IMAGE CO-SALIENCY DETECTION: NOVEL APPROACHES WITH CONVEX OPTIMIZATION AND DEEP NEURAL NETWORKS

A Dissertation

by

CHUNG-CHI TSAI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Xiaoning Qian
Committee Members,	Raffaella Righetti
	Tie Liu
	Jianhua Huang
Head of Department,	Miroslav M. Begovic

August 2018

Major Subject: Electrical Engineering

ABSTRACT

The advance of digital technologies has endowed people with easier access to massive collections of image or video data than ever before. For its capability to borrow signal strengths across images or video frames, image co-saliency detection has become an active research topic to help address many advanced computer vision applications, such as image retrieval and object tracking. Co-saliency detection is to distill the essential content of an image group by locating the eye-catching objects commonly present in multiple images. In this dissertation, we present several novel approaches based on convex optimization and deep neural networks for accurate image co-saliency detection. First, we introduce a region-wise saliency map fusion approach to amend the inherent flaw of traditional map-wise fusion methods. The effectiveness of region-wise fusion motivates us to the improved model by integrating segmentation revealed objectness in the following work. Second, we explore the complementary relationship between image co-saliency detection and co-segmentation for higher quality performance on both tasks with scalability to multiple input images. Third, we improved our region-wise fusion scheme by exploring the power of unsupervised deep learning methods by stacked auto-encoder to relax the underlying foreground consistency assumption in most state-of-the-art fusion models. To achieve the desired practical significance, we further combine our stacked autoencoder-enabled fusion with the convolutional neural networks (CNNs). Our proposed two-stage co-saliency detection model can retain the highly discriminative power from the CNNs without the requirement of massive human labeling. Comprehensive experimental results demonstrate its state-of-the-art performance on the publicly available benchmark data sets. We expect solving these issues in image co-saliency detection can lead to significant contributions to the computer vision community for better image understanding and consequent decision making based on that.

DEDICATION

To my wife, mother, father, grandmother, and grandfather.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Prof. Xiaoning Qian and co-advisor Prof. Yen-Yu Lin from Academia Sinica, for their guidance and support on my research over the years. Their mentoring is like the light at the end of the tunnel for my Ph.D. studies by pointing out these critical research topics. During these days, the professional communication with Prof. Lin leads me to the depth of computer vision area, more importantly, Prof. Qian guidance on my technical writing and logical thinking of any machine learning details from his broad knowledge dramatically complement me with the breadth of research to any related fields. My sincere appreciation also extends to my committee members: Prof. Raffaella Righetti, Prof. Tie Liu and Prof. Jianhua Huang for my attended course from them that helps to lay my knowledge foundation and their advice contributing to my thesis. Last but not least, I would like to thank my collaborator, Mr. Kuang-Jui Hsu, and my colleagues at Academia Sinica and Texas A&M University on the technical discussion as well as my family for their listening and encouragement. Without the support from anyone mentioned above, it would be impossible for me to accomplish the Ph.D. degree.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of my Ph.D. advisor, Prof. Xiaoning Qian and Prof. Raffaella Righetti, Prof. Tie Liu of the Department of Electrical and Computer Engineering and Prof. Jianhua Huang of the Department of Statistics.

Mr. Weizhi Li, from the Department of Electrical and Computer Engineering at Texas A&M University, helped conduct the statistical evaluation on the state-of-the-art image co-segmentation models for Section 4. The project contributing to Section 5 was co-worked with Mr. Kuang-Jui Hsu from the Research Center for Information Technology at Academia Sinica, who designed the self-trained convolutional neural networks (`STCNN`) to enhance the model performance in the pioneering joint fusion-learning based co-saliency detection model.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

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.

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Background

Salient object detection is one of the fundamental research topics in computer vision that stems from the idea of our visual perception system to extract the important objects automatically from an image scene. In 1998, Itti *et al*. [4] established the foundation of saliency detection to use center-surround difference on multi-scale feature maps to predict our eye fixation points on an image. After then, people extended this idea to a "soft" segmentation problem for eye-catching objects since many advanced computer vision applications benefit from the highlighted information about the region of interests. Compared to saliency detection in single images, a group of images may contain more useful information since the repeatedly occurring foregrounds can be utilized to represent the principal message of the image group. This dissertation focuses on co-saliency detection to detect regions of interest across multiple images, which can enhance the performance on single image saliency detection.

The term "Co-saliency" was introduced in 2010 by Jacob *et al*. [32] to help explore the local structural changes in human pose or object orientation between two images for the photographic triage task. The pioneering effort to address image co-saliency detection problems was contributed by Li *et al*. [2] who decided to combine the intra-image saliency maps with inter-image correspondence evidence to get the regions of co-salient objects in paired images. Later, Fu *et al*. [9] extended the above idea to search for visually salient objects repetitively appearing across multiple given images. With its scalability and computational efficiency, image co-saliency detection has become another emerging research topic. Until today, there have been tremendous efforts trying to leverage the intra-image saliency evidence as well as the inter-image reoccurrence to better highlight common regions of interest. Despite the difference in these approaches, strategies for co-saliency detection can be naturally categorized into the bottom-up, fusion-based, and learning-based approaches. For more detailed definitions, please refer to [33].

At the early developing stage of co-saliency detection, the bottom-up methods that rely on predefined saliency measure based on the feature contrast, the spatial location, and the correspondence were the most popular way of co-saliency detection approaches. Despite their efficiency, they are typically too subjective since there are no designed saliency metrics suitable for all various scenarios encountered in practice, mainly due to the lack of thorough understanding of the human visual perception system, thus leading to suboptimal performance. Even though learning-based methods can usually achieve more promising result than the bottom-up approaches, their supervised setting contradicts many applications requirement. Rather than discovering more sophisticated informative cues for representing co-salient objects, the fusion-based strategy, which has become one of a trend in image co-saliency detection, has been proposed to overcome the limitation of bottom-up approach by borrowing the strengths of different saliency models while easing their potential bias due to their corresponding model assumptions [3].

1.2 Problem Statement

The primary fusion formula for co-saliency detection can be abstracted by (1.1) below and the task is to adaptively search for the optimal linear combination w_m of each candidate saliency proposal S_m based on individual map quality [3]. This dissertation centers around the topic of "adaptive fusion-based framework" by providing the solutions to several identified critical issues for more accurate and robust co-saliency detection with multiple images.

$$\text{Co-saliency} = \sum_m w_m \times S_m. \tag{1.1}$$

Specifically, we first extend the traditional global fusion framework to more flexible local saliency proposal fusion by allowing different fusion weights at the superpixel scale so that the fusion can capture important saliency information that may have region-wise variation. We develop graph optimization methods to solve this new adaptive region-wise fusion problem. Based on this region-wise fusion framework, we enhance the co-saliency detection performance by taking advantage of different region-wise feature representations. For example, we explore the segmen-

tation revealed objectness evidence and integrate co-segmentation and co-saliency detection tasks for multiple input images into a unified model with the scalable optimization algorithm to attain better performance for both tasks, due to their bidirectional complementary relationships. Motivated by the recent successes of deep neural networks in computer vision, we further replace the components of human-designed features for region-wise fusion by deep data-driven features for more accurate co-saliency detection.

1.3 Overview

The remainder of the dissertation is organized as follows:

Beginning with Section 2, we first address one crucial but universal issue in the traditional map-wise (global) fusion-based approaches mentioned above. We observe the performance of co-saliency detection substantially relies on the explored visual cues. However, the optimal cues typically vary from region to region. To address this issue, we amend the fusion formula to Equation (1.2) that detects co-salient objects by region-wise saliency map fusion. Specifically, by constructing a superpixel graph and considering individual superpixel (denoted as v_i) as a basic fusion unit, we aim to compute the optimal local fusion weights $w_m(v_i)$ of i-th superpixel for the m-th saliency proposal on superpixel region v_i (denoted as $S_m(v_i)$). Our approach jointly takes intra-image coherence, inter-image correspondence, and spatial consistency into account, and accomplishes saliency detection via solving an energy optimization problem over a graph. We evaluate this proposed method on the Image-Pair [2] benchmark data set and compare to the state-of-the-art techniques. Promising results demonstrate its effectiveness and superiority.

$$\text{Co-saliency}(v_i) = \sum_{m=1}^{M} w_m(v_i) \times S_m(v_i). \tag{1.2}$$

Motivated by the success of our proposed region-wise fusion, we address two additional issues hindering existing image co-saliency detection methods in Section 3. 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. This approach is also evaluated on the Image-Pair [2] benchmark data set and compared to the state-of-the-art methods. Promising results demonstrates that it can perform more favorably than the previous methods without any post-processing step.

Sections 2 and 3 show the improvement brought by the local proposal fusion, and the advantages by further integrating co-segmentation derived objectness into the model. However, those models are only analyzed in the paired image cases. To make it scalable with the number of input images, and also to achieve the state-of-the-art performance, we extend our previous segmentation-guided region-wise fusion framework in Section 3 to be cooperative with multiple input images as well as enhancing the original design of the co-segmentation unary term. In Section 4, 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. Our developed method addresses co-saliency detection and co-segmentation jointly via solving an energy minimization problem over a graph. Specifically, our method iteratively carries out the region-wise adaptive saliency map fusion and object segmentation to transfer useful information between the two complementary tasks elegantly. Through the optimization iterations, sharp saliency maps are gradually obtained to recover entire salient objects by referring to object segmentation, while these segmentation are progressively improved owing to the better saliency prior. We comprehensively evaluate this method on two public benchmark data sets [2, 34] while comparing it to the state-of-the-art techniques. Extensive experiments demonstrate that our approach can provide consistently higher-quality results on both co-saliency detection and co-segmentation.

Besides the analysis of our proposed "segmentation guided local proposal fusion" and "saliency guided co-segmentation" based on the conventional convex optimization presented in Section 4, there are still a few issues worthwhile to be solved to achieve the desired practical significance. For fusion-based methods, we notice that final results often highly depend on the quality and coherence of individual proposals since they combine saliency proposals using a majority voting rule. To alleviate such a limitation whenever adopted saliency proposals become diverse or unreliable, we aim to design a new unsupervised scheme to be able to achieve more favorable results by the demonstrated scene understanding power of deep learning models with the hope of leveraging learning-based methods in our fusion framework to improve the final fusion results. However, learning-based methods typically require ground-truth annotations for training, which are not available for co-saliency detection. Indeed, we invent a fusion-learning integrated two-stage approach and present it in Section 5. At the first stage, an unsupervised deep learning model based on stacked autoencoder (SAE) is proposed to evaluate the quality of saliency proposals. It employs latent representations for image foregrounds, and auto-encodes foreground consistency and foreground-background distinctiveness in a discriminative way. The resultant model, SAE-enabled fusion (`SAEF`), can combine multiple saliency proposals to yield a more reliable saliency map. At the second stage, motivated by the fact that fusion often leads to over-smoothed saliency maps, we develop self-trained convolutional neural networks (`STCNN`) to overcome this negative effect. `STCNN` takes the saliency maps produced by `SAEF` as inputs. It propagates information from regions of high confidence to those of low confidence. During propagation, feature representations are distilled, resulting in sharper and better co-saliency maps. We evaluate this approach comprehensively on three benchmarks, including MSRC [23], iCoseg [34], and the recently proposed Cosal2015 [24]. Extensive experiments show this model, which inherits the advantages of all our previously proposed models as well as the discriminative power by CNNs, performs the most favorably against the state-of-the-arts, even surpassing the supervised deep learning methods.

2. IMAGE CO-SALIENCY DETECTION VIA LOCALLY ADAPTIVE SALIENCY MAP FUSION (CSLA)*

2.1 Introduction

Saliency detection attempts to unsupervisedly identify the salient pixels in an image. It is an active and fundamental topic in image processing, since it can help automate many applications such as image segmentation [35] and video compression [36]. Despite the significant progress, e.g. [1, 4–6, 8, 11, 37], the performance of single-image saliency detection is still restricted by its unsupervised nature, especially when with complex image content. Co-saliency detection, e.g. [2, 9, 38], is introduced to address the difficulties inherent in single-image saliency detection. It aims to locate the common salient objects. The information used in most approaches for co-saliency detection can be divided into two categories, i.e. *intra-image* and *inter-image* evidences. The former is extracted based on appearance contrast and spatial cues in a single image. The latter is obtained by detecting the correspondences between a group of images.

A single type of evidences in general is insufficient for handling complex co-saliency detection problems. Most modern approaches carry out co-saliency detection by fusing multiple saliency maps. For instance, the approaches in [2, 38] adopt *fixed-weight summation* for map fusion, while the one in [9] uses *fixed-weight multiplication*. Cao *et al.* [3] instead proposed a *self-adaptive* framework where the weights for map fusion are dynamically generated according to the input images.

The aforementioned approaches [2,3,9,38] fuse saliency maps in a *map-wise* manner. Namely, a weight is given for the whole-image saliency map. These approaches neglect the phenomenon that the goodness of a saliency map is often *region-dependent*. As an illustration, Figure 2.1 shows an image pair and the saliency maps generated by using the intra-image evidence [1], the inter-image evidence [2], the method of self-adaptive fusion [3], and our proposed method. It can be

*Reprinted with permission from "Image Co-saliency Detection Via Locally Adaptive Saliency Map Fusion" by Chung-Chi Tsai, Xiaoning Qian, and Yen-Yu Lin, 2017. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Pages=1897–1901, Copyright [2017] by IEEE.

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

Figure 2.1: Image co-saliency detection. (a) & (b) An image pair and the ground truth. (c) $\sim$ (f) Saliency maps produced by using (c) the intra-image evidence [1], (d) the inter-image evidence [2], (e) the method in [3], and (f) ours (`CSLA`).

observed that using a single type of evidences doesn't suffice for this case. While using only the intra-image evidence [1] leads to the false alarm in the text part of the second image, using only the inter-image evidence [2] fails to detect a penguin in the first image. The method [3] combines both types of evidences. It gives better results, but it also inherits both the shortcomings of false alarms and misses.

To tackle these challenges of co-saliency detection, we propose an approach that can jointly consider both intra-image and inter-image evidences, and carry out region-wise saliency map fusion. As shown in Figure 2.1(f), our approach effectively alleviates the unfavorable effects of false alarms and misses, and results in the saliency maps of higher quality.

2.2 Related Work

The literature of saliency detection is extensive. Most of them target at *human eye fixation prediction* [4, 5] or *salient object detection* [1, 6, 8, 11, 37]. Approaches to eye fixation prediction are inspired by the primitive human visual system. For example, Itti *et al.* [4] computed center-surround differences across multi-scale image features for detecting saliency. Despite the novelty, this method poorly detects object borders. On the contrary, Hou *et al.* [5] defined the saliency through the residual on the log-frequency domain. Although their method is computationally efficient, it mostly discovers object boundaries rather than the whole salient regions. Both methods [4,5] involve image resizing process, which probably causes the loss of frequency content.

In the category of salient object detection, Achanta *et al.* [6] devised a full resolution method by which more uniformly highlighted salient regions as well as more precise object boundaries can be obtained. However, their method neglected the spatial layout of objects in images, so it tends to predict background regions as salient. Perazzi *et al.* [8] improved Achanta *et al.*'s model by further considering the appearance contrast and the spatial distribution in saliency detection. In addition to the low-level features, Shen and Wu [37] further integrated higher level prior knowledge, such as the center or semantic prior, into detecting salient objects. Yang *et al.* [11] used the background priors inferred from object boundaries as well as the foreground proposals to rank the saliency degrees of superpixels. Following [11], Zhu *et al.* [1] proposed a more robust method for background prior generation. Their method coupled with other contrast cues achieves the state-of-the-art performance in the single-image saliency detection.

Stemming from the unsupervised nature, the performance of the aforementioned approaches to single-image saliency detection is still restricted. Co-saliency detection is introduced to further improve the performance. The shared visual cues obtained across images facilitate foreground location and background removal. For instance, Li and Ngan [2] utilized the *SimRank* algorithm on a co-multilayer superpixel tree, and detected the color and texture similarity between superpixels across images. Meng *et al.* [38] improved the SimRank matching method by further taking geometric constraints into account. Fu *et al.* [9] proposed a clustering based process to learn inter-image correspondence. To effectively integrate multiple cues, Cao *et al.* [3, 39] employed a low-rank constraint on the salient regions of multiple saliency map proposals, and adaptively determined the fusion weight of each map proposal. Inspired by the fact that the optimal saliency map proposal is often region-dependent, our approach adaptively seeks the weights for saliency fusion in a region-wise manner, thus leading to more favorable results.

2.3 The Proposed Approach

Given a pair of images I_1 and I_2 for co-saliency detection, we apply M existing (co-)saliency detection algorithms, e.g. [2, 4–6], and get M saliency maps for each image. For locally adaptive saliency map fusion, images I_1 and I_2 are respectively decomposed into N_1 and N_2 *superpixels*,

which serve as the domain of region-wise fusion. Our approach aims to seek a weight vector $\mathbf{y}_i = [y_{i,1}\ y_{i,2}\ \ldots\ y_{i,M}]^\top \in \mathbb{R}^M$ for each superpixel i, where $i \in \{1, 2, .., N_1 + N_2\}$. The co-saliency detection is accomplished by superpixel-wise fusing the M saliency maps. Our approach formulates this task of region-wise fusion as an energy minimization problem over a graph. In the following, the image pre-processing and the graph construction are introduced first. The proposed energy function and its optimization are then described.

2.3.1 Image Pre-processing

The *SLIC* algorithm [40] is used for deriving superpixels, because it effectively preserves inherent structures while abstracts unnecessary details. We set the numbers of superpixels to $N_1 = N_2 = 200$ in this work.

Two types of visual features, color and texture, are extracted for each superpixel. For color features, each pixel in the three color spaces, `RGB`, `L*a*b*`, and `YCbCr`, is represented by a 9-dimensional vector. Using the *bag-of-words* model, all pixels in the image pair are quantized into clusters by using the k-means algorithm. Each superpixel is then represented as a $k = 100$-dimensional histogram. For texture features, Gabor filter responses with eight orientations, three scales and two phase offsets are extracted for each pixel. The texture features of a superpixel are similarly encoded as a 100-dimensional histogram by using the bag-of-words model.

Let $\mathbf{p}_i$ and $\mathbf{q}_i$ denote the color and texture representations of superpixel i respectively. The similarity between superpixel i and superpixel j is defined as

$$A(i,j) = \exp(-\frac{d(\mathbf{p}_i, \mathbf{p}_j)}{\sigma_c} - \gamma\frac{d(\mathbf{q}_i, \mathbf{q}_j)}{\sigma_g}), \tag{2.1}$$

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

2.3.2 Graph Construction

We construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2)$. In $\mathcal{G}$, each vertex $v_i \in \mathcal{V}$ corresponds to superpixel i, thus $|\mathcal{V}| = N_1 + N_2$. In this dissertation, we use v_i and i interchangeably when referring to either the vertex on this superpixel graph or superpixel itself. The edge $e_{ij} \in \mathcal{E}_1$ is added to link v_i and v_j if superpixels i and j are spatially connected in an image. The edge $e_{ij} \in \mathcal{E}_2$ is included to connect v_i and v_j if superpixel j is one of the ℓ nearest neighbors of superpixel i in the opposite image according to the similarity in Eq. (2.1). We set $\ell = 1$ to simulate the one-to-one superpixel matching scenario. Edge weights for both types of edges are assigned by (2.1) to get the affinity matrix A for $\mathcal{G}$. We also construct the corresponding Laplacian matrix $L \in \mathbb{R}^{N \times N}$, where $N = N_1 + N_2$.

2.3.3 Energy Function

We seek the optimal weights $Y = [\mathbf{y}_1 \ \mathbf{y}_2 \ \ldots \ \mathbf{y}_N] \in \mathbb{R}^{M \times N}$, where M is the number of saliency maps, and N is the total number of superpixels of I_1 and I_2, for superpixel-wise map fusion by minimizing the proposed energy function

$$\begin{aligned} \min_{Y} \quad & \lambda_1 \sum_{i:v_i \in \mathcal{V}} U(\mathbf{y}_i) + \lambda_2 \sum_{i:v_i \in \mathcal{V}} V(\mathbf{y}_i) \\ & + \lambda_3 \sum_{e_{ij} \in \mathcal{E}} B(\mathbf{y}_i, \mathbf{y}_j) + ||Y||_2^2 \\ \text{s.t.} \quad & \|\mathbf{y}_i\|_1 = 1, \mathbf{y}_i \geq \mathbf{0}, \text{ for } 1 \leq i \leq N, \end{aligned} \tag{2.2}$$

where $\mathbf{0}$ is a vector whose elements are zero, and λ_1, λ_2 and λ_3 are three positive constants. There are four terms introduced in Eq. (2.2). The first two unary terms, $U(\mathbf{y}_i)$ and $V(\mathbf{y}_i)$, respectively leverage intra-image and inter-image evidences to estimate the power of each saliency map on superpixel i. The pairwise term $B(\mathbf{y}_i, \mathbf{y}_j)$ encourages the smoothness of the derived weights on superpixel pairs connected in the graph $\mathcal{G}$. The last term $||Y||_2^2$ is included for regularization.

2.3.3.1 *On Designing Unary Term $U(\mathbf{y}_i)$*

We intend to assign a higher weight to a saliency map that is consistent with other saliency maps on superpixel i. It helps exclude distinct biases in individual maps. Inspired by [41], we employ a low-rank constraint for this task, but we further generalize the method in [41] to *locally* estimate the goodness of each saliency map. For superpixel i, we find its n spatially nearest superpixels. Let $\mathbf{x}_{i,m} \in \mathbb{R}^{256}$ be a 256-dimensional histogram representing the intensity 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* [42] is adopted to decompose X_i into a low-rank approximation L_i plus a residual matrix E_i by solving

$$\min_{L_i,E_i}(||L_i||_* + \lambda||E_i||_1), \quad \text{s.t. } X_i = L_i + E_i, \tag{2.3}$$

where $||L_i||_*$ is the nuclear norm of L_i, and λ is a constant. After solving Eq. (2.3), higher weights are assigned to saliency maps with lower residual errors $E_i = [\mathbf{e}_{i,1}\ \ldots\ \mathbf{e}_{i,M}]$, i.e. ,

$$w_{i,m} = \frac{\exp(-||\mathbf{e}_{i,m}||_2^2)}{\sum_{j=1}^{M}\exp(-||\mathbf{e}_{i,j}||_2^2)},\ \text{for } 1 \le m \le M. \tag{2.4}$$

The above procedure is repeated for each superpixel i. A penalty variable $z_{i,m} = \exp(1 - w_{i,m})/\sum_{j=1}^{M}\exp(1 - w_{i,j})$ is introduced to construct the first term in Eq. (2.2) by letting

$$\sum_{i:v_i\in\mathcal{V}} U(\mathbf{y}_i) = \sum_{i=1}^{N}\mathbf{z}_i^\top\mathbf{y}_i = tr(Z^\top Y), \tag{2.5}$$

where $\mathbf{z}_i = [z_{i,1}\ \ldots\ z_{i,M}]^\top$ and $Z = [\mathbf{z}_1\ \ldots\ \mathbf{z}_N]$.

2.3.3.2 *On Designing Unary Term $V(\mathbf{y}_i)$*

This term is designed to reduce the false saliency detection by exploring inter-image correspondences. Let e_i represent the similarity between superpixel i and its most similar superpixel

in the other image. Let $s_{i,m}$ denote the mean saliency value of saliency map m on superpixel i. The larger the value of e_i is, the more likely superpixel i has a correspondence in the other image. Thus, we prefer saliency map m if the value of $s_{i,m}$ is proportional to that of e_i.

This unary term penalizes the case where only one of e_i and $s_{i,m}$ has large values, encouraging salient regions with matched regions in the other image. Penalizing variable $r_{i,m}$ is defined as

$$r_{i,m} = \frac{\exp((1-e_i)s_{i,m} + e_i(1-s_{i,m}))}{\sum_{j=1}^{M} \exp[(1-e_i)s_{i,j} + e_i(1-s_{i,j})]}. \tag{2.6}$$

The denominator in Eq. (2.6) is for normalization. By considering all superpixels, the second term in Eq. (2.2) becomes

$$\sum_{i:v_i \in \mathcal{V}} V(\mathbf{y}_i) = \sum_{i=1}^{N} \mathbf{r}_i^\top \mathbf{y}_i = tr(R^\top Y), \tag{2.7}$$

where $\mathbf{r}_i = [r_{i,1} \ \ldots \ r_{i,M}]^\top$ and $R = [\mathbf{r}_1 \ \ldots \ \mathbf{r}_N]$.

2.3.3.3 On Designing Pairwise Term $B(\mathbf{y}_i, \mathbf{y}_j)$

We impose this pairwise term to encourage the smoothness of the weight distribution Y between connected superpixels in the graph $\mathcal{G}$. The formulation of this term is defined as

$$\sum_{e_{ij} \in \mathcal{E}} B(\mathbf{y}_i, \mathbf{y}_j) = \sum_{e_{ij} \in \mathcal{E}} A(i,j) \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 = tr(YLY^\top), \tag{2.8}$$

where L is the Laplacian matrix of $\mathcal{G}$.

2.3.4 Optimization Process and Spatial Refinement

With the definitions of the unary and pairwise terms in Eqs. (2.5), (2.7), and (2.8), the constrained optimization problem in Eq. (2.2) is a *quadratic programming* (QP) problem, and has a globally optimal solution. The asymptotic worst-case time complexity using the interior-point method for the convex QP is $\mathcal{O}((NM)^3)$ [43]. We adopt the CVX solver [44] on MATLAB to solve it, and the average running time for each image pair is around 13 seconds on a PC with an Intel i7 2.5GHz CPU and 16G RAM. After optimization, the saliency detection results can be compiled by superpixel-wise fusing the saliency maps with the solution Y. To further improve

the performance, the spatial refinement process [3, 39] is applied to the yielded saliency map. It re-scales the saliency values by a combination of thresholding and normalization.

2.4 Experimental Results

This approach CSLA is evaluated on the *Image Pair* data set [2], which consists of 105 image pairs with manually labeled ground truth.

2.4.1 Experimental Setup

Following [2], we compute five saliency map proposals by three saliency detection algorithms, IT [4], SR [5], and FT [6], and one co-saliency detection algorithm [2] with two features, color CC and texture CP. Except the five proposals, our approach is compared with two fusion-based approaches to co-saliency detection, including LI [2] and SACS [3]. Note that the approaches, LI, SACS, and ours, work by fusing the same five map proposals.

The performance of each evaluated approach is measured by the *precision-recall* (PR) curve, which is obtained by varying the saliency threshold. PR curves tend to favor methods that successfully detect the salient regions over methods that precisely locate the non-salient regions. Thus, *receiver operating characteristics* (ROC) curves and *mean absolute error* (MAE) based on the given ground truth are also included for performance evaluation. In our experiments, we have set $\lambda_1 = 8$, $\lambda_2 = 4$, $\lambda_3 = 1$ in (2.2) and $\lambda = 0.05$ in (2.3).

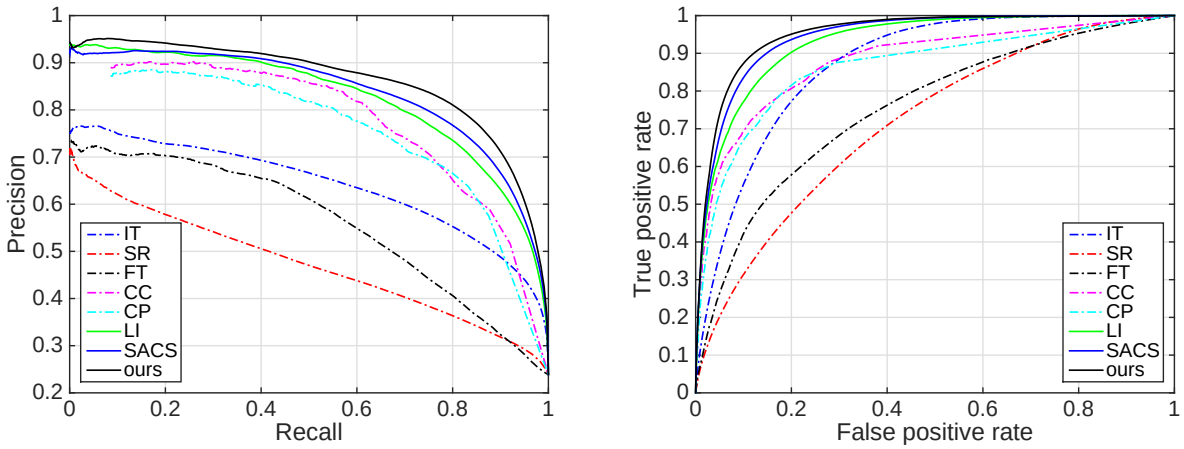
2.4.2 Result Analysis

The PR curves and the ROC curves from our approach and seven competing approaches are shown in Figure 2.2(a) and Figure 2.2(b), respectively. We also report the area under the PR and ROC curves, namely AP (*averaged precision*) and AUC, respectively, as well as the MAE of these approaches in Table 2.1. It can be observed in Figure 2.2 and Table 2.1 that the methods LI [2] and SACS [3] can effectively leverage the mutual signal strengths among the five saliency proposals, IT, SR, FT, CC, and CP, and remarkably outperform all the five proposals. Our approach takes region-wise fusion into account, and can make the most of the five *locally complementary* saliency maps. As shown in Figure 2.2, our approach consistently achieves better performance than all

Method	IT [4]	SR [5]	FT [6]	CC [2]	CP [2]	LI [2]	SACS [3]	ours
AP	0.640	0.471	0.559	0.702	0.681	0.824	0.836	**0.861**
AUC	0.872	0.718	0.756	0.881	0.865	0.930	0.944	**0.952**
MAE	0.259	0.269	0.253	0.163	0.173	0.173	0.172	**0.163**

Table 2.1: The performance of various approaches in 1) AP, 2) AUC, and 3) MAE. The higher the better in the first two measures. The lower the better in MAE.

the competing approaches. In Table 2.1, the performance gain over SACS, the best competing approach, is significant, including 2.5% in AP, 0.9% in the MAE and 0.8% in the AUC.



(a) PR curve (b) ROC curve

Figure 2.2: The performance of various approaches in (a) PR curves and (b) ROC curves.

To gain insight into the quantitative results, Figure 2.3 shows the detected saliency maps on two image pairs by using the seven competing approaches and ours. The saliency proposals that use intra-image evidences, including IT, SR and FT, produce many severe false salient regions. Meanwhile, the saliency proposals that use inter-image evidences, such as CC and CP, detect salient regions with lower confidence. Methods LI and SACS indeed give better results by fusion. Our approach with the aid of region-wise fusion complies the saliency maps that are *perceptually* the

closest to the ground truth. Furthermore, the saliency maps by our approach are sharper, namely detection with higher confidence.

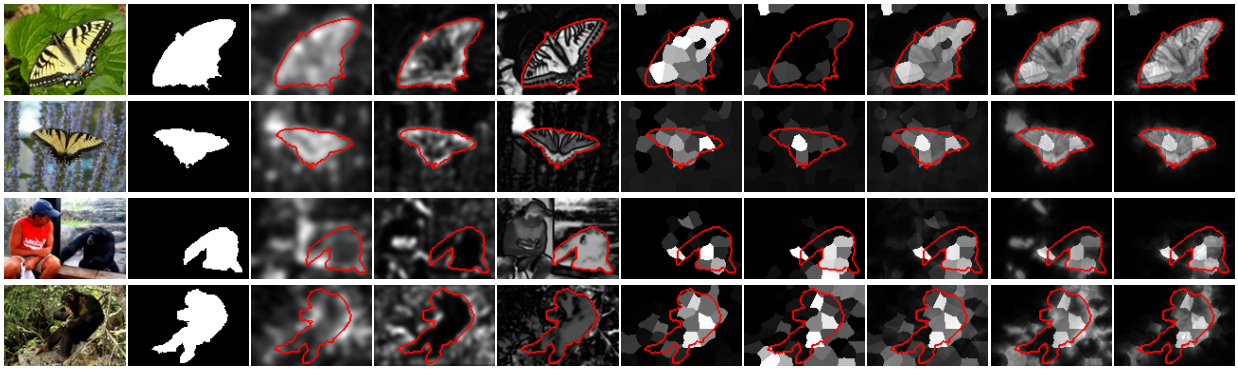


Figure 2.3: (a) & (b) Two exemplar image pairs and the ground truth. (c) $\sim$ (j) Saliency maps generated by different approaches including (c) `IT` [4], (d) `SR` [5], (e) `FT` [6], (f) `CC` [2], (g) `CP` [2], (h) `LI` [2], (i) `SACS` [3], and (j) ours (`CSLA`).

2.5 Conclusions

In this section, we have presented a saliency detection approach that carries out locally adaptive saliency map fusion. Our method is formulated as a quadratic programming problem and can be efficiently optimized by off-the-shelf solvers. Also, our experiments showed it can find a more plausible combination of multiple locally complementary saliency proposals to generate both quantitatively and perceptually high-quality saliency maps. In the following section, we keep improving our approach (`CSLA`) for the robustness of locally adaptive fusion by integrating the objectness evidence revealed by jointly formulated image co-segmentation.

3. SEGMENTATION GUIDED LOCAL PROPOSAL FUSION FOR CO-SALIENCY DETECTION (SGCS)*

3.1 Introduction

Image co-saliency detection [2, 3, 9, 28, 35, 38, 45–49] can help a broad range of image content analysis applications, such as co-segmentation [18, 28, 35] and co-localization [50]. However, its performance is still restricted on extracting representative salient object cues in practical imaging scenarios due to illumination and viewing angle variation. Many 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 [2, 9, 38] or adaptive-weight map fusion methods [3, 49]; 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 [51]. Second, region-wise fusion can be sensitive to noise and content variation in images, and fusion by weighted combinations of saliency proposal typically leads to *blurred* results, especially near the object boundary regions.

Figure 3.1 shows an image pair example, the ground truth for co-saliency detection, the saliency maps generated by using intra-image evidence [6] and inter-image evidence [2] 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 Figure 3.1(e), the adaptive fusion method [3] 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

*Reprinted with permission from "Segmentation Guided Local Proposal Fusion for Co-saliency Detection, Chung-Chi Tsai, Xiaoning Qian, and Yen-Yu Lin, 2017. IEEE International Conference on Multimedia and Expo (ICME), Pages=523–528, Copyright [2017] by IEEE.

remove less certain but real object regions.

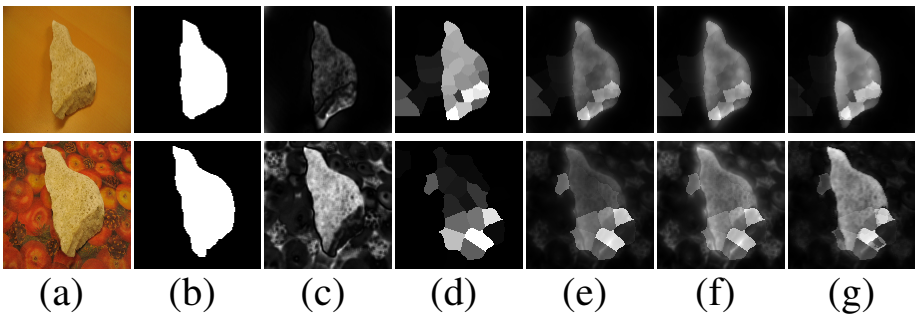


Figure 3.1: Image co-saliency detection. (a) Images. (b) Ground truth. (c) $\sim$ (g) Saliency maps produced by (c) [6] with intra-image evidence, (d) [2] with inter-image evidence, (e) [3] for map fusion, (f) `SGCS` w/o co-segmentation, and (g) ours (`SGCS`).

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 Figure 3.1(f), the variant of our approach where segmentation is turned off recovers the missed foreground regions. Namely, the whole stone in the bottom image is more homogeneously highlighted. In Figure 3.1(g), our approach with the aid of segmentation further suppresses the noise in Figure 3.1(f) and produce sharper co-saliency maps.

3.2 Related Work

Most saliency detection methods target on *human eye fixation prediction* [4,5] or *salient object prediction* [1, 6–8]. 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*. [4] 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*. [6]

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 [2, 35, 45, 46] is introduced to utilize the extra information from inter-image evidence to help salient region localization and background removal. For example, Chang *et al.* proposed a model based on the multiplication of intra-image saliency and inter-image repeatedness [35]. Li and Ngan [2] 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 [4–6]. Meng *et al*. [38] improved the *SimRank* matching method by further taking geometric constraints into account. Fu *et al*. [9] 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*. [45] 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*. [46] 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*. [28] 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*. [3] employed a low-rank constraint to seek the weights for an adaptive combination of multiple saliency proposals. Huang *et al*. [49] constructed a multiscale superpixel tree. Fusion is accomplished by using low-rank analysis to take the saliency results of each scale into account. Methods [3,49] 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

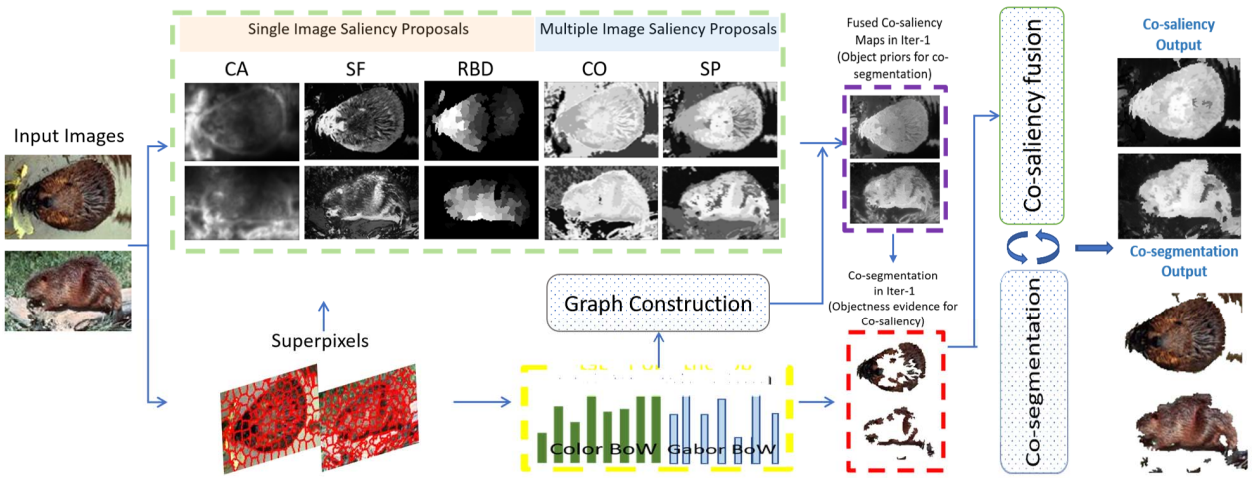


Figure 3.2: The proposed framework of segmentation guided local proposal fusion for co-saliency detection. Given a paired images, we process the input images by compiling their superpixel representation, extracting features from the superpixels, and computing a set of saliency proposals. The proposed approach takes the processed data as input and performs the first-round saliency proposal fusion. The initial co-saliency maps (purple box) provides essential prior knowledge to generate the first-round co-segmentation masks (red box). Afterwards, alternating co-saliency detection and co-segmentation are operated until convergence.

.

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 [3] may consider parts of salient areas with lower saliency confidences as background. If the background priors are incorrectly established [49], 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 foreground-background models derived from the better saliency maps. Thus, post-processing is not further required to obtain good results. Figure 3.2 summarizes our proposed segmentation-guided co-saliency detection based on the local fusion framework.

3.3 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, 2, 4–9] 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 $\mathbf{y}_i = [y_{i,1}\ y_{i,2}\ \ldots\ y_{i,M}]^\top \in \mathbb{R}^M$ for each superpixel i, and use it to accomplish co-saliency detection by region-wise fusing the M saliency maps. We formulate 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.

3.3.1 Image Pre-processing

$N_1 = N_2 = 200$ superpixels are extracted by the *SLIC* algorithm with both *color* and *texture* bag-of-words representations [40]. 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

$$A(i,j) = \exp(-\frac{d(\mathbf{p}_i, \mathbf{p}_j)}{\sigma_c} - \gamma\frac{d(\mathbf{q}_i, \mathbf{q}_j)}{\sigma_g}), \tag{3.1}$$

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.

3.3.2 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. (3.1). The *graph Laplacian* $L \in \mathbb{R}^{N \times N}$ is then obtained.

3.3.3 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:

$$\begin{aligned} J(Y, Z) = \alpha_1 \sum_{i:v_i \in \mathcal{V}} U_1(\mathbf{y}_i) + \alpha_2 \sum_{i:v_i \in \mathcal{V}} U_2(z_i) + \alpha_3 \sum_{i:v_i \in \mathcal{V}} U_3(\mathbf{y}_i, z_i) \\ + \beta_1 \sum_{e_{ij} \in \mathcal{E}} B_1(\mathbf{y}_i, \mathbf{y}_j) + \beta_2 \sum_{e_{ij} \in \mathcal{E}} B_2(z_i, z_j) + |Y\|_2^2 \\ \text{s.t.} \quad \|\mathbf{y}_i\|_1 = 1, \mathbf{y}_i \geq \bar{\mathbf{0}}, z_i \in \{0, 1\}, \text{ for } 1 \leq i \leq N, \end{aligned} \tag{3.2}$$

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 (3.2), $U_1(\mathbf{y}_i)$ 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.

3.3.3.1 On Designing 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. The two parts are

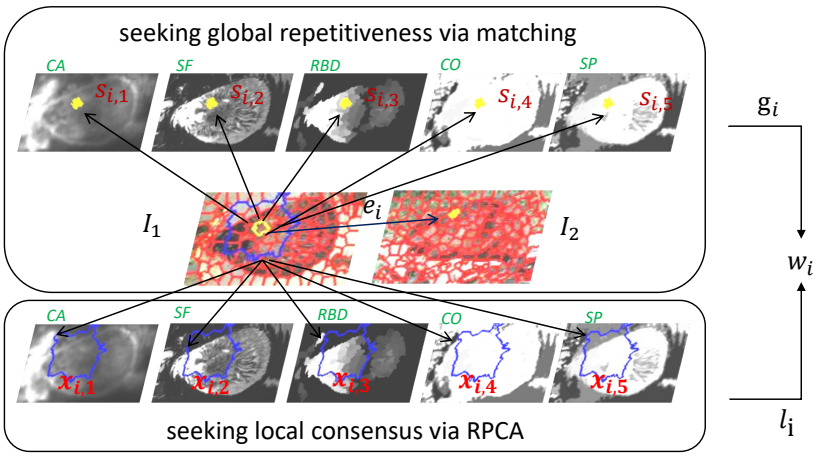


Figure 3.3: Design of the unary term U_1 for regional confidence over saliency proposals in co-saliency detection. See the text for the details.

respectively corresponding to the two diagrams shown in the Figure 3.3.

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 [41], 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

$$\min_{L_i,E_i}(||L_i||_* + \lambda||E_i||_1), \quad \text{s.t. } X_i = L_i + E_i, \tag{3.3}$$

where $||L_i||_*$ is the nuclear norm of L_i, and λ is a constant. After solving (3.3), we compute the normalized residuals by referring to errors $E_i = [\mathbf{e}_{i,1}\ \ldots\ \mathbf{e}_{i,M}]$ via

$$b_{i,m} = \frac{\exp(-||\mathbf{e}_{i,m}||_2^2)}{\sum_{j=1}^{M}\exp(-||\mathbf{e}_{i,j}||_2^2)}, \text{ for } 1 \le m \le M. \tag{3.4}$$

For energy minimization, the associated penalty variable is then defined as

$$l_{i,m} = \exp(1 - b_{i,m}) / \sum_{j=1}^{M} \exp(1 - b_{i,j}). \tag{3.5}$$

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 (3.1). 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

$$g_{i,m} = \frac{\exp((1 - e_i)s_{i,m} + e_i(1 - s_{i,m}))}{\sum_{j=1}^{M} \exp[(1 - e_i)s_{i,j} + e_i(1 - s_{i,j})]}. \tag{3.6}$$

The denominator in (3.6) 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

$$w_{i,m} = \frac{\exp(l_{i,m} + g_{i,m})}{\sum_{j=1}^{M} \exp(l_{i,j} + g_{i,j})}. \tag{3.7}$$

Considering all superpixels, the unary term becomes

$$\sum_{i:v_i \in \mathcal{V}} U_1(\mathbf{y}_i) = \sum_{i=1}^{N} \mathbf{w_i}^\top \mathbf{y_i} = \mathbf{tr}(\mathbf{W}^\top \mathbf{Y}), \tag{3.8}$$

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

3.3.3.2 On Designing Unary Term $U_2(z_i)$

This term estimates the likelihood of superpixel i belonging to the common foreground in co-segmentation. Following [18], 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,

$$\sum_{i:v_i \in \mathcal{V}} U_2(z_i) = \sum_{i=1}^{N} [p(v_i \in \boldsymbol{F}|\mathbf{c}_i)(1 - z_i) + p(v_i \in \boldsymbol{B}|\mathbf{c}_i)z_i]. \tag{3.9}$$

GMM $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_{b,k}$ help predict the probability of superpixel i belonging to the foreground or background. Assuming $p(v_i \in \boldsymbol{F}) = p(v_i \in \boldsymbol{B}) = \frac{1}{2}$, we can get

$$p(v_i \in \boldsymbol{F}|\mathbf{c}_i) = \frac{p(\mathbf{c}_i \in \boldsymbol{F}|\boldsymbol{\theta}_f)p(v_i \in \boldsymbol{F})}{p(\mathbf{c}_i|\boldsymbol{\theta}_f)p(v_i \in \boldsymbol{F}) + \sum_{k=1}^{2} p(\mathbf{c}_i|\boldsymbol{\theta}_{b,k})\delta(v_i \in I_k)p(v_i \in \boldsymbol{B})} \tag{3.10}$$

, where $p(\cdot|\boldsymbol{\theta}_f)$ and $p(\cdot|\boldsymbol{\theta}_{b,k})$ are the Gaussian probability distributions. And, $p(v_i \in \boldsymbol{B}|\mathbf{c}_i)$ is similarly set.

3.3.3.3 On Designing 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

$$s_i = \sum_{m=1}^{M} y_{i,m} s_{i,m} = \mathbf{y}_i^\top \mathbf{s}_i, \tag{3.11}$$

where $\mathbf{y}_i = [y_{i,1} \ \ldots \ y_{i,M}]^\top$ 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

$$\sum_{i:v_i \in \mathcal{V}} U_3(\mathbf{y}_i, z_i) = \sum_{i=1}^{N} s_i(1 - z_i) + (1 - s_i)z_i. \tag{3.12}$$

3.3.3.4 *On Designing 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

$$\sum_{e_{ij}\in\mathcal{E}} B(\mathbf{y}_i, \mathbf{y}_j) = \sum_{e_{ij}\in\mathcal{E}} A(i,j)\|\mathbf{y}_i - \mathbf{y}_j\|_2^2 = tr(YLY^\top), \tag{3.13}$$

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

3.3.3.5 *On Designing 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

$$\sum_{e_{ij}\in\mathcal{E}} B_2(z_i, z_j) = \sum_{e_{ij}\in\mathcal{E}} A(i,j)\|z_i - z_j\|_2^2 = tr(ZLZ^\top). \tag{3.14}$$

3.4 Optimization Process

An iterative strategy is adopted to optimize (3.2). 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.

3.4.1 On Optimizing Y

By fixing Z, the optimization problem in (3.2) becomes

$$\begin{aligned} J(Y) = \alpha_1 \sum_{i:v_i\in\mathcal{V}} U_1(\mathbf{y}_i) + \beta_1 \sum_{e_{ij}\in\mathcal{E}} B_1(\mathbf{y}_i, \mathbf{y}_j) \\ + \alpha_3 \sum_{i:v_i\in\mathcal{V}} U_3(\mathbf{y}_i, z_i) + ||Y||_2^2 \\ \text{s.t.} \quad \|\mathbf{y}_i\|_1 = 1, \mathbf{y}_i \geq \bar{\mathbf{0}}, \text{ for } 1 \leq i \leq N. \end{aligned} \tag{3.15}$$

The above constrained optimization problem is a *quadratic programming* problem. We solve it by using the CVX [44].

3.4.2 On Optimizing Z

By fixing Y, the optimization task in (3.2) becomes

$$\begin{aligned} J(Z) = \alpha_2 \sum_{i:v_i \in \mathcal{V}} U_2(z_i) + \beta_2 \sum_{e_{ij} \in \mathcal{E}} B_2(z_i, z_j) \\ + \alpha_3 \sum_{i:v_i \in \mathcal{V}} U_3(\mathbf{y}_i, z_i) \\ \text{s.t.} \quad z_i \in \{0, 1\}, \text{ for } 1 \leq i \leq N. \end{aligned} \tag{3.16}$$

The energy function in (3.16) is graph representable and regular. Thus it can be efficiently minimized via graph cuts.

3.4.3 Implementation Details

For initialization, we solve the weights Y for saliency map fusion via (3.15) 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 (3.9) and enable the optimization of (3.16) at the first iteration. Following [6], 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 (3.15) and (3.16) iteratively.

3.5 Experimental Results

This proposed approach `SGCS` is evaluated on the *Image Pair* data set [2], which consists of 105 image pairs with manually labeled ground truth.

3.5.1 Experimental Setup

We evaluate our approach, and compare it with the state-of-the-art methods on the *Image Pair data set* [2], 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 [2] and get five saliency proposals consisting of three single-image

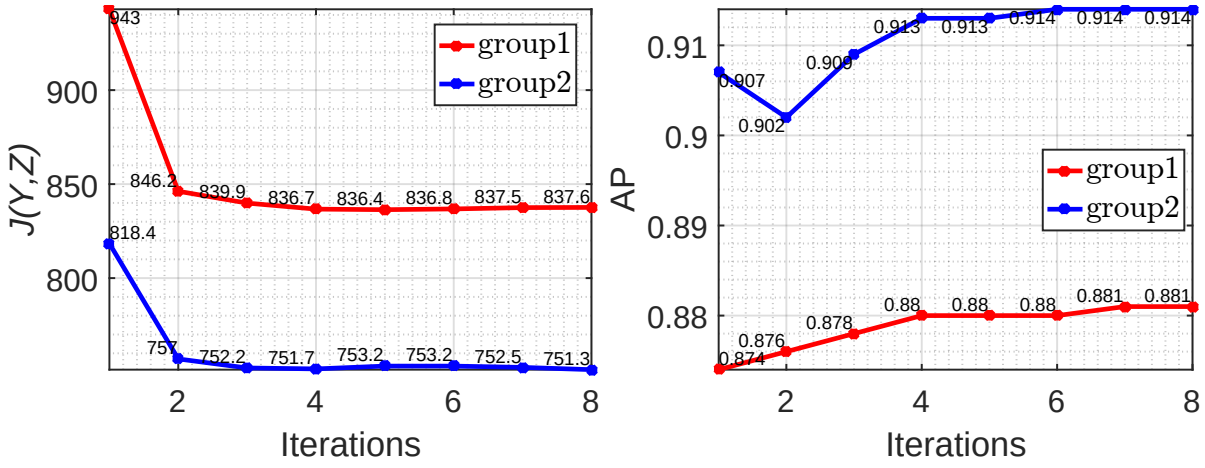


Figure 3.4: (a) The energy curves of the proposed optimization function (b) The AP curves, versus iterations, in two different saliency proposal groups.

saliency maps (SISM) obtained by methods IT [4], SR [5], and FT [6] and two multi-image saliency maps (MISM) by using the algorithm in [2] with two different features, color CC and texture CP. The second group contains three SISMs by using methods CA [7], SF [8], and RBD [1], and two MISMs obtained by using the detection algorithm in [9] with two different features, spatial cues SP and correspondence cues CO. Our approach is also compared with two fusion-based methods for co-saliency detections, including the fixed-weighted summation method CSM [2] and the self-adaptive fusion method SACS [3]. 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 (3.2) and $\lambda = 0.05$ in (3.3). The objective function values in (3.2) and the performance AP by our approach through alternating optimization are shown in Figure 3.4(a) and (b), respectively. Both of them converge rapidly. We report the results of our approach at iteration 7.

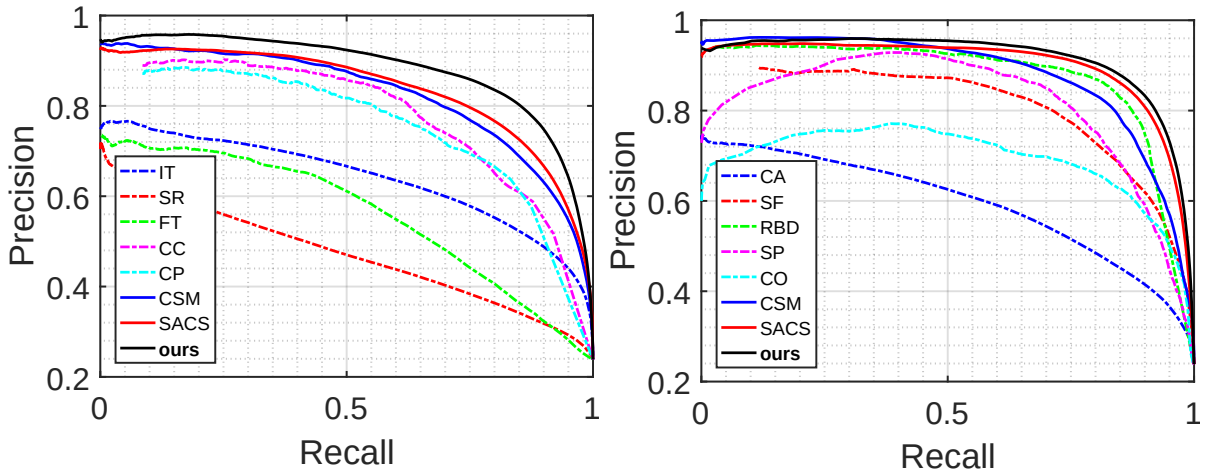


Figure 3.5: The PR curves of the evaluated approaches with the saliency proposals in (a) group 1 and (b) group 2.

3.5.2 Result Analysis

The PR curves of the evaluated approaches with saliency proposal groups 1 and 2 are drawn in Figure 3.5(a) and (b) respectively. The performances in AP and AUC are also reported in Tables 3.1 and 3.2. With saliency proposal group 1, it can be observed in Table 3.1 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 Figure 3.5(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 Figure 3.5 and Table 3.2 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.

method	IT [4]	SR [5]	FT [6]	CC [2]	CP [2]	CSM [2]	SACS [3]	**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 3.1: The performance of various approaches in AP (average precision) and AUC (area under the ROC curve) on saliency proposal group 1.

method	CA [7]	SF [8]	RBD [1]	SP [9]	CO [9]	CSM [2]	SACS [3]	**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 3.2: The performance of various approaches in AP (average precision) and AUC (area under the ROC curve) on saliency proposal group 2.

To gain insight into the quantitative results, Figure 3.6 displays the detected saliency maps on two image pairs, when saliency proposal group 1 is adopted. The saliency proposals, i.e. those in Figure 3.6(c) $\sim$ 3.6(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. Figure 3.7 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.

3.6 Conclusions

In this section, 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 desirable com-

bination 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 the following section, we discuss the bilateral relationship of image co-saliency and co-segmentation based on this segmentation guided fusion model while making it scalable to multiple input images.

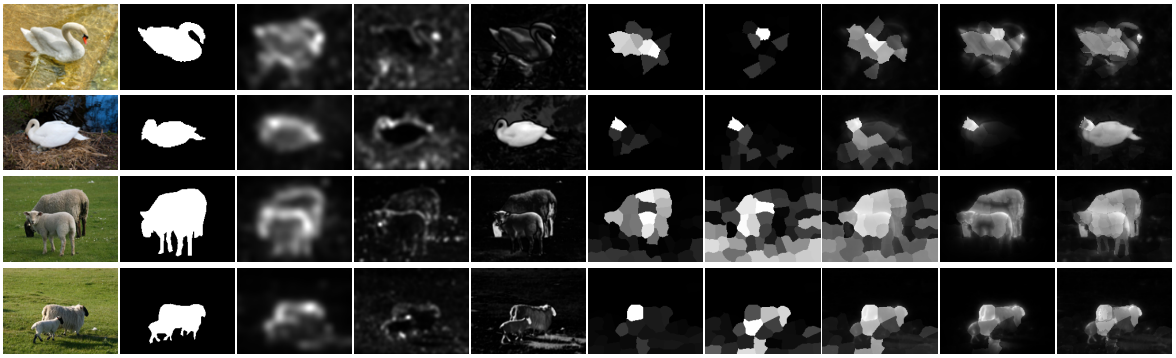


Figure 3.6: (a) & (b) Two image pairs for co-saliency detection and the ground truth. (c) $\sim$ (j) Saliency maps generated by different approaches including (c) IT [4], (d) SR [5], (e) FT [6], (f) CC [2], (g) CP [2], (h) CSM [2], (i) SACS [3], and (j) ours (SGCS).

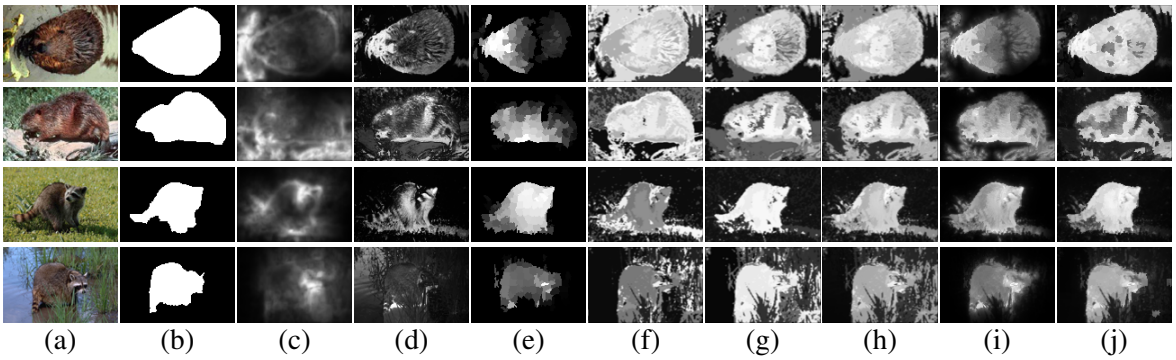
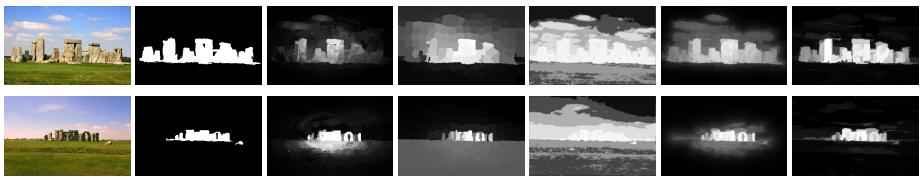


Figure 3.7: (a) & (b) Two image pairs for co-saliency detection and the ground truth. (c) $\sim$ (j) Saliency maps generated by different approaches including (c) CA [7], (d) SF [8], (e) RBD [1], (f) CO [9], (g) SP [9], (h) CSM [2], (i) SACS [3], and (j) ours (SGCS).

4. IMAGE CO-SALIENCY AND CO-SEGMENTATION VIA PROGRESSIVE JOINT OPTIMIZATION (CSCS)

4.1 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 co-localization [50, 52] and video compression [36]. 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 [53] and image matching [54]. In this section, we investigate the strengths and weaknesses of co-saliency detection and co-segmentation given multiple input images. 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)

Figure 4.1: (a) A pair of images for co-saliency detection. (b) The ground truth. (c) $\sim$ (e) Three saliency proposals generated by DSR [10], MR [11], and SpC [9] respectively. (f) The detection results by the fusion-based method SACS [3]. (g) The detection results by our method (CSCS).

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

Figure 4.2: Our approach enables the progressive improvement of co-saliency detection and co-segmentation. (a) & (f) Two images and the ground truth (top row) for co-saliency detection. (b) $\sim$ (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) $\sim$ (j) The results for the image in (b).

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* [2, 45, 55], *fixed-weight multiplication* [9, 55] or *adaptive-weight summation* [3, 39]. Figure 4.1(c) $\sim$ Figure 4.1(e) show different saliency proposals generated by the method `DSR` [10], the method `MR` [11], and using the multi-image spatial cue (`SpC`) [9], respectively. None of them gives satisfactory results. The algorithm `SACS` [3] implements adaptive weighted summation of the three proposals, and significantly improves the detection results as shown in Figure 4.1(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, 51] or previous sections. Secondly, weighted combinations of different saliency proposals typically lead to blurred results, especially near the surrounding areas of objects; also, without holistic object boundary information, region-wise fusion might be interfered by intra-image noise and inter-image content variation. Fortunately, the evidence of *objectness* from co-segmentation can guide region-wise saliency proposal fusion and help recover sharp object boundaries [14]. Our approach can integrate co-segmentation into co-saliency detection, and achieves the superior results displayed in Figure 4.1(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. Figure 4.2 shows an example of the progressive improvement of co-saliency detection and co-segmentation by our method. Given a pair of images in Figs. 4.2(a) and 4.2(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. 4.2(e) and 4.2(j).

4.2 Related Work

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

4.2.1 Saliency Detection

The literature of saliency detection is extensive. Methods for saliency detection can be roughly sorted into human *visual attention prediction* [4, 5, 56–58] and *salient object location* [1, 6, 8, 10–12, 15, 37, 49, 55, 59–65]. 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*. [4] presented a pioneering saliency detection model based on local contrast computed from the center-surround differences across multiple scales. Borgi and Itti [56] 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 [5] defined the saliency through the residual on the Fourier domain of an input image; and Xia *et al*. [57] 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 aims to spotlight entire salient objects, instead of merely their boundaries or discriminative parts in visual attention prediction. Achanta *et al*. [6] approximated saliency based on the deviation between a low-pass filtered image and the average color of the whole image. Perazzi *et al*. [8] 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. [1, 10–12, 15, 37, 55, 59–63], were developed to reduce the computation load and ease the influence of image noise. In addition to low-level features, Shen and Wu [37] further took high-level knowledge, such as face locations and center priors, into account. Some approaches to saliency detection, such as [10, 11, 15], 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*. [1] further integrated global contrast with the improved background priors to achieve better perfor-

mance. Moreover, methods based on graph-based clustering, e.g. [60–63], were proposed to better locate the potential objects.

Recent research efforts, e.g. [64–67], have been made to use *convolutional neural networks* (CNNs) for saliency detection. The features learned by CNNs usually give better performance than handcrafted features. However, CNN-based methods rely on labeled training data or extra information sources for tuning the deep models, which are generally unavailable in single-image saliency detection. Stemming from the unsupervised nature, the performance of these methods based on either the learned features or handcrafted features for single-image saliency detection is still limited.

4.2.2 Co-saliency Detection

Co-saliency detection is a weakly supervised extension of salient detection as it further explores the visual cues shared across multiple images to better identify salient objects. Chang *et al*. [35] formulated co-saliency as a combination of intra-image saliency and inter-image repetitiveness. Fu *et al*. [9] 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.

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*. [2] and Fu *et al*. [9] 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*. [3, 39] sought adaptive fusion weights based on a low-rank constraint on different salient foreground color content. Huang *et al*. [49] 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*. [51]

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*. [68] 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. [3, 49, 51, 68], is that the resultant saliency maps are typically blurred, especially near the object boundaries. Thus, post-processing is often required; but it is often *ad-hoc* and may degenerate the performance.

Segmentation has been integrated into co-saliency detection e.g. [14, 28, 45, 46, 69] to enhance the performance. Li *et al*. [45] applied *GrabCut* [70] to multi-scale initialization windows, and found the common segmented objects for intra-image saliency estimation. Jerripothula *et al*. [28] 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.

4.2.3 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*. [71] introduced the pioneering work of co-segmentation by minimizing the unnormalized foreground histogram dissimilarity in *Markov random field* (MRF). Hochbaum and Singh [72] used a sub-modular rewarding term to encourage similar pixels having same labels and efficiently solved it by *graph-cut*. Joulin *et al*. [17, 20] 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*. [35] introduced a co-saliency guided method for co-segmentation by taking into account foreground similarity and figure-ground dissimilarity. Yu *et al*. [18] 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*. [22] used depth enhanced co-saliency maps for co-segmentation. Meng *et al*. [19] 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*. [21] designed several energy terms constructed by using saliency and correspondence information for co-segmentation.

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. [14, 28, 45, 46, 69], or leverage co-saliency detection to help co-segmentation, e.g. [18,19,21,22,35]. Our approach instead enables simultaneous co-saliency saliency and co-segmentation. It *bidirectionally* links the two tasks in the domain of superpixels whose pair-wise 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 a unidirectional approach, our previous model in Section 3 [14] 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 in Section 3 [14] for co-saliency detection. More importantly, this work further improves co-segmentation with the integration of co-saliency detection, and make extension to two data sets.

4.3 The Proposed Approach

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.

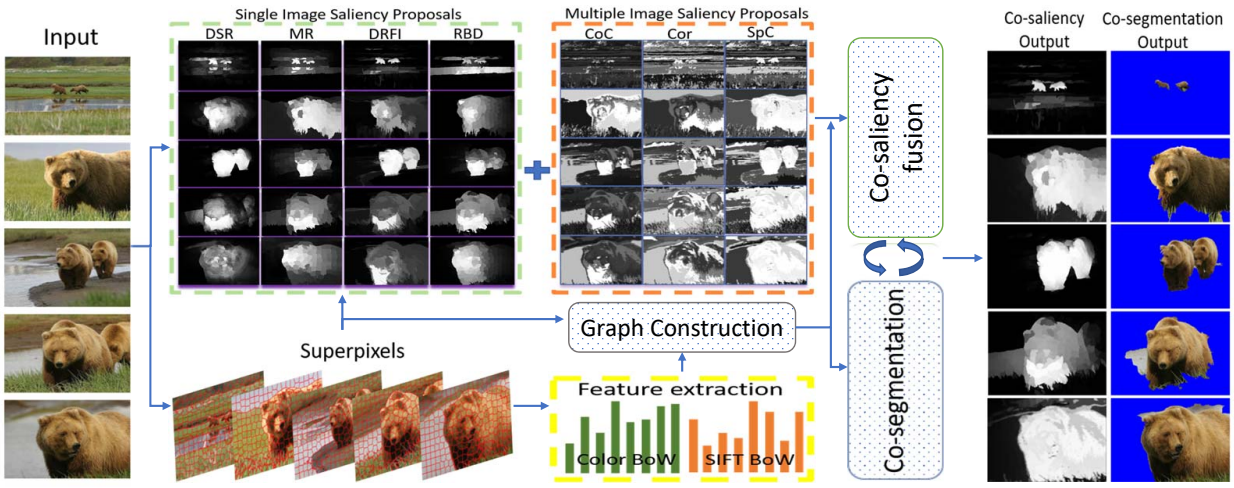


Figure 4.3: The proposed framework for joint co-saliency detection and co-segmentation. Given images of a particular object category, we process the input images by compiling their superpixel representation, extracting features from the superpixels, and computing a set of saliency proposals. The proposed approach takes the processed data as input, and performs alternating co-saliency detection and co-segmentation until convergence.

4.3.1 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. [1, 9, 10, 12, 73], 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 $\mathbf{y}_i = [y_{i,1}\ y_{i,2}\ \ldots\ y_{i,M}]^\top \in [0, 1]^M$ 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 $z_i \in \{0, 1\}$, $i \in \{1, 2, \cdots, N\}$.

Figure 4.3 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.

4.3.2 Superpixel and Feature Extraction

Each image I_j is decomposed into $N_j \approx 200$ superpixels by using the *SLIC* algorithm [40]. We extract both the *color* and *SIFT* [74] features and use the *bag-of-words* (BoWs) model for superpixel representation. The k-means clustering algorithm is applied to pixels in three color spaces, i.e. `RGB`, `L*a*b*`, and `YCbCr`, and generates 200 *visual words*. The color BoWs representation of a superpixel is then a 200-dimensional histogram. The SIFT BoWs representation is similarly set.

Let $\mathbf{h}_i^c$ and $\mathbf{h}_i^s$ denote the color and SIFT histograms of superpixel i, respectively. The similarity between two superpixels i and $\hat{i}$ is defined by

$$s(i,\hat{i}) = \exp(-\frac{d(\mathbf{h}_i^c, \mathbf{h}_{\hat{i}}^c)}{\sigma_c} - \frac{d(\mathbf{h}_i^s, \mathbf{h}_{\hat{i}}^s)}{\sigma_s}), \tag{4.1}$$

where $d(\cdot,\cdot)$ is the χ^2 distance. Constant σ_c is set to the average pair-wise distance between all superpixels under the color features. Constant σ_s is similarly set.

4.3.3 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

$$A(i,\hat{i}) = s(i,\hat{i}) * b(i,\hat{i}), \tag{4.2}$$

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 (4.2), the associated *graph Laplacian* $L \in \mathbb{R}^{N\times N}$ can be computed.

4.3.4 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:

$$\begin{aligned} J(Y,Z) = \|Y\|_2^2 + \alpha_1 \sum_{i:v_i\in\mathcal{V}} U(\mathbf{y}_i) + \alpha_2 \sum_{1\le j<\hat{j}\le n} D(\mathbf{z}_j, \mathbf{z}_{\hat{j}}) \\ + \alpha_3 \sum_{i:v_i\in\mathcal{V}} C(\mathbf{y}_i, z_i) + \alpha_4 \sum_{e_{i\hat{i}}\in\mathcal{E}} B_1(\mathbf{y}_i, \mathbf{y}_{\hat{i}}) + \alpha_5 \sum_{e_{i\hat{i}}\in\mathcal{E}} B_2(z_i, z_{\hat{i}}) \\ \text{s.t.} \quad \|\mathbf{y}_i\|_1 = 1, \mathbf{y}_i \ge \bar{\mathbf{0}}, z_i \in \{0,1\}, \text{ for } 1\le i\le N, \end{aligned} \tag{4.3}$$

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 (4.3), $U(\mathbf{y}_i)$ 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.

4.3.4.1 On Designing Unary Term $U(\mathbf{y}_i)$ for Co-saliency Detection

We follow the co-saliency formula

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

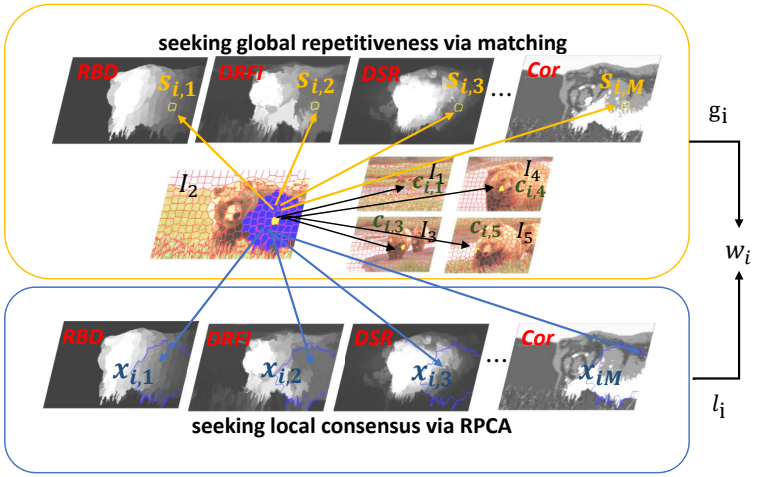


Figure 4.4: Illustration of the unary term U term for regional confidence over saliency proposals in co-saliency detection. See the text for the details.

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 Figure 4.4.

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 [41], 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 Figure 4.4 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 Figure 4.4. 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) [42] is adopted to decompose X_i into a low-rank approximation R_i and a residual matrix E_i by solving

$$\min_{R_i,E_i}(||R_i||_* + \lambda||E_i||_1), \quad \text{s.t. } X_i = R_i + E_i, \tag{4.4}$$

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 (4.4), we convert normalized errors $E_i = [\mathbf{e}_{i,1} \ \ldots \ \mathbf{e}_{i,M}]$ to *belief*:

$$b_{i,m} = \frac{\exp(-||\mathbf{e}_{i,m}||_2^2)}{\sum_{k=1}^{M} \exp(-||\mathbf{e}_{i,k}||_2^2)}, \text{ for } 1 \leq m \leq M. \tag{4.5}$$

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

$$l_{i,m} = \frac{\exp(1 - b_{i,m})}{\sum_{k=1}^{M} \exp(1 - b_{i,k})}. \tag{4.6}$$

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 (4.1), 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 Figure 4.4 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

$$c_i = \frac{\texttt{mean}(\{c_{i,j} | 1 \leq j \leq n\})}{\texttt{var}(\{c_{i,j} | 1 \leq j \leq n\}) + 1}. \tag{4.7}$$

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.

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 Figure 4.4. 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.

$$g_{i,m} = \frac{\exp((1-c_i)s_{i,m} + c_i(1-s_{i,m}))}{\sum_{k=1}^{M} \exp((1-c_i)s_{i,k} + c_i(1-s_{i,k}))}. \tag{4.8}$$

The denominator in (4.8) is used for normalization.

The intra- and inter-image cues on superpixel i and proposal m, i.e. $l_{i,m}$ in (4.6) and $g_{i,m}$ (4.8), are combined via

$$w_{i,m} = \frac{\exp(l_{i,m} + g_{i,m})}{\sum_{k=1}^{M} \exp(l_{i,k} + g_{i,k})} \times \texttt{size}(i), \tag{4.9}$$

where $\texttt{size}(i)$ is the size of superpixel i.

Considering all superpixels, this unary term becomes

$$\sum_{i:v_i \in \mathcal{V}} U(\mathbf{y}_i) = \sum_{i=1}^{N} \mathbf{w}_i^\top \mathbf{y}_i = tr(W^\top Y), \tag{4.10}$$

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

4.3.4.2 On Designing 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 $\mathbf{h}_i = [\mathbf{h}_i^c \ \mathbf{h}_i^s]$, a concatenation of the color and SIFT BoWs representation. 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 [35] to discriminatively assess figure-ground separation for a

pair of images I_j and $I_{\hat{j}}$. This discriminative term is designed below

$$
\begin{aligned}
D(\mathbf{z}_j, \mathbf{z}_{\hat{j}}) &= \|H_j^f - H_{\hat{j}}^f\|_2^2 - \sum_{k\in\{j,\hat{j}\}} \gamma_1 \|H_k^f - \gamma_2 H_k^b\|_2^2 \\
&= R - 2 \sum_{z_i\in\mathbf{z}_j, z_{\hat{i}}\in\mathbf{z}_{\hat{j}}} \langle\, \mathbf{h}_i, \mathbf{h}_{\hat{i}}\rangle z_i z_{\hat{i}} \\
&\quad + 2\gamma_1\gamma_2(1+\gamma_2) \sum_{k\in\{j,\hat{j}\}} \sum_{z_i\in\mathbf{z}_k} \langle \mathbf{h}_i, H_k^f + H_k^b\rangle z_i \\
&\quad + (1 - \gamma_1(1+\gamma_2)^2) \sum_{k\in\{j,\hat{j}\}} \sum_{z_i, z_{\hat{i}}\in\mathbf{z}_k} \langle \mathbf{h}_i, \mathbf{h}_{\hat{i}}\rangle z_i z_{\hat{i}},
\end{aligned} \tag{4.11}
$$

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 [75] can be adopted, this term must satisfy the regularity condition [75]. Namely, the coefficient $(1 - \gamma_1(1+\gamma_2)^2)$ must not be larger than 0. Following [35], 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

$$
\begin{aligned}
D(\mathbf{z}_j, \mathbf{z}_{\hat{j}}) &= R - 2 \sum_{z_i\in\mathbf{z}_j, z_{\hat{i}}\in\mathbf{z}_{\hat{j}}} \langle \mathbf{h}_i, \mathbf{h}_{\hat{i}}\rangle z_i z_{\hat{i}} \\
&\quad + 2\gamma \sum_{k\in j,\hat{j}} \sum_{z_i\in\mathbf{z}_k} \langle \mathbf{h}_i, H_k^f + H_k^b\rangle z_i.
\end{aligned} \tag{4.12}
$$

In (4.12), 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.

4.3.4.3 On Designing 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

$$
s_i = \sum_{m=1}^{M} y_{i,m} s_{i,m} = \mathbf{y}_i^\top \mathbf{s}_i, \tag{4.13}
$$

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 $s_i \in [0,1]$. 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

$$\sum_{i:v_i \in \mathcal{V}} C(\mathbf{y}_i, z_i) = \sum_{i=1}^{N} [s_i(1-z_i) + (1-s_i-\pi)z_i] \times \texttt{size}(v_i), \tag{4.14}$$

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. [21, 35], to prevent the trivial solutions that all superpixels are assigned to background. We will discuss how to set its value in the experiments. In (4.14), the sizes of superpixels are also taken into account.

4.3.4.4 On Designing Pairwise Term $B_1(\mathbf{y}_i, \mathbf{y}_{\hat{i}})$ for Co-saliency Detection

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

$$\sum_{e_{i\hat{i}} \in \mathcal{E}} B_1(\mathbf{y}_i, \mathbf{y}_{\hat{i}}) = \sum_{e_{i\hat{i}} \in \mathcal{E}} A(i,\hat{i}) \|\mathbf{y}_i - \mathbf{y}_{\hat{i}}\|_2^2 = tr(YLY^\top), \tag{4.15}$$

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

4.3.4.5 On Designing 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

$$\sum_{e_{i\hat{i}} \in \mathcal{E}} B_2(z_i, z_{\hat{i}}) = \sum_{e_{i\hat{i}} \in \mathcal{E}} A(i,\hat{i}) \|z_i - z_{\hat{i}}\|_2^2 = tr(ZLZ^\top). \tag{4.16}$$

4.3.5 Optimization

Simultaneously solving the two sets of variables Y and Z is hard. An alternating strategy is adopted to optimize the variables in (4.3). 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.

4.3.5.1 On Optimizing Y

By fixing Z, the optimization problem in (4.3) becomes

$$\begin{aligned} J(Y) = \alpha_3 \sum_{i:v_i \in \mathcal{V}} C(\mathbf{y}_i, z_i) + \alpha_4 \sum_{e_{i\hat{i}} \in \mathcal{E}} B_1(\mathbf{y}_i, \mathbf{y}_{\hat{i}}) \\ + \alpha_1 \sum_{i:v_i \in \mathcal{V}} U(\mathbf{y}_i) + \|Y\|_2^2 \\ \text{s.t.} \quad \|\mathbf{y}_i\|_1 = 1, \mathbf{y}_i \geq \bar{\mathbf{0}}, \text{ for } 1 \leq i \leq N. \end{aligned} \tag{4.17}$$

The above constrained optimization problem is a *quadratic programming* problem. We efficiently solve it by using the public software CVX [44].

4.3.5.2 On Optimizing Z

By fixing Y, the optimization problem in (4.3) is reduced to

$$\begin{aligned} J(Z) = \alpha_3 \sum_{i:v_i \in \mathcal{V}} C(\mathbf{y}_i, z_i) + \alpha_5 \sum_{e_{i\hat{i}} \in \mathcal{E}} B_2(z_i, z_{\hat{i}}) \\ + \alpha_2 \sum_{1 \leq j < \hat{j} < n} D(\mathbf{z}_j, \mathbf{z}_{\hat{j}}) \\ \text{s.t.} \quad z_i \in \{0, 1\}, \text{ for } 1 \leq i \leq N, \end{aligned} \tag{4.18}$$

which is a *binary labeling* problem. The energy function in (4.18) is graph representable and regular, and hence can be efficiently minimized by graph-cut [75].

For initialization, we solve the weights Y for saliency proposal fusion via (4.17) 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 (4.11) is determined. It follows that the optimization problems in (4.17) and (4.18) can be iteratively solved. The value of the objective function decreases and converges to a local optimum. To conclude this

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 $\mathcal{I}$; (Sec. 4.3-A)

1: Decompose each image into superpixels; (Sec. 4.3-B)
2: Extract features for each superpixel; (Sec. 4.3-B)
3: Construct graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with affinity matrix A in (4.2);
4: Initialize the saliency maps via (4.17) with term C removed;
5: Set γ in (4.12) based on the foreground-background ratios;
6: Iteration $\leftarrow$ 1; Iteration $\leq$ T Solve Y for co-saliency detection via (4.17); Solve Z for co-segmentation via (4.18); Iteration = Iteration + 1

section, we summarize our approach in Algorithm 1.

4.4 Experimental Results

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

4.4.1 Data Sets for Performance Evaluation

Two benchmarks, the *Image-Pair* [2] and the *iCoseg* [34] data sets, used for performance valuation are described below:

4.4.1.1 Image-Pair Data Set

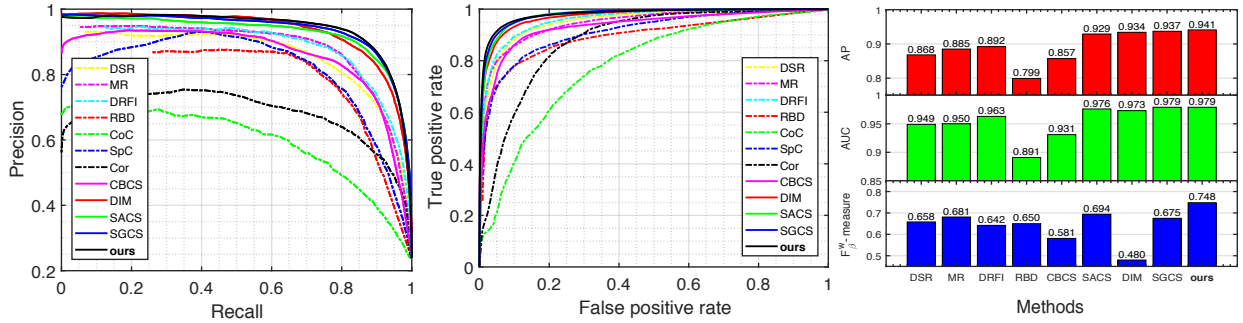
This data set has 105 image pairs with manually labeled ground truth. Images of a pair contain one or multiple common objects appearing on two distinct backgrounds. We use the whole data set for co-saliency detection, and the subset of 30 pairs used in [76] for co-segmentation.

4.4.1.2 iCoseg Data Set

It is a large-scale data set 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 [21,28] using the same 31 groups for co-segmentation. The images of a group contain one single or multiple similar objects with

(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_β^w	0.307	0.772	F_β^w	0.320	0.855

Figure 4.5: Deficiency of AP and AUC. (a) & (d) The ground truth of two examples. (b) & (c) Two saliency proposals for (a). (e) & (f) Two saliency proposals for (d).



(a) (b) (c)

Figure 4.6: Co-saliency evaluation on Image-Pair in (a) PR curves, (b) ROC curves, and (c) Overall quantitative scores. The models adopted to generate our fusion proposals are plotted in dash lines, while the state-of-the-art models are in solid lines.

various poses and sizes on complex backgrounds. Therefore, this benchmark is more challenging than the Image-Pair data set for both co-saliency detection and co-segmentation.

4.4.2 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 binary map. The precision $\mathcal{P}$, recall $\mathcal{R}$, and the false positive rate (FPR) are respectively defined by

$$\mathcal{P} = \frac{TP}{TP+FP}, \ \mathcal{R} = \frac{TP}{TP+FN}, \text{ and } FPR = \frac{FP}{TN+FP}. \tag{4.19}$$

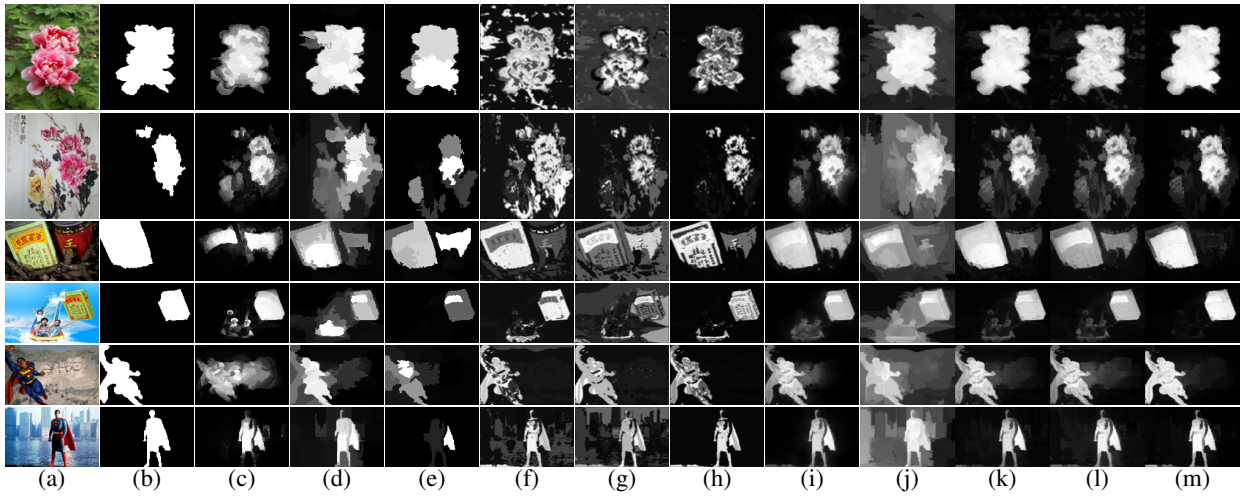


Figure 4.7: (a) & (b) Three image pairs for co-saliency detection and the ground truth. (c) $\sim$ (j) Saliency maps generated by different approaches including (c) `DSR` [10], (d) `DRFI` [12], (e) `RBD` [1], (f) `Cor` [9], (g) `CoC` [9], (h) `CBCS` [9], (i) `SACS` [3], (j) `DIM` [13], (k) `SGCS` [14], (l) our approach without referencing co-segmentation evidence `CSCS-iter1`, and (m) ours `(CSCS)` .

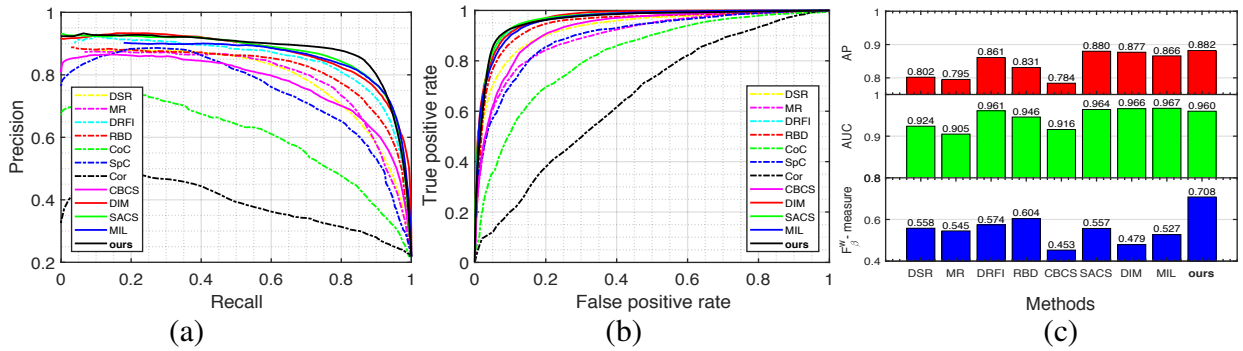


Figure 4.8: Co-saliency evaluation on iCoseg in (a) PR curves, (b) ROC curves, and (c) Overall quantitative scores. The models adopted to generate our fusion proposals are plotted in dash lines, while the state-of-the-art models are in solid lines.

To evaluate the performance of co-saliency detection, we first consider two universally-agreed criteria, i.e. *average precision* (AP), and *area under the ROC curve* (AUC). AUC can be considered the aggregated statistics from the *receiver operating characteristic* (ROC) curve for true positive rate (or recall $\mathcal{R}$) and false positive rate (FPR). AP is the score computed as the area under the precision ($\mathcal{P}$) and recall ($\mathcal{R}$) curve (PR curve). The PR and ROC curves are generated by

thresholding the pixels in the predicted co-saliency maps with a series of integers from 0 to 255. 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.

As pointed out in [77], AP and AUC are less discriminative in some circumstances. Two such examples are shown in Figure 4.5. The saliency proposals in Figures 4.5(c) & 4.5(f) are perceptually closer to the respective ground truth in Figures 4.5(a) & 4.5(d). However due to the *interpolation flaw* [77], the proposals in Figures 4.5(b) & 4.5(e) have the AP and AUC scores higher than those in Figures 4.5(c) & 4.5(f), respectively. To address this issue for more comprehensive evaluation, we also adopt the generalized F_β-measure, i.e. weighted F_β^w measure [77], defined as:

$$F_\beta^w = \frac{(1+\beta^2)\mathcal{P}^w \cdot \mathcal{R}^w}{\beta^2 \cdot \mathcal{P}^w + \mathcal{R}^w} \tag{4.20}$$

which alleviates the hidden flaws of AP and AUC, including the interpolation flaw, dependency flaw, and equal importance flaw, for more reliable evaluation of the detected saliency maps. In the experiment, we set $\beta = 1$ by following the original setting in the source code of [77] that equally weighs the importance of *weighted precision* ($\mathcal{P}^w$) and *weighted recall* ($\mathcal{R}^w$) based on the similar definitions in (4.19) with four weighted basic quantities, i.e. , TP^w, TN^w, FP^w and FN^w, defined as:

$$\begin{aligned} TP^w &= (1 - E^w) \cdot G \\ TN^w &= (1 - E^w) \cdot (1 - G) \\ FP^w &= (E^w) \cdot (1 - G) \\ FN^w &= (E^w) \cdot G, \end{aligned} \tag{4.21}$$

where 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 column-stack representation of the predicted saliency map) by considering the individual pixel error according to

their relative location and neighborhood information by referencing 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 pixels in ground-truth. The two criteria are defined as follows:

$$\mathcal{A} = \frac{TP + TN}{TP + TN + FP + FN} \text{ and } \mathcal{J} = \frac{TP}{TP + FP + FN}. \tag{4.22}$$

4.4.3 Implementation Details

For saliency proposal fusion, we choose four single-image saliency proposals (SISP), i.e. `DSR` [10], `MR` [73], `DRFI` [12] and `RBD` [1], 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 [9]. We select these proposals by jointly considering their performances, popularity and complementary effect on proposal fusion. To generate the saliency proposals on the Image-Pair and iCoseg data sets, we run the source code from the corresponding publications with the default settings.

We evaluate our approach in two different perspectives/tasks on each data set, namely, *co-segmentation guided co-saliency detection* and *co-saliency detection guided co-segmentation* with the same optimization model (3.2). 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 data set (Image-Pair or iCoseg), 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 data set.

We search for the proper values of the parameters in the following order based on their importance to the model (3.2), i.e. α_1, α_2, α_3, π, α_4, 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 method [21] by adjusting the background shift per image group in the iCoseg data set.

Methods	Alaskan-Bear		Red-Sox		Stongehenge		Salisbury		Liverpool		Ferrari		LeszekZadlo		Inde-du		Egypt		Elephants	
	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w
DSR [10]	0.753	0.544	0.742	0.498	0.734	0.464	0.748	0.425	0.623	0.484	0.907	0.684	0.756	0.540	0.527	0.362	0.549	0.520	0.922	0.666
MR [73]	0.765	0.523	0.843	0.492	0.826	0.432	0.798	0.477	0.632	0.476	0.851	0.639	0.722	0.480	0.505	0.419	0.530	0.467	0.884	0.508
DRFI [12]	0.808	0.493	0.920	0.563	0.942	0.673	0.815	0.468	0.592	0.459	0.820	0.596	0.700	0.477	0.558	0.381	0.511	0.464	0.885	0.549
RBD [1]	0.755	0.547	0.906	0.648	0.908	0.629	0.850	0.610	0.693	0.519	0.876	0.660	0.863	0.541	0.401	0.379	0.590	0.575	0.850	0.588
CoC [9]	0.350	0.219	0.754	0.275	0.254	0.179	0.300	0.193	0.583	0.282	0.481	0.329	0.833	0.351	0.143	0.102	0.143	0.113	0.221	0.160
SpC [9]	0.666	0.315	0.744	0.196	0.840	0.351	0.883	0.621	0.650	0.194	0.891	0.475	0.835	0.366	0.518	0.264	0.683	0.300	0.793	0.273
Cor [9]	0.548	0.247	0.667	0.148	0.158	0.200	0.540	0.462	0.470	0.173	0.469	0.322	0.755	0.284	0.334	0.266	0.592	0.282	0.443	0.245
CBCS [9]	0.544	0.309	0.875	0.566	0.501	0.299	0.606	0.223	0.711	0.429	0.879	0.593	0.899	**0.588**	0.407	0.267	0.335	0.212	0.563	0.324
SACS [3]	0.799	0.505	**0.948**	0.555	0.938	0.553	0.849	0.543	0.740	0.438	0.909	0.626	**0.902**	0.526	0.498	0.381	0.521	0.475	0.942	0.554
DIM [13]	0.807	0.417	0.932	0.422	0.940	0.567	**0.940**	0.620	0.749	0.336	**0.958**	0.532	0.868	0.460	0.517	0.293	0.398	0.321	0.902	0.438
MIL [16]	0.848	0.478	0.918	0.450	0.928	0.489	0.873	0.603	**0.866**	0.442	0.949	0.595	0.894	0.398	**0.695**	0.391	0.500	0.343	0.920	0.403
ours (CSCS)	**0.862**	**0.654**	0.939	**0.695**	**0.955**	**0.724**	0.894	**0.726**	0.752	**0.571**	0.866	**0.706**	0.874	0.582	0.544	**0.461**	**0.707**	**0.608**	**0.950**	**0.706**

Methods	Goose		Pandas-Tai		Helicopter		Planes		Huntsville		Cheetah		Pandas		Brighton-kite		Kitekid		Margate-kite	
	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w
DSR [10]	0.686	0.377	0.841	0.496	0.904	0.809	0.710	0.440	0.944	0.631	0.822	0.564	0.876	0.467	0.963	0.812	0.913	0.547	0.876	0.618
MR [73]	0.914	0.497	0.842	0.526	0.930	0.805	0.572	0.474	0.688	0.567	0.848	0.578	0.849	0.498	0.905	0.807	0.832	0.469	0.810	0.448
DRFI [12]	0.883	0.551	0.913	0.551	0.962	0.744	0.779	0.442	0.941	0.562	0.914	0.574	0.893	0.525	0.961	0.655	0.938	0.572	0.959	0.706
RBD [1]	0.931	0.550	0.833	0.508	0.900	0.774	0.676	0.493	0.838	0.634	0.776	0.516	0.902	0.553	0.957	0.781	0.905	0.546	0.884	0.648
CoC [9]	0.947	0.515	0.656	0.426	0.907	0.427	0.785	0.202	0.937	0.122	0.403	0.302	0.791	0.504	0.854	0.317	0.750	0.461	0.538	0.339
SpC [9]	0.811	0.504	0.861	0.572	0.821	0.214	0.563	0.152	0.775	0.048	0.885	0.555	0.817	0.602	0.917	0.239	0.917	0.549	0.913	0.500
Cor [9]	0.507	0.328	0.756	0.473	0.363	0.146	0.516	0.062	0.609	0.042	0.528	0.395	0.730	0.571	0.323	0.116	0.518	0.432	0.600	0.398
CBCS [9]	0.937	0.432	0.873	0.444	0.903	0.676	0.771	0.441	0.910	0.443	0.805	0.387	0.885	0.497	0.985	0.622	0.875	0.428	0.908	0.500
SACS [3]	0.878	0.532	0.927	0.596	0.968	0.645	**0.800**	0.356	**0.966**	0.411	0.910	0.608	0.923	0.595	**0.987**	0.628	**0.966**	0.604	0.958	0.632
DIM [13]	**0.978**	0.634	0.945	0.580	0.953	0.508	0.790	0.278	0.942	0.247	0.923	0.542	0.916	0.595	0.959	0.501	0.896	0.496	0.970	0.611
MIL [16]	0.969	0.753	**0.978**	0.699	0.930	0.572	0.654	0.265	0.646	0.212	**0.966**	0.682	**0.955**	0.674	0.954	0.466	0.929	0.543	0.974	0.606
ours (CSCS)	0.932	**0.785**	0.906	**0.700**	**0.977**	**0.855**	0.703	**0.571**	0.950	**0.723**	0.881	**0.700**	0.929	**0.764**	0.980	**0.857**	0.953	**0.743**	**0.977**	**0.836**

Methods	Colt-Park		Gymnastics-1		Gymnastics-2		Gymnastics-3		Skating-Rich		Skating-ISU		Woman-Socc1		Woman-Socc2		Monks		Hot-Balloons	
	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w
DSR [10]	0.950	0.702	0.936	0.638	0.893	0.667	0.958	0.712	0.882	0.511	0.975	0.829	0.816	0.620	0.417	0.332	0.821	0.582	0.965	0.689
MR [73]	0.942	0.685	0.857	0.467	0.867	0.678	0.841	0.571	0.787	0.453	0.952	0.855	0.833	0.595	0.413	0.334	0.919	0.753	0.928	0.500
DRFI [12]	0.948	0.600	0.980	0.671	0.930	0.647	0.973	0.741	0.860	0.586	0.982	0.702	0.793	0.603	0.658	0.498	0.861	0.555	0.977	0.685
RBD [1]	0.890	0.632	0.982	0.764	0.931	0.675	0.970	0.792	0.878	0.584	0.885	0.722	0.833	0.602	0.519	0.474	0.896	0.652	0.902	0.652
CoC [9]	0.882	0.259	0.869	0.412	0.771	0.411	0.891	0.508	0.855	0.512	0.711	0.382	0.774	0.366	0.383	0.241	0.802	0.461	0.665	0.255
SpC [9]	0.841	0.182	0.717	0.169	0.750	0.226	0.906	0.370	0.806	0.516	0.871	0.144	0.914	0.380	0.692	0.317	0.911	0.607	0.933	0.246
Cor [9]	0.095	0.098	0.062	0.066	0.078	0.073	0.093	0.123	0.669	0.413	0.264	0.090	0.339	0.200	0.397	0.202	0.334	0.330	0.260	0.153
CBCS [9]	0.946	0.596	0.881	0.398	0.817	0.427	0.913	0.524	0.925	0.451	0.980	0.702	0.855	0.555	0.462	0.342	0.943	0.646	0.894	0.388
SACS [3]	**0.988**	0.587	**0.992**	0.619	0.945	0.616	**0.977**	0.684	0.922	0.602	**0.989**	0.643	0.875	0.582	0.503	0.415	0.955	0.694	0.991	0.584
DIM [13]	0.972	0.359	0.980	0.521	**0.962**	0.539	0.972	0.563	**0.968**	0.622	0.971	0.465	0.894	0.488	**0.741**	0.396	0.973	0.644	0.930	0.411
MIL [16]	0.886	0.451	0.979	0.578	0.952	0.530	0.969	0.622	0.962	0.613	0.955	0.660	**0.936**	0.560	0.446	0.376	**0.978**	0.689	0.931	0.552
ours (CSCS)	0.975	**0.768**	**0.992**	**0.795**	0.953	**0.772**	0.976	**0.798**	0.885	**0.709**	0.984	**0.903**	0.840	**0.645**	0.608	**0.506**	0.933	**0.763**	**0.994**	**0.782**

Methods	EricaJoy		Christ		Speedskating		Track		Windmill		Kendo-Kendo		Kendo-EKC		Brown-Bear		**Average**	
	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w	AP	F_β^w
DSR [10]	0.888	0.627	0.889	0.581	0.523	0.336	0.646	0.344	0.454	0.195	0.857	0.595	0.968	0.744	0.860	0.530	0.802	0.558
MR [73]	0.938	0.639	0.830	0.606	0.663	0.309	0.643	0.362	0.541	0.324	0.926	0.732	0.951	0.718	0.830	0.529	0.795	0.545
DRFI [12]	0.977	0.682	0.936	0.637	**0.868**	0.357	**0.778**	**0.492**	**0.670**	**0.439**	0.960	0.678	0.972	0.712	0.902	0.529	0.861	0.574
RBD [1]	0.978	0.707	**0.959**	0.706	0.807	0.416	0.666	0.427	0.512	0.382	0.956	0.728	0.983	0.851	0.755	0.489	0.831	0.604
CoC [9]	0.955	0.504	0.793	0.468	0.235	0.161	0.415	0.301	0.395	0.163	**0.984**	0.665	0.961	0.567	0.499	0.294	0.644	0.336
SpC [9]	0.861	0.503	0.846	0.460	0.318	0.097	0.410	0.257	0.275	0.105	0.673	0.358	0.679	0.336	0.816	0.514	0.763	0.344
Cor [9]	0.531	0.317	0.836	0.491	0.078	0.083	0.343	0.262	0.058	0.067	0.719	0.306	0.431	0.293	0.533	0.355	0.435	0.250
CBCS [9]	0.968	0.523	0.861	0.484	0.500	0.257	0.524	0.268	0.522	0.257	0.977	0.773	0.971	0.615	0.666	0.321	0.784	0.453
SACS [3]	0.979	0.683	0.950	0.680	0.852	0.312	0.716	0.429	0.607	0.283	0.979	0.713	0.991	0.719	0.892	0.554	0.880	0.557
DIM [13]	0.978	0.629	0.883	0.603	0.707	0.221	0.693	0.314	0.518	0.206	0.980	0.623	0.992	0.710	0.921	0.501	0.877	0.479
MIL [16]	**0.988**	0.783	0.873	0.737	0.707	0.198	0.665	0.363	0.494	0.302	0.975	0.730	0.984	0.648	0.877	0.553	0.866	0.527
ours (CSCS)	0.983	**0.851**	0.901	**0.789**	0.768	**0.484**	0.683	0.445	0.569	0.399	**0.984**	**0.888**	**0.994**	**0.906**	**0.954**	**0.734**	**0.882**	**0.708**

Table 4.1: Group-wise co-saliency evaluation results on the iCoseg data set. The best performance is marked in bold. With segmentation guidance, our unsupervised approach can even surpass the learning-based models, i.e. , MIL and DIM, that either requires high-quality pseudo ground truth or transferred knowledge from an auxiliary data set, respectively.

When solving the optimization for (3.2), alternating optimization scheme of co-saliency detection and co-segmentation is repeated for a few iterations until the energy in (3.2) converges. Our model converges rapidly, so we set the maximum number of iterations $T = 5$ in Algorithm 1.

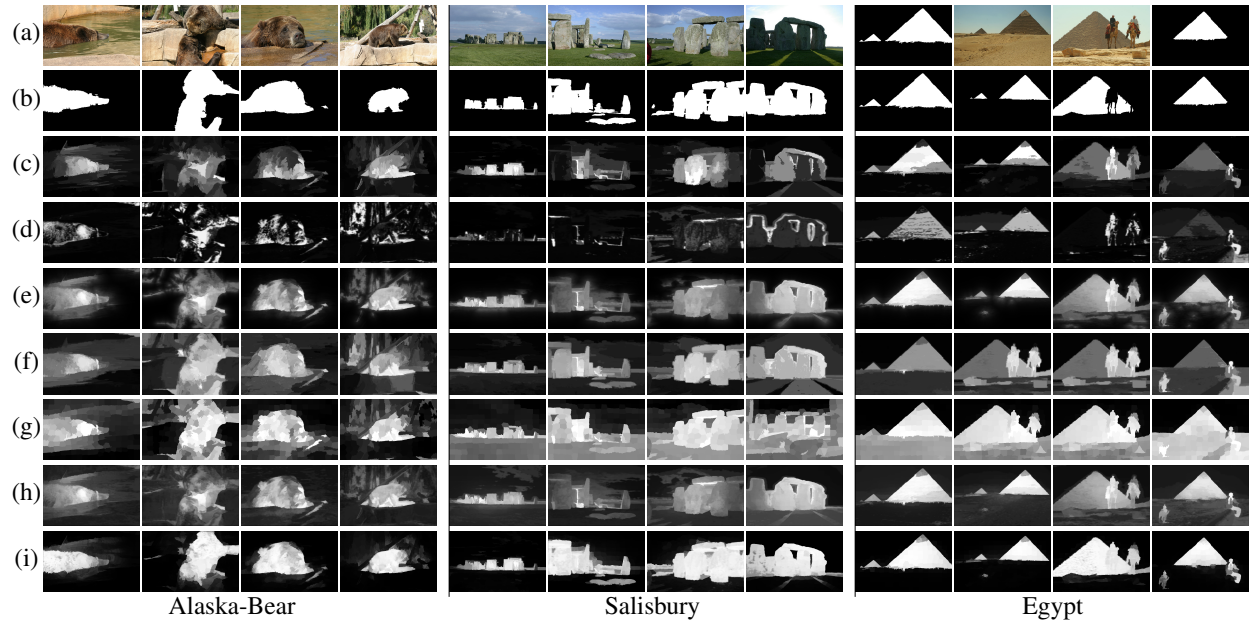


Figure 4.9: (a) & (b) Three image groups for co-saliency detection and the ground truth. (c) $\sim$ (i) Saliency maps generated by different approaches including (c) `DRFI` [15], (d) `CBCS` [9], (e) `SACS` [3], (f) `DIM` [13], (g) `MIL` [16], (h) `CSCS-iter1`: our approach without referring to the co-segmentation evidence, and (i) ours (`CSCS`).

4.4.4 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 and iCoseg data sets in the following.

4.4.4.1 Image-Pair Data Set

We compare our approach with seven adopted saliency proposals and other co-saliency detection methods, including the bottom-up based co-saliency model `CBCS` [9], the map-wise fusion-based co-saliency model `SACS` [3], and our prior work `SGCS` [14] based on the same set of the saliency proposals. We further include the deep learning-based approach `DIM` [13] 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 [6] data set. We either reproduce the co-saliency detection results from the released code [3, 9, 14] or directly get the results from their Websites [13].

Method	Joul10 [17]	Yu14 [18]	Gao13 [76]	Meng13 [19]	ours (CSCS)
Jaccard ($\mathcal{J}$)	59.1	55.6	—	77.7	**82.5**
Accuracy ($\mathcal{A}$)	79.0	86.2	92.4	92.8	**95.1**

Table 4.2: Co-segmentation evaluation result on the 30 image pairs from the Image-Pair data set. The best performance is marked in bold

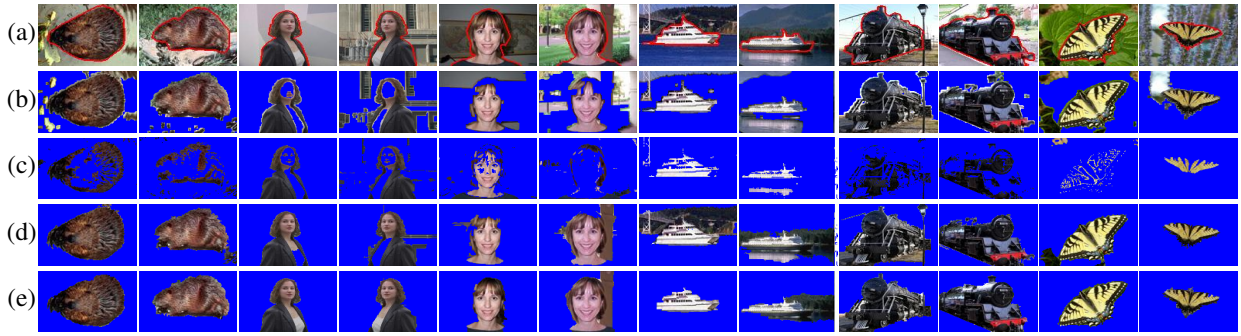


Figure 4.10: (a) Six image pairs for co-segmentation with the ground truth marked by the red contours. (b) $\sim$ (e) Segmentation results generated by different approaches including (b) Joul10 [17], (c) Yu14 [18], (d) Meng13 [19], and (e) ours (CSCS).

Figure 4.6 displays the PR and ROC curves, and the overall scores in different performance measures. We find that fusion based approaches consistently improve the bottom-up saliency proposals based on certain visual cues. Our method further addresses the issues of map-wise fusion in SACS by using region-wise fusion. Meanwhile, it enhances the co-segmentation strength 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 0.7% in AP and 26.8% in F_β^w.

Figure 4.7 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 [10], DRFI [12], and RBD [1], can get the dominating performance as they either produce some unfavorable false alarms or miss some object parts. The proposal Cor [9] searches the corresponding regions and the proposal CoC [9] looks for the contrast regions across images. They

Methods	Alaskan-Bear		Red-Sox		Salisbury		Liverpool		Ferrari		Taj Mahal		Egypt		Elephants	
	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.
Joul12 [20]	49.3	76.8	19.3	61.3	64.1	78.2	39.0	77.8	38.7	65.7	38.0	68.1	30.7	56.8	35.7	68.7
Rubi13 [21]	65.3	88.6	65.7	87.7	59.5	78.9	**54.1**	89.4	72.4	**92.7**	46.0	78.9	**61.1**	87.3	68.8	90.7
Fu15 [22]	**75.4**	**93.5**	**71.6**	96.5	65.4	83.5	48.4	**92.1**	**72.5**	91.7	43.2	82.7	51.1	87.8	69.4	90.4
ours (CSCS)	69.9	92.2	70.7	**96.9**	**74.2**	**84.3**	47.5	91.9	69.2	91.5	**51.0**	**84.4**	60.7	**88.2**	**79.2**	**95.8**
Methods	Goose		Pandas-Tai		Helicopter		Cheetah		Pandas		Gymnastics-1		Gymnastics-2		Gymnastics-3	
	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.
Joul12 [20]	81.8	95.1	79.8	**90.1**	15.3	52.9	37.4	59.0	36.9	53.3	17.3	57.0	31.4	68.2	39.1	73.5
Rubi13 [21]	74.2	91.4	75.9	88.5	80.3	98.5	69.7	85.6	62.5	74.7	**94.8**	**99.4**	**84.0**	**96.1**	**89.6**	**98.2**
Fu15 [22]	47.6	82.6	62.1	82.1	**85.7**	**98.8**	62.8	84.2	44.7	64.6	59.2	95.4	55.2	93.3	53.0	92.0
ours (CSCS)	**86.8**	**95.5**	**80.1**	89.7	80.5	98.5	**74.8**	**87.5**	**73.1**	**82.7**	88.7	98.7	73.7	94.2	76.9	95.7
Methods	Skating-Rich		Skating-ISU		Woman-Soccer1		Woman-Soccer2		Monks		Hot-Balloons		EricaJoy		Christ	
	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.
Joul12 [20]	62.5	80.2	15.9	58.2	28.9	58.4	15.7	53.0	26.2	47.7	32.5	71.2	84.8	95.2	82.7	**93.6**
Rubi13 [21]	**73.5**	**87.9**	**91.1**	**99.3**	**66.1**	**90.1**	**53.0**	**86.9**	68.1	87.0	65.7	90.5	79.9	93.5	77.0	89.7
Fu15 [22]	56.9	81.7	84.5	98.7	51.6	90.0	29.2	82.9	57.4	83.8	90.3	96.5	77.7	92.7	61.5	82.3
ours (CSCS)	62.1	82.6	89.6	99.2	56.4	88.3	49.6	**86.9**	**73.6**	**90.9**	**92.1**	**99.2**	**86.7**	**96.3**	**82.8**	**93.6**
Methods	Speedskating		Track		Windmill		Kendo-Kendo		Kendo-EKC		Brown-Bear				**Average**	
	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.	Jacc.	Accu.			Jacc.	Accu.
Joul12 [20]	15.1	61.5	45.9	74.2	12.9	51.9	86.9	96.4	67.1	88.7	48.6	73.7			42.6	70.2
Rubi13 [21]	44.9	87.5	**51.9**	**86.3**	**49.2**	**90.8**	77.8	92.3	82.6	94.8	73.6	92.0			69.3	89.8
Fu15 [22]	**51.4**	91.6	32.4	84.8	23.0	85.5	61.0	89.7	60.3	88.9	**78.7**	**94.8**			59.4	88.5
ours (CSCS)	46.9	**92.3**	45.0	81.7	31.5	84.5	**89.1**	**97.2**	**94.8**	**98.6**	**78.7**	93.7			**71.2**	**91.8**

Table 4.3: Group-wise and average co-segmentation evaluation results on the iCoseg data set. The best performance is marked in bold.

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 [9] 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 CSCS-iter1, which shows the saliency 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.

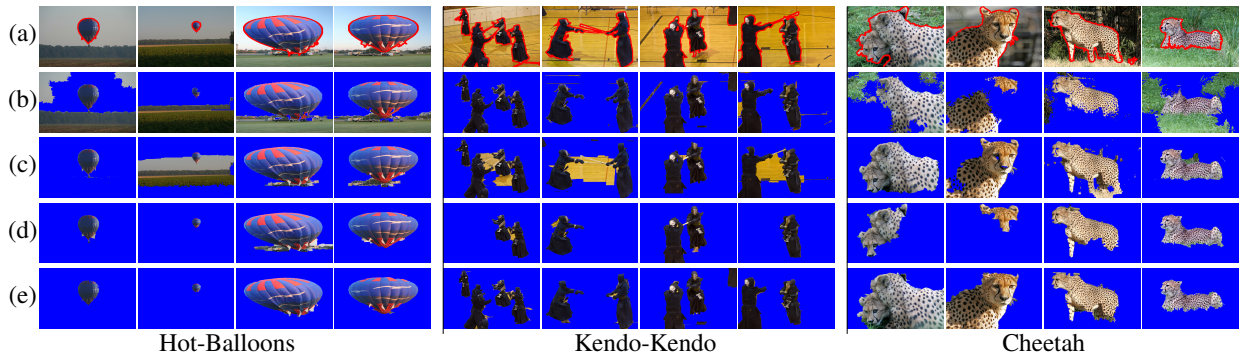


Figure 4.11: (a) Three image groups for co-segmentation with their ground truth marked by the red contours. (b) $\sim$ (e) Segmentation results generated by different approaches including (b) `Joul12` [20], (c) `Rubi13` [21], (d) `Fu15` [22], and (e) ours (`CSCS`).

4.4.4.2 iCoseg Data Set

We further evaluate our method for co-segmentation guided co-saliency detection on the iCoseg data set. Table 4.1 shows the group-wise AP and F_β^w evaluation and Figure 4.8 shows the overall evaluation. Our approach results in a large performance gain over seven adopted saliency proposals in all evaluation metrics. We further compare our approach with the state-of-the-art co-saliency detection models, including the conventional unsupervised approaches, i.e. `CBCS` [9] and `SACS` [3], and the recent deep learning models, `DIM` [13] and `MIL` [16], with more complex initialization taking advantage of the deep networks pre-trained on other data sets. As an unsupervised method, our method outperforms these methods and achieves the best performance in average AP and F_β^w as well as comparable performance in average AUC on this more challenging data set.

Visual comparison is shown in Figure 4.9. 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 Alaska-Bear and Stonehenge. 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.

4.4.5 Co-saliency Detection Guided Co-segmentation

In the following, we evaluate our model for co-segmentation with the integration of co-saliency detection.

4.4.5.1 Image-Pair Data Set

We first evaluate the co-segmentation performance on the Image-Pair data set. Table 4.2 reports the performances of our method and three powerful co-segmentation methods, including Joul10 [17], Meng13 [19], Yu14 [18]. The performance gain of our approach over Meng13, the best competing approach tailored for paired image co-segmentation, is significant, i.e. 4.8% gain in Jaccard index and 2.3% gain in accuracy.

In Joul10 [17], 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 Figure 4.10, 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 second boat image. In addition, Joul10 [17] usually requires more images to derive a good hyperplane separating foreground instances from background.

The MRF-based model Yu14 [18] 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 [9]. We reproduce their results with the recommended settings. As shown in the third row of Figure 4.10, this method has fewer false positives compared to Joul10, but it suffers from large object variations across images. In fact, it has the lowest Jaccard indices in Table 4.2.

The method Meng13 [19] 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 Figure 4.10, 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 4.2 and Figure 4.10 show that our method remarkably outperforms the competing methods.

4.4.5.2 iCoseg Data Set

We further evaluate our method for co-segmentation on the iCoseg data set. Table 4.3 reports the quantitative results and Figure 4.11 shows visual comparison among our method and the existing methods for co-segmentation of more than two images, including Joul12 [20], Rubi13 [21], and Fu15 [22]. We download each of their co-segmentation masks from the original authors' Websites.

The method Joul12 [20] extends their previous work [17] 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 Figure 4.11(b).

The method Rubi13 [21] addresses the issues for images with noisy background or irrele-

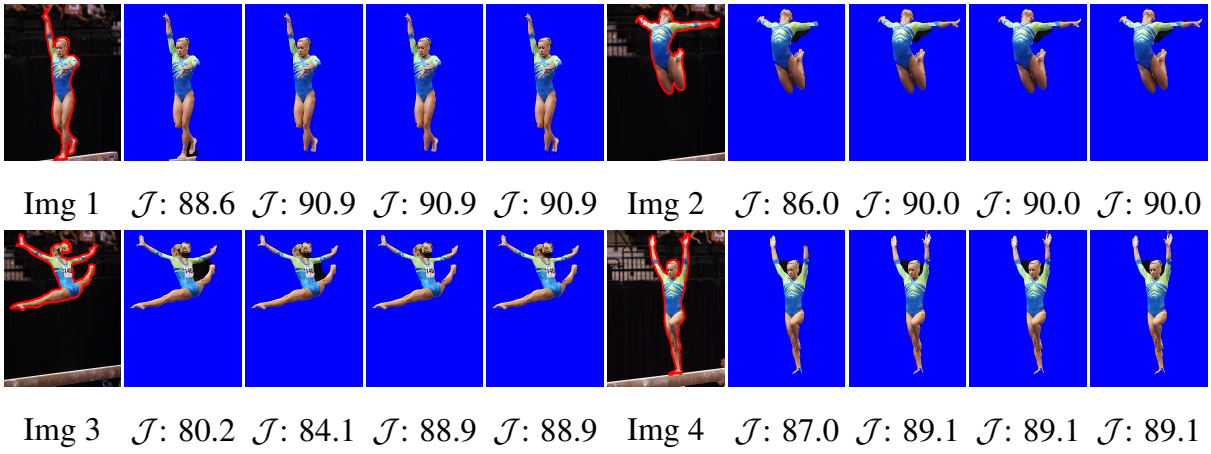


Figure 4.12: Progressively improved co-segmentation results in terms of both visualization and Jaccard indices.

vant 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 `Joul12`. In Figure 4.11(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` [22] 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 [3]. 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 Figure 4.11(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 the last row of Figure 4.11. We also illustrate the progressive improvement of co-segmentation in Figure 4.12. It is clear that at the first round, the segmented regions of the dancer have some deficiency or contain false positives. After a few iterations, the background regions in Img 3 and the saddle in Img 1 are excluded. Meanwhile, the dancer's hands in Img 2 and Img 4 are restored.

4.5 Conclusions

In this section, we have presented an unsupervised learning framework that simultaneously accomplishes co-saliency detection and co-segmentation with multiple input images. On 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 prior. The benefits of the joint optimization formulation are evident as it produces 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 prior. Moreover, unlike existing co-saliency models relying on additional post-processing to smooth their model outputs, our formulation has already inherently merged this step into the unified optimization process and generates even superior results in both tasks evaluated on two different data sets under the same evaluation metrics. Motivated by the success of our efforts, we further identify the remaining challenge of fusion-based (co-)saliency detection in the following section and solve it by using deep learning to achieve practical significance in real-world applications that require high-quality co-saliency maps.

5. DEEP CO-SALIENCY DETECTION VIA STACKED AUTOENCODER-ENABLED FUSION AND SELF-TRAINED CNNS

5.1 Introduction

Co-salient object detection simulates human visual systems to search for visually attracting objects repetitively appearing across images. As an essential component of visual content understanding, it has become an inherent part in many applications, such as image/video co-segmentation [22, 28] and content-aware compression [78]. Despite the significant progress on co-saliency detection [2, 3, 9, 13, 14, 16, 24, 46, 51, 69, 79–83], the general conclusion is still that no single model is sufficient for handling increasingly complex saliency detection of broad object categories.

To overcome this issue, saliency detection via proposal fusion has been a trend since it can combine the strengths of diverse saliency models while easing individual bias as we have witnessed in the previous sections. Advanced fusion methods, e.g. [3, 14, 28, 51], often adaptively rank the proposal quality before determining the weights for fusion. However, these methods judge the proposals' quality by measuring the degree of consistency with the other proposals. In other words, they assume the foreground regions from different saliency proposals have a high correlation; and thus they consider a proposal more reliable if its corresponding predictions agree with the group consensus. However, such an assumption may not hold if the adopted saliency proposals are not reliable or have substantial variations.

Another research trend is to employ deep convolutional models to automatically learn the discriminative features for salient object detection [27, 67, 81, 84–88]. However, most off-the-shelf models require large-scale manual supervision for the ground truth and cannot address the task of co-saliency detection due to its unsupervised nature. We believe the concepts of fusion-based and deep-learning-based approaches can well complement each other if we can design a unified method such that their particular advantages can be transferred to help each other.

We confront this challenge by proposing a two-stage approach for robust co-saliency detection.

Figure 5.1: Two examples of co-saliency detection. (a) Input images of the category car from the MSRC data set (**left**) [23] and the Cosal2015 data set (**right**) [24]. (b) $\sim$ (d) Three saliency proposals generated by GP [25], MST [26], and SVFSal [27] respectively. (e) Results generated by the map-wise proposal fusion method SACS [3]. (f) Results generated by the region-wise proposal fusion method CSSCF [28]. (g) Results generated by the proposed SAEF. (h) Refined results by the proposed STCNN.

At the first stage, we develop an unsupervised deep learning model, called stacked autoencoder-enabled fusion (SAEF), to evaluate and fuse multiple saliency proposals. The idea behind SAEF is simple: A saliency proposal for an image is considered good if its foreground can be well reconstructed by using object-like regions of other images while its background cannot. Specifically, SAEF learns a stacked autoencoder to reconstruct the object-like regions of an image, and apply the learned autoencoders across images to estimate not only foreground consistency but also foreground-background distinctiveness. In addition to image-level proposal evaluation, SAEF achieves better fusion by further exploring the complementary, co-saliency likelihood for region-level proposal evaluation. In brief, SAEF resolves the limitations of fusion-based and deep-learning-based methods. As an unsupervised model, it does not require supervisory data for training. It evaluates the quality of saliency proposals via discriminative reconstruction, and does not suffer from the difficulties caused by substantial variations or unreliable proposals.

Saliency maps generated by fusing multiple proposals are prone to be over-smoothed, and may inherit noise from proposals. At the second stage, we design self-trained convolutional neural networks (STCNN) to address the two issues. STCNN refines the saliency maps produced by SAEF. It propagates information from high-confidence regions to low-confidence ones in an iterative fashion while keeping the refined saliency maps as similar to the original ones as possible. The refined

saliency maps preserve object boundaries, and hence are sharper with their noise removed.

Figure 5.1 shows two examples of co-saliency detection. The input images of the category car are displayed in Figure 5.1(a). Three saliency proposals by using GP [25], MST [26], and SVFSal [30] are given in Figure 5.1(b) $\sim$ (d), respectively. The proposals by SVFSal are of higher quality but are not consistent with the other two proposals. Figure 5.1(e) and (f) show the co-saliency maps detected by map-wise [3] and region-wise [28] proposal fusion, respectively. The two fusion-based methods work based on majority consensus, and fail to assign a higher weight to the better proposal by SVFSal. Thus, their results in Figure 5.1(e) and (f) are not satisfactory, and are even worse than the proposal by SVFSal. The proposed SAEF performs discriminative reconstruction for proposal evaluation. It derives a more plausible combination of the proposals, yielding much better results in Figure 5.1(g). The proposed STCNN refines the saliency maps by using self-paced learning. As illustrated in Figure 5.1(h), the refined saliency maps homogeneously highlight the whole objects and the background noise is greatly suppressed.

The main message of this work is two-fold. First, we propose stacked autoencoder-enabled fusion (SAEF) to tackle the limitations of fusion-based and learning-based co-saliency detection. SAEF carries out discriminative reconstruction for reliably measuring the quality of saliency proposals in an unsupervised manner. Thus, it does not need manual annotations for training data and does not suffer from the problems caused by proposals with substantial variations. Second, the proposed STCNNs refine saliency maps by propagating information in a self-taught fashion. The resultant saliency maps adhere to object boundaries. Hence, they are sharper and can highlight the whole salient objects. The proposed method is evaluated on three representative and large-scale benchmarks, including the MSRC [23], iCoseg [34], and Cosal2015 [24] data sets. The results show that our method performs favorably against the state-of-the-arts.

5.2 Related Work

This section reviews a few research topics relevant to our method.

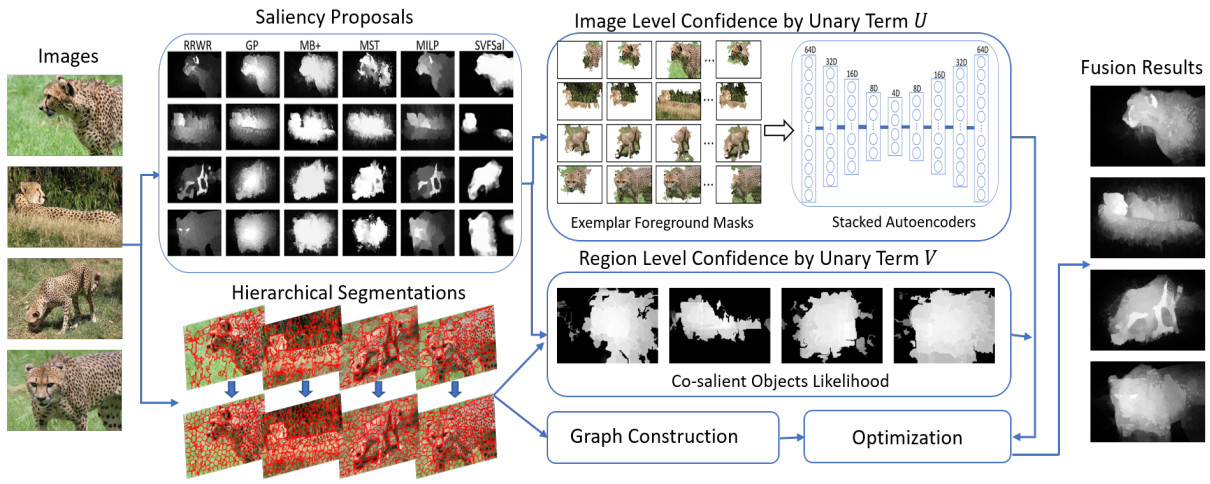


Figure 5.2: Overview of the proposed SAEF. SAEF collects multiple saliency proposals, extract superpixels, and construct a graph structure for the input images. It formulates co-saliency detection as an optimization problem with an objective function considering both image- and region-level confidence. After completing the optimization, saliency maps are produced.

5.2.1 Saliency Detection

Conventional methods for single-image saliency detection, e.g. [11, 25, 26, 29, 89–92], aim to distinguish salient objects in an image from the background based on various low-level features and assumptions. Despite the efficiency, the performance of these methods is limited due to the unsupervised nature. To address this issue, some supervised methods [27, 67, 84–87] are delivered. They use machine learning methodologies to better accomplish salient object detection. However, these methods rely on supervisory data annotations, which are costly and not available in general for saliency detection.

5.2.2 Co-saliency Detection

Co-saliency detection is a weakly supervised extension to saliency detection. It leverages not only intra-image appearance evidence but also inter-image co-occurrence to locate common salient objects. Different strategies have been proposed for this task. *Bottom-up* methods utilize contrast hypothesis and different prior knowledge by either handcrafted features [2, 3, 9, 14, 46, 51, 79, 80] or learned features [16, 24] to catch intra-image saliency as well as inter-image consistency. To

further improve the performance, *fusion-based* methods merge several saliency models to exclude individual prediction bias while retaining the shared information. To this end, methods of this category fuse the saliency proposals generated by different models via fixed weight fusion [2], adaptive weight fusion [3], or region-wise adaptive fusion [14, 28, 51]. Fusion-based methods typically work based on the assumption that plausible proposals are those sharing higher similarity with other proposals. Their performance drops when the assumption does not hold: the adopted saliency proposals have common prediction errors or large variations. *Deep-learning-based* methods [13, 81, 82, 88] are effective in distilling semantic object information in complex scenes, and have greatly enhanced co-saliency detection. However, these methods work in a supervised manner and require either a pre-trained deep model or labeled training data. Furthermore, the supervised setting also reduces their generalizability of handling objects of unseen categories.

Our `SAEF` tackles the cross-cutting issues of fusion-based and deep-learning-based methods. SAEF employs an unsupervised deep model to estimate the quality of each saliency proposal via auto-encoding both foreground consistency and foreground-background separation. It can more accurately identify the plausible proposals, and does not suffer from the unfavorable effects caused by fusion using majority voting. Besides, it does not rely on annotated training data and can detect salient objects of unseen categories.

5.2.3 Self-paced Learning

Kumar *et al*. [93] proposed self-paced learning (SPL) to imitate humans' learning behavior, namely starting to learn easier parts of a task and gradually considering more complex parts. Specifically, SPL associates each data sample with a weight. A self-spaced regularizer is attached to determine each weight value. Through sequential optimization, gradually increasing penalty on the regularizer includes more samples from easy to complex in training in a self-paced way. SPL has been widely used in various applications, such as matrix factorization [94], multimedia search [95], object tracking [96], image deblurring [97], action understanding [98], and co-saliency detection [16].

The method by Zhang *et al*. [16] is the most relevant to ours. Their method also adopts self-

paced learning for co-saliency detection. However, their method treats feature extraction and co-saliency detection as separate steps, leading to suboptimal performance. In contrast, our proposed SPL module STCNN is built on CNNs so that CNNs can jointly learn the relevant features and refine co-saliency detection in a self-paced fashion. The quality of the resultant saliency maps is hence greatly improved.

5.3 SAEF for Proposal Fusion

The proposed approach is composed of two components, i.e. stacked autoencoder-enabled fusion (SAEF) and self-trained convolutional neural networks (STCNN). The former fuses candidate saliency proposals and generates plausible saliency maps with the aid of unsupervised deep learning. The latter takes the saliency maps produced by SAEF as pseudo ground truth, and implements self-paced learning for saliency map refinement. The two components are described in the following two sections, respectively. We give the formulation, design the objective function, and the optimization of SAEF in the rest of this section. The flowchart of SAEF is shown in the Figure 5.2.

5.3.1 Problem Formulation

Given a set of N images $\mathcal{I} = \{I_n\}_{n=1}^{N}$ covering salient objects of the same category, we aim at detecting the salient objects in $\mathcal{I}$. As a fusion-based method, SAEF applies M existing saliency detection models, including [25,26,29,30,89,92] in this work, to $\mathcal{I}$, and gets M saliency proposals $\{S_{n,m}\}_{m=1}^{M}$ for each image I_n. To abstract unnecessary details and extract the intrinsic structures at different scales, we hierarchically decompose each image I_n into K_n *segments* and T_n *superpixels*. Specifically, we derive initial coarse-level segments based on the algorithm in [99], and then group pixels into fine-level superpixels that can adhere to the boundary of the segments at the coarse level. In our experiments, we set the number of superpixels in each image to 200 and the number of pixels within each segment to be greater than 200. It follows that set $\mathcal{I}$ contains $K = \sum_n K_n$ segments and $T = \sum_n T_n$ superpixels in total. For proposal fusion, SAEF optimizes plausible weights $Y = [\mathbf{y}_1 \ \cdots \ \mathbf{y}_i \ \cdots \mathbf{y}_T] \in [0,1]^{M \times T}$, where vector $\mathbf{y}_i = [y_{i,1} \ y_{i,2} \ \ldots \ y_{i,M}]^\top \in [0,1]^M$ corresponds to superpixel i, to fuse the M saliency proposals in the domain of superpixels.

SAEF formulates the task of optimizing Y as an energy minimization problem over a graph $\mathcal{G} = (\mathcal{V} = \cup\, \mathcal{V}_n, \mathcal{E} = \cup\, \mathcal{E}_n)$, which encodes the spatial relationships among superpixels. Set $\mathcal{V}_n$ contains T_n nodes, one for each superpixel in image I_n. Edge e_{ij} is added to $\mathcal{E}_n$ for linking nodes v_i and v_j if superpixels i and j are spatially connected in image I_n. Edge e_{ij} is associated with a weight $a_{ij} = \exp\left(-\|\mathbf{v}_i - \mathbf{v}_j\|^2\right)$, where $\mathbf{v}_i$ and $\mathbf{v}_j$ are the deep features of superpixels i and j, respectively. How $\mathbf{v}_i$ and $\mathbf{v}_j$ are extracted will be given later. *Graph Laplacian* $L \in \mathbb{R}^{T\times T}$ of $\mathcal{G}$ is then computed based on the affinity matrix $A = [a_{ij}] \in \mathbb{R}^{T\times T}$.

Before designing the objective function for optimizing Y, we investigate the potential foreground areas of each image I_n. To this end, B object proposals $\{f_{n,b}\}_{b=1}^{B}$ are generated by applying the scheme in [100] to I_n, where B is set to 350 here. To further explore the object mask corresponding to each proposal $f_{n,b}$, we consider superpixel v_i belongs to $f_{n,b}$ if 1) it is fully covered by $f_{n,b}$ or 2) it is partially covered by $f_{n,b}$ and the area ratio $|v_i \cap f_{n,b}|/|f_{n,b}|$ is larger than $|f_{n,b}|/|I_n|$. The corresponding mask of $f_{n,b}$ is defined to be composed of all the superpixels belonging to $f_{n,b}$. The feature representation of the mask is denoted by $\mathbf{f}_{n,b}$ and is yielded by max-pooling the feature vectors of all superpixels it covers. The procedure is repeated for each object proposal. The collected foreground masks of image I_n are $\mathcal{F}_n = \{\mathbf{f}_{n,b}\}_{b=1}^{B}$, which represent our initial estimation of the salient object in I_n.

5.3.2 Objective Function

SAEF seeks the optimal weights $Y = [\mathbf{y}_1 \ \cdots \ \mathbf{y}_T] \in \mathbb{R}^{M\times T}$ for superpixel-wise saliency proposal fusion by minimizing the following objective function defined over $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

$$\begin{aligned} &\min_{Y} \sum_{i:v_i\in\mathcal{V}} \left(U(\mathbf{y}_i) + \lambda_1 V(\mathbf{y}_i)\right) + \lambda_2 \sum_{e_{ij}\in\mathcal{E}} B(\mathbf{y}_i, \mathbf{y}_j) + \lambda_3 \|Y\|_2^2 \\ &\text{s.t.} \quad \|\mathbf{y}_i\|_1 = 1, \mathbf{y}_i \geq \mathbf{0}, \text{ for } 1 \leq i \leq T, \end{aligned} \tag{5.1}$$

where $\mathbf{0}$ is an all-zero vector, λ_1, λ_2 and λ_3 are three positive constants. The unary term $U(\mathbf{y}_i)$ is the primary element in SAEF. It refers to the proposals' reconstruction errors from SAE (stacked

autoencoder) and image-wisely determines the quality of each proposal. The auxiliary unary term $V(\mathbf{y}_i)$ takes the *co-salient object likelihood* into account and can superpixel-wisely refine the weights for fusion. Pairwise term $B(\mathbf{y}_i, \mathbf{y}_j)$ encourages the spatial smoothness of the derived weights. Lastly, $||Y||_2^2$ is a regularization term. The terms $U(\mathbf{y}_i)$, $V(\mathbf{y}_i)$, and $B(\mathbf{y}_i, \mathbf{y}_j)$ are detailed as follows.

5.3.2.1 On Designing Unary Term $U(\mathbf{y}_i)$

This term evaluates the quality of each saliency proposal for the image covering superpixel i based on a stacked autoencoder (SAE) [101] representation, encoding both foreground consistency and foreground-background distinctiveness, to determine a plausible weight vector $\mathbf{y}_i$ for fusion.

Recall that we collect B potential object masks for each image I_n, extract and denote their features by $\mathcal{F}_n = \{\mathbf{f}_{n,b}\}_{b=1}^{B}$. For I_n, we learn an SAE H_{θ_n} by minimizing the cross-entropy between the inputs in $\mathcal{F}_n$ and the reconstructed outputs, where θ_n is the learned parameter set. In this way, this SAE can reconstruct the estimated foreground masks of I_n. The procedure is repeated for every image. A total of N SAEs $\{H_{\theta_n}\}_{n=1}^{N}$ are obtained.

Considering image I_n and its mth proposal $S_{n,m}$, I_n can be partitioned into the foreground and the background sub-images by Otsu's thresholding, denoted as $I_{n,m}^f$ and $I_{n,m}^b$. We use the same way to represent the sub-image $I_{n,m}^f$. Namely, the feature representation $\mathbf{x}_{n,m}^f$ of $I_{n,m}^f$ is yielded by max-pooling the feature vectors of the superpixels belonging to $I_{n,m}^f$. The feature representation $\mathbf{x}_{n,m}^b$ of $I_{n,m}^b$ is obtained similarly.

Assume that the mth proposal for image I_n is of high quality. The reconstruction error by applying SAE $H_{\theta_{n'}}$ to the detected foreground of I_n, i.e. $\|\mathbf{x}_{n,m}^f - H_{\theta_{n'}}(\mathbf{x}_{n,m}^f)\|$, is expected to be low since images I_n and $I_{n'}$ have common foreground objects. In addition, the reconstruction error $\|\mathbf{x}_{n,m}^b - H_{\theta_{n'}}(\mathbf{x}_{n,m}^b)\|$ is probably high when we feed SAE $H_{\theta_{n'}}$ with the estimated background of I_n. To jointly consider inter-image foreground similarity and foreground-background distinctiveness, we compute the ratio between the foreground and background reconstruction errors

$$\hat{g}_{n,m} = \frac{\sum_{n'=1}^{N} \|\mathbf{x}_{n,m}^b - H_{\theta_{n'}}(\mathbf{x}_{n,m}^b)\|^2}{\sum_{n'=1}^{N} \|\mathbf{x}_{n,m}^f - H_{\theta_{n'}}(\mathbf{x}_{n,m}^f)\|^2}. \tag{5.2}$$

To take other proposals into account, the image-level score $g_{n,m}$ of the mth proposal for image I_n is calculated by

$$g_{n,m} = \frac{\exp(\hat{g}_{n,m})}{\sum_{m'=1}^{M} \exp(\hat{g}_{n,m'})}. \tag{5.3}$$

A penalty variable $r_{i,m} = (1 - g_{n,m})$ is introduced if superpixel i belongs to image I_n. The first term in Eq. (5.1) is defined by

$$\sum_{i:v_i \in \mathcal{V}} U(\mathbf{y}_i) = \sum_{i=1}^{T} \mathbf{r}_i^\top \mathbf{y}_i = tr(R^\top Y), \tag{5.4}$$

where $\mathbf{r}_i = [r_{i,1} \ \cdots \ r_{i,M}]^\top$ and $R = [\mathbf{r}_1 \ \cdots \ \mathbf{r}_T]$.

5.3.2.2 On Designing Unary Term $V(\mathbf{y}_i)$

This term refines the fusion weights $\mathbf{y}_i$ on superpixel i locally. It is designed based on the formula of co-salient object likelihood

$$\textit{Co-saliency} = \textit{Saliency} \times \textit{Correspondence}. \tag{5.5}$$

Suppose superpixel i belongs to image I_n. For the *saliency* part, we transfer the objectness score $\psi_{n,b}$ suggested by [100] from every object mask $\mathbf{f}_{n,b}$ covering superpixel i to superpixel i, i.e.

$$O(v_i) = \sum_{b=1}^{B} \psi_{n,b} \delta(v_i \in \mathbf{f}_{n,b}), \tag{5.6}$$

where δ is the indicator function.

We also explore the location information by using the functional properties of coarse-level segments. Unlike fine-level superpixels that have grid alike structure, coarse-level segments adhere better to the image content variation; and are usually long boundary connected in the background area while having smaller fragmented regions on the object area. Since segments near the image center, Ctr_n, more likely belong to the foreground while those overlapping with the set of the image boundary pixels, Bou_n, tend to be covered by background. Suppose that superpixel i is

covered by the kth segment u_k. The location prior of superpixel i is defined as

$$L(v_i) = \mathcal{N}(\|\texttt{cord}(u_k) - Ctr_n\|^2 \mid 0, \sigma^2) \times \exp\big(\frac{-2|u_k \cap Bou_n|}{\texttt{per}(u_k)}\big), \tag{5.7}$$

where $\texttt{cord}(u_k)$ and $\texttt{per}(u_k)$ are the center and the perimeter of segment u_k, respectively. $\mathcal{N}$ is the normal distribution with σ set to the geometric mean of the image width and height.

For $O(v_i)$ in Eq. (5.6) and $L(v_i)$ in Eq. (5.7), we linearly scale each of them to $[0, 1]$ by taking all superpixels in the same image into account. Then, the *saliency* score of superpixel i is yielded by averaging the corresponding scaled values.

For the *correspondence* part, we examine if there are strong correspondences of superpixel i in other images. To this end, we apply Otsu's thresholding method to divide the superpixels of each image I_n into foreground and background according to their *saliency* parts estimated above. Recall that all superpixels are represented by the deep features. A *Gaussian mixture model* (GMM) $\boldsymbol{\theta}_f$ with five components is fit to the foreground superpixels of all images. Meanwhile, a five-component GMM $\boldsymbol{\theta}_{b,n}$ is fit to the background superpixels of I_n, for $n = 1, 2, ..., N$. The *correspondence* score of superpixel i is defined as

$$C(v_i) = \frac{p(\mathbf{v}_i|\boldsymbol{\theta}_f)}{p(\mathbf{v}_i|\boldsymbol{\theta}_f) + \sum_{n=1}^{N} p(\mathbf{v}_i|\boldsymbol{\theta}_{b,n})\delta(v_i \in I_n)}, \tag{5.8}$$

where $\mathbf{v}_i$ is the deep features of superpixel i, $p(\mathbf{v}_i|\boldsymbol{\theta}_f)$ and $p(\mathbf{v}_i|\boldsymbol{\theta}_{b,n})$ are the probabilities estimated by GMMs $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_{b,n}$, respectively.

$C(v_i)$ is also linearly scaled to $[0, 1]$ and then multiplied by the *saliency* score to compute Eq. (5.5) as the co-saliency prior $CS(v_i)$ of superpixel i. Let $s_{i,m}$ be the mean saliency value of the mth saliency proposal on superpixel i. We prefer a saliency proposal consistent with the co-saliency prior. Let ϕ_n be the Otsu's threshold over the co-saliency prior of all superpixels in I_n. The score of saliency proposal m on superpixel i is defined as

$$l_{i,m} = \frac{\exp(-\|\delta(CS(v_i) \geq \phi_n) - s_{i,m}\|^2)}{\sum_{m'=1}^{M} \exp(-\|\delta(CS(v_i) \geq \phi_n) - s_{i,m'}\|^2)}. \tag{5.9}$$

With $\mathbf{l}_i = [l_{i,1} \cdots l_{i,M}]^\top$, the second term in Eq. (5.1) is set to

$$\sum_{i:v_i \in \mathcal{V}} V(\mathbf{y}_i) = \sum_{i=1}^{T} (1 - \mathbf{l}_i)^\top \mathbf{y}_i. \tag{5.10}$$

5.3.2.3 On Designing Pairwise Term $B(\mathbf{y}_i, \mathbf{y}_j)$

This pairwise term is added to encourage the spatial smoothness of Y on $\mathcal{G}$:

$$\sum_{e_{ij} \in \mathcal{E}} B(\mathbf{y}_i, \mathbf{y}_j) = \sum_{e_{ij} \in \mathcal{E}} a_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 = tr(YLY^\top), \tag{5.11}$$

where a_{ij} is the weight of e_{ij} and L is the graph Laplacian of $\mathcal{G}$.

5.3.3 Optimization and Implementation Details

With the terms $U(\mathbf{y}_i)$ in Eq. (5.4), $V(\mathbf{y}_i)$ in Eq. (5.10), and $B(\mathbf{y}_i, \mathbf{y}_j)$ in Eq. (5.11), the constrained optimization problem in Eq. (5.1) can be solved by *quadratic programming* (QP). We optimize it with the CVX solver [44], and get the weights for fusion Y^*. The saliency maps $\{\hat{S}_n\}_{n=1}^N$ for images $\{I_n\}_{n=1}^N$ are then produced.

In our implementation of the stacked autoencoders (SAE), we adopt the 5-layer network architecture used in [101], but reduce the numbers of neurons in the five layers to 64, 32, 16, 8, and 4 respectively due to the limited training data. For each image, its features are generated by applying ResNet50 [102] to it. Specifically, we up-sample and concatenate the feature maps in layers conv1_relu, res2c_relu, res3d_relu, res4f_relu, and res5c_relu to yield the 3904-d *hypercolumn* representation. The feature vector of each superpixel is calculated by max-pooling over the region it covers.

5.4 `STCNN` for Saliency Map Refinement

We introduce self-trained CNNs (`STCNN`) for saliency map refinement in this section.

5.4.1 Problem Formulation

The saliency maps produced by SAEF are prone to being over-smoothed and may contain inherited noise from proposals. STCNN addresses these issues by introducing self-paced learning. It propagates information from high-confidence regions to low-confidence ones, and progressively refines the saliency maps.

STCNN is a CNN-based model with two network streams, f_g and f_l. Both take an image as input and predict its saliency map. While f_g approximates the saliency maps that SAEF produces as the pseudo ground truth, f_l carries out self-paced learning for iterative saliency map refinement. The objective for training STCNN is

$$\ell(\mathbf{w}_g, \mathbf{w}_l; \mathcal{I}) = \ell_g(\mathbf{w}_g; \mathcal{I}) + \ell_l(\mathbf{w}_l; \mathcal{I}), \tag{5.12}$$

where loss functions ℓ_g and ℓ_l guide the training of f_g and f_l respectively, and will be detailed later. Sets $\mathbf{w}_g$ and $\mathbf{w}_l$ cover the learnable parameters of f_g and f_l, respectively. After optimizing Eq. (5.12), the refined saliency map S_n of image I_n is produced via $S_n = f_g(I_n; \mathbf{w}_g) \times f_l(I_n; \mathbf{w}_l) = S_n^g \times S_n^l$.

5.4.1.1 On Designing Loss ℓ_g

This term aims to detect the common salient objects by approximating the saliency maps $\{\hat{S}_n\}$, treated as the pseudo ground truth, generated by SAEF, and ℓ_g is defined as

$$\ell_g(\mathbf{w}_g; \mathcal{I}) = \sum_{n=1}^{N} \sum_{p \in I_n} Q_n(p) |S_n^g(p) - \hat{S}_n(p)|^2, \tag{5.13}$$

where p is the index of the pixels in I_n and $S_n^g = f_g(I_n; \mathbf{w}_g)$ is the saliency map generated by f_g. $S_n^g(p)$ and $\hat{S}_n(p)$ are the saliency values of S_n^g and $\hat{S}_n$ at pixel p, respectively. $Q_n(p)$ indicates the importance of pixel p. We partition the pixels in $\hat{S}_n$ into two categories, salient and non-salient, by using the mean value of $\hat{S}_n$ as the threshold. $Q_n(p)$ is introduced to address the potential size unbalance between the two categories. Let ρ be the ratio between salient pixels and all pixels.

$Q_n(p)$ is set to $1-\rho$ if pixel p is categorized as salient, and ρ otherwise. In this way, the pixels in the two categories contribute equally in Eq. (5.13).

The loss function in Eq. (5.13) is optimized by considering all images $\{I_n\}_{n=1}^N$ simultaneously. Compared with SAEF, f_g can better learn the visual properties shared among salient objects while excluding the individual backgrounds, to achieve better performance.

5.4.1.2 On Designing Loss ℓ_l

The term ℓ_l in Eq. (5.12) leverages self-paced learning (SPL) to iteratively identify and learn from high-confidence regions, and propagate the information to better predict low-confidence regions in saliency maps. It is defined as

$$\begin{aligned}\ell_l(\mathbf{w}_l,&\{\mathbf{M}_n,\mathbf{V}_n\}_{n=1}^N;\mathcal{I}) = \\ &\sum_{n=1}^{N}\sum_{p\in I_n}\mathbf{V}_n(p)|S_n^l(p)-\mathbf{M}_n(p)|^2-\gamma\mathbf{V}_n(p), \\ s.t.\ &\mathbf{V}_n(p)\in[0,1],\mathbf{M}_n(p)\in\{0,1\},\forall n,p,\end{aligned} \tag{5.14}$$

where $S_n^l = f_l(I_n;\mathbf{w}_l)$ and the constant γ controls the learning pace. For image I_n, the auxiliary variable $\mathbf{M}_n$ denotes the estimated co-saliency mask. Each pixel p is associated with a latent weight variable $\mathbf{V}_n(p)$ weighting the corresponding loss. The first term in Eq. (5.14) measures the consistency between the predicted saliency maps and the estimated masks while the second term favors selecting easy over complex samples (pixels here). Namely, a sample with less loss is considered *easy*, so it is learned with a higher priority and vice versa. In sum, minimizing ℓ_l in Eq. (5.14) decreases the weighted training loss together with the negative ℓ_1-norm regularizer.

Eq. (5.14) consists of three sets of optimization variables, $\mathbf{w}_l$, $\{\mathbf{M}_n\}_{n=1}^N$, and $\{\mathbf{V}_n\}_{n=1}^N$. Because directly optimizing Eq. (5.14) is difficult, we instead adopt an alternating iterative strategy to optimize $\mathbf{w}_l$, $\{\mathbf{M}_n\}_{n=1}^N$, and $\{\mathbf{V}_n\}_{n=1}^N$. At each iteration, one of the three variables is optimized while keeping the others fixed in an alternating fashion. The iterative procedure is repeated until convergence.

5.4.1.2.1 On Optimizing $\mathbf{w}_l$: We fix $\{\mathbf{M}_n, \mathbf{V}_n\}_{n=1}^N$. The optimization problem in Eq. (5.14) is reduced to

$$\ell_l(\mathbf{w}_l, \{\mathbf{M}_n, \mathbf{V}_n\}_{n=1}^N; \mathcal{I}) = \sum_{n=1}^{N} \sum_{p \in I_n} \mathbf{V}_n(p) |S_n^l(p) - \mathbf{M}_n(p)|^2. \tag{5.15}$$

Stochastic gradient descent (SGD) is adopted to optimize the parameters $\mathbf{w}_l$ of CNNs f_l.

5.4.1.2.2 On Optimizing $\{\mathbf{M}_n\}_{n=1}^N$: By fixing $\mathbf{w}_l$ and $\{\mathbf{V}_n\}_{n=1}^N$, the optimization problem in Eq. (5.14) becomes

$$\begin{aligned} &\ell_l(\mathbf{w}_l, \{\mathbf{M}_n, \mathbf{V}_n\}_{n=1}^N; \mathcal{I}) = \sum_{n=1}^{N} \sum_{p \in I_n} \mathbf{V}_n(p) |S_n^l(p) - \mathbf{M}_n(p)|^2, \\ &s.t.\ \mathbf{M}_n(p) \in \{0, 1\}, \forall n, p. \end{aligned} \tag{5.16}$$

It is obvious that the optimal $\mathbf{M}_n(p)$ takes value 0 if $S_n^l(p) \leq 0.5$, and 1 otherwise.

5.4.1.2.3 On Optimizing $\{\mathbf{V}_n\}_{n=1}^N$: When fixing $\mathbf{w}_l$ and $\{\mathbf{M}_n\}_{n=1}^N$, as shown in [93], the global optimum $\{\mathbf{V}_n\}_{n=1}^N$ can be obtained via

$$\mathbf{V}_n(p) = \begin{cases} 1, & \text{if } |S_n^l(p) - \mathbf{M}_n(p)|^2 < \gamma, \\ 0, & \text{otherwise.} \end{cases} \tag{5.17}$$

5.4.2 Optimization and Implementation Details

Consider the optimization of `STCNN` in Eq. (5.12). We first optimize Eq. (5.13) with backward propagation to obtain optimum $\mathbf{w}_g^*$ for f_g, and then optimum $\mathbf{w}_l^*$ for f_l is obtained by optimizing Eq. (5.14) iteratively. Prior to running alternating optimization, we initialize $\{\mathbf{M}_n, \mathbf{V}_n\}$ with saliency maps $\{\hat{S}_n\}$ `SAEF` produces as follows

$$\mathbf{V}_n(p) = \begin{cases} 1, & \text{if } \hat{S}_n(p) > \mu_n + \sigma_n, \\ 1, & \text{if } \hat{S}_n(p) < \mu_n - \frac{\sigma_n}{4}, \\ 0, & \text{otherwise,} \end{cases} \text{ and } \mathbf{M}_n(p) = \begin{cases} 1, & \text{if } \hat{S}_n(p) > \mu_n + \sigma_n, \\ 0, & \text{if } \hat{S}_n(p) < \mu_n - \frac{\sigma_n}{4}, \\ \times, & \text{otherwise,} \end{cases}$$

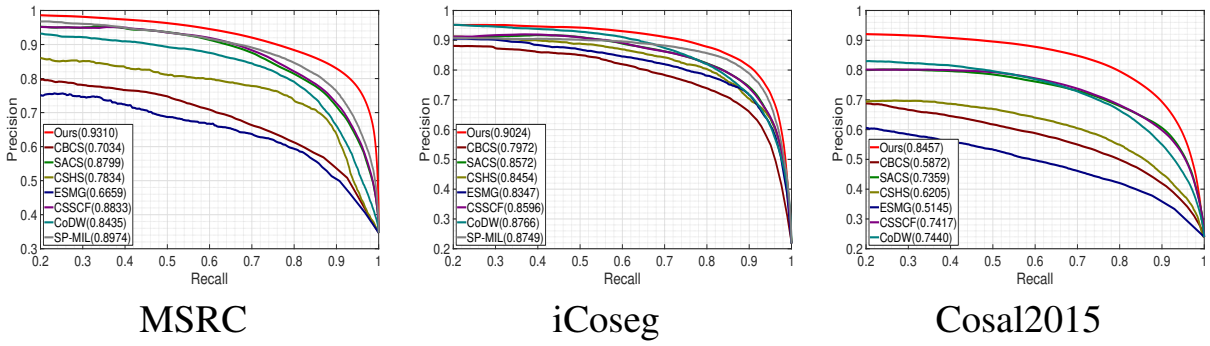
Algorithm 2 Optimization Procedure

Input: Image set, $\{I^n\}_{n=1}^N$; Saliency proposals, $\{S_{j,m}\}_{m,n=1}^{m=M,n=n}$, Max epochs, T;
$\{Y_{n,m}^*\}$ are generated via the optimization of Eq. (5.1);
$\{\hat{S}_n\}$ are generated with the optimum fusion weights $\{Y_{n,m}^*\}$;
$\mathbf{w}_g^*$ are generated via optimizing Eq. (5.13) with backward propagation;
Initialize $\mathbf{V}$ and $\mathbf{M}$;
for $i \leftarrow 1, 2, \ldots, T$ **do**
  $\mathbf{w}_l \leftarrow \mathbf{w}_l^*$, where $\mathbf{w}_l^*$ is optimized via
  Update $\{S_n^l = f_l(I_n; \mathbf{w}_l)\}$ with forward propagation;
  Update $\{\mathbf{M}_n\}$ via binarizing $\{S_n^l\}$ with a threshold 0.5;
  Update $\{\mathbf{V}_n\}$ via Eq. (5.17);
end for
$\{S_n^g = f_g(I_n; \mathbf{w}_g^*)\}$ are generated via the optimum $\mathbf{w}_g^*$;
$\{S_n^l = f_l(I_n; \mathbf{w}_l^*)\}$ are generated via the optimum $\mathbf{w}_l^*$;
S_n are obtained via $S_n = S_n^g \times S_n^l$ for each image;
Post-precess $\{S_n\}$ via DenseCRFs;
Output: Co-saliency maps, $\{S_n\}$

where $\times$ denotes *don't-care*. μ_n and σ_n are the mean and standard deviation of the saliency values in $\hat{S}_n$.

Pixel p with $V_n(p) = 1$ represents that it can be confidently assigned to either the salient regions ($\mathbf{M}_n(p) = 1$) or the background ($\mathbf{M}_n(p) = 0$). It is taken into account at the current epoch. Others with $V_n(p) = 0$ are ambiguous, so they are currently ignored. ADAM [103] is chosen as the optimization solver for its rapid convergence. In practice, for each image I_n at each epoch, the mask $\mathbf{M}_n$ and the latent varialbes $\mathbf{V}_n$ are updated only when $\mathbf{w}_l$ is stable enough, namely the squared error between the predicted saliency map and the estimated mask less than 0.1^2 in our cases. The maximum number of epochs is set to 60. The gradient derivation with respect to the optimization variables is straightforward, so it is omitted here. With $\mathbf{w}_g^*$ and $\mathbf{w}_l^*$, the refined saliency map S_n by `STCNN` is then calculated by $S_n = S_n^g \times S_n^l = f_g(I_n; \mathbf{w}_g^*) \times f_l(I_n; \mathbf{w}_l^*)$. The whole optimization is summarized in Algorithm 2.

`STCNN` is implemented using `MatConvNet` [104]. The same network architecture, i.e. VGG-16 [105] setting of FCN [106], is adopted for both network streams, f_g and f_l. We replace the activation function *softmax* in the last layer with the *sigmoid* function. The learning rate is fixed as



MSRC iCoseg Cosal2015

Figure 5.3: The performance of various methods in PR curves on different data sets. The numbers in parentheses denote AP.

10^{-5}. The weight decay, momentum, and batch size are set to 0.0005, 0.9, 5, respectively.

5.4.2.1 Post-processing by DenseCRFs

Spatial coherence and object boundary preservation of the saliency maps generated by `STCNN` can be further enhanced. Following the previous work [67, 86], DenseCRFs [31] is adopted to post-process each saliency map. In our cases, the unary and the pairwise terms in DenseCRFs are set to S_n and bilateral filtering, respectively. Please refer to Li and Yu's paper [67, 86] for the definitions of the two terms in detail. After post-processing, the inferred posterior probabilities of being salient yield the final saliency map. In this work, the public DenseCRFs code implemented by Li and Yu [67] is used.

5.5 Experimental Results

In this section, we first describe the data sets and evaluation metrics. Then, we compare our method with state-of-the-art methods, and investigate contributions of individual components by conducting ablation studies.

5.5.1 Data Sets

We evaluated the proposed approach on three most widely used public benchmark data sets for co-saliency detection: *iCoseg* [34], *MSRC* [23], and *Cosal2015* [24]. *iCoseg* consists of 38 groups of total 643 images. The images of *iCoseg* contain single or multiple similar objects in

Method	Year	Setting	MSRC			iCoseg			Cosal2015		
			AP	F_β	S_α	AP	F_β	S_α	AP	F_β	S_α
DSS [86]	CVPR2017	SI+S	0.8700	0.8313	0.7435	0.8802	0.8386	0.8483	0.7745	0.7510	0.7582
UCF [27]	ICCV2017	SI+S	0.9217	0.8114	0.8175	0.9292	0.8261	0.8754	0.8080	0.7197	0.7797
Amulet [87]	ICCV2017	SI+S	0.9219	0.8159	0.8162	0.9395	0.8381	0.8937	0.8201	0.7387	0.7863
MSC-NET [88]	MM2017	SI+S	0.9035	0.8419	0.7673	0.8845	0.8378	0.8518	0.8328	0.7683	0.7994
DIM [13]	TNNLS2016	CS+S	-	-	-	0.8773	0.7918	0.7583	-	-	-
UMLBF [82]	TCSVT2017	CS+S	0.9160	0.8410	-	-	-	-	0.8210	0.7120	-
RRWR [89]	CVPR2015	SI+US	0.8127	0.7534	0.6653	0.7986	0.7784	0.7022	0.6647	0.6636	0.6628
GP [25]	ICCV2015	SI+US	0.8200	0.7422	0.6844	0.7821	0.7495	0.7198	0.6851	0.6580	0.6721
MB+ [92]	ICCV2015	SI+US	0.8367	0.7817	0.7200	0.7868	0.7706	0.7272	0.6715	0.6693	0.6732
MST [26]	CVPR2016	SI+US	0.8057	0.7491	0.6460	0.8019	0.7659	0.7292	0.7099	0.6672	0.6681
MILP [29]	TIP2017	SI+US	0.8334	0.7776	0.6871	0.8182	0.7883	0.7514	0.6802	0.6737	0.6757
SVFSal [30]	ICCV2017	SI+US	0.8669	0.7934	0.7688	0.8376	0.8056	0.8271	0.7467	0.7123	0.7607
CBCS [9]	TIP2013	CS+US	0.7034	0.5910	0.4801	0.7972	0.7408	0.6580	0.5872	0.5583	0.5444
SACS [3]	TIP2014	CS+US	0.8799	0.8027	0.7341	0.8572	0.8048	0.7783	0.7359	0.7089	0.7170
CSHS [46]	SPL2014	CS+US	0.7834	0.7118	0.6661	0.8454	0.7549	0.7502	0.6205	0.6186	0.5918
ESMG [80]	SPL2015	CS+US	0.6659	0.6245	0.5804	0.8347	0.7766	0.7677	0.5145	0.5120	0.5454
CSSCF [28]	TMM2016	CS+US	0.8833	0.8136	0.7626	0.8596	0.7929	0.7686	0.7417	0.6997	0.6950
CoDW [24]	IJCV2016	CS+US	0.8435	0.7724	0.7129	0.8766	0.7985	0.7500	0.7440	0.7048	0.6482
SP-MIL [16]	TPAMI2017	CS+US	0.8974	0.8029	0.7687	0.8749	0.8143	0.7715	-	-	-
MVSRC [83]	TIP2017	CS+US	0.8530	0.7840	-	0.8680	0.8100	-	-	-	-
SAEF	/	CS+US	0.8850	0.8110	0.7758	0.8561	0.7967	0.7808	0.7401	0.7052	0.7269
ours	/	CS+US	0.9310	0.8397	0.8062	0.9024	0.8452	0.8216	0.8457	0.7814	0.7703

Table 5.1: Quantitative comparison with 20 methods on three benchmark data sets. "SI" and "CS" denote the single-image saliency and co-saliency methods, respectively. "S" and "US" indicate the supervised and unsupervised methods, respectively. Among the "US" methods, the top three results are marked in red, green and blue, in the order. Our fusion method SAEF mostly outperforms the other two fusion methods SACS and CSSCF. With self-training CNNs (STCNN), the final result (ours) leads all the competing unsupervised methods in most cases and has comparable performance with the supervised approaches.

various poses and sizes with complex backgrounds. *MSRC* contains 7 groups of total 240 images. Compared to *iCoseg*, co-salient objects in *MSRC* exhibit less pose or viewing angle variation; however, it contains different colors and shapes. Thus, the *MSRC* appears to be almost equally difficult as the *iCoseg* data set. Lastly, *Cosal2015* is a more recent and the most challenging data set among three so far. It has 50 groups and a total of 2015 images containing significant poses and sizes, appearance variations and even more complex backgrounds.

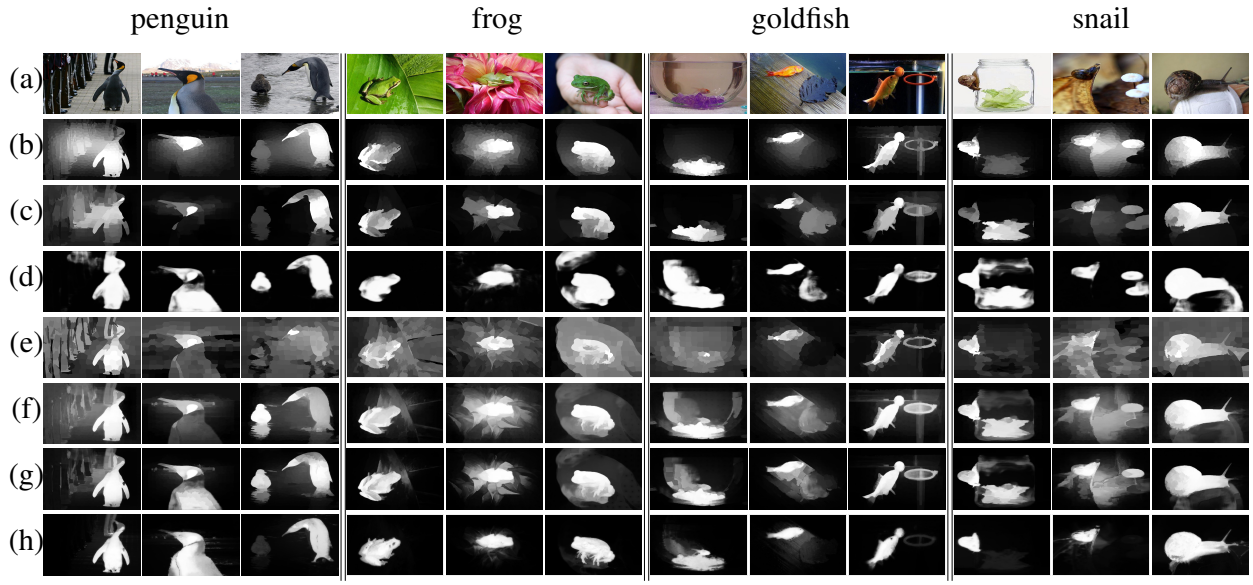


Figure 5.4: Visual comparison with state-of-the-art methods. (a) Images from four image groups of Cosal2015 data set for co-saliency detection. (b)∼(h) Saliency maps generated by different approaches, including (b) GP [25], (c) MILP [29], (d) SVFSal [30], (e) CoDW [24], (f) CSSCF [28], (g) SAEF, and (h) ours.

5.5.2 Evaluation Metrics

To evaluate the performance of co-saliency detection, we adopt two commonly used metrics: *average precision* (AP) and *F-measure* (F_β), as well as a newly proposed metric: *structure measure* (S_α) [107]. AP is the area under the Precision-Recall (PR) curve by comparing ground truth with the binary masks produced by varying the saliency map threshold continuously in the range of $[0, 1]$. Meanwhile, with a self-adaptive threshold $T = \mu + \sigma$, where μ and σ denote the mean and standard deviation of the saliency map respectively, *F_β-measure* is computed by the harmonic mean of the precision and recall values: $F_\beta = \frac{(1+\beta^2)\times precision \times recall}{\beta^2 \times precision + recall}$, with the imposed weight $\beta^2 = 0.3$ to emphasize more on recall as suggested in [11, 16, 24, 108]. In addition to the aforementioned pixel-based metrics, a region-based image quality measure, *structure measure* (S_α) [107], is adopted to evaluate the spatial structure similarity of saliency maps based on both region-aware structural similarity S_r and object-aware structural similarity S_o, defined as $S_\alpha = \alpha * S_r + (1 - \alpha) * S_o$, where $\alpha = 0.5$ following [107]. Specifically, to evaluate the region-

aware structural similarity measure, the full saliency map is first divided into K non-overlapping blocks. Then the region similarity of each block $ssim(k)$ is computed by comparing with the ground truth based on the product of three components: luminance comparison, contrast comparison, and structure comparison. For each component,the similarity measure is defined similarly as Pearson correlation [107]. With $ssim(k)$, a different weight w_k is assigned to each block based on the foreground region each block covers, and it is formulated as: $S_r = \sum_{k=1}^{K} w_k \times ssim(k)$. Meanwhile, the object-aware structural similarity is designed with respect to two characteristics: *sharp foreground-background contrast* and *uniform saliency distribution* by measuring the mean pixel values of the final saliency map in foreground ($\bar{x}_{FG}$) & background ($\bar{x}_{BG}$) regions and the corresponding standard deviation values of foreground ($\sigma_{x_{FG}}$) & background ($\sigma_{x_{BG}}$) regions (defined by the ground truth), respectively. Specifically, $S_o = (O_{FG} + O_{BG})/2$, $O_{FG} = \frac{2\bar{x}_{FG}}{(\bar{x}_{FG})^2+1+2\lambda\times\sigma_{x_{FG}}}$, and O_{BG} is similarly computed. This metric is proposed to alleviate the flaw of widely used pixel-based measures, for example, AP, AUC, F_β, or even the recently introduced generalized F-measure F_β^w [77], as they observe that any foreground map that more preserves the entire object structure can help the machine to learn a more complete object information, more importantly, it can order the performance more consistently to both the application ranking and their user studies.

5.5.3 Comparison with the State-of-the-Art Methods

To have a thorough comparison with state-of-the-art methods, we divide them into four groups, i.e. the unsupervised saliency [25, 26, 29, 30, 89, 92] and co-saliency [3, 9, 16, 24, 28, 46, 80, 83] detection methods as well as supervised saliency [27, 86–88] and co-saliency [13, 82] detection methods. The overall performance statistics are summarized in Table 5.1 and Figure 5.3. Please note that all compared supervised single-image saliency detection methods are CNN-based. We reproduced the experimental results using the publicly available source code with default parameters provided by the authors. For methods without releasing source code, we either evaluated on their released results (`SP-MIL` [16], `CoDW` [24] and `DIM` [13]), or directly reported the numbers in their papers (`UMLBF` [82] and `MVSRC` [83]).

The precision-recall (PR) curves by our method and seven competing co-saliency detection

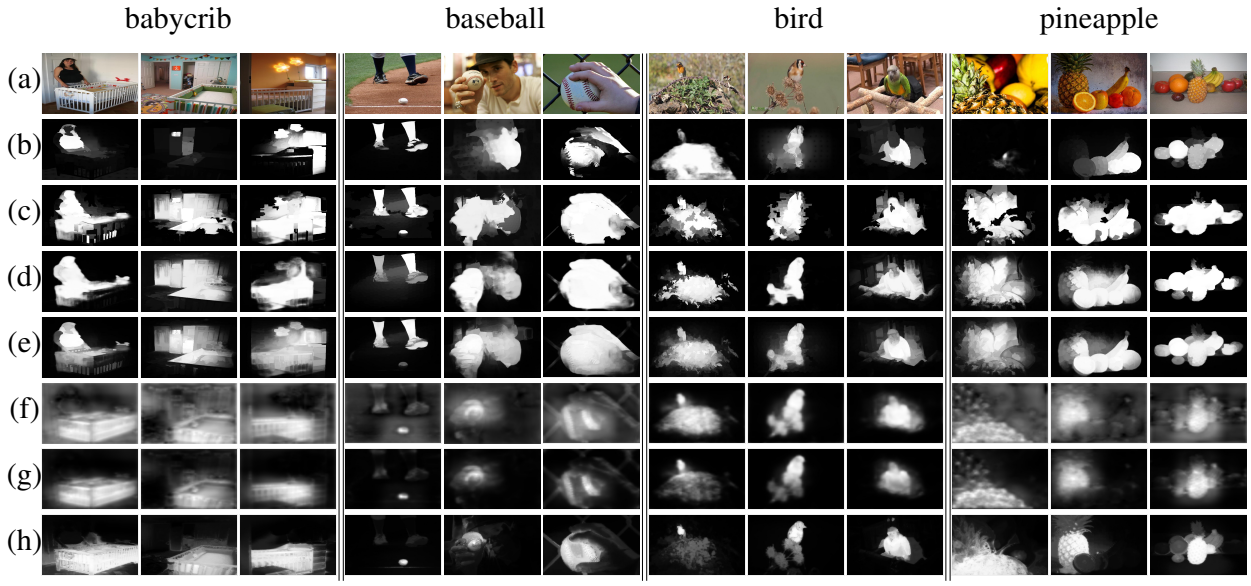


Figure 5.5: Visual illustration of ablation studies. (a) Images from four groups for co-saliency detection. (b)∼(h) Co-saliency maps generated by different combinations of components, including (b) U, (c) $U + V$, (d) $U + V + B$, (e) $U + V + B + Reg$ (SAEF), (f) SAEF+CNN, (g) SAEF+CNN+SPL, (h) SAEF+CNN+SPL+denseCRFs (ours), respectively.

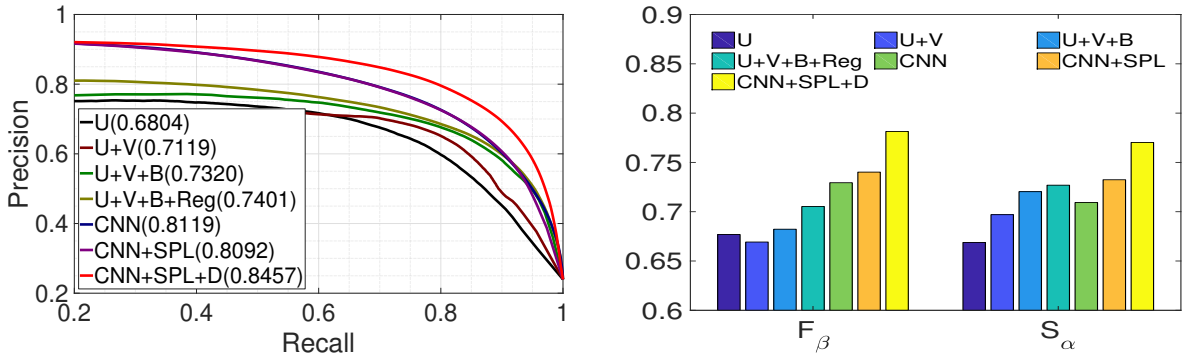
methods on three different data sets are shown in Figure 5.3. The overall quantitative result is reported in Table 5.1. With the same unsupervised setting, our method leads both the single-image saliency detection and co-saliency detection methods by a large margin. Moreover, by leveraging unsupervised deep learning and self-paced learning, our method even surpasses many supervised CNN-based single-image saliency methods that exploit object annotations. Last but not least, compared with the supervised co-saliency method DIM [13] that employs stack denoising autoencoder (SDAE) and UMLBF [82] that similarly applies adaptive feature learning for co-saliency detection, our method outperforms them without requiring expensive object annotations.

To gain insights into the quantitative results, Figure 5.4 shows example saliency maps on four groups from the most challenging co-saliency detection data set: Cosal2015 (please refer to APPENDIX A for more comprehensive visual comparison). Single-image saliency detection methods generally cannot give satisfactory results. For instance, methods GP and MILP relying on specific handcrafted cues inevitably produce many false positives in the first image of the Penguin class

and miss the majority of penguin's body in the second image. Furthermore, without jointly exploiting the common objects in multiple images, single-image saliency detection methods cannot exclude the visually salient objects that do not repetitively appear in other images. For instance, although the CNN-based single-image saliency detection method SVFSal can better delineate object boundaries, it often includes unrelated regions. As an example, the bird on the left-hand side of the third penguin image is wrongly taken as part of the co-salient object. Next, results of CoDW show that significant intra- and inter-object variations can sometimes mislead co-saliency detection and lead to results even worse than the single-image saliency detection methods. Though more relevant images bring more prosperous and shared information to explore in co-saliency detection, the problem is also more challenging as it needs to cope with potential variations across images. The fusion-based approach CSSCF deals with large inter-object variations by fusing the saliency proposals from the methods GP, MILP and SVFSal. It boosts the performance and surpasses the method CoDW. However, as mentioned above, it relies on the group consensus and can not discriminatively put more weight to the best saliency proposal. Our proposed SAEF generates better results than CSSCF by overcoming the inherent group biasing issue. Finally, our two-stage approach elegantly integrates a self-trained CNN guided by SAEF and gives sharper and more homogeneous saliency detection results by successfully filtering out the background noise and recovering the omissions.

5.5.4 Ablation Studies

Figure 5.6 reports ablation studies with different metrics to investigate contributions from individual energy terms of SAEF and from each component in the proposed STCNN network i.e. CNN (f_g), self-paced learning (SPL, f_l), and DenseCRFs (D). From AP and F_β scores, it is clear that the results improve progressively by adding individual energy terms, U, V, and B, into the objective function in Eq. (5.1). STCNN further boosts the detection results by combining CNN's object recognition capability with SPL dealing with the limited quantity of pseudo ground truth under the unsupervised learning setting. By integrating the merit of DenseCRFs, our method achieves superior results. The progressive improvement is not as evident for the metric S_α that measures the



(a) PR-curves and AP

(b) F_β and S_α

Figure 5.6: Ablation studies on Cosal2015 data set in terms of (a) the PR curves and AP (in parentheses), and (b) F_β and S_α. "D" represents the `denseCRFs` [31] postprocessing.

local structure similarity of the detected objects to the ground truth. The major reason is that some background regions badly predicted by SAEF will lead to the fuzzier maps generated by CNNs, which makes S_α lower.

Figure 5.5 shows the co-saliency maps that visually illustrate the ablation study. Initially, by using only the image-wise confidence computed from the stacked autoencoder, the results (Figure 5.5(b)) tend to bias toward a saliency map that indicates only apparent objects. By adding the region-wise confidence computed from the co-salient object likelihood, many missed regions are recovered (Figure 5.5(c)). Furthermore, by adding the pairwise term that promotes smoothness, the resultant saliency maps are smoothed out (Figure 5.5(d)). Lastly, after adding the regularization term, we obtain the best fusion result (Figure 5.5(e)). However, as mentioned above, the drawback of fusion is that the outcome is limited by the adopted saliency proposals. Fortunately, after further integration with CNNs by propagating information from regions with high confidence, as shown in Figure 5.5(f), the fusion results are gradually improved by emphasizing the common salient regions, but still left some blur background regions in some images. Figure 5.5(g) shows that self-paced learning improves the results by reducing irrelevant backgrounds. Finally, DenseCRFs help yield sharper and more complete co-saliency maps as shown in Figure 5.5(h).

5.6 Conclusions

We have presented an unsupervised framework for co-saliency detection. Our fusion-learning integrated model is composed of two stages. First, we propose `SAEF` to carry out the saliency proposal fusion via jointly exploring the image-level confidence based on the reconstruction error of SAE and the region-level confidence from co-salient object likelihood. Afterwards, our proposed `STCNN` can gradually learn co-salient objects in a self-taught fashion. The benefits of integrating both the fusion-based and deep-learning-based methods are evident as it produces the co-saliency maps of high quality via making the most of multiple locally complementary saliency proposals. Moreover, unlike existing fusion methods relying on the low-rank assumption of salient foreground regions, we propose a novel idea that takes advantage of the unsupervised SAE into our unified optimization process and generates even better results.

6. CONCLUSIONS AND FUTURE WORK

In this dissertation, we first presented our preliminary unsupervised fusion-based framework [51] that carries out region-wise saliency proposal fusion for more accurate and robust image (co-)saliency detection in Section 2. After then, we further improved the locally adaptive fusion model by exploring the advantages of the common object evidence generated from image co-segmentation in Section 3. Then, we additionally extended our effective segmentation-guided region-wise fusion framework [14] to be compatible with multiple images and analyzed the bilateral relationship between image co-saliency detection and co-segmentation in Section 4. Last but not least, we notice the learning-based methods using deep learning models have the potential to overcome the limitations in the proposed fusion models that depends on traditional group foreground coherence assumption and their performance can even surpass the quality of the proposal fused maps; unfortunately, current deep learning models often require significant human-labeled ground truth for training, which is usually unavailable in the co-saliency detection setup. For this reason, we proposed another solution by developing a more advanced unsupervised method in Section 5 that integrated the deep learning models into a joint fusion-learning integrated framework. With the comparison between our developed traditional convex optimization methods and the recently popular deep learning based method, both operated in an unsupervised manner, we have gained a better understanding of the co-saliency detection problem. More importantly, we released our implementation source code to be publicly available, which can be obtained at the following link (`http://github.com/chungchi`).

In future, we aim to further unify our two-stage model into a straightforward end-to-end learning model based on the fully convolutional network (FCN) [106]. Namely, with raw images as the input, we expect the context information can be incorporated in the successive deep convolutional layers to enable co-saliency detection at the output. The main motivation is that even though our proposed `STCNN` can learn to correct the previous errors from the `SAEF`, leading to better results, the quality of the adopted saliency proposals in `SAEF` still plays an important role in the final

co-saliency detection performance. Along the same direction of unsupervised preference as our previous works, our plan is to decompose co-saliency detection into two sub-streams, *object detection* and *correspondence discovery* based on the object recognition capability of the pre-trained FCN and novel correlation layers inspired by the recent work [109,110], with output constrained by two novel unsupervised losses, *the foreground consistency loss* and *foreground-background seperation loss*. The two losses can again be modeled on a graphical model where the former and the latter act as the unary and pairwise terms, respectively. By modeling the whole task as an energy minimization problem over a graph as we have done in this dissertation, we expect the two losses can be jointly optimized for generating co-saliency maps of high quality. We hope that our continuing research efforts can contribute to the future vision applications where saliency maps of high quality are appreciated, such as object tracking, image feature extraction, dense image alignment and scene understanding.

REFERENCES

[1] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2014.

[2] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. on Image Processing*, 2011.

[3] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Trans. on Image Processing*, 2014.

[4] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1998.

[5] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2007.

[6] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2009.

[7] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2012.

[8] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[9] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. on Image Processing*, 2013.

[10] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2013.

[11] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2013.

[12] 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. Computer Vision and Pattern Recognition*, 2013.

[13] D. Zhang, J. Han, J. Han, and L. Shao, "Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining," *IEEE Trans. on Neural Networks and Learning Systems*, 2016.

[14] C.-C. Tsai, X. Qian, and Y.-Y. Lin, "Segmentation guided local proposal fusion for co-saliency detection," in *Proc. Int'l Conf. Multimedia and Expo*, 2017.

[15] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing markov chain," in *Proc. Int'l Conf. Computer Vision*, 2013.

[16] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2017.

[17] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2010.

[18] H. Yu, M. Xian, and X. Qi, "Unsupervised co-segmentation based on a new global gmm constraint in mrf," in *Proc. Int'l Conf. Image Processing*, 2014.

[19] F. Meng, H. Li, G. Liu, and K. N. Ngan, "Image cosegmentation by incorporating color reward strategy and active contour model," *IEEE Trans. on Systems, Man, and Cybernetics*, 2013.

[20] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[21] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2013.

[22] H. Fu, D. Xu, S. Lin, and J. Liu, "Object-based rgbd image co-segmentation with mutex constraint," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2015.

[23] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. Int'l Conf. Computer Vision*, 2005.

[24] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int'l J. Computer Vision*, 2016.

[25] P. Jiang, N. Vasconcelos, and J. Peng, "Generic promotion of diffusion-based salient object detection," in *Proc. Int'l Conf. Computer Vision*, 2015.

[26] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016.

[27] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. Int'l Conf. Computer Vision*, 2017.

[28] K. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *IEEE Trans. on Multimedia*, 2016.

[29] F. Huang, J. Qi, H. Lu, L. Zhang, and X. Ruan, "Salient object detection via multiple instance learning," *IEEE Trans. on Image Processing*, 2017.

[30] D. Zhang, J. Han, and Y. Zhang, "Supervision by fusion: Towards unsupervised learning of deep salient object detector," in *Proc. Int'l Conf. Computer Vision*, 2017.

[31] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Proc. Neural Information Processing Systems*, 2011.

[32] D. E. Jacobs, D. B. Goldman, and E. Shechtman, "Cosaliency: Where people look when comparing images," in *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 2010.

[33] D. Zhang, H. Fu, J. Han, A. Borji, and X. Li, "A review of co-saliency detection algorithms: Fundamentals, applications, and challenges," *ACM TIST*, 2018.

[34] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2010.

[35] 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. Computer Vision and Pattern Recognition*, 2011.

[36] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *J. Image and Vision Computing*, 2011.

[37] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[38] F. Meng, H. Li, and G. Liu, "A new co-saliency model via pairwise constraint graph matching," in *Proc. Int'l Symp. Intelligent Signal Processing and Communication Systems*, 2012.

[39] X. Cao, Z. Tao, B. Zhang, H. Fu, and X. Li, "Saliency map fusion based on rank-one constraint," in *Proc. Int'l Conf. Multimedia and Expo*, 2013.

[40] 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. on Pattern Analysis and Machine Intelligence*, 2012.

[41] J. Li, J. Ding, and J. Yang, "Visual salience learning via low rank matrix recovery," in *Proc. Asian Conf. on Computer Vision*, 2014.

[42] 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 Information Processing Systems*, 2009.

[43] F. A. Potra and S. J. Wright, "Interior-point methods," *J. Computational and Applied Mathematics*, 2000.

[44] M. Grant, S. Boyd, and Y. Ye, "Cvx: Matlab software for disciplined convex programming," 2008.

[45] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE Trans. on Multimedia*, 2013.

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

[47] C.-R. Huang, Y.-J. Chang, Z.-X. Yang, and Y.-Y. Lin, "Video saliency map detection by dominant camera motion removal," *IEEE Trans. on Circuits and Systems for Video Technology*, 2014.

[48] 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 *Proc. Int'l Conf. Multimedia and Expo*, 2014.

[49] R. Huang, W. Feng, and J. Sun, "Saliency and co-saliency detection by low-rank multiscale fusion," in *Proc. Int'l Conf. Multimedia and Expo*, 2015.

[50] K. R. Jerripothula, J. Cai, and J. Yuan, "Cats: Co-saliency activated tracklet selection for video co-localization," in *Proc. Euro. Conf. Computer Vision*, 2016.

[51] C.-C. Tsai, X. Qian, and Y.-Y. Lin, "Image co-saliency detection via locally adaptive saliency map fusion," in *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, 2017.

[52] K. R. Jerripothula, J. Cai, and J. Yuan, "Quality-guided fusion-based co-saliency estimation for image co-segmentation and co-localization," *IEEE Trans. on Multimedia*, 2018.

[53] J. Guo, Z. Li, L.-F. Cheong, and S. Z. Zhou, "Video co-segmentation for meaningful action extraction," in *Proc. Int'l Conf. Computer Vision*, 2013.

[54] H.-Y. Chen, Y.-Y. Lin, and B.-Y. Chen, "Co-segmentation guided hough transform for robust feature matching," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2015.

[55] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proc. Int'l Conf. Internet Multimedia Computing and Service*, 2014.

[56] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2012.

[57] C. Xia, P. Wang, F. Qi, and G. Shi, "Nonlocal center-surround reconstruction-based bottom-up saliency estimation," in *Proc. Int'l Conf. Image Processing*, 2013.

[58] C. Xia, F. Qi, and G. Shi, "An iterative representation learning framework to predict the sequence of eye fixations," in *Proc. Int'l Conf. Multimedia and Expo*, 2017.

[59] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2011.

[60] 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. British Conf. Machine Vision*, 2011.

[61] Ç. Aytekin, E. C. Ozan, S. Kiranyaz, and M. Gabbouj, "Visual saliency by extended quantum cuts," in *Proc. Int'l Conf. Image Processing*, 2015.

[62] Y. Xie, H. Lu, and M.-H. Yang, "Bayesian saliency via low and mid level cues," *IEEE Transactions on Image Processing*, 2013.

[63] K. Fu, C. Gong, I. Y.-H. Gu, and J. Yang, "Normalized cut-based saliency detection by adaptive multi-level region merging," *IEEE Trans. on Image Processing*, 2015.

[64] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[65] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep cnn features," *IEEE Trans. on Image Processing*, 2016.

[66] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. on Image Processing*, 2016.

[67] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016.

[68] R. Huang, J. Sun, and W. Feng, "Color feature reinforcement for co-saliency detection without single saliency residuals," *Signal Processing Letters*, 2017.

[69] L. Ye, Z. Liu, J. Li, W.-L. Zhao, and L. Shen, "Co-saliency detection via co-salient object discovery and recovery," *Signal Processing Letters*, 2015.

[70] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM transactions on graphics*, 2004.

[71] 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. Computer Vision and Pattern Recognition*, 2006.

[72] D. S. Hochbaum and V. Singh, "An efficient algorithm for co-segmentation," in *Proc. Int'l Conf. Computer Vision*, 2009.

[73] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2013.

[74] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l J. Computer Vision*, 2004.

[75] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2004.

[76] Z. Gao, P. Shi, H. R. Karimi, and Z. Pei, "A mutual grabcut method to solve co-segmentation," *J. Image and Video Processing*, 2013.

[77] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2014.

[78] J. Xue, C. Li, and N. Zheng, "Proto-object based rate control for jpeg2000: an approach to content-based scalability," *IEEE Trans. on Image Processing*, 2011.

[79] X. Cao, Y. Cheng, Z. Tao, and H. Fu, "Co-saliency detection via base reconstruction," in *Proc. ACM Int'l Conf. Multimedia*, 2014.

[80] Y. Li, K. Fu, Z. Liu, and J. Yang, "Efficient saliency-model-guided visual co-saliency detection," *Signal Processing Letters*, 2015.

[81] L. Wei, S. Zhao, O. E. F. Bourahla, X. Li, and F. Wu, "Group-wise deep co-saliency detection," in *Proc. Int'l Joint Conf. Artificial Intelligence*, 2017.

[82] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based framework for co-saliency detection," *IEEE Trans. on Circuits and Systems for Video Technology*, 2017.

[83] 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. on Image Processing*, 2017.

[84] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *CVPR*, 2015.

[85] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016.

[86] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2017.

[87] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. Int'l Conf. Computer Vision*, 2017.

[88] X. Li, F. Yang, H. Cheng, J. Chen, Y. Guo, and L. Chen, "Multi-scale cascade network for salient object detection," in *Proc. ACM Int'l Conf. Multimedia*, 2017.

[89] C. Li, Y. Yuan, W. Cai, Y. Xia, D. D. Feng, *et al.*, "Robust saliency detection via regularized random walks ranking.," in *CVPR*, 2015.

[90] P. Hu, W. Wang, and K. Lu, "Detecting salient objects via spatial and appearance compactness hypotheses," in *Proc. ACM Int'l Conf. Multimedia*, 2015.

[91] Y. Tang, X. Wu, and W. Bu, "Saliency detection based on graph-structural agglomerative clustering," in *Proc. ACM Int'l Conf. Multimedia*, 2015.

[92] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in *ICCV*, 2015.

[93] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. Neural Information Processing Systems*, 2010.

[94] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, and A. G. Hauptmann, "Self-paced learning for matrix factorization.," in *AAAI*, 2015.

[95] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann, "Easy samples first: Self-paced reranking for zero-example multimedia search," in *Proc. ACM Int'l Conf. Multimedia*, 2014.

[96] J. S. Supancic III and D. Ramanan, "Self-paced learning for long-term tracking," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2013.

[97] D. Gong, M. Tan, Y. Zhang, A. van den Hengel, and Q. Shi, "Self-paced kernel estimation for robust blind image deblurring," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2017.

[98] I. Lillo, J. C. Niebles, and A. Soto, "A hierarchical pose-based approach to complex action understanding using dictionaries of actionlets and motion poselets," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016.

[99] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int'l J. Computer Vision*, 2004.

[100] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2017.

[101] C. Xia, F. Qi, and G. Shi, "Bottom–up visual saliency estimation with deep autoencoder-based sparse reconstruction," *IEEE Trans. on Neural Networks and Learning Systems*, 2016.

[102] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2016.

[103] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int'l Conf. Learning Representations*, 2014.

[104] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proc. ACM Int'l Conf. Multimedia*, 2015.

[105] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int'l Conf. Learning Representations*, 2014.

[106] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional models for semantic segmentation," in *CVPR*, 2015.

[107] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. Int'l Conf. Computer Vision*, 2017.

[108] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. on Image Processing*, 2015.

[109] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. Int'l Conf. Computer Vision*, 2015.

[110] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. Conf. Computer Vision and Pattern Recognition*, 2017.

APPENDIX A

MORE VISUAL RESULTS FOR OUR MOST ADVANCED CO-SALIENCY DETECTION MODEL PRESENTED IN SECTION 5

Figure A.1: 5 categories, *Alaskan Brown Bear*, *Red Sox Players*, *Agra Taj Mahal-Inde du Nord*, *Elephants* and *Pandas* from the iCoseg benchmark data set.

Figure A.2: 2 categories, *House* and *Face* from the MSRC benchmark data set.

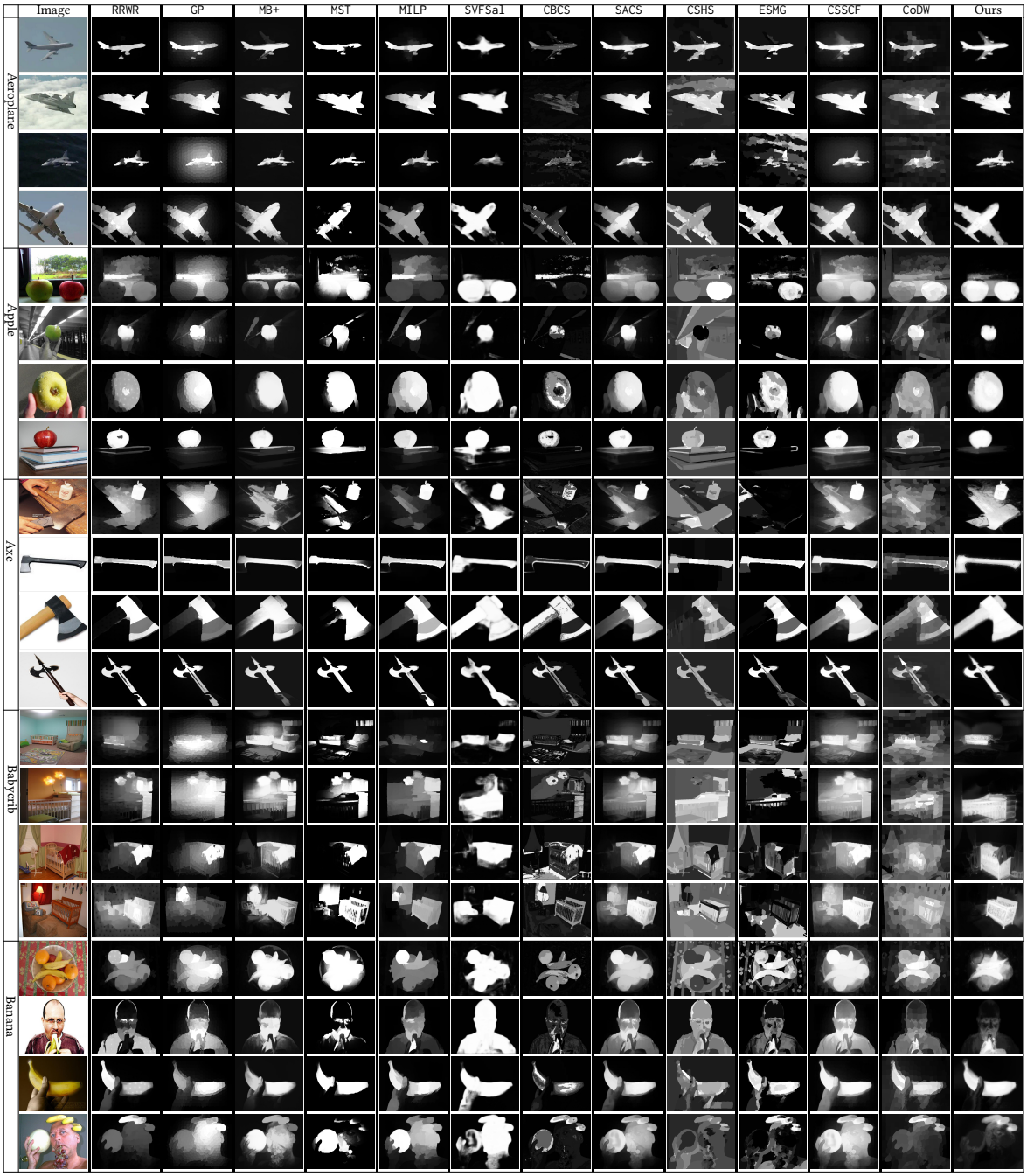


Figure A.3: 5 categories, *Aeroplane*, *Apple*, *Axe*, *Babycrib* and *Banana* from the Cosal2015 benchmark data set.