SEGMENTATION GUIDED LOCAL PROPOSAL FUSION FOR CO-SALIENCY DETECTION

Chung-Chi Tsai[1,2] *Xiaoning Qian*[1] *Yen-Yu Lin*[2]

[1] Texas A&M University [2]Academia Sinica

ABSTRACT

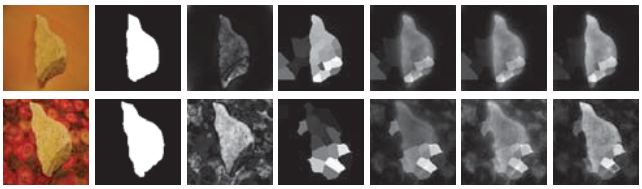
We address two issues hindering existing image co-saliency detection methods. First, it has been shown that object boundaries can help improve saliency detection; But segmentation may suffer from significant intra-object variations. Second, aggregating the strength of different saliency proposals via fusion helps saliency detection covering entire object areas; However, the optimal saliency proposal fusion often varies from region to region, and the fusion process may lead to blurred results. Object segmentation and region-wise proposal fusion are complementary to help address the two issues if we can develop a unified approach. Our proposed segmentation-guided locally adaptive proposal fusion is the first of such efforts for image co-saliency detection to the best of our knowledge. Specifically, it leverages both object-aware segmentation evidence and region-wise consensus among saliency proposals via solving a joint co-saliency and co-segmentation energy optimization problem over a graph. Our approach is evaluated on a benchmark dataset and compared to the state-of-the-art methods. Promising results demonstrate its effectiveness and superiority.

***Index Terms*—** Co-saliency, co-segmentation, adaptive fusion, energy minimization, alternating optimization

1. INTRODUCTION

Image co-saliency detection [1–11] aims to identify the common salient pixels in a set of images. It can help a broad range of image content analysis applications, such as co-segmentation [10–12] and co-localization [13]. Unlike single image saliency detection [14–19], co-saliency detection leverages not only *intra-image* appearance evidence but also *inter-image* correspondences to locate common salient regions. However, its performance is still restricted on extracting informative object cues in practical imaging scenarios due to illumination and viewing angle variation.

Many modern methods enhance co-saliency detection by fusing multiple, complementary saliency maps as candidate saliency proposals, each of which is based on using a particular detection algorithm. There exist fixed-weight [1–3] or adaptive-weight map fusion methods [8,9]; however, they still suffer from two major drawbacks. First, the fusing weights are derived from the whole image but the goodness of different saliency map proposals often varies from region to region [20]. Second, fusion by weighted combinations of saliency proposal typically leads to *blurred* results, especially near the object boundary regions.



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

Fig. 1: Image co-saliency detection. (a) Images. (b) Ground truth. (c) $\sim$ (g) Saliency maps produced by (c) [16] with intra-image evidence, (d) [1] with inter-image evidence, (e) [8] for map fusion, (f) ours w/o co-segmentation, and (g) ours.

Fig. 1 shows an image pair example, the ground truth for co-saliency detection, the saliency maps generated by using intra-image evidence [16] and inter-image evidence [1] in the first four columns, respectively. Neither intra-image evidence nor inter-image evidence can achieve satisfactory results individually. The former fails to detect the stone in the top image and has many false alarms in the bottom image, while the latter misses the stone in the bottom image. As shown in Fig. 1(e), the adaptive fusion method [8] combines intra- and inter-image evidence, and generates better results; but such a global whole-image fusion cannot make the most of the two *region-wise* complementary saliency proposals, failing to yield a homogeneously highlighted foreground or to reduce false alarms. Further spatial refinement by enforcing the co-saliency distribution compactness may help address the aforementioned drawback; but without the object information, it may remove less certain but real object regions.

We propose to improve these drawbacks by conducting segmentation for object information to guide co-saliency fusion, so that the fused maps can well preserve the object boundaries. Specifically, it leverages both object-aware segmentation evidence and region-wise consensus among saliency proposals via solving a joint co-saliency and co-segmentation energy optimization problem. Through alternating optimization, saliency maps of higher quality are generated. As shown in Fig. 1(f), the variant of our approach where segmentation is turned off recovers the missed fore-

This work was partially supported by Award #1547557 from the National Science Foundation, and Grants MOST 104-2628-E-001-001-MY2, MOST 105-2221-E-001-030-MY2 from the Ministry of Science and Technology.

978-1-5090-6067-2/17/$31.00

ground regions. Namely, the whole stone in the bottom image is more homogeneously highlighted. In Fig. 1(g), our approach with the aid of segmentation further suppresses the noise in Fig. 1(f) and produce sharper co-saliency maps.

2. RELATED WORK

Most saliency detection methods target on *human eye fixation prediction* [14,15] or *salient object prediction* [16–19]. Methods for the former are inspired by the primitive human visual system to predict human eye gaze patterns, such as the pioneering work by Itti *et al*. [14] with the center-surround differences across multi-scale image features to simulate the human eye visual system for saliency detection. Methods for the latter include the representative method by Achanta *et al*. [16] that defines pixel saliency based on the color differences from the average color of the whole image. Stemming from the unsupervised nature, the performance of these methods for single-image saliency detection is limited.

Co-saliency detection [1, 4, 5, 10] is introduced to utilize the extra information from inter-image evidence to help salient region localization and background removal. For example, Chang *et al*. [10] proposed a model based on the multiplication of intra-image saliency and inter-image repeatedness. Li and Ngan [1] utilized the *SimRank* algorithm on a co-multilayer superpixel tree to detect the inter-image similarity, and combined saliency maps produced by three existing algorithms [14–16]. Meng *et al*. [2] improved the *SimRank* matching method by further taking geometric constraints into account. Fu *et al*. [3] proposed a clustering-based co-saliency detection using the likelihood of pixels belonging to clusters.

To further improve the performance, high-level knowledge such as "objectness" obtained via segmentation is integrated into co-saliency detection. For example, Li *et al*. [4] chose multi-scale segmentation voting to locate the intra-image salient objects with enhanced local descriptors to determine the concurrence of salient objects across images. Liu *et al*. [5] computed region-wise co-saliency based on the local contrast and global similarity on the fine-scale segmentation together with the border connectivity based object priors in the coarse-scale segmentation. Jerripothula *et al*. [11] exploited saliency detection to enhance the performance of co-segmentation. We note that these methods derive segmentation and saliency detection in separated steps.

A research trend in saliency detection is to fuse a set of saliency proposals, each of which focuses on different aforementioned image properties. The fused saliency map is derived to share the most information with these proposals while excluding their individual biases. Cao *et al*. [8] employed a low-rank constraint to seek the weights for an adaptive combination of multiple saliency proposals. Huang *et al*. [9] constructed a multiscale superpixel tree. Fusion is accomplished by using low-rank analysis to take the saliency results of each scale into account. Methods [8,9] using proposal fusion often give better results. However, these methods adopt map-wise fusion to have global fusing weights for the whole images, and ignore the fact that the optimal saliency proposal is often region-dependent. Moreover, fusion-based methods often couple with post-processing to further refine the fusion results. However, post-processing may also lead to unfavorable effects. For instance, the spatial compactness post-processing [8] may consider parts of salient areas with lower saliency confidences as background. If the background priors are incorrectly established [9], they may misguide the refinement process to generate unfavorable fused co-saliency maps.

Our proposed method addresses these issues via performing a coupled co-saliency and co-segmentation optimization problem through an alternating optimization process. It adaptively seeks the weights for saliency proposal fusion in a region-wise manner. Meanwhile, the high-level priors generated from co-segmentation are iteratively refined and fed back to guide the fusion process. In this way, saliency maps of higher quality are detected owing to the object-aware evidence revealed by segmentation, while the performance of segmentation is progressively improved by using the figure-ground models derived from the better saliency maps. Thus, post-processing is not further required to obtain good results.

3. THE PROPOSED APPROACH

Given a pair of images I_1 and I_2 for co-saliency detection, we apply M existing saliency detection algorithms [1, 3, 14–19] and obtain M saliency maps with values normalized to $[0, 1]$ for each image. Images I_1 and I_2 are respectively decomposed into N_1 and N_2 *superpixels* as image regions, which preserve the intrinsic structures of the images while abstract unnecessary details. We aim to seek a plausible weight vector $\mathbf{y}_i = [y_{i,1}\ y_{i,2}\ \dots\ y_{i,M}]^\top \in \mathbb{R}^M$ for each superpixel i, and use it to accomplish co-saliency detection by region-wise fusing the M saliency maps. We formulates this task as a co-segmentation guided energy minimization problem over a graph. In the following, image pre-processing, graph construction, and the proposed energy function are described.

3.1. Image pre-processing

$N_1 = N_2 = 200$ superpixels are extracted by the *SLIC* algorithm with both *color* and *texture* bag-of-words representations. The color bag-of-words representations are based on clustering pixels in the three color spaces, `RGB`, `L*a*b*`, and `YCbCr` into 100 *visual words*, then each superpixel is represented as a histogram using the bag-of-words model. Similarly the texture bag-of-words representations are derived based on Gabor filter responses with eight orientations, three scales, and two phase offsets. A superpixel is similarly represented by a 100-dimensional histogram. Let $\mathbf{p}_i$ and $\mathbf{q}_i$ denote the color and texture histograms of superpixel i respectively. The similarity between two superpixels i and j is defined as

$$A(i,j) = \exp\left(-\frac{d(\mathbf{p}_i, \mathbf{p}_j)}{\sigma_c} - \gamma\frac{d(\mathbf{q}_i, \mathbf{q}_j)}{\sigma_g}\right), \qquad (1)$$

where $d(\cdot,\cdot)$ is the χ^2 distance. We set $\gamma = 1.5$ to put more emphasis on the texture features. Constant σ_c is set to the average pair-wise distance between all superpixels under the color features. Constant σ_g is similarly set.

3.2. Graph construction

We construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to encode the relationships among superpixels. Each vertex $v_i \in \mathcal{V}$ corresponds to superpixel i, thus $|\mathcal{V}| = N = N_1 + N_2$. A *2-ring graph* is employed to enhance connectivity. Namely, edge $e_{ij} \in \mathcal{E}$ is added for linking v_i and v_j if superpixels i and j are spatially connected or they are both connected to the same superpixel. The edge set $\mathcal{E}$ is associated with the weight matrix $A \in \mathbb{R}^{N\times N}$ in Eq. (1). The *graph Laplacian* $L \in \mathbb{R}^{N\times N}$ is then obtained.

3.3. Energy 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 map fusion by minimizing the following co-segmentation energy function:

$$\begin{aligned} J(Y,Z) = {} & \alpha_1 \sum_{v_i\in\mathcal{V}} U_1(\mathbf{y}_i) + \alpha_2 \sum_{v_i\in\mathcal{V}} U_2(z_i) + \alpha_3 \sum_{v_i\in\mathcal{V}} U_3(\mathbf{y}_i, z_i) \\ & + \beta_1 \sum_{e_{ij}\in\mathcal{E}} B_1(\mathbf{y}_i,\mathbf{y}_j) + \beta_2 \sum_{e_{ij}\in\mathcal{E}} B_2(z_i,z_j) + |Y\|_2^2 \qquad (2) \\ \text{s.t.} \quad & \|\mathbf{y}_i\|_1 = 1, \mathbf{y}_i \geq \bar{\mathbf{0}}, z_i \in \{0,1\}, \text{ for } 1 \leq i \leq N, \end{aligned}$$

where $\bar{\mathbf{0}}$ is a zero vector, and α_1, α_2, α_3, β_1 and β_2 are five positive constants. $Z = [z_1\ z_2\ \ldots\ z_N] \in \mathbb{R}^N$ denotes the figure-ground configuration of co-segmentation. 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 nice properties from co-segmentation, e.g. object-aware contours and sharp foreground, can be transferred to facilitate co-saliency detection. In (2), $U_1(\mathbf{y}_i)$ and $B_1(\mathbf{y}_i,\mathbf{y}_j)$ are the unary and pairwise terms for co-saliency detection, respectively. $U_2(z_i)$ and $B_2(z_i,z_j)$ are the unary and pairwise terms for co-segmentation, respectively. The coupling term $U_3(\mathbf{y}_i, z_i)$ is included to encourage the coherence between the co-saliency map and the figure-ground segmentation. Lastly, the term $||Y\|_2^2$ is introduced for regularization. These terms are detailed in the following sections.

3.3.1. Unary term $U_1(\mathbf{y}_i)$

This unary term for saliency detection contains two parts, which respectively leverage the intra- and inter-image cues to infer the goodness of each saliency map on superpixel i.

For the intra-image cue, we intend to assign a higher weight to a saliency map that is consistent with other saliency maps. Inspired by [21], we employ a low-rank constraint to realize this task, but we further generalize it to *locally* estimate the goodness of saliency maps. For superpixel i, we find its n $(= 50)$ spatially nearest superpixels. Let $\mathbf{x}_{i,m} \in \mathbb{R}^{256}$ be a histogram denoting the 256-bin distribution of saliency values of saliency map m on these n superpixels. By stacking the M different vectors for all 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 consistent part by seeking a low-rank surrogate of X_i. Specifically, *robust PCA* is adopted to decompose X_i into a low-rank approximation L_i plus a residual matrix E_i by solving

$$\min_{L_i,E_i} (||L_i||_* + \lambda||E_i||_1), \quad \text{s.t. } X_i = L_i + E_i, \qquad (3)$$

where $||L_i||_*$ is the nuclear norm of L_i, and λ is a constant. After solving (3), we compute the normalized residuals by referring to errors $E_i = [\mathbf{e}_{i,1}\ \ldots\ \mathbf{e}_{i,M}]$ via

$$b_{i,m} = \frac{\exp(-||\mathbf{e}_{i,m}||_2^2)}{\sum_{j=1}^{M}\exp(-||\mathbf{e}_{i,j}||_2^2)}, \text{ for } 1 \leq m \leq M. \qquad (4)$$

For energy minimization, the associated penalty variable is then defined as $l_{i,m} = \exp(1 - b_{i,m}) / \sum_{j=1}^{M}\exp(1 - b_{i,j})$.

For the inter-image cue, we reduce the false alarms in saliency detection by exploring inter-image correspondences. Let $e_i \in [0,1]$ represent the similarity between superpixel i and its most similar superpixel in the other image. The similarity of all v_i in the image pair are initially measured via (1). We concatenate the e_i in the same image into a vector and normalize it, such that $e_i = 1$ represents the highest likelihood of superpixel i having a correspondence in another image. Let $s_{i,m}$ denote the mean saliency value of saliency map m on superpixel i. We prefer saliency map m if the value of $s_{i,m}$ is proportionate to that of e_i. By designing a variable $g_{i,m}$ penalizing the case where just one of e_i and $s_{i,m}$ is large, we get

$$g_{i,m} = \frac{\exp((1-e_i)s_{i,m} + e_i(1-s_{i,m}))}{\sum_{j=1}^{M}\exp[(1-e_i)s_{i,j} + e_i(1-s_{i,j})]}. \qquad (5)$$

The denominator in (5) is used for normalization.

The intra- and inter-image cues on superpixel i and map m, i.e. $l_{i,m}$ and $g_{i,m}$, are combined via

$$w_{i,m} = \frac{\exp(l_{i,m} + g_{i,m})}{\sum_{j=1}^{M}\exp(l_{i,j} + g_{i,j})}. \qquad (6)$$

Considering all superpixels, the unary term becomes

$$\sum_{v_i\in\mathcal{V}} U_1(\mathbf{y}_i) = \sum_{i=1}^{N} \mathbf{w_i}^\top \mathbf{y_i} = \mathbf{tr}(\mathbf{W}^\top \mathbf{Y}), \qquad (7)$$

where $\mathbf{w}_i = [w_{i,1}\ \ldots\ w_{i,M}]^\top$ and $W = [\mathbf{w_1}\ \ldots\ \mathbf{w_N}]$.

3.3.2. Unary term $U_2(z_i)$

This term estimates the likelihood of superpixel i belonging to the common foreground in co-segmentation. Following [12], we represent each superpixel i by its mean RGB color, i.e. $\mathbf{c}_i \in \mathbb{R}^3$. During the iterative optimization that will be introduced later, a *Gaussian mixture model* (GMM) with five components and the corresponding model parameters $\boldsymbol{\theta}_f$, is fit to

the superpixels that are currently labeled as foreground (F). Meanwhile, another five-component GMM $\boldsymbol{\theta}_{b,k}$ is fit to the background (B) superpixels of $I_k, k \in \{1, 2\}$. Specifically,

$$\sum_{v_i \in \mathcal{V}} U_2(z_i) = \sum_{i=1}^{N} [p(v_i \in F|\mathbf{c}_i)(1 - z_i) + p(v_i \in B|\mathbf{c}_i)z_i]. \quad (8)$$

GMM $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_{b,k}$ help predict the probability of superpixel i belonging to the foreground or background. Assuming $p(v_i \in F) = p(v_i \in B) = \frac{1}{2}$, we can get $p(v_i \in F|\mathbf{c}_i) = \frac{p(\mathbf{c}_i \in F|\boldsymbol{\theta}_f)p(v_i \in F)}{p(\mathbf{c}_i|\boldsymbol{\theta}_f)p(v_i \in F) + \sum_{k=1}^{2} p(\mathbf{c}_i|\boldsymbol{\theta}_{b,k})\delta(v_i \in I_k)p(v_i \in B)}$, where $p(\cdot|\boldsymbol{\theta}_f)$ and $p(\cdot|\boldsymbol{\theta}_{b,k})$ are the Gaussian probability distributions. And, $p(v_i \in B|\mathbf{c}_i)$ is similarly set.

3.3.3. Coupling term $U_3(\mathbf{y}_i, z_i)$

$U_3(\mathbf{y}_i, z_i)$ encourages the coherence between the co-saliency maps and the co-segmentation result. For measuring the degree of coherence on superpixel i, we compute the mean saliency value of the fused map on this superpixel by

$$s_i = \sum_{m=1}^{M} y_{i,m} s_{i,m} = \mathbf{y}_i^\top \mathbf{s}_i, \quad (9)$$

where $\mathbf{y}_i = [y_{i,1} \ \dots \ y_{i,M}]^\top$ is the weight vector for saliency map fusion on superpixel i, and $s_{i,m}$ is again the mean saliency value of map m on superpixel i. Note that both the values of $\mathbf{y}_i$ and $\{s_{i,m}\}_{m=1}^{M}$ are in $[0, 1]$, thus $s_i \in [0, 1]$. 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 as

$$\sum_{v_i \in \mathcal{V}} U_3(\mathbf{y}_i, z_i) = \sum_{i=1}^{N} s_i(1 - z_i) + (1 - s_i)z_i. \quad (10)$$

3.3.4. Binary term $B_1(\mathbf{y}_i, \mathbf{y}_j)$

This term encourages smooth weights Y between the connected superpixels in graph $\mathcal{G}$. Its formulation is given below

$$\sum_{e_{ij} \in \mathcal{E}} B(\mathbf{y}_i, \mathbf{y}_j) = \sum_{e_{ij} \in \mathcal{E}} A(i, j)\|\mathbf{y}_i - \mathbf{y}_j\|_2^2 = tr(YLY^\top), \quad (11)$$

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

3.3.5. Binary term $B_2(z_i, z_j)$

This binary term is imposed to enforce the spatial smoothness of co-segmentation results. It is defined as

$$\sum_{e_{ij} \in \mathcal{E}} B_2(z_i, z_j) = \sum_{e_{ij} \in \mathcal{E}} A(i, j)\|z_i - z_j\|_2^2 = tr(ZLZ^\top). \quad (12)$$

4. OPTIMIZATION PROCESS

An iterative strategy is adopted to optimize (2). At each iteration, one set of variables Y and Z is optimized while keeping the other fixed, and then their roles are switched. The alternating optimization procedure is iterated until convergence of the energy values.

4.1. On optimizing Y

By fixing Z, the optimization problem in (2) becomes

$$\begin{aligned} J(Y) = \alpha_1 \sum_{v_i \in \mathcal{V}} U_1(\mathbf{y}_i) + \beta_1 \sum_{e_{ij} \in \mathcal{E}} B_1(\mathbf{y}_i, \mathbf{y}_j) \\ + \alpha_3 \sum_{v_i \in \mathcal{V}} U_3(\mathbf{y}_i, z_i) + ||Y||_2^2 \\ \text{s.t.} \quad \|\mathbf{y}_i\|_1 = 1, \mathbf{y}_i \geq \bar{\mathbf{0}}, \text{ for } 1 \leq i \leq N. \end{aligned} \quad (13)$$

The above constrained optimization problem is a *quadratic programming* problem. We solve it by using the CVX [22].

4.2. On optimizing Z

By fixing Y, the optimization task in (2) becomes

$$\begin{aligned} J(Z) = \alpha_2 \sum_{v_i \in \mathcal{V}} U_2(z_i) + \beta_2 \sum_{e_{ij} \in \mathcal{E}} B_2(z_i, z_j) \\ + \alpha_3 \sum_{v_i \in \mathcal{V}} U_3(\mathbf{y}_i, z_i) \\ \text{s.t.} \quad z_i \in \{0, 1\}, \text{ for } 1 \leq i \leq N. \end{aligned} \quad (14)$$

The energy function in (14) is graph representable and regular. Thus it can be efficiently minimized via graph cuts.

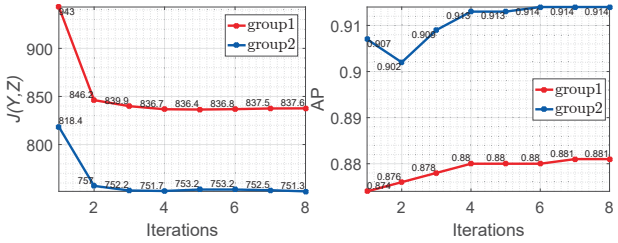
4.3. Implementation details

For initialization, we solve the weights Y for saliency map fusion via (13) with the coupling term U_3 removed. Then, the fused co-saliency maps are binarized into foregrounds and backgrounds to initialize GMMs $\boldsymbol{\theta}_f$, $\boldsymbol{\theta}_{b,1}$ and $\boldsymbol{\theta}_{b,2}$ in (8) and enable the optimization of (14) at the first iteration. Following [16], an adaptive image-dependent threshold for binarization is set to $2m$, where m the mean saliency value of the fused map. In the alternating optimization process, the value of the objective function decreases and converges to a local optimum when solving (13) and (14) iteratively.

5. EXPERIMENTAL RESULTS

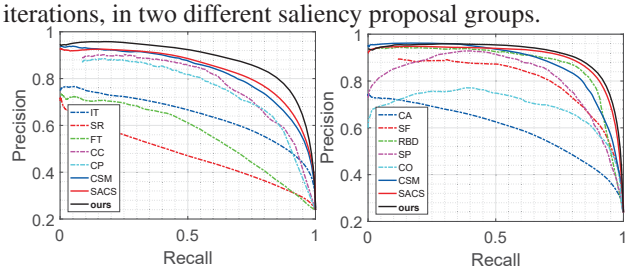
5.1. Experimental setup

We evaluate our approach, and compare it with the state-of-the-art methods on the *Image Pair dataset* [1], which is composed of 105 image pairs with manually labeled ground truth. We choose two groups of saliency map proposals to have comprehensive studies of co-saliency detection. For the first group, we follow [1] and get five saliency proposals consisting of three single-image saliency maps (SISM) obtained by methods `IT` [14], `SR` [15], and `FT` [16] and two multi-image saliency maps (MISM) by using the algorithm in [1] with two different features, color `CC` and texture `CP`. The second group contains three SISMs by using methods `CA` [17], `SF` [18], and `RBD` [19], and two MISMs obtained by using the detection algorithm in [3] with two different features, spatial cues `SP` and correspondence cues `CO`. Our approach is also compared with



(a) (b)

Fig. 2: (a) The energy curves of (2) (b) The AP curves, versus iterations, in two different saliency proposal groups.



(a) (b)

Fig. 3: The PR curves of the evaluated approaches with the saliency proposals in (a) group 1 and (b) group 2.

two fusion-based methods for co-saliency detections, including the fixed-weighted summation method CSM [1] and the self-adaptive fusion method SACS [8]. While our approach (ours) and SACS adaptively determine the weights for proposal fusion, the weight in CSM is set to 0.0167 for each SISM and 0.4 for each MISM.

The performance is measured by *precision-recall* (PR) curves, which are obtained by varying the saliency thresholds. We also distill the overall performance of the PR and *receiver operating characteristics* (ROC) curves into the areas under the curves. They are denoted by AP and AUC respectively. In all the experiments, we set $\alpha_1 = 5$, $\alpha_2 = 2$, $\alpha_3 = 5$, $\beta_1 = 1$, and $\beta_2 = 0.1$ in (2) and $\lambda = 0.05$ in (3). The objective function values in (2) and the performance AP by our approach through alternating optimization are shown in Fig. 2(a) and (b), respectively. Both of them converge rapidly. We report the results of our approach at iteration 7.

5.2. Result Analysis

The PR curves of the evaluated approaches with saliency proposal groups 1 and 2 are drawn in Fig. 3(a) and (b) respectively. The performances in AP and AUC are also reported in Tables 1 and 2. With saliency proposal group 1, it can be observed in Table 1 that the proposal CC gives the best performance among the five proposals. The fusion-based methods CSM and SACS can exploit the five proposals to remarkably improve the performance. Our approach integrates co-segmentation into co-saliency detection so that high-level object-aware information can guide the region-wise proposal fusion. As shown in Fig. 3(a), it consistently outperforms all the competing methods. Its performance gain over method SACS, the best competing approach, is significant, i.e. 4.5% in AP and 1.3% in AUC. Similar observations can be found in

method	IT [14]	SR [15]	FT [16]	CC [1]	CP [1]	CSM [1]	SACS [8]	**ours**
AP	0.640	0.471	0.559	0.702	0.681	0.824	0.836	**0.881**
AUC	0.872	0.718	0.756	0.881	0.865	0.930	0.944	**0.958**

Table 1: Performance in AP (average precision) and AUC (area under the ROC curve) on saliency proposal group 1.

method	CA [17]	SF [18]	RBD [19]	SP [3]	CO [3]	CSM [1]	SACS [8]	**ours**
AP	0.595	0.701	0.847	0.813	0.692	0.879	0.900	**0.914**
AUC	0.843	0.922	0.936	0.915	0.886	0.948	0.970	**0.974**

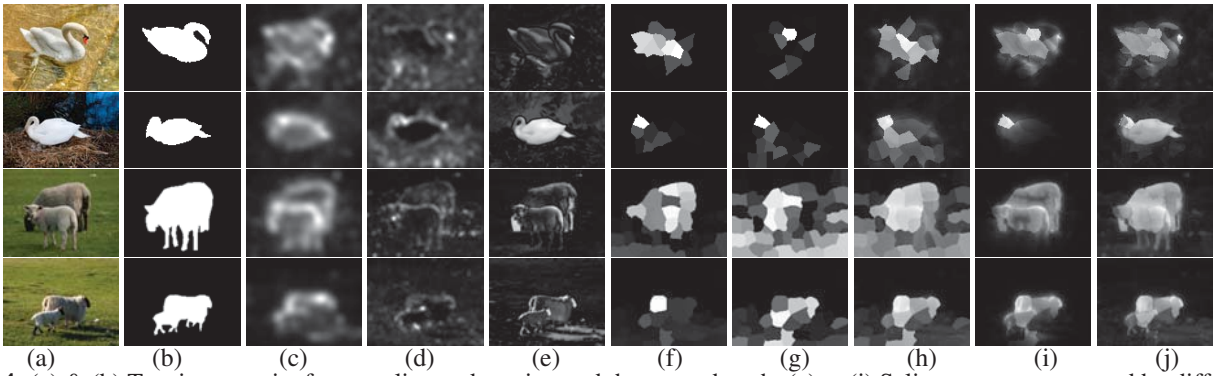
Table 2: Performance in AP (average precision) and AUC (area under the ROC curve) on saliency proposal group 2.

Fig. 3 and Table 2 for the approaches with saliency proposal group 2, though the performance gain of the fusion-based approaches, including CSM, SACS, and ours, becomes less significant. The main reason is that the proposal RBD individually gives satisfactory results. Thus, the proposals in group 2 are not as complementary as those in group 1. Nevertheless, our approach still achieves more favorable performance than all the competing approaches thanks to the adaptive region-wise fusion.

To gain insight into the quantitative results, Fig. 4 displays the detected saliency maps on two image pairs, when saliency proposal group 1 is adopted. The saliency proposals, i.e. those in Fig. 4(c) $\sim$ 4(g), do not perform well individually. They contain many false alarms and misses. Methods CSM and SACS indeed get better results via proposal fusion. Our approach with the aid of co-segmentation carries out region-wise fusion, and can generate the saliency maps perceptually closest to the ground truth. Fig. 5 shows another two examples when saliency proposal group 2 is used. We observed that fusion-based methods CSM and SACS can only give comparable or even worse maps than the saliency proposal RBD, since the proposals in group 2 are less complementary. Our approach fuses these proposals in a region-wise fashion, so it does not suffer from this problem. More importantly, our approach gives sharper and more homogeneously highlighted result without any additional post-processing.

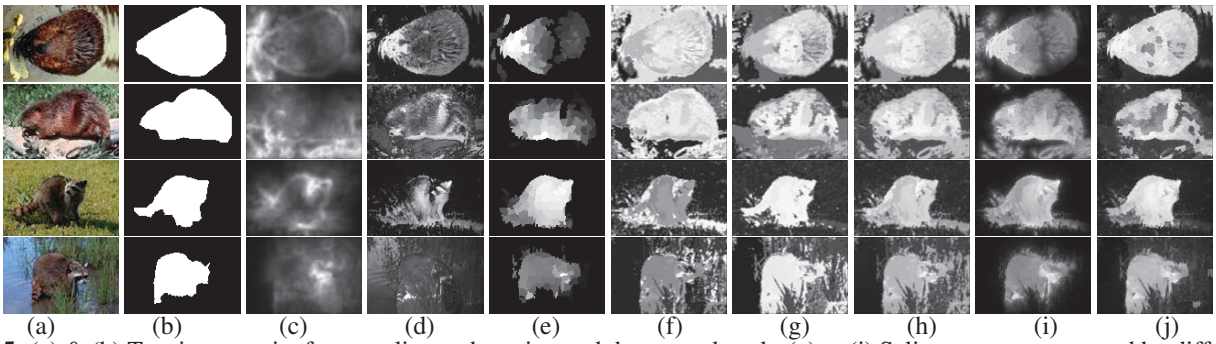
6. CONCLUSIONS

In this paper, we have presented an unsupervised learning framework that carries out saliency proposal fusion via jointly exploring the common object evidence generated from co-segmentation and the consensus among various saliency proposals. The benefits of its joint optimization formulation are evident as it produces the fused maps of high quality via making the most of multiple locally complementary saliency proposals. Moreover, unlike existing models relying on additional post-processing to smooth the fused maps, our framework has already merged the advantages of such post-processing into our unified optimization process, and generates even better results. In future, we plan to apply our algorithm to vision applications where saliency maps of high quality are appreciated, such as object recognition, image feature extraction, and scene understanding.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 4: (a) & (b) Two image pairs for co-saliency detection and the ground truth. (c) $\sim$ (j) Saliency maps generated by different approaches including (c) `IT` [14], (d) `SR` [15], (e) `FT` [16], (f) `CC` [1], (g) `CP` [1], (h) `CSM` [1], (i) `SACS` [8], and (j) `ours`.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 5: (a) & (b) Two image pairs for co-saliency detection and the ground truth. (c) $\sim$ (j) Saliency maps generated by different approaches including (c) `CA` [17], (d) `SF` [18], (e) `RBD` [19], (f) `CO` [3], (g) `SP` [3], (h) `CSM` [1], (i) `SACS` [8], and (j) `ours`.

7. REFERENCES

[1] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *TIP*, 2011.

[2] F. Meng, H. Li, and G. Liu, "A new co-saliency model via pairwise constraint graph matching," in *ISPACS*, 2012.

[3] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *TIP*, 2013.

[4] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *TMM*, 2013.

[5] Z. Liu, W. Zou, L. Li, L. Shen, and O. Le Meur, "Co-saliency detection based on hierarchical segmentation," *SPL*, 2014.

[6] C.-R. Huang, Y.-J. Chang, Z.-X. Yang, and Y.-Y. Lin, "Video saliency map detection by dominant camera motion removal," *TCSVT*, 2014.

[7] L. Li, Z. Liu, W. Zou, X. Zhang, and O. Le Meur, "Co-saliency detection based on region-level fusion and pixel-level refinement," in *ICME*, 2014.

[8] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *TIP*, 2014.

[9] R. Huang, W. Feng, and J. Sun, "Saliency and co-saliency detection by low-rank multiscale fusion," in *ICME*, 2015.

[10] 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 *CVPR*, 2011.

[11] K. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *TMM*, 2016.

[12] H. Yu, M. Xian, and X. Qi, "Unsupervised co-segmentation based on a new global gmm constraint in mrf," in *ICIP*, 2014.

[13] K. R. Jerripothula, J. Cai, and J. Yuan, "Cats: Co-saliency activated tracklet selection for video co-localization," in *ECCV*, 2016.

[14] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *TPAMI*, 1998.

[15] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *CVPR*, 2007.

[16] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *CVPR*, 2009.

[17] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *TPAMI*, 2012.

[18] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *CVPR*, 2012.

[19] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *CVPR*, 2014.

[20] C.-C. Tsai, X. Qian, and Y.-Y. Lin, "Image co-saliency detection via locally adaptive saliency map fusion," in *ICASSP*, 2017.

[21] J. Li, J. Ding, and J. Yang, "Visual salience learning via low rank matrix recovery," in *ACCV*, 2014.

[22] M. Grant and S. Boyd, "CVX users guide for CVX version 1.22," 2012.