((1-c_i)s_{i,m} + c_i(1-s_{i,m}))}{\sum_{k=1}^{M}\exp((1-c_i)s_{i,k} + c_i(1-s_{i,k}))}. \quad (8)$$

The denominator in (8) is used for normalization.

The intra- and inter-image cues on superpixel i and proposal m, i.e. $l_{i,m}$ in (6) and $g_{i,m}$ (8), are combined via

$$w_{i,m} = \frac{\exp(l_{i,m} + g_{i,m})}{\sum_{k=1}^{M}\exp(l_{i,k} + g_{i,k})} \times \mathtt{size}(i), \quad (9)$$

where $\mathtt{size}(i)$ is the size of superpixel i. It can be observed that a lower penalty $w_{i,m}$ implies that the m-th saliency proposal on superpixel i is more reliable, so a higher fusion weight $y_{i,m}$ should be assigned to minimize the energy cost. Considering all superpixels, this unary term becomes

$$\sum_{v_i \in \mathcal{V}} U(\mathbf{y}_i) = \sum_{i=1}^{N} \mathbf{w}_i^\top \mathbf{y}_i = tr(W^\top Y), \quad (10)$$

where $\mathbf{w}_i = [w_{i,1} \ \ldots \ w_{i,M}]^\top$ and $W = [\mathbf{w}_1 \ \ldots \ \mathbf{w}_N]$.

2) Discriminative Term $D(\mathbf{z}_j, \mathbf{z}_{\hat{j}})$ for Co-Segmentation: This term estimates the quality of figure-ground separation of images I_j and $I_{\hat{j}}$, which is parametrized by $\mathbf{z}_j$ and $\mathbf{z}_{\hat{j}}$, in a discriminative manner. Two attributes for being high-quality figure-ground separation are considered. First, the foreground appearances of images I_j and $I_{\hat{j}}$ need to be similar. Second, the foreground and background regions of each image should be dissimilar.

The feature representation of superpixel i is expressed by $\mathbf{h}_i = [\mathbf{h}_i^c \ \mathbf{h}_i^s \ \mathbf{h}_i^d]$, a concatenation of the BoWs representation from color, SIFT and deep features. Let H_j^f denote the estimated foreground of image I_j. Since H_j^f is a collection of superpixels, we represent it by summing the feature representation of all superpixels that it covers, i.e. $H_j^f = \sum_{z_i \in \mathbf{z}_j} \mathbf{h}_i z_i$, where $\mathbf{z}_j$ is figure-ground configuration of image I_j. The estimated background of image I_j is similarly defined as $H_j^b = \sum_{z_i \in \mathbf{z}_j} \mathbf{h}_i(1-z_i)$. We adopt the global energy term in [41] to discriminatively assess figure-ground separation for a pair of images I_j and $I_{\hat{j}}$. This discriminative term is designed below

$$\begin{aligned} D(\mathbf{z}_j, \mathbf{z}_{\hat{j}}) &= \|H_j^f - H_{\hat{j}}^f\|_2^2 - \sum_{k\in\{j,\hat{j}\}} \gamma_1 \|H_k^f - \gamma_2 H_k^b\|_2^2 \\ &= R - 2 \sum_{z_i \in \mathbf{z}_j, z_{\hat{i}} \in \mathbf{z}_{\hat{j}}} \langle \mathbf{h}_i, \mathbf{h}_{\hat{i}} \rangle z_i z_{\hat{i}} \\ &\quad + 2\gamma_1\gamma_2(1+\gamma_2) \sum_{k\in\{j,\hat{j}\}} \sum_{z_i \in \mathbf{z}_k} \langle \mathbf{h}_i, H_k^f + H_k^b \rangle z_i \\ &\quad + (1 - \gamma_1(1+\gamma_2)^2) \sum_{k\in\{j,\hat{j}\}} \sum_{z_i, z_{\hat{i}} \in \mathbf{z}_k} \langle \mathbf{h}_i, \mathbf{h}_{\hat{i}} \rangle z_i z_{\hat{i}}, \end{aligned} \quad (11)$$

where R is a constant and is irrelevant to optimization. γ_1 controls the relative importance of foreground-background dissimilarity. γ_2 is set to the ratio between the foreground and background regions and is not a tuneable parameter. To make sure that the *graph-cut* algorithm [64] can be adopted, this term must satisfy the regularity condition [64]. Namely, the coefficient $(1-\gamma_1(1+\gamma_2)^2)$ must not be larger than 0. Following [41], we set γ_1 to $\frac{1}{(1+\gamma_2)^2}$ and let $\gamma = \frac{\gamma_2}{(1+\gamma_2)}$. This discriminative term D becomes

$$\begin{aligned} D(\mathbf{z}_j, \mathbf{z}_{\hat{j}}) = R &- 2 \sum_{z_i \in \mathbf{z}_j, z_{\hat{i}} \in \mathbf{z}_{\hat{j}}} \langle \mathbf{h}_i, \mathbf{h}_{\hat{i}} \rangle z_i z_{\hat{i}} \\ &+ 2\gamma \sum_{k\in j,\hat{j}} \sum_{z_i \in \mathbf{z}_k} \langle \mathbf{h}_i, H_k^f + H_k^b \rangle z_i. \end{aligned} \quad (12)$$

In (12), the value of γ depends only on γ_2, which is set to the area ratio between the foreground and background. We will discuss how to determine the value γ_2 later.

3) Coupling Term $C(\mathbf{y}_i, z_i)$: This term encourages the coherence between the co-saliency and co-segmentation results. For measuring the degree of coherence on superpixel i, we first compute its mean saliency value by

$$s_i = \sum_{m=1}^{M} y_{i,m} s_{i,m} = \mathbf{y}_i^\top \mathbf{s}_i, \quad (13)$$

where $\mathbf{y}_i = [y_{i,1} \ \ldots \ y_{i,M}]^\top \in [0,1]^M$ is the weight vector for saliency proposal fusion on superpixel i. $s_{i,m} \in [0,1]$ is the mean saliency value of proposal m on superpixel i. Note that the values of $\{s_{i,m}\}_{m=1}^M$ are in $[0,1]$ and vector $\mathbf{y}_i$ is a distribution, thus $s_i \in [0,1]$. A higher mean saliency value s_i implies the higher likelihood of superpixel i belonging to foreground. To enhance the consistency between co-saliency detection and co-segmentation, this term, penalizing the cases where one of s_i and z_i is large while the other is small, is defined by

$$\sum_{v_i \in \mathcal{V}} C(\mathbf{y}_i, z_i) = \sum_{i=1}^{N} [s_i(1-z_i) + (1-s_i-\pi)z_i] \times \mathtt{size}(v_i), \quad (14)$$

where $\pi \in [0,1]$, called the *background shift*, is introduced to adjust the likelihood of background superpixels. It is often used in co-saliency detection, e.g. [41], [57], to prevent the trivial solutions that all superpixels are assigned to background. We will discuss how to set its value in the experiments. In (14), the sizes of superpixels are also taken into account.

4) Pairwise Term $B_1(\mathbf{y}_i, \mathbf{y}_{\hat{i}})$ for Co-Saliency Detection: We observe that different saliency proposals have individual strengths and weaknesses. For instance, proposals based on the background prior may not work well for objects with considerable overlap with image boundaries. To address this issue, this term encourages the neighboring superpixels in graph $\mathcal{G}$ to use similar subsets of saliency proposals. Its formulation is given below

$$\sum_{e_{i\hat{i}}\in\mathcal{E}} B_1(\mathbf{y}_i, \mathbf{y}_{\hat{i}}) = \sum_{e_{i\hat{i}}\in\mathcal{E}} A(i,\hat{i})\|\mathbf{y}_i - \mathbf{y}_{\hat{i}}\|_2^2 = tr(YLY^\top), \quad (15)$$

where L is the graph Laplacian of $\mathcal{G}$ with affinity matrix A.

5) Pairwise Term $B_2(z_i, z_{\hat{i}})$ for Co-Segmentation: This binary term is imposed to enforce the spatial smoothness of co-segmentation results. It is defined by

$$\sum_{e_{i\hat{i}}\in\mathcal{E}} B_2(z_i, z_{\hat{i}}) = \sum_{e_{i\hat{i}}\in\mathcal{E}} A(i,\hat{i})\|z_i - z_{\hat{i}}\|_2^2 = tr(ZLZ^\top). \quad (16)$$

E. Optimization

Simultaneously solving the two sets of variables Y and Z is hard. An alternating strategy is adopted to optimize the variables in (3). At each iteration, one set of the variables is optimized while keeping the other fixed, and then their roles are switched. Iterations are repeated until the convergence of the energy function values.

1) On Optimizing Y: By fixing Z, the optimization problem in (3) becomes

$$\begin{aligned} J(Y) = {} & \alpha_3 \sum_{i:v_i\in\mathcal{V}} C(\mathbf{y}_i, z_i) + \alpha_4 \sum_{e_{i\hat{i}}\in\mathcal{E}} B_1(\mathbf{y}_i, \mathbf{y}_{\hat{i}}) \\ & + \alpha_1 \sum_{i:v_i\in\mathcal{V}} U(\mathbf{y}_i) + \|Y\|_2^2 \\ & \text{s.t. } \|\mathbf{y}_i\|_1 = 1, \mathbf{y}_i \geq \bar{\mathbf{0}}, \quad \text{for } 1 \leq i \leq N. \end{aligned} \quad (17)$$

The above constrained optimization problem is a *quadratic programming* problem. We efficiently solve it by using the public software CVX [65].

2) On Optimizing Z: By fixing Y, the optimization problem in (3) is reduced to

$$\begin{aligned} J(Z) = {} & \alpha_3 \sum_{i:v_i\in\mathcal{V}} C(\mathbf{y}_i, z_i) + \alpha_5 \sum_{e_{i\hat{i}}\in\mathcal{E}} B_2(z_i, z_{\hat{i}}) \\ & + \alpha_2 \sum_{1\leq j<\hat{j}<n} D(\mathbf{z}_j, \mathbf{z}_{\hat{j}}) \\ & \text{s.t. } z_i \in \{0,1\}, \quad \text{for } 1 \leq i \leq N, \end{aligned} \quad (18)$$

which is a *binary labeling* problem. The energy function in (18) is graph representable and regular, and hence can be efficiently minimized by graph-cut [64].

For initialization, we solve the weights Y for saliency proposal fusion via (17) with the coupling term C removed. The saliency maps are generated via region-wise fusing the saliency proposals with optimized Y. We binarize each saliency map into foreground-background segmentation via Otsu's thresholding method. With the binary maps, the averaged area ratios of the foreground and the background of images can be measured, then, $\gamma = \frac{\gamma_2}{(1+\gamma_2)}$ in (11) is determined. It follows that the optimization problems in (17) and (18) can be iteratively solved. The value of the objective function decreases and converges to a local optimum. To conclude this section, we summarize our approach in Algorithm 1.

Algorithm 1 The Optimization Procedure of Our Method

Input: Images $\mathcal{I} = \{I_1, I_2, \cdots, I_n\}$, Max Iteration T;
Output: Co-saliency maps Y and co-segmentation masks Z;

```
Generate M saliency proposals for I; (Sec. III-A)
Decompose each image into superpixels; (Sec. III-B)
Extract features for each superpixel; (Sec. III-B)
Construct graph G = (V, E) with affinity matrix A in (2);
Initialize the saliency maps via (17) with term C removed;
Set γ in (12) based on the foreground-background ratios;
Iteration ← 1;
while Iteration ≤ T do
    Solve Y for co-saliency detection via (17);
    Solve Z for co-segmentation via (18);
    Iteration = Iteration + 1
end while
```

IV. Experimental Results

We evaluate the proposed method in this section. The benchmark datasets used for evaluation are described first. The adopted evaluation metrics and implementation details are then given. Finally, the qualitative and quantitative results are reported, analyzed, and discussed.

A. Datasets

We evaluate our method for co-saliency detection and co-segmentation on four benchmark datasets. Two benchmarks, the *Image-Pair* [10] and *iCoseg* [66] datasets, are used for performance evaluation on both tasks. The challenging *Cosal2015* [42] and *MSRC* [67] datasets serve as the testbeds for co-saliency detection and co-segmentation, respectively.

1) Image-Pair: This dataset has 105 image pairs with manually labeled ground truth. Each image pair contain one or multiple common objects appearing on two distinct backgrounds. We use the whole dataset for co-saliency detection, and the subset of 30 pairs used in [68] for co-segmentation.

2) iCoseg: It is a large-scale dataset for both co-saliency detection and co-segmentation. It contains 38 groups of total 643 images with manually labeled ground truth. Each group has 4 $\sim$ 42 images. We use the whole 38 groups for co-saliency detection and follow [48], [57] using the same 31 groups for co-segmentation. The images of a group contain single or multiple similar objects with various poses and sizes on complex backgrounds. Therefore, this benchmark is more challenging than the *Image-Pair* dataset for both co-saliency detection and co-segmentation.

3) Cosal2015: It is the largest and the most challenging dataset for co-saliency detection. It has 50 image groups, each of which contains 26 $\sim$ 52 images, with a total of 2015 images. Images of a group contain objects of a specific category. Variations caused by different object poses and scales, background clutters, and uncorrelated objects make this dataset quite challenging.

4) MSRC: This dataset is widely used for image co-segmentation. It consists of 14 groups with 418 images. Each group has about 30 images. Compared with the *iCoseg*, instances in each group of the *MSRC* dataset have higher appearance variations and less regular object boundaries, such as the thin branch of a tree. Thereby, this dataset is more challenging than *iCoseg*.

(a)	(b)	(c)	(d)	(e)	(f)
AP	0.975	0.970	AP	0.991	0.991
AUC	0.991	0.991	AUC	0.997	0.995
F-measure	0.939	0.935	F-measure	0.937	0.983
F_β^w	0.307	0.772	F_β^w	0.320	0.855

Fig. 5. Deficiency of AP, AUC, and F-measure. (a) & (d) The ground truth of two examples. (b) & (c) Two saliency proposals for (a). (e) & (f) Two saliency proposals for (d). Instead, F_β^w can successfully discriminate the map quality.

B. Evaluation Metrics

Let TP, TN, FP and FN respectively denote the numbers of true positives, true negatives, false positives and false negatives when evaluating a predicted binary map with respect to the ground truth figure-ground segmentation. The precision ($\mathcal{P}$), recall ($\mathcal{R}$), and the false positive rate (FPR) are respectively defined by

$$\mathcal{P} = \frac{TP}{TP+FP}, \quad \mathcal{R} = \frac{TP}{TP+FN}, \quad \text{and } FPR = \frac{FP}{TN+FP}. \tag{19}$$

To evaluate the performance of co-saliency detection, we first consider three widely used criteria, i.e. *average precision* (AP), *area under the ROC curve* (AUC) and F-measure (F_β). AUC can be considered as the aggregated statistics from the *receiver operating characteristic* (ROC) curve for the true positive rate (or recall) and false positive rate. AP is the score computed as the area under the precision and recall curve (PR curve). The PR and ROC curves are generated by thresholding the pixels in the predicted co-saliency maps with 256 levels from 0 to 1. Note that the number of non-salient pixels is often much larger than the number of salient pixels in saliency detection. Therefore, AP is more informative than AUC since AUC is often over-optimistic. Meanwhile, with a self-adaptive threshold $T = \mu + \sigma$, where μ and σ denote the mean and standard deviation of saliency values in a saliency map respectively, F-measure, defined as

$$\text{F-measure} = \frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}, \tag{20}$$

is obtained by the harmonic mean of the precision and recall, with $\beta^2 = 0.3$ to emphasize more on recall as suggested in [42], [70], and [71].

As pointed out in [72], the traditional measures mentioned above are less discriminative in some circumstances. Two such examples are shown in Fig. 5. The saliency proposals in Figs. 5(c) & 5(f) are perceptually closer to the respective ground truth in Figs. 5(a) & 5(d). However due to the combination of *interpolation*, *dependency*, and *equal importance* flaws introduced in [72], the proposals in Figs. 5(b) & 5(e) may have higher AP, AUC and F-measure scores than those in Figs 5(c) & 5(f), respectively. To address this issue, we also adopt a generalized F-measure, i.e. F_β^w [72], defined as

$$F_\beta^w = \frac{(1+\beta^2)\mathcal{P}^w \cdot \mathcal{R}^w}{\beta^2 \cdot \mathcal{P}^w + \mathcal{R}^w}, \tag{21}$$

which alleviates the hidden flaws of AP, AUC and F-measure for more objective evaluation of the detected saliency maps. In the experiment, we set $\beta = 1$ in (21) by following the original setting in [72] that equally weighs the importance of *weighted precision* ($\mathcal{P}^w$) and *weighted recall* ($\mathcal{R}^w$) based on the similar definitions in (19) with four weighted basic quantities, i.e. , TP^w, TN^w, FP^w and FN^w, defined as:

$$TP^w = (1-E^w) \cdot G \tag{22}$$
$$TN^w = (1-E^w) \cdot (1-G) \tag{23}$$
$$FP^w = (E^w) \cdot (1-G) \tag{24}$$
$$FN^w = (E^w) \cdot G, \tag{25}$$

where $\cdot$ denotes the inner product, and G and E^w respectively denote the column-stack representation of the binary ground truth, and the column-stack weighted error map (defined as $|G-D|$, with D being the column-stack representation of the predicted saliency map) by considering the individual pixel error according to their relative location and neighborhood information by referring to the ground truth.

For co-segmentation, we adopt two widely used criteria, i.e. *accuracy* ($\mathcal{A}$) and *jaccard index* ($\mathcal{J}$). Accuracy is the percentage of pixels that are correctly predicted in co-segmentation. Jaccard index, also named as "IoU", is the ratio of the intersection to the union of the segmented object and the foreground in ground truth. The two criteria are defined below

$$\mathcal{A} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{and } \mathcal{J} = \frac{TP}{TP+FP+FN}. \tag{26}$$

C. Implementation Details

For saliency proposal fusion, we choose four single-image saliency proposals (SISP), i.e. `DSR` [6], `MR` [7], `DRFI` [27] and `RBD` [29], together with three multiple-image saliency proposals (MISP) by distinct co-saliency evidences, i.e. `SpC` (Spatial cue), `Cor` (Corresponding cue) and `CoC` (Contrast cue), extracted from the `CBCS` model [8]. In general, methods `DST`, `MR`, `DRFI` and `RBD` measure the saliency based on feature distinctness between the predicted foreground and the surrounding superpixels. Thus, they may not perform well on objects that are connected on image boundaries. `DRFI` in comparison gives better results than the other SISPs since it additionally utilizes the supervised learning approach to map the local feature vector to a saliency score. Merely using SISPs may focus on only the salient objects that do not repeatedly appear across images, and neglect the low-contrast co-salient objects in images. Therefore, three independent co-saliency evidences from methods `CBCS` are used to complement the deficiency of SISPs. For instance, the correspondence evidence `CoC` can detect the co-occurring regions across images. In short, we select these proposals by jointly considering their performances, popularity and complementary effect on proposal fusion. Generally, the more accurate the saliency proposals, the better the co-saliency detection, which further benefits the joint co-segmentation task. To generate the saliency proposals on the respective dataset, we run the source

TABLE I

QUANTITATIVE RESULTS FOR CO-SALIENCY DETECTION ON THREE BENCHMARK DATASETS. "SI" AND "CS" DENOTE THE METHODS FOR SINGLE-IMAGE SALIENCY DETECTION AND MULTI-IMAGE CO-SALIENCY DETECTION, RESPECTIVELY. THE BEST RESULT IS HIGHLIGHTED IN BOLD, AND "-" MEANS NO REPORTED RESULT ON THAT DATASET

Method	Year	Setting	*Image-Pair*				*iCoseg*				*Cosal2015*			
			AP	AUC	F_β	F_β^w	AP	AUC	F_β	F_β^w	AP	AUC	F_β	F_β^w
DSR [6]	CVPR2013	SI	0.859	0.951	0.811	0.658	0.787	0.920	0.748	0.557	0.680	0.892	0.657	0.464
MR [7]	CVPR2013	SI	0.882	0.948	0.849	0.681	0.798	0.901	0.780	0.544	0.662	0.871	0.659	0.458
DRFI [27]	CVPR2013	SI	0.887	0.960	0.838	0.642	0.846	0.964	0.789	0.574	0.703	0.918	0.681	0.452
RBD [29]	CVPR2014	SI	0.809	0.893	0.794	0.650	0.821	0.944	0.779	0.604	0.667	0.890	0.659	**0.487**
SpC [8]	TIP2013	CS	0.812	0.912	0.802	0.460	0.757	0.909	0.677	0.345	0.541	0.762	0.534	0.299
Cor [8]	TIP2013	CS	0.674	0.879	0.659	0.484	0.431	0.718	0.420	0.251	0.345	0.625	0.253	0.253
CoC [8]	TIP2013	CS	0.594	0.783	0.601	0.394	0.643	0.849	0.622	0.336	0.321	0.543	0.320	0.204
CBCS [8]	TIP2013	CS	0.859	0.934	0.788	0.581	0.800	0.938	0.741	0.452	0.594	0.823	0.568	0.318
CSHS [46]	SPL2014	CS	0.896	0.952	0.869	0.622	0.845	0.957	0.755	0.487	0.621	0.851	0.619	0.340
SACS [9]	TIP2014	CS	0.926	0.977	0.873	0.694	0.871	0.966	0.796	0.557	0.724	**0.919**	0.692	0.446
CoDW [42]	IJCV2016	CS	-	-	-	-	0.877	0.957	0.799	0.476	**0.744**	0.913	**0.705**	0.385
DIM [69]	TNNLS2016	CS	0.933	0.973	0.862	0.480	0.877	**0.968**	0.792	0.479	-	-	-	-
SGCS [15]	ICME2017	CS	0.937	0.979	0.884	0.675	-	-	-	-	-	-	-	-
MIL [70]	TPAMI2017	CS	-	-	-	-	0.875	0.964	0.814	0.527	-	-	-	-
Ours	/	CS	**0.945**	**0.980**	**0.898**	**0.741**	**0.878**	**0.968**	**0.820**	**0.627**	0.722	0.910	0.696	0.462

code from the corresponding publications with the default settings.

We evaluate our approach in two different perspectives/tasks on each dataset, namely, *co-segmentation guided co-saliency detection* and *co-saliency detection guided co-segmentation* with the same optimization model (3). For fair comparison with the state-of-the-art methods each of which tunes its parameters for one specific task (co-saliency detection or co-segmentation) on a dataset (*Image-Pair*/*iCoseg*/*Cosal2015*/*MSRC*), we also tune the parameters of our approach in a task-dataset centric manner, namely seeking a set of optimal parameter values for each task on each dataset.

We search for the proper values of the parameters in the order based on their importance to our model. In addition, we search for the parameters regarding co-saliency detection first to provide a proper initialization for co-segmentation. The resultant order is, α_1, α_3, π, α_4, α_2, and α_5. One parameter is tuned while the others are fixed. The tuning process is done sequentially in the above order and iteratively until the performance of the task, co-saliency detection or co-segmentation, no longer improves. We follow the competing methods such as [57] by adjusting the background shift per image group on the *iCoseg* and *MSRC* datasets for co-segmentation evaluation. When solving the optimization for (3), alternating optimization scheme of co-saliency detection and co-segmentation is repeated for a few iterations until the energy in (3) converges. Our model converges rapidly, so we set the maximum number of iterations $T = 4$ in Algorithm 1.

D. Co-Segmentation Guided Co-Saliency Detection

We evaluate the effectiveness of the proposed model for co-segmentation guided co-saliency detection on the *Image-Pair*, *iCoseg* and *Cosal2015* datasets in the following.

1) Image-Pair Dataset: We compare our approach with seven adopted saliency proposals and other co-saliency detection methods, including the bottom-up based co-saliency model CBCS [8] and CSHS [46], the adaptive weight map-wise fusion-based co-saliency model SACS [9], and our prior work SGCS [15] based on the same set of the saliency proposals. We further include the deep learning-based approach DIM [69] for comparison which uses the stacked denoising autoencoder to learn the intra- and inter-saliency information with a supervised training phase on the auxiliary ASD [21] dataset. We either reproduce the co-saliency detection results from the released code [8], [9], [15] or directly get the results from their Websites [69].

TABLE I shows the overall performance of the evaluated approaches in different metrics and Fig. 6(a) displays the PR and ROC curves. We find that fusion-based approaches consistently improve their saliency proposals by properly combining these proposals. SACS addresses the inherent issues of the traditional fixed-weight linear fusion model via adaptively emphasizing the higher-quality saliency proposals. Our method further addresses the problems of map-wise fusion in SACS by using region-wise fusion. Meanwhile, it enhances the co-segmentation strength by additionally considering the figure-background distinctness in (11) besides encouraging only the foreground coherence in SGCS, thus achieving the best results in all evaluation metrics. Our model even surpasses the state-of-the-art supervised deep learning approach DIM with the gains of about 1.2% in AP and 26.1% in F_β^w.

Fig. 7 visualizes the saliency maps generated by different approaches on two image pairs. Taking the second pair as an example, none of the single-image saliency detection methods, i.e. DSR [6], MR [7], DRFI [27], and RBD [29], can get the dominating performance as they either produce some unfavorable false alarms or miss some object parts. The proposal Cor [8] searches the corresponding regions and the proposal CoC [8] looks for the contrast regions across images. They give relatively clean results in these examples. However, the saliency maps in the object regions are not sharp enough and there is noise in background. Method CBCS [8] jointly takes into account the intra-image CoC, SpC cues and inter-image Cor, CoC and the SpC cues from the paired images, which helps suppress the false positives. Method SACS instead exploits a map-wise fusion of multiple proposals to yield the final saliency maps. We observe that it often uniformly spotlights the co-salient regions. We also consider a variant of our model Ours-iter1, which shows the saliency

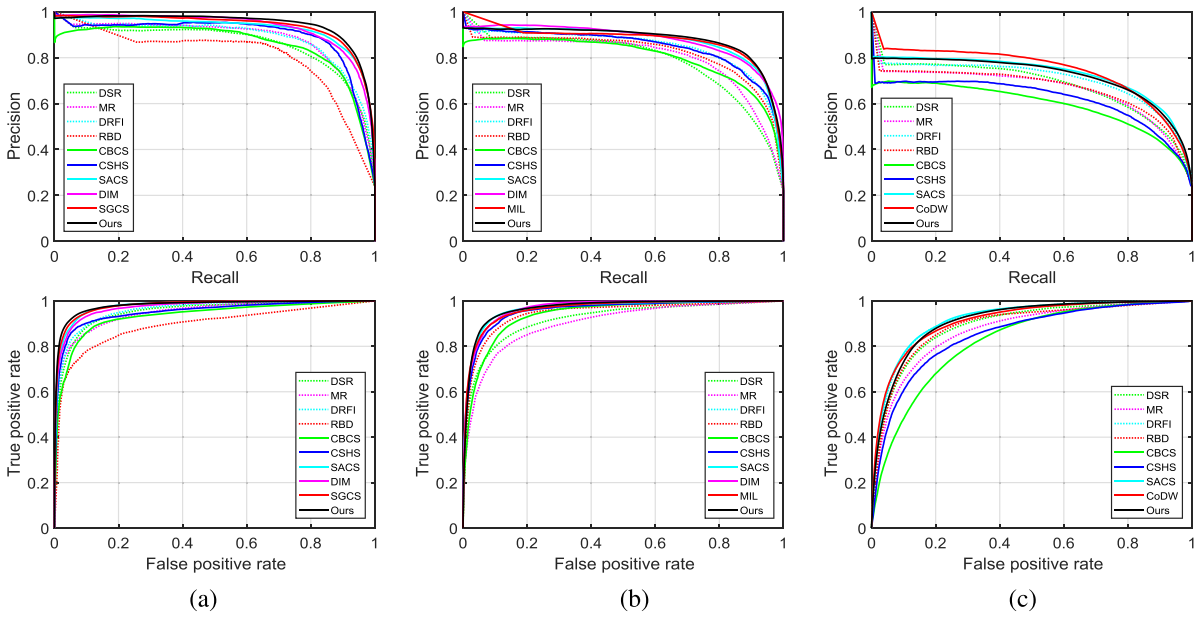


Fig. 6. Performance of co-saliency detection in PR and ROC curves on three benchmark datasets including the (a) Image-Pair, (b) iCoseg, and (c) Cosal2015 datasets. The models adopted to generate our fusion proposals are plotted in dash lines, while the state-of-the-art co-saliency detection methods are in solid lines.

maps produced by our model at the first iteration, namely without the aid of co-segmentation. This variant combines the locally complementary signal strengths from different saliency proposals, and produces comparable results with the SACS, which needs an additional post-processing refinement step. However, without higher level objectness information, some false positives are present. After turning on the coupling term, our regional fusion additionally seeks consensus with the co-segmentation results, and it further tackles the limit of region-wise fusion SGCS. Our model yields the saliency maps superior to those generated by all the competing methods.

2) iCoseg Dataset: Next, we evaluate our co-segmentation guided co-saliency detection on the iCoseg dataset. Fig. 6(b) displays the PR and ROC curves, and TABLE I shows the overall evaluation scores. Our approach results in a large performance gain over seven adopted saliency proposals in all evaluation metrics. We further compare our approach with several powerful co-saliency detection models, including the conventional unsupervised approaches, i.e. CBCS [8], CSHS [46] and SACS [9], and the learning-based models, DIM [69], CoDW [42] and MIL [70], with more complex initialization, such as taking advantage of the deep networks pre-trained on other datasets or taking negative samples for additional background images from other groups. It is worthwhile to note that SACS applying an adaptive combination of the adopted saliency proposals already generates comparable or superior quantitative results to the aforementioned learning-based models, especially on the weighted F-measure scores. As a fusion-based method, our method achieves even better performance in all evaluation metrics by integrating the segmentation guidance into saliency map fusion.

Visual comparison is shown in Fig. 7. It can be observed that method CBCS is insufficient to handle the cases of size-varying co-salient objects, changing backgrounds, and different illumination conditions, especially in the groups of Cheetah and Salisbury. Single-image saliency detection DRFI, by training a random forest regressor based on the extracted over-complete features, gives more preferable results than CBCS. However, many salient parts are still missing. SACS achieves significant improvement over CBCS and DRFI because of its model of self-adaptive weighted fusion. Our co-segmentation guided region-wise fusion approach collaboratively captures the objectness cues and estimates the region-wise goodness of different proposals, thus yielding higher-quality co-saliency maps. The good properties of our co-saliency maps include uniformly highlighted objects and less false positives. More importantly, our method gives clearer borders between the salient objects and background regions, which is favorable for the co-segmentation task.

3) Cosal2015 Dataset: We compare our method with existing co-saliency detection approaches and report the overall statistics in TABLE I and Fig. 6(c). Likewise, our method is compared with the conventional bottom-up approaches, i.e. CBCS [8], CSHS [46], the fusion-based approach with the same set of saliency proposals, SACS [9], as well as the state-of-the-art method CoDW [42] proposed by the authors who established this dataset. In this dataset, our method slightly falls behind or is comparable to SACS [9] and CoDW [42] and performs favorably against the other competing approaches. Compared with SACS [9] and CoDW [42], our method has slightly lower performance in AP, AUC, and F_β but has better results in F_β^w. As pointed out in Fig. 5 by [72], the measures, including AP, AUC, and F_β, have some limitations and may lead to inaccurate evaluation; instead, the measure F_β^w gives judgment more closed to human perception. Furthermore, CoDW [42] requires a set of object proposals to pre-train the restricted Boltzmann machines as the feature extractor, but our method does not require those pre-processing steps. In SACS [9], a post-processing step is used to refine the object boundary by suppressing the false positive regions. In contrast, our method achieve higher F_β^w scores without an extra post-processing step.

Fig. 7. (a) Six image groups and object ground truth with left two groups from the Image-Pair dataset, middle two groups from the iCoseg dataset, and right two groups from the Cosal2015 dataset. (b) ∼ (h) The adopted saliency proposals produced by approaches including (b) DSR [6], (c) MR [7], (d) DRFI [26], (e) RBD [29], (f) SpC [8], (g) Cor [8] and (h) CoC [8]. (i) ∼ (k) Results by co-saliency detection methods including (i) CBCS [8], (j) CSHS [46], and (k) DIM [69] on left two groups, MIL [70] on middle two groups, and CoDW [42] on right two groups. (l) ∼ (n) Results by fusion-based approaches including (l) SACS [9], (m) Ours-iter1: our approach without referring to the co-segmentation evidence, and (n) Ours.

To gain insight into the quantitative results, Fig. 7 visualizes the detected saliency maps with different approaches. Taking the right image of class "Deer" as an example, we see the results by the adopted proposals, i.e. DSR [6], MR [7], DRFI [27], RBD [29], SpC [8], Cor [8] and CoC [8] have their respective strength even though they do not give satisfactory results in overall. Specifically, the results by MR and SpC successfully highlight the deer body but missing the head region. On the contrary, results by DRFI successfully delineate the object's region, but with low figure-background contrast. Our method adaptively selects the reliable proposals region-wisely to form a better co-saliency map than the adopted proposals as well as the SACS by map-wise fusion manner. Furthermore, we observe the results by the unsupervised state-of-the-art methods, i.e. CBCS [8], CSHS [46], SACS [9] and CoDW [42] have many false positives in the background because the common objects in the cases are similar to the background. With the segmentation guidance, our method can more effectively remove the false positives due to the low figure-background contrast issues in those images.

TABLE II
CO-SEGMENTATION RESULTS IN JACCARD INDEX ($\mathcal{J}$) AND ACCURACY ($\mathcal{A}$) ON THE IMAGE-PAIR DATASET

Method	$\mathcal{J}$	$\mathcal{A}$
Jou10 [52]	59.1	79.0
Yu14 [54]	55.6	86.2
Gao13 [68]	-	92.4
Meng13 [56]	77.7	92.8
Ours	**81.9**	**94.8**

E. Co-Saliency Detection Guided Co-Segmentation

In the following, we evaluate our model for co-segmentation with the integration of co-saliency detection on the *Image-Pair*, *iCoseg* and *MSRC* datasets.

1) Image-Pair Dataset: We first evaluate the co-segmentation performance on the *Image-Pair* dataset. TABLE II reports the performances of our method and four powerful co-segmentation methods, including Jou10 [52],

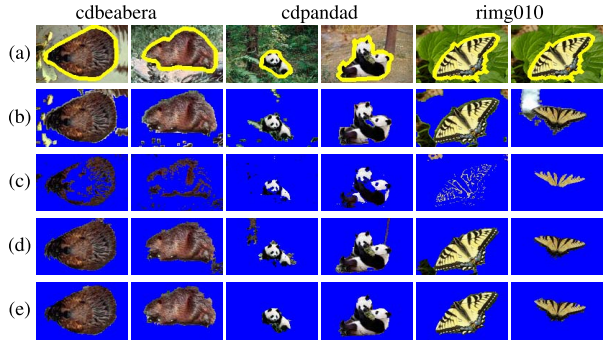


Fig. 8. (a) Three image pairs from the Image-Pair dataset for co-segmentation with the ground truth marked by the contours. (b) ∼ (e) Segmentation results generated by different approaches including (b) Jou10 [52], (c) Yu14 [54], (d) Meng13 [56], and (e) Ours.

Yu14 [54], Gao13 [68], and Meng13 [56]. Except for Jou10, methods mentioned above similarly require the prior knowledge of foreground by either bounding boxes or saliency information as in our model. Overall, the performance gain of our method over Meng13, the best competing method tailored for paired image co-segmentation, is significant, i.e. 4.2% gain in Jaccard index and 2% gain in accuracy.

In Jou10 [52], both spatial and color features are used to train a maximum margin classifier with a formulation combining discriminative clustering and spectral clustering. An important parameter μ weighs the influence of spatial and color consistency in the discriminative cost function. To obtain better results of this model, we tune μ for each of the 30 image pairs while keep the other settings adopted in the released code. As shown in the second row of Fig. 8, this method can identify the common regions, but the results are noisy because of the complex image appearance. For instance, due to the lake reflection, this method incorrectly classifies the reflection as parts of the foreground in the first cdbeabera image. In addition, Jou10 [52] usually requires more images to derive a good hyperplane separating foreground instances from background.

The MRF-based model Yu14 [54] considers individual image segmentation with the constraints of high foreground similarity by using the Gaussian mixture models (GMMs). In this model, image segmentation is similarly initialized via the co-saliency priors by CBCS [8]. We reproduce their results with the recommended settings. As shown in the third row of Fig. 8, this method has fewer false positives compared to Jou10, but it suffers from the object variations across images. In fact, it has the lowest Jaccard indices in TABLE II.

The method Meng13 [56] combines the active contour method with a rewarding strategy based on both the foreground similarity and background consistency. We also reproduce their results with the default settings. As shown in the fourth row of Fig. 8, this method is more preferable compared to the previous two competing methods. However, the active contour segmentation requires extra initial bounding boxes for the objects of interest. In a different manner, we estimate the initial object regions via saliency priors obtained by jointly solving co-saliency detection and co-segmentation. Not only the foreground similarity but also background consistency constraints in the perspectives of co-saliency detection and co-segmentation are taken into account. Both TABLE II and Fig. 8 show that our method remarkably outperforms the competing methods.

TABLE III
CO-SEGMENTATION RESULTS IN JACCARD INDEX ($\mathcal{J}$) AND ACCURACY ($\mathcal{A}$) ON THE ICOSEG AND MSRC DATASETS

	iCoseg		MSRC	
Method	$\mathcal{J}$	$\mathcal{A}$	$\mathcal{J}$	$\mathcal{A}$
Jou10 [52]	39.5	61.0	45.2	70.8
Kim11 [73]	39.8	70.3	34.2	54.7
Jou12 [53]	42.6	70.2	50.7	73.6
Rub13 [57]	69.3	89.8	**68.1**	**87.7**
Fu15 [55]	59.4	88.5	-	-
Ours	**72.3**	**90.8**	67.8	86.5

2) iCoseg Dataset: Next, we evaluate our method for co-segmentation on the *iCoseg* dataset. TABLE III reports the quantitative results and the left part of Fig. 9 shows visual comparison among our method and the existing methods for co-segmentation of more than two images, including Jou12 [53], Rubi13 [57], and Fu15 [55]. We download each of their co-segmentation masks from the authors' Websites.

The method Jou12 [53] extends their previous work [52] to co-segment multiple images that consist of the multiple objects by an iterative EM algorithm. However, without proper saliency information, there are still similar issues that background regions with similar image appearance across multiple images tend to be considered as objects of interest, leading to false positives as displayed in Fig. 9(b).

The method Rubi13 [57] addresses the issues for images with noisy background or irrelevant objects. Using the SIFT flow and visual saliency, it separates the common objects from the noisy signals by alternating dense pixel correspondence inference and foreground estimation. This method with the aid of saliency information greatly improves the figure-ground separation compared to Jou12. In Fig. 9(c), we observe that many background regions in Hot-Balloons and Cheetah images are successfully excluded. However, single-image saliency information rather than co-saliency priors obtained from the image sets may not be sufficient to handle large intra-objects shape variations as illustrated in the images of Kendo-Kendo.

The method Fu15 [55] solves an energy minimization problem that integrates the depth cue to help capture common object regions while excluding complex backgrounds by fusing several existing RGB-based co-saliency maps via a low-rank representation [9]. This method works well on removing the background regions from the foreground; however, it sometimes misses significant foreground regions, as shown in the group Kendo-Kendo of Fig. 9(d).

Compared to these methods, our model considering co-saliency and co-segmentation simultaneously achieves the improved co-segmentation performance. Unlike the competing method Fu15 with the map-wise integration of multiple saliency proposals to derive the co-saliency priors, our region-wise fusion method better integrates locally complementary saliency proposals, and hence guides and facilitates the following co-segmentation. Along the process of alternating optimization, better results are achieved with the iteratively refined co-saliency priors and the guided co-segmentation as illustrated in Fig. 9(e).

3) MSRC Dataset: We further evaluate our method for co-segmentation on the *MSRC* dataset and compare it with the

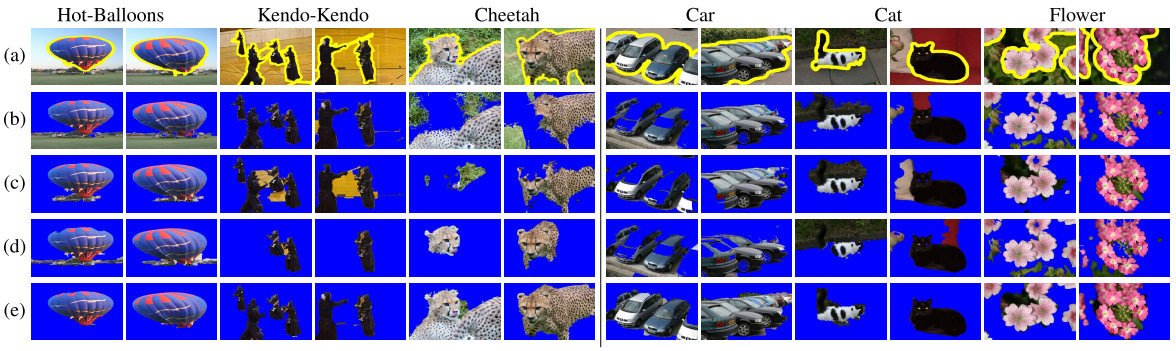


Fig. 9. (a) Six image groups from the iCoseg (left three groups) and MSRC (right three groups) for co-segmentation with their ground truth marked by the yellow contours. (b) ∼ (e) Segmentation results generated by different approaches including (b) Jou12 [53], (c) Rub13 [57], (d) Fu15 [55] on the iCoseg dataset and Jou10 [52] on the MSRC dataset, and (e) Ours.

models mentioned above that also have reported their results on the *MSRC* dataset except for method Fu15. As summarized in TABLE III, Kim11 yields relatively lower performance because it decomposes the multi-image co-segmentation problem into several paired-image graph bi-partition sub-problems to facilitate parallel computation. The *MSRC* dataset contains images of different instances sharing only the same class information; moreover, many *MSRC* images have similar backgrounds, e.g. similar airports frequently appear in the "Plane" class. These issues make it more challenging for the task of separating foreground from background on the *MSRC* dataset. Differently, our method considers the figure-background distinctness from all images to ease the potential difficulties in this situation. In general, our method performs reasonably well compared with other superpixel-based methods, i.e. Jou10, Kim11, and Jou12 which may suffer from the superpixel segmentation error due to the highly complex object content and its boundary. In contrast, Rub13 is a pixel-based method using its computed saliency map combined with single-image Grabcut for co-segmentation, which derives better segmentation results. In fact, our method is also superpixel-based but already generates comparable results to the pixel-based method Rub13, even without additional Grabcut post-processing.

To gain better insight, we display the visual results on the right side of Fig. 9. Taking the group "Car" for example, which is relatively more challenging than the other two cases since it exhibits various view angles on the same types of objects closely neighboring each other. However, without proper saliency information, many objects of interest are mistakenly regarded as the background regions, as shown in Fig. 9(b) and (d) by the method Jou12 [53] and their previous work Jou10 [52]. By leveraging saliency information, Rub13 is expected to identify objects' locations more precisely. However, we observe many background regions are also included in their results. In comparison, we derive co-saliency priors from the image sets to target the objects with large intra-object shape variations on the *MSRC* dataset. Our co-saliency priors enable our method to better localize objects. However, some background is also segmented out, as shown in the case "Flower." We expect fine superpixel extraction can be helpful to reduce the error.

F. Model Analysis

In the following, we evaluate the contribution of each energy term in (3), conduct the convergence analysis, and discuss the limitations of our method.

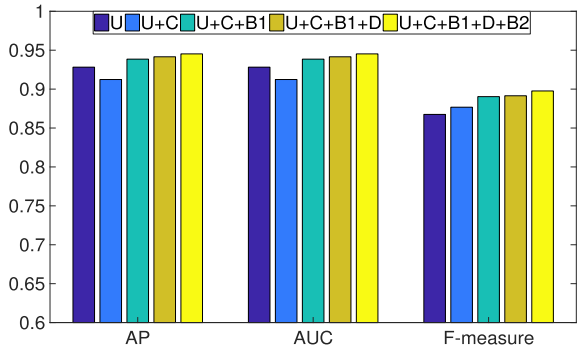


Fig. 10. Ablation studies on the Image-Pair dataset in AP, AUC, and F-measure (F_β).

1) Ablation Studies: Fig. 10 reports ablation studies on the Image-Pair dataset to investigate the contribution from each individual energy term to the proposed model. Fig. 11 shows the corresponding visual results. In general, we can see that adding the energy terms, U, C, B_1, D, and B_2 to the objective function in (3) step by step can progressively improve the results. Initially, we generate the original co-saliency map by applying the unary term U to locally search for the proper saliency proposals to fuse. As the energy terms responsible for segmentation are turned off at this stage, the corresponding segmentation mask in Fig. 11(b) is shown as the whole background by default. Next, by turning on the coupling term C, the co-saliency map in Fig. 11(b) is treated as object priors for the co-segmentation; meanwhile, it allows the segmentation mask to guide the region-wise saliency proposal fusion. By associating the information between the co-saliency map and co-segmentation mask, we observe high coherence between the fused saliency map and the segmentation result. Furthermore, after combining the co-saliency smoothness term, B_1, as illustrated in Fig. 11(d), the quality of the saliency map is improved. Meanwhile, scores in Fig. 10 are also elevated. Next, by adding the discriminative term D, which helps remove the potential background and recover the common foregrounds in images, the performance scores are improved. Finally, the best performance is obtained by encouraging the smoothness of the segmentation labeling using B_2, as illustrated in Fig. 11(f).

2) Convergence Analysis: The objective function values in (17) and (18) corresponding to co-saliency map fusion and co-segmentation in our iteration scheme are plotted

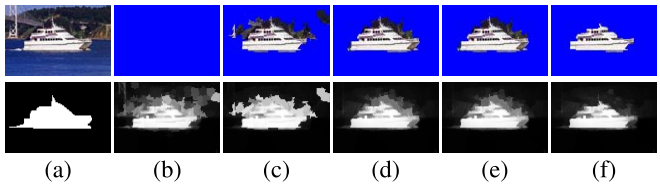


Fig. 11. Visual illustration of ablation studies. (a) An image from the Image-Pair dataset and its ground truth for joint co-saliency detection and co-segmentation. (b)∼(f) Results generated by using different combinations of the energy terms, including (b) $Reg{+}U$, (c) $Reg{+}U{+}C$, (d) $Reg{+}U{+}C{+}B_1$, (e) $Reg + U + C + B_1 + D$, and (f) $Reg + U + C + B_1 + D + B_2$.

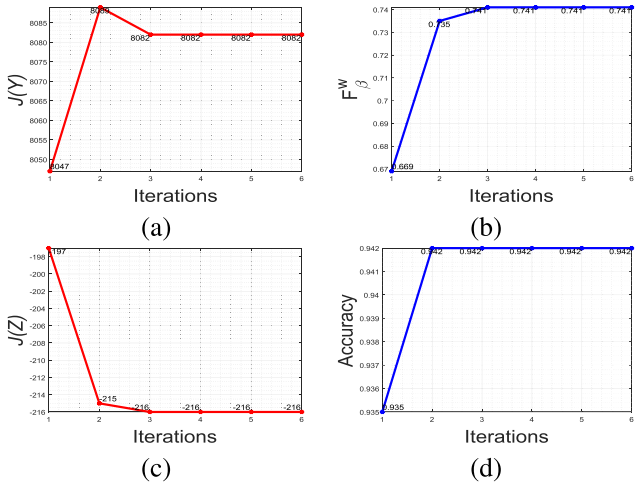


Fig. 12. Learning curves during optimizing (3) on the Image Pair dataset. (a) The objective values of (17). (b) The weighted F-measure scores of co-saliency detection. (c) The objective values of (18). (d) The accuracy scores ($\mathcal{A}$) of co-segmentation.

in Fig. 12. Initially, the adopted saliency proposals are integrated to produce a baseline for co-saliency detection, which provides a good initialization for image co-segmentation. By conducting the iteration scheme, the co-saliency detection and co-segmentation results are continually optimized according to Figs. 12(b) and (d), meanwhile, the energy curves in Figs. 12(a) and (c) converge rapidly. Since both the energy curves no longer appear obvious changes after 3 iterations, thus setting the maximum iteration number to 4 is reasonable in our experiments. Note that the reason why (17) has the lowest energy at the first iteration is that the coupling term C is off at that iteration. The term C is then turned on after the first-round co-segmentation masks are obtained.

3) Limitations: Our method is to combine the advantages of different saliency proposals for more accurate co-saliency detection; however, if the majority of the adopted saliency proposals lose their discriminative power toward salient objects, our fusion might fail due to the group voting scenario embedded in the unary term U. For instance, compared with the ground truth in Fig. 13(a), the saliency proposal from the method `MR` gives more favorable results than the other three SISM; however, the group voting scenario implicitly forces our method to trust more on the other methods, leading to the degenerated fused saliency map shown in the last image of Fig. 13 row (a). Therefore, how to independently emphasize the better saliency proposal to fuse can be one interesting

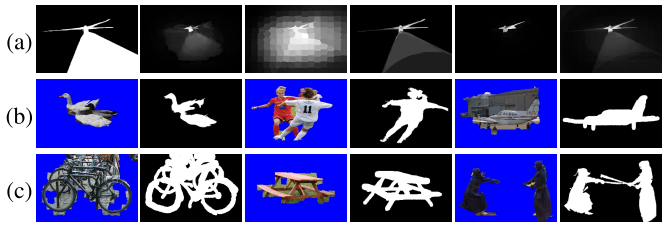


Fig. 13. Some challenging cases/examples where our method fails. (a) Most of the saliency proposals do not perform well. From left to right, the ground truth, four adopted SISMs from `DSR`, `MR`, `DRFI` and `RBD`, and our fused map are shown. (b) Multiple similar objects are present and/or the goal is to extract a particular object instance. (c) Objects of interest have complex shapes or long-thin boundaries.

direction to pursue. Next, Fig. 13 row (b) shows that three of our co-segmentation results include all the salient objects instead of the most commonly appearing objects. We find this is a prevalent co-segmentation problem to the other works as well since the soccer player appears in 30 out of 31 images in that group, and the dark-green background and the background airport also frequently show up among the whole images in that group. Lastly, Fig. 13 row (c) shows the inherent difficulty in segmentation, which typically encourages segments with short boundaries since the penalty we pay is the length of the cut as mentioned in [74]; however, not all natural objects have short boundaries, as in the example "Bike." Furthermore, some foreground parts can also be mistakenly excluded if they are assigned to the same superpixels with the background, like the long-thin Kendo sword in the Kendo group. In other words, the superpixel size may need to be set for the appropriate trade-off between segmentation accuracy and computational complexity, which is beyond the focus of this work.

V. Conclusions

In this paper, we have presented an unsupervised learning framework that simultaneously accomplishes co-saliency detection and co-segmentation. On the one hand, our method carries out saliency proposal fusion via jointly exploring the common object evidence generated from co-segmentation and the consensus among various saliency proposals. On the other hand, we take advantage of this joint optimization framework for an enhanced co-segmentation mask from the improved co-saliency priors. The benefits of the joint optimization formulation are evident as it produces the high-quality saliency maps by region-adaptive fusion of multiple locally complementary saliency proposals, and generates accurate co-segmentation masks with the aid of the iteratively refined co-saliency priors. Moreover, unlike existing co-saliency models relying on additional post-processing to smooth their model outputs, our formulation has already merged such advantages into the unified optimization process and generates even superior results in both tasks evaluated on three respective datasets under the same evaluation metrics. In future, we plan to apply deep learning techniques to the proposed segmentation guided fusion framework for category-specific object detection that can benefit specific applications where saliency maps or segmentation masks of high quality are appreciated, such as the passenger-specific salient object detection for the development of autonomous driving.

REFERENCES

[1] H. Fu, D. Xu, B. Zhang, S. Lin, and R. K. Ward, "Object-based multiple foreground video co-segmentation via multi-state selection graph," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3415–3424, Nov. 2015.

[2] K. R. Jerripothula, J. Cai, and J. Yuan, "Quality-guided fusion-based co-saliency estimation for image co-segmentation and colocalization," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2466–2477, Sep. 2018.

[3] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *J. Image Vis. Comput.*, vol. 29, no. 1, pp. 1–14, 2011.

[4] J. Guo, Z. Li, L.-F. Cheong, and S. Z. Zhou, "Video co-segmentation for meaningful action extraction," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2232–2239.

[5] H.-Y. Chen, Y.-Y. Lin, and B.-Y. Chen, "Co-segmentation guided Hough transform for robust feature matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 2388–2401, Dec. 2015.

[6] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2013, pp. 2976–2983.

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

[8] 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.

[9] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4175–4186, Sep. 2014.

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

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

[12] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, 2014, p. 23.

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

[14] C.-C. Tsai, X. Qian, and Y.-Y. Lin, "Image co-saliency detection via locally adaptive saliency map fusion," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2017, pp. 1897–1901.

[15] C.-C. Tsai, X. Qian, and Y.-Y. Lin, "Segmentation guided local proposal fusion for co-saliency detection," in *Proc. Int. Conf. Multimedia Expo*, Jul. 2017, pp. 523–528.

[16] 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.

[17] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[18] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 478–485.

[19] C. Xia, P. Wang, F. Qi, and G. Shi, "Nonlocal center-surround reconstruction-based bottom-up saliency estimation," in *Proc. Int. Conf. Image Process.*, Sep. 2013, pp. 206–210.

[20] C. Xia, F. Qi, and G. Shi, "An iterative representation learning framework to predict the sequence of eye fixations," in *Proc. Int. Conf. Multimedia Expo*, Jul. 2017, pp. 1530–1535.

[21] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

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

[23] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior," in *Proc. Brit. Conf. Mach. Vis.*, 2011, pp. 110.1–110.12.

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

[25] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 853–860.

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

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

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

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

[30] R. Huang, W. Feng, and J. Sun, "Saliency and co-saliency detection by low-rank multiscale fusion," in *Proc. Int. Conf. Multimedia Expo*, Jun./Jul. 2015, pp. 1–6.

[31] K. Fu, C. Gong, I. Y.-H. Gu, and J. Yang, "Normalized cut-based saliency detection by adaptive multi-level region merging," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5671–5683, Dec. 2015.

[32] K. Simonyan, A. Vedaldi, and A. Zisserman. (2013). "Deep inside convolutional networks: Visualising image classification models and saliency maps." [Online]. Available: https://arxiv.org/abs/1312.6034

[33] X. Li *et al.*, "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, Aug. 2016.

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

[35] 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.

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

[37] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Weakly supervised saliency detection with a category-driven map generator," in *Proc. Brit. Conf. Mach. Vis.*, 2017, pp. 1–13.

[38] D. Zhang, D. Meng, L. Zhao, and J. Han, "Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3538–3544.

[39] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[40] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Jun. 2015, pp. 3431–3440.

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

[42] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 215–232, 2016.

[43] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 568–579, Feb. 2018.

[44] X. Yao, J. Han, D. Zhang, and F. Nie, "Revisiting co-saliency detection: A novel approach based on two-stage multi-view spectral rotation co-clustering," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3196–3209, Jul. 2017.

[45] R. Huang, J. Sun, and W. Feng, "Color feature reinforcement for cosaliency detection without single saliency residuals," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 569–573, May 2017.

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

[47] L. Ye, Z. Liu, J. Li, W.-L. Zhao, and L. Shen, "Co-saliency detection via co-salient object discovery and recovery," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 2073–2077, Nov. 2015.

[48] K. R. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1896–1909, Sep. 2016.

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

[50] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching—Incorporating a global constraint into MRFs," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 993–1000.

[51] D. S. Hochbaum and V. Singh, "An efficient algorithm for co-segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 1–8.

[52] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1943–1950.

[53] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 542–549.

[54] H. Yu, M. Xian, and X. Qi, "Unsupervised co-segmentation based on a new global GMM constraint in MRF," in *Proc. Int. Conf. Image Process.*, Oct. 2014, pp. 4412–4416.

[55] H. Fu, D. Xu, S. Lin, and J. Liu, "Object-based RGBD image co-segmentation with mutex constraint," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4428–4436.

[56] F. Meng, H. Li, G. Liu, and K. N. Ngan, "Image cosegmentation by incorporating color reward strategy and active contour model," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 725–737, Apr. 2013.

[57] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in Internet images," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1939–1946.

[58] J. Han, R. Quan, D. Zhang, and F. Nie, "Robust object co-segmentation using background prior," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1639–1651, Apr. 2018.

[59] 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.

[60] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[61] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. (2014). "Return of the devil in the details: Delving deep into convolutional nets." [Online]. Available: https://arxiv.org/abs/1405.3531

[62] J. Li, J. Ding, and J. Yang, "Visual salience learning via low rank matrix recovery," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 112–127.

[63] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. Neural Inf. Process. Syst.*, 2009, pp. 2080–2088.

[64] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.

[65] M. Grant and S. Boyd. (2014). *CVX: MATLAB Software for Disciplined Convex Programming, Version 2.1*. [Online]. Available: http://cvxr.com/cvx

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

[67] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2005, pp. 1800–1807.

[68] Z. Gao, P. Shi, H. R. Karimi, and Z. Pei, "A mutual GrabCut method to solve co-segmentation," *J. Image Video Process.*, vol. 2013, no. 1, pp. 1–11, Apr. 2013.

[69] D. Zhang, J. Han, J. Han, and L. Shao, "Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1163–1176, Jun. 2016.

[70] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.

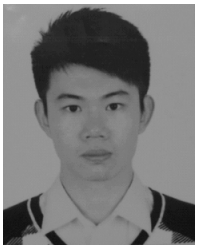
[71] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[72] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.

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

[74] S. Jegelka and J. Bilmes, "Submodularity beyond submodular energies: Coupling edges in graph cuts," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1897–1904.

Chung-Chi Tsai (S'16) received the B.S. degree from National Tsing-Hua University, Hsinchu, Taiwan, the M.S. degree from the University of California at Santa Barbara, Santa Barbara, CA, USA, and the Ph.D. degree from Texas A&M University, College Station, TX, USA, in 2009, 2012, and 2018, respectively, and all in electrical engineering. During his senior year in college, he attended a one-year exchange program at the University of New Mexico, Albuquerque, NM, USA. He has also participated in the Summer Internship with MediaTek Inc. in the summer of 2013, 2015, and 2016, respectively. He is currently a Senior Engineer with Qualcomm Technologies, Inc., San Diego, CA, USA. His primary research interests include computer vision and pattern recognition in object detection, tracking, and visual attention modeling.



Weizhi Li received the B.S. degree from the Department of Electronic Engineering, Shandong University, Weihai, China, and the M.S. degree from the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree with Arizona State University. His research interests include machine learning, computer vision, and speech processing.

Kuang-Jui Hsu received the B.S. degree from the Department of Electrical Engineering, National Sun Yat-sen University in 2011, and the M.S. degree from the Graduate Institute of Networking and Multimedia, National Taiwan University in 2013, where he is currently pursuing the Ph.D. degree with the Department of Computer Science and Information Engineering. He is also a Research Assistant with the Research Center for Information Technology Innovation, Academia Sinica. His research interests include computer vision, machine learning, deep learning, and image processing.

Xiaoning Qian (S'01–M'07–SM'17) received the Ph.D. degree in electrical engineering from Yale University, New Haven, CT, USA. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. He is also with the TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering (CBGSE) and the Center for Translational Environmental Health Research (CTEHR), Texas A&M University, College Station, TX, USA. His research interests include machine learning and Bayesian computation with their applications in computational network biology, genomic signal processing, and biomedical signal and image analysis. He was a recipient of the National Science Foundation CAREER Award, the Texas A&M Engineering Experiment Station Faculty Fellow, and the Montague-Center for Teaching Excellence Scholar at Texas A&M University. His recent work on computational network biology has received the Best Paper Award at the 11th Asia Pacific Bioinformatics Conference in 2013 and the Best Paper Award in the International Conference on Intelligent Biology and Medicine in 2016.

Yen-Yu Lin (M'12) received the B.B.A. degree in information management, and the M.S. and Ph.D. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2001, 2003, and 2010, respectively. He is currently an Associate Research Fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei. His current research interests include computer vision, machine learning, and artificial intelligence.