A Feature-Adaptive Semi-Supervised Framework for Co-saliency Detection

Xiaoju Zheng
Institute of Intelligent Machines, Chinese Academy of Sciences
University of Science and Technology of China
xiaoju@mail.ustc.edu.cn

Zheng-Jun Zha*
University of Science and Technology of China
zhazj@ustc.edu.cn

Liansheng Zhuang
University of Science and Technology of China
lszhuang@ustc.edu.cn

ABSTRACT

Co-saliency detection, which refers to the discovery of common salient foreground regions in a group of relevant images, has attracted increasing attention due to its widespread applications in many vision tasks. Existing methods assemble features from multiple views toward a comprehensive representation, however overlook the efficacy disparity among various features in detecting co-saliency. This paper proposes a novel feature-adaptive semi-supervised (FASS) framework for co-saliency detection, which seamlessly integrates multi-view feature learning, graph structure optimization and co-saliency prediction in a unified solution. In particular, the FASS exploits the efficacy disparity of multi-view features at both view and element levels by a joint formulation of view-wise feature weighting and element-wise feature selection, leading to an effective representation robust to feature noise and redundancy as well as adaptive to the task at hand. It predicts co-saliency map by optimizing co-saliency label prorogation over a graph of both labeled and unlabeled image regions. The graph structure is optimized jointly with feature learning and co-saliency prediction to precisely characterize underlying correlation among regions. The FASS is thus able to produce satisfactory co-saliency map based on the effective exploration of multi-view features as well as inter-region correlation. Extensive experiments on three benchmark datasets, *i.e.*, iCoseg, Cosal2015 and MSRC, have demonstrated that the proposed FASS outperforms the state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Interest point and salient region detections**;

*Corresponding Author: Zheng-Jun Zha (zhazj@ustc.edu.cn)

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240648

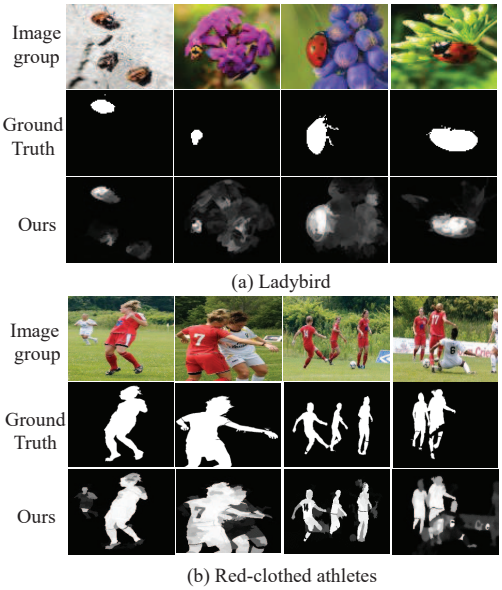


Figure 1: Illustration of the image groups and co-saliency maps of "ladybird" and "red-clothed athletes", respectively. While low-level color feature is effective for detecting co-saliency corresponding to, high-level semantic cue is more useful for co-saliency detection of "ladybird".

KEYWORDS

Co-saliency detection, multi-view feature, graph optimization, semi-supervised learning

ACM Reference Format:
Xiaoju Zheng, Zheng-Jun Zha, and Liansheng Zhuang. 2018. A Feature-Adaptive Semi-Supervised Framework for Co-saliency Detection. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3240508.3240648

1 INTRODUCTION

The rapid advancements of image acquisition devices have increased the volume of digital image collections significantly. These big image collections occupy huge amounts of storage space and require huge computational resources for processing. There is an intense demand to selectively store, deliver and

process region-of-interest rather than original images in many applications. Co-saliency detection, which imitates human vision system to detect common salient foregrounds (CFs) within relevant images, has attracted increasing attention in recent years [8, 9, 11, 26, 27].

In the past decades, many researches focus on discovering effective image representation for co-saliency detection. Conventional methods primarily use low-level features, such as color, texture or SIFT descriptors, to represent image regions [3–5, 7, 15, 16, 25], assuming that the co-salient objects within relevant images share certain low-level visual consistency. Recently, high-level semantic features have been exploited in co-saliency detection [26, 31, 32]. The high-level features provide semantic cues and are relatively robust to the variations in viewpoints, shapes, and luminance *etc.* To explore the complementarity among features from multiple views, some approaches have been developed to assemble multi-view features towards a comprehensive representation [10, 22, 33, 34]. However, they treat various features equally important and overlook their capacity disparity in discovering co-saliency. Actually, as shown in Figure 1, various features possess different abilities in representing CFs and distinguishing CFs from background and such ability changes with the tasks at hand.

Traditional co-saliency detection methods mainly search for CFs within an image group based on some human-designed co-saliency priors in unsupervised bottom-up and fusion-based manners [2–4, 6, 7, 15, 16, 26, 30]. However, human-designed priors are typically subjective and not generalizable to various cases of co-saliency, resulting in unsatisfactory performance. Recently, a few of preliminary supervised co-saliency detection has been proposed to learn co-saliency maps using pixel-level ground truth based on either traditional learning model [10] or deep neural networks [29]. However, it is highly time-assuming and labor-intensive to manually annotate co-salient CFs within images with pixel-level masks. As a result, supervised learning methods usually suffer from the lack of sufficient labeled samples.

Motivated by the above observations, we propose a novel feature-adaptive semi-supervised (FASS) framework for co-saliency detection. The FASS seamlessly integrates multi-view feature learning, graph structure optimization and co-saliency detection in a unified semi-supervised solution. In particular, the FASS exploits the efficacy disparity of multi-view features at both view and element levels and formulates a joint learning of view-wise feature weighting and element-wise feature selection. In this way, FASS is able to concentrate more on features from the important views and filter out redundant and noisy elements within original features. The resultant representation is effective and adaptive to the tasks at hand. FASS explores the abundant unlabeled images to boost co-saliency detection in a semi-supervised manner. It predicts co-saliency map by optimizing co-saliency label prorogation over a graph of both labeled and unlabeled image regions. The graph structure is optimized jointly with feature learning and co-saliency prediction to precisely characterize underlying correlation among regions. By the joint exploration of multi-view feature learning, graph structure optimization and co-saliency label inferring, FASS is able to generate accurate co-saliency maps. We have conducted extensive experiments on three widely used co-saliency detection datasets, *i.e.*, i-Coseg, Cosal2015 and MSRC datasets. Experimental results have shown that the proposed FASS framework outperforms the state-of-the-art methods.

2 RELATED WORK

Co-saliency detection. A wealth of approaches have been proposed for co-saliency detection in recent years. Existing methods can be roughly grouped into three main categories: bottom-up, fusion-based, and supervised learning methods. The bottom-up methods [4, 7, 15, 16, 26] scored each region in an image group by using human designed co-saliency priors. Generally, it consists of image preprocessing, feature extraction, single prior exploration, and multi-prior combination stages. For example, Faktor *et al.* [6] utilized a co-segmentation detection prior that common regions contained by the given image group should be the objects of interest. Fu *et al.* [7] proposed a cluster-based co-saliency detection approach, by using three bottom-up saliency cues, *i.e.*, the contrast, spatial and correspondence cues. The final co-saliency prediction was obtained based on the combination of the explored bottom-up cues. The fusion-based solution [2, 3, 30] was to aggregate detection results from multiple methods. For example, Cao *et al.* [2] proposed to combine multiple co-saliency/saliency maps by a rank constraint with self-adaptive weights. Cao *et al.* [3] further proposed a reconstruction-based fusion approach based on the ensemble of results from multiple existing saliency detection methods. Huang *et al.* [30] fused multi-scale saliency maps, which were generated based on super-pixels at multiple scales. However, The fusion-based methods heavily rely on the fused methods and usually suffer from the imprecise results of the fused methods.

Recently, supervised co-saliency detection approaches have been proposed [10, 29, 33, 34]. For example, Zhang *et al* [33, 34] proposed a self-paced multiple-instance method to gradually learn the patterns of co-salient objects from confident image regions to ambiguous ones. It used the image-level labels as weak supervision, indicating whether an image contained the to-be-detected co-salient objects. Wei *et al.* [29] proposed an end-to-end group-wise deep co-saliency detection approach by combing single visual feature maps and the common semantic feature map in an image group. Han *et al.* [10] proposed a unified metric learning framework for improving co-saliency detection. Most supervised methods often require a large amount of training samples with pixel-level co-saliency ground truth. However, manual annotation for pixel-level ground truth is very labor-intensive, resulting in the lack of training data.

Features for co-saliency. As a basic yet critical factor, visual features adopted to represent image pixels or regions significantly affect the performance of co-saliency detection

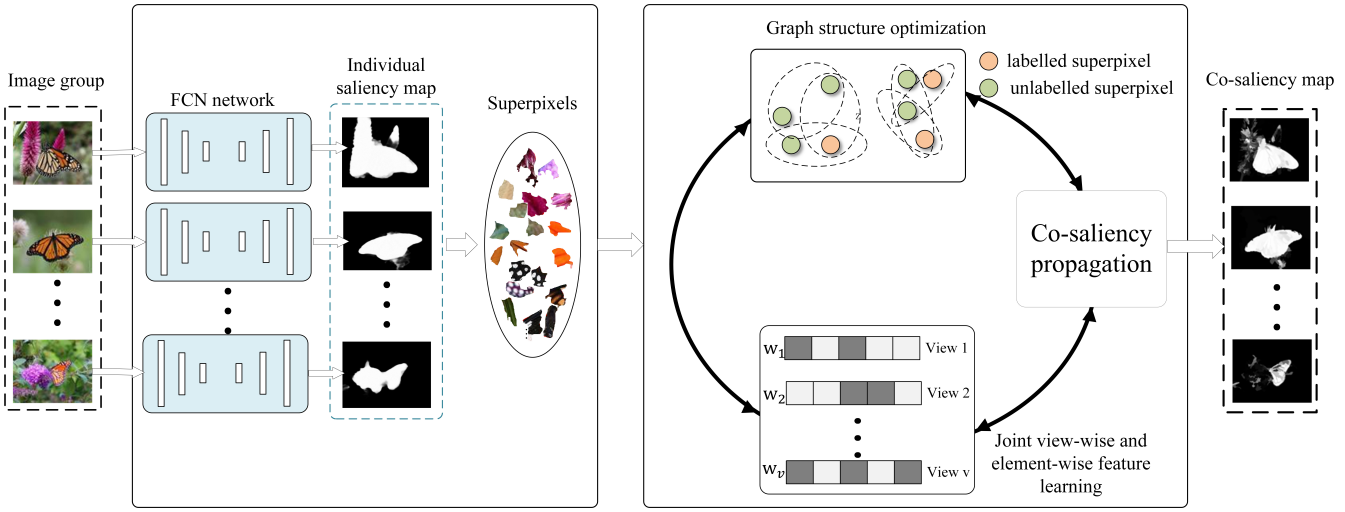


Figure 2: Overview of the proposed Feature-Adaptive Semi-Supervised (FASS) framework for co-saliency detection.

models. Existing works mainly exploited the low-level, high-level and hypercolumn features for co-saliency detection. Here, following existing research [10, 33, 34], the high-level features referred to the representations learned by deep neural networks.

The low-level features [3–5, 7, 15, 16, 25], such as color histograms, Gabor filters, or SIFT descriptors have widespread applications in co-saliency detection tasks. However, low-level features were often not robust to the variations in viewpoints, shapes, and luminance *etc.* and sometimes were instable for various cases of co-saliency. Moreover, low-level features lacked of the abstraction of semantic cues. Recently, high-level semantic features learned by deep neural networks have been introduced for co-saliency detection [10, 33, 34]. The deep neural networks were pre-trained on auxilary image datasets,*e.g.*, ImageNet [12] and in turn used to extract features from co-saliency image collections. Generally, neither low-level nor high-level features can individually handle all the cases in co-saliency detection. Zhang *et al.* [33, 34] proposed a hypercolumn representation, which is a combination of feature maps from different CNN layers, towards capturing both low-level and and high-level cues. Recently, Han *et al.* [10] learned a feature transformation matrix by embedding a metric learning term into support vector machine(SVM). The original hypercolumn features were projected to a new feature space by using the learnt transformation matrix.

3 THE PROPOSED APPROACH

3.1 Overview

Co-saliency detection aims at extracting common salient regions in relevant images. The common salient regions are not only salient in each individual image but also commonly appear in a group of relevant images. Hence, "salient" and "common" are two crucial attributes that together reflect the definition of co-saliency. It is straightforward to recast co-saliency detection as a classification task, classifying each region/super-pixel within images as co-salient or not. Figure 2 illustrates the proposed feature-adaptive semi-supervised (FASS) framework for co-saliency detection. As aforementioned, the FASS simultaneously optimizes view-wise feature weighting, element-wise feature selection, graph structure as well as co-saliency label prorogation in a unified semi-supervised solution. These components enhance each other mutually, together facilitating accurate co-saliency detection.

The FASS starts with constructing a collection of candidate salient super-pixels from image groups. It first produces individual saliency map $\overline{T}$ for each image by using a pre-trained deep saliency detection model [17]. Other saliency detection models are also feasible. The final co-saliency result is not sensitive to the saliency detection model. Then, the simple linear iterative clustering (SLIC) algorithm is applied to partition each image within $\{I_m\}_{m=1}^{M}$ into a set of super-pixels. We define T_k^m as the individual saliency value of k-th super-pixel from m-th image. T_k^m is calculated by average pooling of the saliency scores of the corresponding pixels in $\overline{T}$. We set a saliency threshold θ ($\theta = 0.3$) and select the super-pixels with saliency score $T_k^m \geq \theta$ to form the collection $\mathcal{V}$ of candidate salient super-pixels, which are the processing units for the subsequent modules.

3.2 The FASS Algorithm

In this section, we elaborate the proposed FASS algorithm including multi-view feature learning, graph structure optimization and co-saliency prediction. We first introduce the graph to be optimized to characterize the underlying correlation among various super-pixels. Let $\mathcal{G}^k = (\mathcal{V}, \mathcal{E}^k, \mathbf{A}^k)$ denote the graph consisting of labeled and unlabeled super-pixels constructed based on k-th view feature. A node in $\mathcal{G}^k$ corresponds to a candidate salient super-pixel in $\mathcal{V}$ and an

edge e_{ij} in $\mathcal{E}^k = [e_{ij}]^k_{n\times n}$ represents the affinity between two related super-pixels. $\mathbf{A}^k$ denotes the affinity matrix. The similarity a^k_{ij} between two super-pixels is calculated as $a^k_{ij} = \exp(-\left\|\tilde{\mathbf{x}}^k_i - \tilde{\mathbf{x}}^k_j\right\|^2)$, where $\tilde{\mathbf{x}}^k$ is the to-be-optimized k-th view feature. $\mathbf{A}^k$ is initialized using the original feature $\mathbf{x}^k_i$ and updated with the optimized feature iteratively. Let $\mathbf{L}^k$ denote the Laplacian matrix of graph $\mathcal{G}^k$. $\mathbf{L}^k = \begin{bmatrix} \mathbf{L}^k_{ll} & \mathbf{L}^k_{lu} \\ \mathbf{L}^k_{ul} & \mathbf{L}^k_{uu} \end{bmatrix}$, where l and u refer to the index of matrix block corresponding to labeled and unlabeled vertices.

In order to characterize the underlying correlation among super-pixels precisely and comprehensively, we design a global graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{S})$ based on multi-view features, where $\mathbf{S}$ is the affinity matrix of the graph. $\mathcal{G}$ is optimized through an appropriate ensemble of $\mathcal{G}^k$ from multiple views as follows:

$$\begin{aligned} &\min \left\| \mathbf{S} - \sum_{k=1}^{v} \mu_k \mathbf{A}^k \right\|_F^2 \\ &\text{s.t. } \forall i,\ {\mathbf{S}_i}^T \mathbf{1} = 1,\ 0 \le S_{ij} \le 1, \operatorname{rank}(\mathbf{L}_s) = n-2, \mu^{(t)}_l = w^{(t-1)}_k \end{aligned} \tag{1}$$

where μ_k is the ensemble weight of k-th view graph, representing the importance of k-th view feature. μ_k is initialized as a uniform weight as $\frac{1}{v}$ and updated as the value of w_k at last iteration during optimization. $\operatorname{rank}(\mathbf{L}_s) = n-2$ is the constraint on the rank of matrix $\mathbf{L}_s$. If it is satisfied, the $\mathbf{S}$ becomes an ideal neighbor assignment and the data points are partitioned into 2 clusters.

In order for a comprehensive and effective representation from multiple views, we formulate a multi-view feature learning consisting of view-wise feature weighting and element-wise feature selection. It simultaneously considers the efficacy disparity of various features at both view and element levels. The multi-view feature learning is formulated as follows:

$$\begin{aligned} &\min \sum_{k=1}^{v} w_k \sum_{i,j} \left\| (\boldsymbol{\beta}^k)^T \mathbf{x}^k_i - (\boldsymbol{\beta}^k)^T \mathbf{x}^k_j \right\|_2^2 S_{ij} + \gamma \sum_{k=1}^{v} \left\| \boldsymbol{\beta}^k \right\|_2^1 \\ &\text{s.t. } \sum_{k=1}^{v} w_k = 1, \end{aligned} \tag{2}$$

where $\mathbf{x}^k_i$ is the k-th view feature of i-th super-pixel. $\{w_k\}^v_1$ refers to the weighting parameters of different views. $\boldsymbol{\beta}^k \in \mathbb{R}^{m^k \times d^k}$ is the projection matrix for k-th view feature. It is used to find a discriminative subspace for the original features and select important feature elements via the $l_{2,1}$-norm regularization $\boldsymbol{\beta}^k$. m^k and d^k are the original and projected feature dimensionality, respectively.

We conduct semi-supervised co-saliency learning by prorogating co-saliency labels over labeled and unlabeled super-pixels. The assumption here is that the similar super-pixels are likely to have the same co-saliency label. The co-saliency proportion is formulated as follows:

$$\min \sum_{i,j} \|f_i - f_j\|_2^2 S_{ij} \tag{3}$$

where $f_i \in \{0, 1\}$ is the co-saliency label of i-th super-pixel. For any labeled super-pixel, f_i is generated by average pooling the co-saliency groundtruth of pixels within the super-pixel and binarilizing the average co-saliency score via a threshold of 0.5. Let $\mathbf{F}_l$ denote the co-saliency label vector of labeled super-pixels. $\mathbf{F}_u$ is the to-be-inferred labels for unlabeled ones. $\mathbf{F} = [\mathbf{F}_l, \mathbf{F}_u]^T$.

The above co-saliency prorogation is jointly optimized with the graph structure optimization in Eq. (1), and multi-view feature learning in Eq. (2) via a unified formulation as the following Eq. (4), so that the three learning tasks can enhance each other mutually and together lead to a optimal solution to infer labels of super-pixels, which is jointly learnt with the graph structure optimization and multi-view feature learning. Then, we predict indicator unlabeled matrix $\mathbf{F}_u$ by the following unified formulation:

$$\begin{aligned} &\min \sum_{k=1}^{v} w_k \sum_{i,j} \left\| (\boldsymbol{\beta}^k)^T \mathbf{x}^k_i - (\boldsymbol{\beta}^k)^T \mathbf{x}^k_j \right\|_2^2 S_{ij} + \alpha \sum_{i,j} \|f_i - f_j\|_2^2 S_{ij} \\ &+ \gamma \sum_{k=1}^{v} \left\| \boldsymbol{\beta}^k \right\|_2^1 + \delta \left\| \mathbf{S} - \sum_{k=1}^{v} \mu_k \mathbf{A}_k \right\|_F^2 \\ &\text{s.t. } \forall i,\ {\mathbf{S}_i}^T \mathbf{1} = 1,\ 0 \le S_{ij} \le 1, \operatorname{rank}(\mathbf{L}_s) = n-2, \\ &\sum_{k=1}^{v} w_k = 1, \mu^{(t)}_l = w^{(t-1)}_k, \end{aligned} \tag{4}$$

3.3 Optimization

We introduce an alternative optimization strategy for optimizing $\{w_k\}^v_1, \{\boldsymbol{\beta}^k\}^v_1, \mathbf{S}, \mathbf{F}$ in Eq. (4).

(1) **Fixing S, F and $\{w_k\}^v_1$, update $\{\boldsymbol{\beta}^k\}^v_1$.** $\{\boldsymbol{\beta}^k\}^v_1$ is optimized to form a linear transform of the features and element-wise feature selection. The optimization can be equivalently decomposed into sub-problems with respect to each $\boldsymbol{\beta}^k$. By denoting the constant $\mathbf{C}$ as the sum of the related items of $\boldsymbol{\beta}^i (i \neq k)$, we can optimize any $\boldsymbol{\beta}^k$ as follows:

$$\min \sum_{i,j} \left\| (\boldsymbol{\beta}^k)^T x^k_i - (\boldsymbol{\beta}^k)^T x^k_j \right\|_2^2 S_{ij} + (\frac{\gamma}{w_k}) \sum_{l=1}^{v} \left\| \boldsymbol{\beta}^k \right\|_2^1 + \mathbf{C} \tag{5}$$

The objective function in Eq. (5) is convex and can be solved by an existing strategy [21].

(2) **Fixing S, F and $\{\boldsymbol{\beta}^k\}^v_1$, update $\{w_k\}^v_1$.** Following [20], w_k is dependent on $\mathbf{S}$, $\boldsymbol{\beta}^k$ and can be updated as follows:

$$w_k = \frac{1}{2\sqrt{\sum_{i,j} \left\| (\boldsymbol{\beta}^k)^T x^k_i - (\boldsymbol{\beta}^k)^T x^k_j \right\|_2^2 S_{ij}}}. \tag{6}$$

(3) **Fixing $\{w_k\}^v_1$, S, $\{\boldsymbol{\beta}^k\}^v_1$, update F.** $\mathbf{F}_u$ can be calculated as according to the following formulation:

$$\mathbf{F}_u = -{\mathbf{L}_{uu}}^{-1} \mathbf{L}_{ul} \mathbf{F}_l. \tag{7}$$

(4) **Fixing $\{w_k\}_1^v$, S,$\{\boldsymbol{\beta}^k\}_1^v$ and F, update S.** The optimization of $\mathbf{S}$ contains two sub-problems including individual graph optimization and global graph alignment.

Individual graph optimization: We take an alternative solution for optimizing the individual graph of a single view. First, $\mathbf{A}^k$ can be optimized by the following objective function:

$$\min \sum_{i,j} \left\| \boldsymbol{\beta}^T x_i - \boldsymbol{\beta}^T x_j \right\|_2^2 a_{ij} + \lambda \sum_{i,j} \|q_i - q_j\|_2^2 a_{ij} + \eta {a_{ij}}^2$$
$$s.t.\ \forall i,\ \mathbf{a}_i^T \mathbf{1} = 1,\ 0 \le a_{ij}, \mathrm{rank}\left(\mathbf{L}_{A^k}\right) = n - 2 \tag{8}$$

where q_i corresponds to the predicted labels for i-th super-pixel under k-th view. The first step is to fix $\mathbf{Q}$ and update $\mathbf{A}^k$. Here, we denote $\left\|\beta^T x_i - \beta^T x_j\right\|_2^2$ as $m_{ij}^{\beta x}$ and $\|q_i - q_j\|_2^2$ as m_{ij}^f. $d_{ij} = m_{ij}^{\beta x} + \lambda m_{ij}^f$ is the feature difference and label disparity between i-th and j-th super-pixels. Each row $\mathbf{a}_i^k$ of $\mathbf{A}^k$ is obtained via solving the following optimization problem:

$$\min_{\forall i\, {\mathbf{a}_i^k}^T \mathbf{1} = 1, \mathrm{rank}\left(\mathbf{L}_{A^k}\right) = n-2,} \left\| \mathbf{a}_i^k - \frac{1}{2\eta} \mathbf{d}_i \right\|_2^2 \tag{9}$$

where $\mathbf{d}_i = [d_{i1}, d_i, ..., d_{in}]^T$. The second step is to fix $\mathbf{A}^k$ and update $\mathbf{Q} = [\mathbf{F}_l, \mathbf{Q}_u]^T$. Similar to Eq. (7), $\mathbf{Q}$ can be solved as *i.e.*, $\mathbf{Q}_u = -{\mathbf{L}_{uu}^k}^{-1} \mathbf{L}_{ul}^k \mathbf{F}_l$. We alternatively optimize $\mathbf{Q}$ and $\mathbf{A}^k$ until the sum of the two smallest eigenvalues of $\mathbf{L}_{A^k}$ becomes zero [21]. η is initialized as 1 and is decreased if the connected component of $\mathbf{A}^k$ is larger than the class number two or otherwise increased during the iteration.

Global graph alignment: The second step is to build an integrated graph. We obtain individual view affinity matrix $\{\mathbf{A}^k\}_1^v$ by Eq. (7), The optimization of the global graph can be rewritten as follows:

$$\min \sum_{k=1}^{v} w_k \sum_{i,j} \left\| {(\boldsymbol{\beta}^k)}^T x_i^k - {(\boldsymbol{\beta}^k)}^T x_j^k \right\|_2^2 S_{ij} + \alpha \sum_{i,j} \|f_i - f_j\|_2^2 S_{ij}$$
$$+ \delta \left\| \mathbf{S} - \sum_{k=1}^{v} \mu_k \mathbf{A}^k \right\|_F^2$$
$$s.t.\ \forall i,\ \mathbf{s}_i^T \mathbf{1} = 1,\ 0 \le s_{ij} \le 1, \mathrm{rank}\left(\mathbf{L}_{A^k}\right) = n - 2, \tag{10}$$

where μ_l is set to the value of w_l at last iteration, *i.e.*, $\mu_l^t = w_l^{t-1}$. The optimization strategy for Eq. (10) is same as that for Eq. (9). After each iteration, we can obtain the co-saliency labels $\mathbf{F}_u$ based on the current learned features and similarity matrix $\mathbf{S}$.

Algorithm 1: The FASS Algorithm

Input: $\mathcal{X} = \{\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^v\}$;
$\mathbf{X}^k = [x_1^l, x_2^k, \ldots, x_n^k] \in \mathbb{R}^{m^k \times n}$;
the trade-off parameters $\alpha,\ \gamma, \delta$; label matrix $\mathbf{F}$.
Output: The predicted label matrix for unlabeled super-pixels.
Initial: The weight for k-th view $w_k = \frac{1}{v}$,
$\boldsymbol{\beta}^k = \mathbf{I}^{(m^k \times d^k)}$ and $\mathbf{S}$ initialized by Euclidean distance.
Repeat:
1. Update $\{\boldsymbol{\beta}^k\}_1^v$ via Eq. (5).
2. Update $\{w_k\}_1^v$ via Eq. (6).
3. Update $\mathbf{F}$ via Eq. (7), where it is formed by the c eigenvectors of $\mathbf{L}_s$ corresponding to the c smallest eigenvalues and $c = 2$.
4. Update $\mathbf{S}$ via Eq. (10).

Until converge

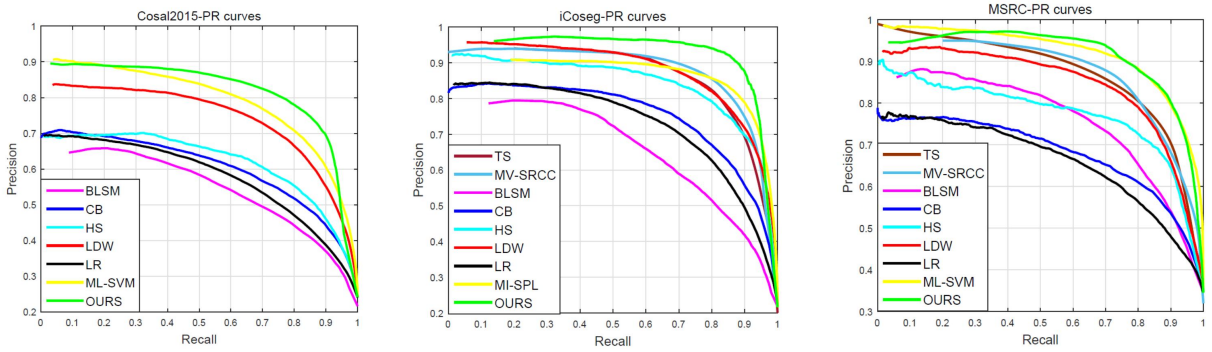
4 EXPERIMENTS

4.1 Experimental Setup

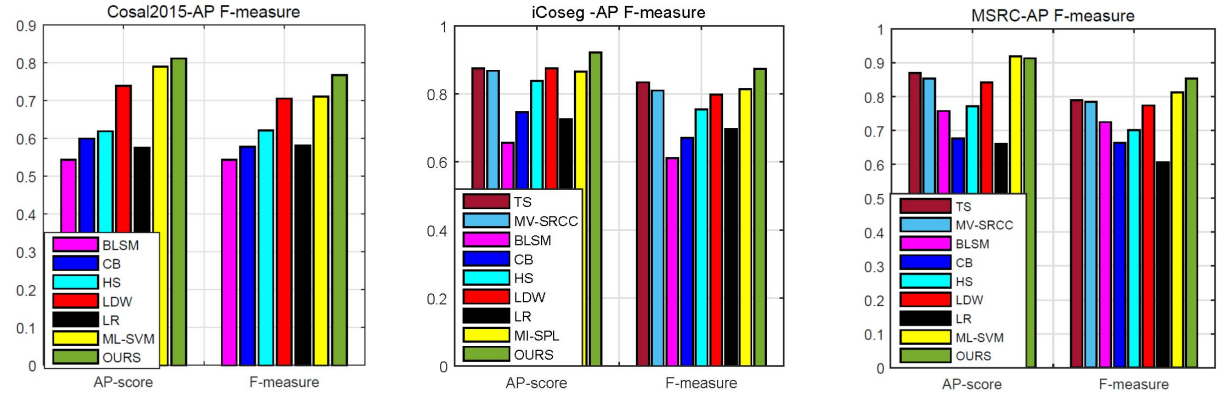
Dataset and Performance Metric. We evaluate the proposed FASS approach on three widely used benchmark datasets: Cosal2015, iCoseg and MSRC datasets. Cosal2015 is the largest and most challenging dataset for co-saliency detection. It contains 50 image groups with a total of 2,015 images. iCoseg consists of 38 image groups with a total of 643 images. MSRC dataset consists of 7 image groups with a total of 240 images. All the images within these datasets have been manually labeled with pixel-level co-saliency ground-truth. We adopt three popular performance metric in the experiments, including the precision–recall (PR) curve, average precision (AP) and F-measure. The PR curve is generated by a series of thresholds T, varying from 0 to 255.

Image Features. We adopt two categories of image features, *i.e.*, traditional low-level features and deep learning features, to represent various characteristics of image content. Six types of low-level features are used, including RGB (3-D), LAB (3-D), HSV (3-D), Texture (15-D), Texture-hist (15-D), and LBP-hist (256-D) descriptors [24]. For the deep learning features, we use off-the-shelf fully connected convolutional network (FCN) as feature extractor due to its impressive performance in image segmentation and saliency detection. "FCN-32s" network [19] is used in our experiments, which contains seven convolutional layers, five pooling layers and one upsampled prediction layer. The network is pre-trained on saliency datasets.The feature maps from the first and fifth convolutional layers are used as the low-level and high-level image representation, respectively. The sizes of the feature maps are $56 \times 256 \times 256$ and $512 \times 17 \times 17$, respectively. We resize the feature maps to the size of input image by bilinear interpolation and then conduct average pooling to produce the low-level and high-level features for each super-pixel of 56 dimensionality and 512 dimensionality, respectively.

Implementation Details. We randomly select 50% images from each image group as labeled samples and use the remaining images as unlabeled ones. We conduct five-fold cross validation and report the average performance. We initialize the supe-rpixel number n of each image as 200 and the precise value of n is determined by the SLIC algorithm [1]. The individual saliency maps are generated by the pre-trained DHS-Net model [17]. For the proposed FASS algorithm, we initialize the parameter $\delta = 1$ in Eq. (4) and decrease δ if the connected component of S is larger than the class number c



(b) The results measured in terms of the PR curves for the iCoseg, Cosal2015 and MSRC datasets.



(c) The results measured in terms of the AP score and the F-measure for the iCoseg, Cosal2015 and MSRC datasets.

Figure 3: The performance comparison between the proposed FASS and nine state-of-the-arts methods in terms of AP curve, AP score and F-measure.

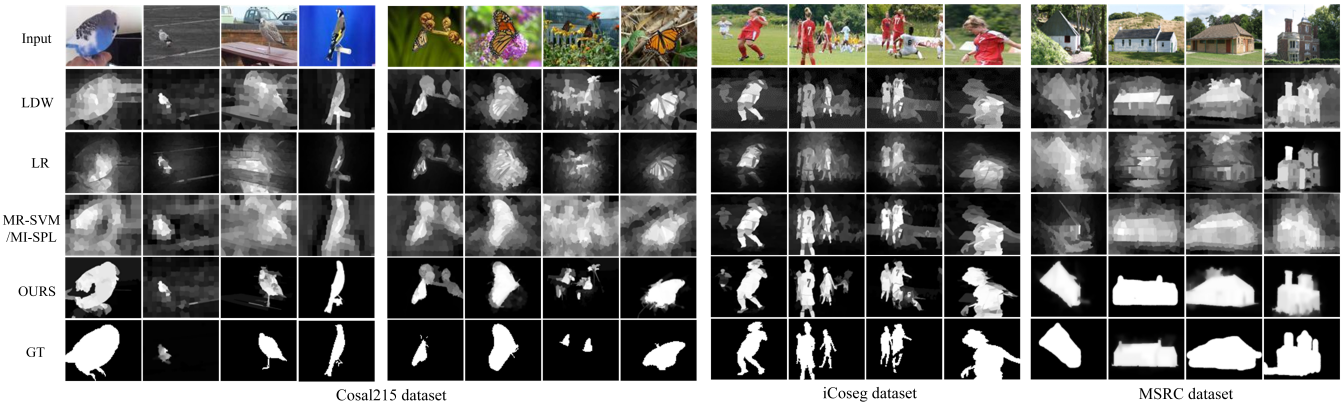


Figure 4: Illustration of sample co-saliency maps by the proposed FASS and multiple representative state-of-the-art methods. Note that the saliency maps on Cosal2015 and MSRC in the fourth row are produced by the ML-SVM approach and the co-saliency maps on iCoseg are from MI-SPL.

or otherwise increase δ during the iteration. The setting for the parameter η in Eq. (9) is in the same way. We set the parameters α and γ in Eq. (4) to 1 and set the dimensionality d^l of $\boldsymbol{\beta}^l$ as $d^l = \frac{2}{3}m^l$ following the work [21].

Running Time. The experiments are run on a PC with an i5-3.3 GHz CPU, 8 GB of memory and a Titan 1080 Ti GPU. Our code is implemented in MATLAB and C without optimisation and the CNN feature extractor is implemented

in Python. Without the extraction of multi-view features and initial saliency map, the average running time per image for the iCoseg, Coseg2015 and MSRC datasets are 10.75 s, 13.28 s and 9.53 s, respectively. This indicates that the proposed FASS is efficient.

4.2 Comparison to State-of-the-Arts

We compare the proposed FASS approach to nine state-of-the-art co-saliency detection methods, ranging from unsupervised methods, *i.e.*, CS [7], HS [18], LR [23], LDW [32], BLSM [30], MV-SRCC [31] and TS [28], to a weakly supervised method MI-SPL [34] using image category label and a supervised method ML-SVM [10].

Figure 3 illustrates the performance comparison between these methods in terms of PR curve, AP score and F-measure. From these results, we can obtain the following observations. (a) The proposed FASS approach performs better than the state-of-the-art methods in terms of all the performance metric on the iCoseg and Cosal2015 datasets. It improves F-measure over the state-of-the-art methods on the MSRC dataset and has a marginal AP degradation compared to the supervised method ML-SVM; (b) It outperforms the best compared method ML-SVM by 2.80% and 7.90% in terms of AP score and F-measure respectively on the challenging Cosal2015 dataset. It achieves 5.34% performance improvement in terms of AP score on the iCoseg dataset compared to the best performed existing method LDW as well as 7.36% F-measure improvement compared to MI-SPL. On the MSRC dataset, it improves the compared method ML-SVM by 5.14% in terms of F-measure and has a marginal performance degradation of 0.58% in terms of AP score.

Figure 4 shows some sample co-saliency maps produced by the proposed FASS approach and four competitive existing methods including LDW [32],LR [23],ML-SVM [10] and MI-SPL [34]. It illustrates the co-saliency maps for the image groups of "bird" and "butterfly" in Cosal2015, "red-clothed athletes" in iCoseg and "house" in MSRC. We can see that the proposed FASS approach generates much more accurate co-saliency maps than the existing methods. Although the CFs in each image group have large variations in pose, shape, color and point-of-view, FASS detects and localizes the co-salient objects accurately. It has almost no background regions wrongly detected as foreground.

4.3 Ablation Study

In this section, we evaluate the effectiveness of the components within FASS and the robustness of it. We conduct evaluation on the most challenging dataset, *i.e.*, Cosal2015.

We first implement two variants of FASS, *i.e.*, "FASS_\MVL" and "FASS_\MVL+GSO". "FASS_\MVL" refers to FASS without the component of multi-view feature learning. It conducts co-saliency prorogation and graph structure optimization based on the concatenation of multi-view features. FASS_\MVL+GSO refers to FASS without the components of both multi-view feature learning and graph structure optimization. It performs semi-supervised co-saliency prediction over a fixed group based on the concatenated multi-view features. For the sake of fair comparison, these two baseline variants have the same experimental settings with FASS. Figure 5 (a) shows the performance comparison on Cosal2015 in teams of AP score and F-measure. We can see that FASS_\MVL+GSO causes heave performance degradation as compared to FASS. "FASS_\MVL" outperforms FASS_\MVL+GSO, indicating the effectiveness of graph structure optimization. FASS performs better than "FASS_\MVL" and achieves the best performance. This demonstrates that the multi-view feature learning indeed improves co-saliency detection.

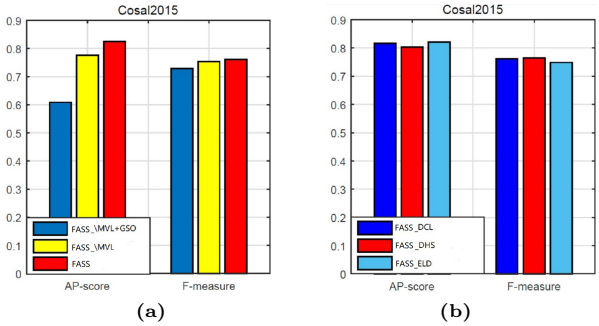


Figure 5: Evaluation of component effectiveness and robustness of FASS.

Moreover, we conduct experiment to investigate the robustness of FASS to the initial single-image saliency maps. We apply several single-image saliency detection methods including DCL[14], DHS[17] and ELD [13] to generate the initial saliency maps for FASS. Figure 5(b) shows the performance comparison and demonstrates the robustness of FASS.

5 CONCLUSIONS

This work proposes a novel feature-adaptive semi-supervised (FASS) framework for co-saliency detection. The FASS is able to predict accurate co-saliency map by a joint learning of multi-view features, graph structure and co-saliency prorogation. Specially, the multi-view feature learning consists of both view-wise feature weighting and element-wise feature selection leads to effective representation robust to feature noise and redundancy as well as adaptive to task at hand. The graph structure optimization offers an optimal graph that represents underlaying inter-region correlation precisely and comprhensively. Thus, our proposed FASS method is able to generate satisfactory co-saliency map based on the effective exploration of multi-view features as well as the correlation among regions. We conduct extensive experiments to evaluate the proposed FASS approach on three widely used co-salience datasets, *i.e.*, Cosal2015, iCoseg and MSRC. Experimental results have shown that the FASS outperforms the state-of-the-art methods.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No.: 61622211, 61472392, and 61620106009 as well as the Fundamental Research Funds for the Central Universities under Grant No.: WK2100100030.

REFERENCES

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, Sabine Süsstrunk, et al. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence* 34, 11 (2012), 2274–2282.

[2] Xiaochun Cao, Yupeng Cheng, Zhiqiang Tao, and Huazhu Fu. 2014. Co-Saliency Detection via Base Reconstruction. (2014), 997–1000.

[3] Xiaochun Cao, Zhiqiang Tao, Bao Zhang, Huazhu Fu, and Wei Feng. 2014. Self-adaptively weighted co-saliency detection via rank constraint. *IEEE Transactions on Image Processing* 23, 9 (2014), 4175–4186.

[4] Kai-Yueh Chang, Tyng-Luh Liu, and Shang-Hong Lai. 2011. From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In *Computer vision and pattern recognition (cvpr), 2011 ieee conference on*. IEEE, 2129–2136.

[5] Hwann-Tzong Chen. 2010. Preattentive co-saliency detection. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 1117–1120.

[6] Alon Faktor and Michal Irani. 2013. Co-segmentation by composition. In *Proceedings of the IEEE International Conference on Computer Vision*. 1297–1304.

[7] Huazhu Fu, Xiaochun Cao, and Zhuowen Tu. 2013. Cluster-based co-saliency detection. *IEEE Transactions on Image Processing* 22, 10 (2013), 3766–3778.

[8] Huazhu Fu, Dong Xu, Stephen Lin, and Jiang Liu. 2015. Object-based RGBD image co-segmentation with mutex constraint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4428–4436.

[9] Huazhu Fu, Dong Xu, Bao Zhang, Stephen Lin, and Rabab Kreidieh Ward. 2015. Object-based multiple foreground video co-segmentation via multi-state selection graph. *IEEE Transactions on Image Processing* 24, 11 (2015), 3415–3424.

[10] Junwei Han, Gong Cheng, Zhenpeng Li, and Dingwen Zhang. 2017. A unified metric learning-based framework for co-saliency detection. *IEEE Transactions on Circuits and Systems for Video Technology* (2017).

[11] Shi-Min Hu, Tao Chen, Kun Xu, Ming-Ming Cheng, and Ralph R Martin. 2013. Internet visual media processing: a survey with graphics and vision applications. *The Visual Computer* 29, 5 (2013), 393–405.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[13] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. 2016. Deep saliency with encoded low level distance map and high level features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 660–668.

[14] Guanbin Li and Yizhou Yu. 2016. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 478–487.

[15] Hongliang Li, Fanman Meng, and King Ngi Ngan. 2013. Co-salient object detection from multiple images. *IEEE Transactions on Multimedia* 15, 8 (2013), 1896–1909.

[16] Hongliang Li and King Ngi Ngan. 2011. A co-saliency model of image pairs. *IEEE Transactions on Image Processing* 20, 12 (2011), 3365–3375.

[17] Nian Liu and Junwei Han. 2016. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 678–686.

[18] Zhi Liu, Wenbin Zou, Lina Li, Liquan Shen, and Olivier Le Meur. 2014. Co-saliency detection based on hierarchical segmentation. *IEEE Signal Process. Lett* 21, 1 (2014), 88–92.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

[20] Feiping Nie, Guohao Cai, and Xuelong Li. 2017. Multi-View Clustering and Semi-Supervised Classification with Adaptive Neighbours.. In *AAAI*. 2408–2414.

[21] Feiping Nie, Wei Zhu, Xuelong Li, et al. 2016. Unsupervised Feature Selection with Structured Graph Optimization.. In *AAAI*. 1302–1308.

[22] Rong Quan, Junwei Han, Dingwen Zhang, and Feiping Nie. 2016. Object co-segmentation via graph optimized-flexible manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 687–695.

[23] Xiaohui Shen and Ying Wu. 2012. A unified approach to salient object detection via low rank matrix recovery. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 853–860.

[24] Hangke Song, Zhi Liu, Yufeng Xie, Lishan Wu, and Mengke Huang. 2016. RGBD co-saliency detection via bagging-based clustering. *IEEE Signal Processing Letters* 23, 12 (2016), 1722–1726.

[25] Zhiyu Tan, Liang Wan, Wei Feng, and Chi-Man Pun. 2013. Image co-saliency detection by propagating superpixel affinities. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2114–2118.

[26] Zhiqiang Tao, Hongfu Liu, Huazhu Fu, and Yun Fu. 2017. Image Cosegmentation via Saliency-Guided Constrained Clustering with Cosine Similarity.. In *AAAI*. 4285–4291.

[27] Wenguan Wang, Jianbing Shen, Xuelong Li, and Fatih Porikli. 2015. Robust video object cosegmentation. *IEEE Trans. Image Processing* 24, 10 (2015), 3137–3148.

[28] Zuyi Wang and Lihe Zhang. 2017. Two-Stage Co-Salient Object Detection. In *2017 10th International Conference on Intelligent Computation Technology and Automation (ICICTA)*. IEEE, 287–290.

[29] Lina Wei, Shanshan Zhao, Omar El Farouk Bourahla, Xi Li, and Fei Wu. 2017. Group-wise deep co-saliency detection. *arXiv preprint arXiv:1707.07381* (2017).

[30] Yulin Xie, Huchuan Lu, and Ming-Hsuan Yang. 2013. Bayesian saliency via low and mid level cues. *IEEE Transactions on Image Processing* 22, 5 (2013), 1689–1698.

[31] Xiwen Yao, Junwei Han, Dingwen Zhang, and Feiping Nie. 2017. Revisiting co-saliency detection: a novel approach based on two-stage multi-view spectral rotation co-clustering. *IEEE Trans. Image Process* 26, 7 (2017), 3196–3209.

[32] Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang, and Xuelong Li. 2016. Detection of co-salient objects by looking deep and wide. *International Journal of Computer Vision* 120, 2 (2016), 215–232.

[33] Dingwen Zhang, Deyu Meng, and Junwei Han. 2017. Co-saliency detection via a self-paced multiple-instance learning framework. *IEEE transactions on pattern analysis and machine intelligence* 39, 5 (2017), 865–878.

[34] Dingwen Zhang, Deyu Meng, Chao Li, Lu Jiang, Qian Zhao, and Junwei Han. 2015. A self-paced multiple-instance learning framework for co-saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 594–602.