Revisiting Co-Saliency Detection: A Novel Approach Based on Two-Stage Multi-View Spectral Rotation Co-clustering

Xiwen Yao, Junwei Han, *Senior Member, IEEE*, Dingwen Zhang, and Feiping Nie

***Abstract*— With the goal of discovering the common and salient objects from the given image group, co-saliency detection has received tremendous research interest in recent years. However, as most of the existing co-saliency detection methods are performed based on the assumption that all the images in the given image group should contain co-salient objects in only one category, they can hardly be applied in practice, particularly for the large-scale image set obtained from the Internet. To address this problem, this paper revisits the co-saliency detection task and advances its development into a new phase, where the problem setting is generalized to allow the image group to contain objects in arbitrary number of categories and the algorithms need to simultaneously detect multi-class co-salient objects from such complex data. To solve this new challenge, we decompose it into two sub-problems, i.e., how to identify subgroups of relevant images and how to discover relevant co-salient objects from each subgroup, and propose a novel co-saliency detection framework to correspondingly address the two sub-problems via two-stage multi-view spectral rotation co-clustering. Comprehensive experiments on two publically available benchmarks demonstrate the effectiveness of the proposed approach. Notably, it can even outperform the state-of-the-art co-saliency detection methods, which are performed based on the image subgroups carefully separated by the human labor.**

***Index Terms*— Co-clustering, multi-class salient object detection.**

I. Introduction

With the goal of discovering the common and salient objects from the given image groups, co-saliency detection has received growing attention in recent years. Compared with the traditional saliency detection (Fig.1 (a)), co-saliency detection (Fig.1 (b)) needs to additionally consider the group-level information among multiple relevant images. Thus, it is more desirable yet challenging than the traditional saliency detection task. Compared with image co-segmentation [1], [2] which considers not only common salient foreground regions but also similar non-salient background areas in images, co-saliency detection focuses on exploring the most important information, i.e., the common foreground regions, among the image group with a reduced computational demand by implying priorities based on human visual attention. Thus, co-saliency tends to be a promising preprocessing step for many high-level visual information understanding tasks such as video foreground extraction [3], image retrieval [4], and object detection [5].

However, most of the existing co-saliency detection methods are performed based on a strong assumption that all the images in the given image group should contain co-salient objects in only one category such as pyramid, cheetah, building as shown in Fig.1 (b). They can hardly be applied in practice, particularly for the large-scale image set obtained from the internet, as such "clean" data can hardly be obtained unless expensive human labor is devoted to manually grouping those relevant images. For example, in the recently established real-world image recognition systems [6]–[8], one of the biggest challenges, i.e., how to leverage massive image data obtained from the social media portals is still under-addressed. This problem is mainly caused by the fact that such unconstrained image collection is so complex that it always contains diverse objects, e.g., "basket", "basketball", "player", "course", that are relevant to the searched entity, e.g., "basketball". Thus, if one can decompose such complex image collection into multiple more compact subgroups and subsequently learn the foreground model of the co-salient objects in each subgroup, it would become much easier to understand the image collection and learn object models for the desirable categories more precisely.

For the sake of the above analysis, this paper revisits the co-saliency detection task and propels its development into a new phase, where the problem setting is generalized to allow the image group to contain objects in arbitrary number of categories and the algorithms need to simultaneously detect multi-class co-salient objects from such complex data (see Fig.1 (c)). As can be seen, co-saliency detection under this weak assumption can better serve for the real-world multimedia and vision tasks whereas it encounters much more challenges. Based on our understanding, these challenges can be summarized into two folds: 1) how to identify subgroups of relevant images, which contain co-salient objects of the same

Manuscript received June 21, 2016; revised November 28, 2016 and March 29, 2017; accepted March 30, 2017. Date of publication April 13, 2017; date of current version May 9, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jing-Ming Guo. This work was supported in part by the National Science Foundation of China under Grant 61522207 and in part by the Excellent Doctorate Foundation of Northwestern Polytechnical University *(Corresponding author: Junwei Han.)*

X. Yao, J. Han, and D. Zhang are with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: yaoxiwen517@gmail.com; junweihan2010@gmail.com; zhangdingwen 2006yyy@gmail.com).

F. Nie is with the School of Computer Science and Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: feipingnie@gmail.com).

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

Digital Object Identifier 10.1109/TIP.2017.2694222

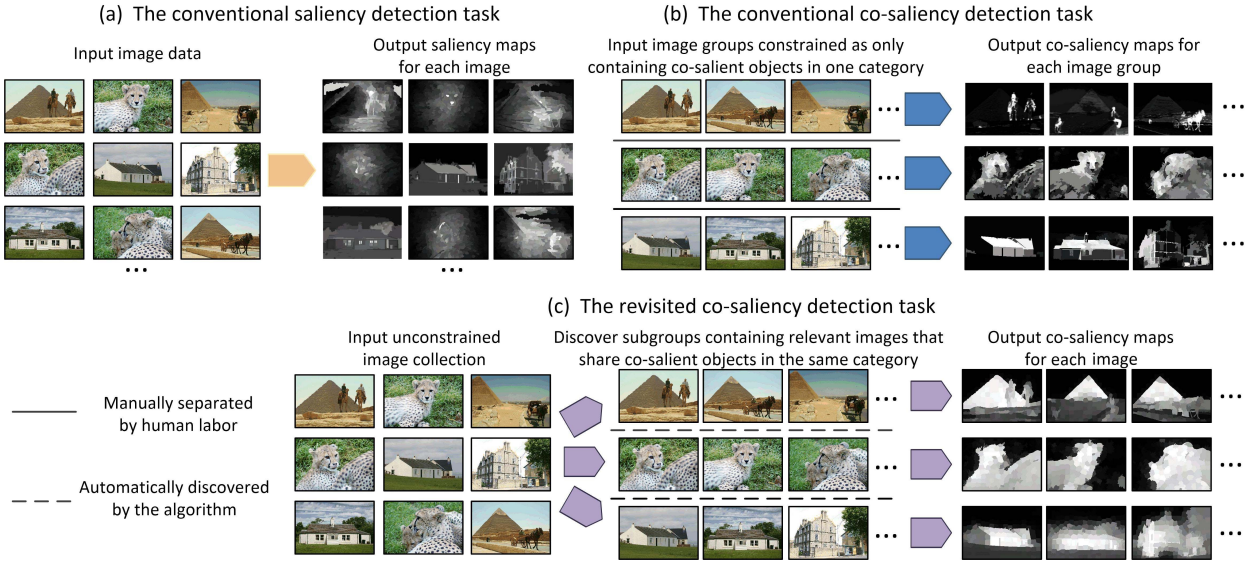

Fig. 1. Illustration of the difference between the revisited co-saliency detection task and the conventional saliency detection and co-saliency detection tasks.

class, from the cluttered image set; and 2) how to discover co-salient objects from the image scenes in each subgroup.

A naïve method that performs clustering in two successive stages may solve these two key problems. At the first stage, image scenes with similar appearances are clustered into the same subgroup by using conventional clustering methods such as k-means or spectral clustering. Then, the same clustering method is further employed on superpixels to separate co-salient objects from backgrounds in each subgroup. However, at the first stage, the naïve method may incorrectly group images with similar background but different foreground objects into the same subgroup. As for the second stage, similar superpixels coming from foreground objects and backgrounds also can be detrimental to discover co-salient objects in images of the same subgroup.

Through a closer look at the benchmark datasets as shown in Fig.1, we find that exploring the co-occurring relationships of object proposals (OP) with image scenes and superpixels in the two successive stages can mitigate the limitations of the naïve method. By introducing object proposals, we hope that 1) in the same cluster of image scenes, the inner object proposals can be grouped together, and 2) superpixels are clustered on the basis of the object proposals in which they co-occur. Unlike one-side clustering used in the naïve method which just groups similar images or superpixels, we here resort to co-clustering which makes use of relationships of OP with images and superpixels to simultaneously group them into image clusters, OP clusters, and superpixel clusters. As far as we know, we make the earliest effort to introduce co-clustering to model useful co-occurring relationship in multi-class co-salient object detection task. We also demonstrate that co-clustering is an effective way for solving the aforementioned two challenging problems.

Although co-clustering has shown impressive performance improvement over traditional one-side clustering on document analysis [9]–[11], the traditional co-clustering often suffers from a severe result deviation from the true discrete solution. The main reason is that the traditional co-clustering method [9] adopts k-means to obtain the final cluster indicator matrix (each column of this matrix gives the cluster to which each item was assigned) from the co-occurring matrix and the potential flaw of k-means objective function cannot guarantee accurate transform from continuous co-clustering solution to discrete cluster indicator matrix. To address this problem, we impose an additional orthonormal constraint to the objective function and thus obtain a spectral rotation invariant property, which can guarantee that the final cluster indicator matrix could best approximate the continuous co-clustering solution. Moreover, instead of using a single feature type to perform co-clustering in those two stages, it is of great interests to combine multiple complementary feature modalities in co-clustering to further improve the clustering result. In this paper, we refer to the modified co-clustering method as multi-view spectral rotation co-clustering method (MV-SRCC) and show that it can effectively solve the problems in each of the co-clustering stage in our framework.

The concrete framework of the proposed algorithm is shown in Fig.2. In the first stage, given a cluttered image set consisting of images from diverse categories, we first extract OPs from each image via objectness measure method [12]. Then, co-occurring matrix is constructed based on the relationship between OPs and the image scenes. After performing MV-SRCC with this matrix, we can obtain the subgroups of the relevant images in this stage. In the second stage, we first extract superpixel regions from each image in the obtained same subgroup by using SLIC method [13] and then construct another co-occurring matrix based on the relationship between superpixel regions and previous extracted OPs. Afterwards, MV-SRCC is performed on this co-occurring matrix to derive clusters of superpixels. To compute the co-saliency of each superpixel cluster, an efficient computation method is proposed by jointly considering the cluster's repetitiveness among

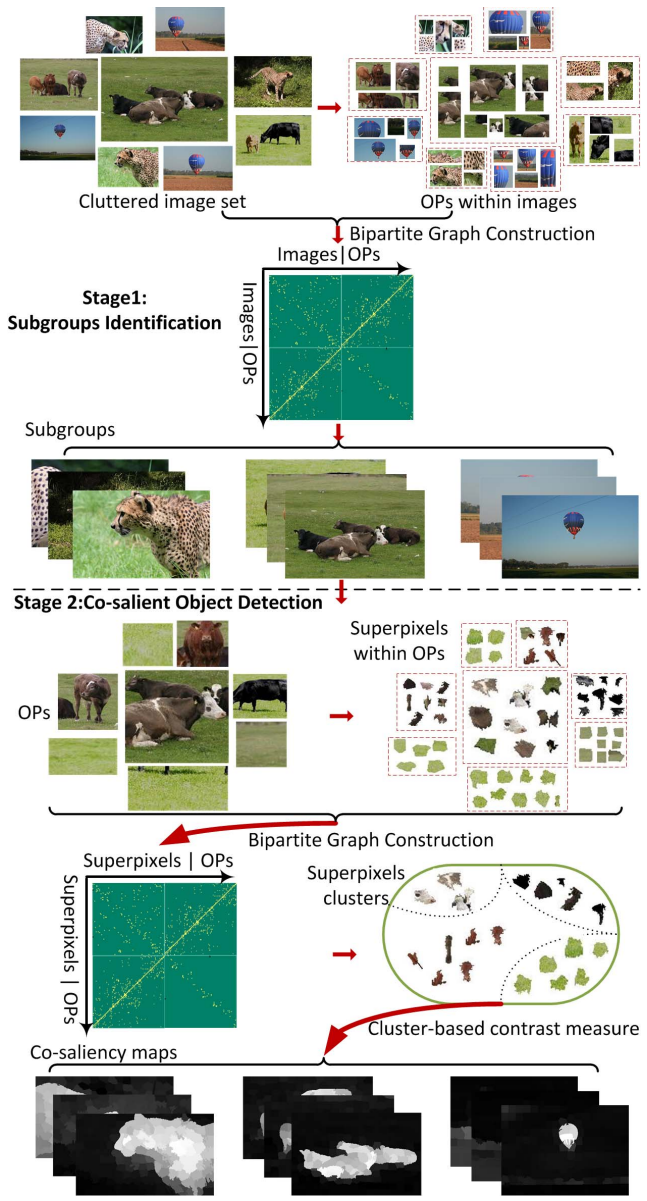

Fig. 2. The framework of the proposed co-saliency detection approach, where "OPs" denotes the object proposals.

multiple images and compactness in an image. The key motivation is based on the observation that superpixels belonging to co-salient foreground cluster are not only widely distributed over the majority of images in the subgroup but also reveals a more compact distribution within an individual image compared with superpixels of background cluster. Finally, the pixel-level co-saliency map is obtained by applying spatial refinement on the raw cluster-level saliency to promote gathering salient pixels together within an image.

To sum up, this paper mainly has four-fold contributions:

1) We revisit the co-saliency detection task and propel its development into a new phase, where the problem setting is generalized to allow the image group to contain objects in arbitrary number of categories and the algorithms need to simultaneously detect multi-class co-salient objects from the unconstrained image collection.

2) We naturally formulate the multi-class co-salient object detection as a problem of exploring the co-occurring relationship among different levels of image data such as image scene, object proposals and superpixel regions, and then propose an effective solution using the novel co-clustering algorithm as described in contribution 3).

3) A novel co-clustering algorithm named multi-view spectral rotation co-clustering (MV-SRCC) is proposed, which takes advantage of spectral rotation invariant property and multiple complementary feature modalities to guarantee good clustering results.

4) Comprehensive experiments on two widely used benchmark datasets are conducted to demonstrate the effectiveness of the proposed co-clustering algorithm as well as the entire framework for the revisited co-saliency detection task. Notably, the performance of our approach can be even better than the conventional methods which need human labor to separate the image subgroup manually.

The rest of this paper is organized as follows. The next section gives a brief review of the related works. Section III describes the proposed MV-SRCC algorithm and Section IV gives the details of our multi-class co-salient object detection framework. Comprehensive experimental results are provided in Section V and Section VI concludes this paper.

II. Related Works

This section provides a brief review of recent works on co-saliency detection and co-clustering.

A. Co-Saliency Detection

Originally, co-saliency detection approaches [14]–[17] focused on detecting common salient objects from image pairs. For example, the earliest co-saliency method [15] was designed to detect common objects in a pair of images with similar backgrounds. However, the requirement of highly similar backgrounds limits its applicability as different backgrounds could be a more general case. In [14], a simple but efficient progressive algorithm which explored the joint information provided by the image pairs was proposed to detect co-salient objects. Tan *et al.* [17] first computed affinity matrix for superpixels of image pairs and then performed propagation on the affinity matrix using SimRank algorithm to derive co-saliency maps. In [16], co-saliency maps were obtained by linearly combining single-image saliency and multi-image saliency based on a co-multilayer graph model.

The aforementioned methods can only deal with the image groups containing two relevant images. In order to extend the co-saliency detection to work on image groups consisting of more than two relevant images, a number of co-saliency detection methods are proposed recently [18]–[30]. Some existing methods concentrate on fusing intra-image saliency and inter-image saliency to generate the final co-saliency maps. The intra-image saliency is used to describe the saliency for individual image in the group and the inter-image saliency is defined as the correspondence among multiple images. In [21], a multi-scale segmentation voting scheme

was exploited to compute intra-image saliency and a pairwise similarity ranking algorithm was proposed to measure inter-image saliency. Then the obtained intra- and inter-image saliency were linearly combined to produce the final co-saliency maps. To mitigate the limitation of method [21] that exploits a fusion of intra- and inter-image saliency with fixed weights, several methods [19], [23] attempted to self-adaptively combine multiple saliency cues via rank constraint. The key idea is that the feature representation of the co-salient regions should be both similar and consistent and thus the feature matrix appears low rank. The main limitation of [19] and [23] is that the performance relies on the power of the elementary saliency maps. Additionally, the hand-designed metrics used in [23] also limited its ability to handle more complex scenes. In [30], intra-image contrast was integrated with inter-image consistency in a principled Bayesian framework to compute co-saliency map. In [26], an unsupervised random forest was used to extract the rough contours of common objects, and then intra- and inter-image saliency map were fused in a way that takes advantage of multiplication. Instead of fusing intra- and inter-image saliency, methods [18], [24], [27] only computed intra-image saliency for each image. The work in [18] considered the salient regions frequently occurring in most images as co-salient. The work in [24] adjusted the intra-image saliency by exploiting the dissimilarities and global consistency of regions to obtain the co-saliency maps. The work in [27] fully exploited the intra-image saliency under a two-stage query scheme to guide co-saliency detection. However, these methods may suffer from accuracy degrade if the intra-image saliency is invalid. For method [22], a candidate set of co-salient regions was first discovered and further employed as reconstruction bases, and then the co-saliency map was obtained by computing the reconstruction error. Along this line of consideration, the candidate co-salient regions were defined as exemplars in [29] and their exemplar saliency was propagated to other matched regions to generate co-saliency maps. To capture more meaningful regions for effective saliency detection, the work in [25] proposed a hierarchical segmentation based co-saliency detection model. Intra-image saliency was measured on the fine segmentation and object prior saliency map was measured on the coarse segmentation, and the two maps were finally integrated to generate the co-saliency map.

Another group of related works are cluster-based co-saliency methods proposed in [20] and [28], which obtain global correspondence among multiple images by performing clustering. In [20], pixels were first clustered into different clusters and three bottom-up saliency cues (i.e., contrast, spatial, and corresponding) were exploited to measure the cluster-level saliency. Then, the cluster-level saliency was fused with intra-image saliency to yield the co-saliency map. The method [20] has been treated as a very simple and effective pre-processing step for many applications. However, as stated in [28], this kind of co-saliency detection method may suffer from the curse of dimensionality. To address this problem, in [28], a structured sparse PCA with feature selection scheme was proposed to improve the performance of cluster-based co-saliency detection. Although our work proposed in this paper is also a kind of cluster-based method, our work differs in the way of using more effective and novel clustering method and performing under more practical and challenging settings. Specially, we naturally cast the co-saliency task as a co-clustering problem by introducing co-occurring relationships among images, object proposals. and superpixels into clustering, which can achieve more reliable correspondence among multiple images. Moreover, we attempt to simultaneously detect multi-class co-salient objects from cluttered image sets instead of a given well-organized image group containing single-class co-saliency objects.

B. Co-Clustering

Compared with traditional one-side clustering algorithms such as popular k-means or spectral clustering, which only consider the information of the to-be-clustered data to perform clustering. Co-clustering algorithm also leverages co-occurring relationship to realize data division. This co-occurring relationship is established based on two different types of data, where one is affiliated to the other. By leveraging the co-occurring relationship from these two types of data, better clustering results can be obtained than the traditional ways. Co-clustering was first proposed in [9] to cluster a collection of unlabeled documents under the guidance of the co-occurring relationship between the documents and the words in them. Recently, co-clustering has shown impressive clustering results in various applications, such as image segmentation [31], video co-summarization [32] and hyperspectral image clustering [33]. The work in [31] constructed a co-occurring matrix to model the relationship between pixels and superpixels and this matrix was further decomposed by transfer cut method to perform clustering. In [32], a video was summarized by exploiting co-clustering to find shots that co-occur most frequently across videos. In [33], an unsupervised co-clustering framework was proposed to incorporate both the pixel spectral and spatial information to improve the clustering performance on hyperspectral image data. In this paper, we attempt to exploit co-clustering to tackle the co-saliency detection problem and introduce the relationship between image scene and object proposals, and the relationship between object proposals and superpixels to better identify subgroup image scenes constraining the same class of co-salient objects and detect the co-salient objects from the identified subgroup images.

However, the above existing co-clustering methods usually formulate the co-clustering task as a bipartite graph partitioning problem and obtain the final clusters by using k-means, which has a potential flaw that the obtained cluster indicator matrix might severely deviate from the real solutions. To remedy this problem, this paper proposes an efficient spectral rotation co-clustering method by taking advantage of the rotation invariant property to establish a better cluster indicator matrix. Moreover, complementary information of multiple feature modalities is also exploited to further improve co-clustering performance. Finally, the proposed co-clustering algorithm can be used in our framework to detect multi-class co-salient object effectively.

III. Multi-View Spectral Rotation Co-Clustering Algorithm

A. Formulation of MV-SRCC

As shown in Fig.2, the proposed MV-SRCC algorithm will be used in two stages of our co-saliency detection framework. In the first stage, MV-SRCC is performed with two multi-view data sets of image scenes and their inside object proposals to identify subgroups of relevant images. In the second stage, two multi-view data sets of object proposals and their inside superpixels are fed to MV-SRCC to obtain superpixels clusters. Next, we will give the detailed formulation about the proposed MV-SRCC algorithm.

We firstly define a view as a feature representation extracted from one convolutional layer of convolutional neural networks (CNNs) and can obtain two multi-view data sets $\mathcal{X} = \{\mathbf{X}^{(1)}, ..., \mathbf{X}^{(V)}\}$ and $\mathcal{Y} = \{\mathbf{Y}^{(1)}, ..., \mathbf{Y}^{(V)}\}$ from V views in each stage. The ith sample of υth view from data $\mathcal{X}$ is denoted as $\mathbf{x}_i^{(\upsilon)} \in \mathrm{R}^{n_\upsilon \times 1}$. Similarly, the jth sample of υth view from data $\mathcal{Y}$ is denoted as $\mathbf{y}_j^{(\upsilon)} \in \mathrm{R}^{n_\upsilon \times 1}$. All the N samples of υth view is denoted as $\mathbf{X}^{(\upsilon)} = [x_1^{(\upsilon)}, x_2^{(\upsilon)}, ..., x_N^{(\upsilon)}] \in \mathrm{R}^{n_\upsilon \times N}$ and all the M samples of υth view is denoted as $\mathbf{Y}^{(\upsilon)} = [y_1^{(\upsilon)}, y_2^{(\upsilon)}, ..., y_M^{(\upsilon)}] \in \mathrm{R}^{n_\upsilon \times M}$ The goal of the proposed MV-SRCC algorithm is to simultaneously group $\mathcal{X}$ and $\mathcal{Y}$ into K co-clusters by fully making use of inter data relationships as well as intra data relationships. The co-clustering result is represented by cluster indicator matrix $\mathbf{G} \in \{0, 1\}^{(N+M)\times K}$, where $\mathbf{G}_{ik} = 1$ if the ith sample is assigned to the kth cluster and 0 otherwise.

Given view υ, we model a pair of $\mathbf{X}^{(\upsilon)}$ and $\mathbf{Y}^{(\upsilon)}$ as a weighted bipartite graph $\mathcal{G}^{(\upsilon)} = (\mathcal{V}^{(\upsilon)}, \mathcal{E}^{(\upsilon)}, \mathbf{W}^{(\upsilon)})$, where $\mathcal{V}^{(\upsilon)} = \mathbf{X}^{(\upsilon)} \cup \mathbf{Y}^{(\upsilon)}$ is the vertex, $\mathcal{E}^{(\upsilon)}$ is the edge set, and $\mathbf{W}^{(\upsilon)}$ is the adjacency matrix. $\mathbf{W}^{(\upsilon)}$ is constructed under the guidance of enforcing similarity smoothness of individual data and co-occurring interaction between different data. The construction is as follows:

$$\mathbf{W}^{(\upsilon)} = \begin{bmatrix} \mathbf{W}_{\mathbf{X}}^{(\upsilon)} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{W}_{\mathbf{Y}}^{(\upsilon)} \end{bmatrix}, \tag{1}$$

where $\mathbf{W}_{\mathbf{X}}^{(\upsilon)} \in \mathrm{R}^{N\times N}$ and $\mathbf{W}_{\mathbf{Y}}^{(\upsilon)} \in \mathrm{R}^{M\times M}$ are similarity matrices of $\mathbf{X}^{(\upsilon)}$ and $\mathbf{Y}^{(\upsilon)}$, respectively, and $\mathbf{C} \in \mathrm{R}^{N\times M}$ is co-occurring matrix to measure the relationship between $\mathbf{X}^{(\upsilon)}$ and $\mathbf{Y}^{(\upsilon)}$.

In these matrices, $\mathbf{W}_{\mathbf{X}}^{(\upsilon)}$ is defined as:

$$\mathbf{W}_{\mathbf{X}}^{(\upsilon)} = \begin{cases} \exp\left(-\rho d\left(\mathbf{x}_i^{(\upsilon)}, \mathbf{x}_j^{(\upsilon)}\right)\right), & if\ \mathbf{x}_i^{(\upsilon)} \in \mathcal{N}_K\left(\mathbf{x}_j^{(\upsilon)}\right) \\ & or\ \mathbf{x}_j^{(\upsilon)} \in \mathcal{N}_K\left(\mathbf{x}_i^{(\upsilon)}\right) \\ 0, & otherwise, \end{cases} \tag{2}$$

where $d(\cdot\ ,\ \cdot)$ is the Euclidean distance measure function, ρ is the bandwidth value which is set to the median of all distance values in our paper, and $\mathcal{N}_K(\mathbf{x})$ represents the K nearest neighbors of $\mathbf{x}$ in the given distance measure function. Similarly, $\mathbf{W}_{\mathbf{Y}}^{(\upsilon)}$ is defined analogously to $\mathbf{W}_{\mathbf{X}}^{(\upsilon)}$.

In [31], the entry $\mathbf{C}_{ij}$ of co-occurring matrix is assigned with a fixed value if $\mathbf{x}_i^{(\upsilon)}$ is included in $\mathbf{y}_j^{(\upsilon)}$ and 0 otherwise. In other words, the elements in $\mathbf{y}_j^{(\upsilon)}$ with the same weights are considered as equally important. In this paper, we attempt to assign the relationship between $\mathbf{x}_i^{(\upsilon)}$ and $\mathbf{y}_j^{(\upsilon)}$ with different values by further analyzing the specific characteristics of $\mathbf{x}_i^{(\upsilon)}$ and $\mathbf{y}_j^{(\upsilon)}$. Thus, the co-occurring matrix $\mathbf{C}$ is defined as:

$$\mathbf{C} = \begin{cases} f(\mathbf{x}_i^{(\upsilon)}, \mathbf{y}_j^{(\upsilon)}), & if\ \mathbf{x}_i^{(\upsilon)} \in \mathbf{y}_j^{(\upsilon)} \\ 0, & otherwise, \end{cases} \tag{3}$$

where $f(\mathbf{x}_i^{(\upsilon)}, \mathbf{y}_j^{(\upsilon)})$ is a function to measure the importance of $\mathbf{x}_i^{(\upsilon)}$ for $\mathbf{y}_j^{(\upsilon)}$. We will give more details about this function in next section.

To get the co-clustering result matrix $\mathbf{G}^{(\upsilon)}$ in view υ, we firstly construct the Laplacian matrix $\mathbf{L}^{(\upsilon)} = \mathbf{D}^{(\upsilon)} - \mathbf{W}^{(\upsilon)}$, where $\mathbf{D}^{(\upsilon)} \in \mathrm{R}^{(N+M)\times(N+M)}$ is the degree matrix with $\mathbf{D}^{(\upsilon)} = diag((\mathbf{W}^{(\upsilon)})^T e_{N+M})$. Then, similar to spectral clustering, the single-view co-clustering problem can be transformed to solve the following objective function:

$$\min_{(\mathbf{G}^{(\upsilon)})^T\mathbf{G}^{(\upsilon)}=\mathbf{I}} Tr((\mathbf{G}^{(\upsilon)})^T\mathbf{L}^{(\upsilon)}\mathbf{G}^{(\upsilon)}). \tag{4}$$

To naturally integrate multiple features, we further propose a unified objective function to simultaneously minimize the co-clustering error of each view and the differences between the multi-view co-clustering result and each single-view co-clustering result. The objective function is defined as follows:

$$\begin{aligned} \min_{\mathbf{G}^T\mathbf{G}=\mathbf{I}, \mathbf{G}^{(\upsilon)}} & \sum_{\upsilon=1}^{V} Tr((\mathbf{G}^{(\upsilon)})^T\mathbf{L}^{(\upsilon)}\mathbf{G}^{(\upsilon)}) \\ & + \alpha Tr((\mathbf{G} - \mathbf{G}^{(\upsilon)})^T(\mathbf{G} - \mathbf{G}^{(\upsilon)})), \end{aligned} \tag{5}$$

where α is the penalty parameter. Thus, given the Laplacian matrix of each view, we try to learn the cluster indicator matrix for each view and cluster indicator matrix for the multiple views simultaneously.

B. Solution of MV-SRCC

Eq. (5) can be equivalent to minimizing the following optimization problem in Eq. (6). For clarity, the detailed derivation process is provided in the appendix.

$$\min_{\mathbf{G}^T\mathbf{G}=\mathbf{I}} Tr(\mathbf{G}^T\mathbf{P}\mathbf{G}), \tag{6}$$

where multi-view Laplacian matrix $\mathbf{P}$ is set to be $\mathbf{P} = \sum_{\upsilon=1}^{V} \mathbf{I} - \alpha(\mathbf{L}^{(\upsilon)} + \alpha\mathbf{I})^{-1}$.

Similar to spectral clustering, the optimal continuous solution $\mathbf{H}$ of the above function is the collection of eigenvectors corresponding to the top smallest K eigenvalues of $\mathbf{P}$. Since $\mathbf{H}$ is in relaxed continuous form, it is a common practice to apply K-Means to $\mathbf{H}$ to obtain the final discrete solution $\mathbf{G}$. However, the potential flaw of such common practice is that the obtained cluster indicator matrix might severely deviate from the real solutions, which would lead to the sub-optimal clustering results of images or superpixels in the two stages. Inspired by [34], we take advantage of the spectral solution invariance property to obtain the final discrete indicator

Algorithm 1 Multi-View Spectral Rotation Co-Clustering Algorithm

Input: two data sets $\mathcal{X} = \{\mathbf{X}^{(1)}, ..., \mathbf{X}^{(V)}\}$ and $\mathcal{Y} = \{\mathbf{Y}^{(1)}, ..., \mathbf{Y}^{(V)}\}$ from V views

Output: Cluster indicator matrix $\mathbf{G}$

1: Construct the adjacency matrix $\mathbf{W}^{(v)}$ for each view $v \in 1, ..., V$ by using Eq.(1)

2: Calculate the Laplacian matrix $\mathbf{L}^{(v)}$ for each view $v \in 1, ..., V$ by $\mathbf{L}^{(v)} = \mathbf{D}^{(v)} - \mathbf{W}^{(v)}$

3: Obtain multi-view Laplacian matrix $\mathbf{P}$ by $\mathbf{P} = \sum_v \mathbf{I} - \alpha(\mathbf{L}^{(v)} + \alpha\mathbf{I})^{-1}$

4: Calculate the K smallest eigenvectors of $\mathbf{P}$ and obtain the continuous solution $\mathbf{H}$

5: Obtain $\mathbf{G}$ by minimizing Eq.(7) via the alternative optimization method proposed in [34]

return $\mathbf{G}$

matrix $\mathbf{G}$ from continuous solution $\mathbf{H}$ by minimizing the following objective function

$$\min_{\mathbf{G},\mathbf{R}} ||\mathbf{HR} - \mathbf{G}||_F^2 \quad s.t. \mathbf{R}^T\mathbf{R} = \mathbf{I}, \tag{7}$$

where $\mathbf{R}$ is an arbitrary orthonormal matrix. Based on the rotation invariant property, $\mathbf{HR}$ is also the demanded continuous solution for any solution $\mathbf{H}$. As can be seen, this function aims to find a $\mathbf{G}$ which can best approximate $\mathbf{HR}$ among all discrete cluster indicator matrices under the orthonormal constraint $\mathbf{R}^T\mathbf{R} = \mathbf{I}$. Compared with directly applying K-Means, the spectral rotation technique can lead to an improvement in clustering accuracy, which motivates us to use it to obtain the co-clustering indicator matrix $\mathbf{G}$ in this paper. We resort to the alternative optimization method proposed in [34] to obtain $\mathbf{G}$ and $\mathbf{R}$. The overall procedure of the proposed co-clustering algorithm is shown in Algorithm 1.

IV. Multi-Class Co-Salient Object Detection

In this section, we will give the detailed descriptions about applying the proposed MV-SRCC algorithm to perform co-salient object detection. Firstly, how to extract multi-view features for image scenes, object proposals and superpixels by using CNN model is introduced. Then, two stages of the proposed co-salient objects detection framework are described in detail.

A. Multi-View Features

Recently, CNN has shown significant performance improvement for many computer vision tasks compared with the traditional hand-designed features. Moreover, combination of multiple layers of convolutional neural networks has proven beneficial in vision tasks such as pedestrian detection [35], fine-grained localization [36] and semantic segmentation [37]. The main reason is that multi-level representation can be obtained from different layers of CNN. For example, the bottom convolutional layers of CNN mainly illustrate the saliency from low-level image representation (e.g. edge information) whereas the middle layers can capture coarse-grained shape information and the top layers can show high-level semantic object information. Inspired by this, in this paper, we consider the multiple convolutional layers as multi-view features and try to exploit the complementary information across multiple levels to boost clustering performance.

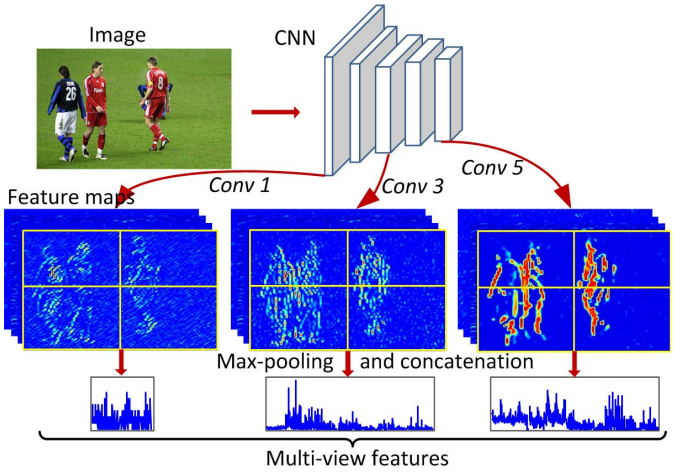

Fig. 3. Illustration of multi-view feature extraction for an image scene.

The structure and model parameters of the network used in our paper are identical to the CNN-S model in [38], which contains five convolutional layers and three fully-connected layers. Here, we use *conv1* (96 channels), *conv2* (256 channels), *conv3* (512 channels), *conv4* (512 channels), *conv5* (512 channels) to denote the convolutional layers, respectively. Of these layers, conv1, conv3 and conv5 are employed to extract multi-view features of the images and the corresponding object proposals and superpixels. Next, we present the details about how to extract multi-view features for image scenes, object proposals and superpixels, respectively.

As shown in Fig.3, when applying the CNN model to an image scene, each convolutional layer can result in a number of feature maps with one corresponding to each channel filter. The size (i.e. height and width) of the feature maps varies with the convolutional layers. We upsampled all the feature maps to the size of the image by using bilinear interpolation.

Given the resized feature maps produced by a certain convolutional layer, we define a compact robust representation to describe image scenes by encoding the global spatial relations between the underlying parts of image scenes. Specifically, a 2×2 spatial grid cell partitioning scheme is employed on the resized feature maps. Then, a channel-wise max pooling is performed on all pixels in each grid cell of feature maps. Afterwards, the max-pooled features in each grid cell are concatenated together to serve as the image-level features of the current convolutional layer. By repeating the above process on the resized feature maps of other convolutional layers, we can finally obtain multiple types of features for an image scene.

As for extracting multi-view features of object proposals and superpixels, the only difference is that we need to perform the above procedure on the masked feature maps of object

proposals and superpixels. The masked feature maps are obtained as follows. Firstly, binary maps are produced according to the position of each object proposal and suprepixel in the image scenes. Then, these binary maps are used to mask the feature maps of image scenes to obtain a set of masked feature maps for object proposals and superpixels.

B. Subgroup Identification

In this subsection, MV-SRCC is performed to group relevant images from a cluttered image set into the same subgroup by using the co-occurring relationship between image scenes and object proposals.

The first step is to extract a number of object proposals from each image by adopting objectness measure method [12], which measures the probability of any image window containing a generic object. The objectness measure is designed based on the general properties of an object as shown in an image, such as well-defined closed boundary, unique and salient appearance. We randomly sample 10000 windows over each image, and assign each window w with a probability score $p(w)$ to indicate its objectness score. Thereafter, we select the top largest N_{op} windows to serve as object proposals.

The second step is to extract multi-view features data of image scenes $\mathcal{X}$ and object proposals $\mathcal{Y}$ as described in Section IV.A. For images feature sets in all V views $\mathcal{X} = \{\mathbf{X}^{(1)}, ..., \mathbf{X}^{(V)}\}$, $\mathbf{X}^{(v)} = [x_1^{(v)}, x_2^{(v)}, ..., x_N^{(v)}] \in \mathrm{R}^{n_v \times N}$ denotes the feature matrix that contains the vth view feature data of all N images, where $\mathbf{x}_i^{(v)} \in \mathrm{R}^{n_v \times 1}$ is the feature vector of the ith image from vth view. For object proposals feature data sets $\mathcal{Y} = \{\mathbf{Y}^{(1)}, ..., \mathbf{Y}^{(V)}\}$, the vth view feature matrix $\mathbf{Y}^{(v)} = [y_1^{(v)}, y_2^{(v)}, ..., y_M^{(v)}] \in \mathrm{R}^{n_v \times M}$ contains the feature data of all $M = N \times N_{op}$ object proposals in the vthview, where $\mathbf{y}_j^{(v)} \in \mathrm{R}^{n_v \times 1}$ denotes the feature vector of the jthobject proposal in the vth view.

Given the two types of feature matrixes$\mathcal{X}$ and $\mathcal{Y}$ in V views, the next step is to construct a bipartite graph for a pair of $\mathbf{X}^{(v)}$ and $\mathbf{Y}^{(v)}$ in each view as formulated in Eq. (1) by jointly considering the co-occurrence relationships between images and their inside object proposals as well as the inter-image relationships and inter-OP relationships. In the construction, the inter-image and inter-OP relationships are represented in the form of similarity matrix $\mathbf{W}_{\mathbf{X}}^{(v)}$ and $\mathbf{W}_{\mathbf{Y}}^{(v)}$ as defined in Eq. (2). For co-occurring matrix $\mathbf{C}$, instead of assigning the relationship between the image and its inside OPs with a fixed value as in [31] to model the co-occurring relationship, we design a function $f(\cdot)$ in Eq. (3) to measure the importance of each object proposal for their image. Intuitively, the higher the objectness score is, the more important the object proposal is for its image. Thus, in this stage, we simply exploit the objectness score calculated by [12] to serve as the output of the importance measure function.

Given the constructed bipartite graph in each view, we can obtain the cluster indicator matrix $\mathbf{G}_1 \in \{0, 1\}^{(N+M)\times K}$for both images and object proposals by following steps 2-5 as listed in Algorithm 1. For this stage, the cluster label for each image is listed in the top N rows of $\mathbf{G}_1$.Thus, the cluttered image set can be divided into subgroups of relevant images.

C. Co-Salient Object Detection

1) Group Superpixels: In this subsection, MV-SRCC is performed to jointly cluster superpixels and object proposals in each subgroup.

Firstly, for image I_q in the subgroup $\mathcal{I} = \{I_q\}_{q=1}^{Q}$ of Q images, we adopt SLIC (Simple Linear Iterative Clustering) method [13] to obtain N_q superpixels and use the same object proposals extracted in the first stage.

Secondly, multi-view feature data $\mathcal{X}$ and $\mathcal{Y}$ are extracted from superpixels and object proposals of each image in the subgroup $\mathcal{I}$ as described in Section IV.A. Here $\mathcal{X}$ denotes the feature sets of superpixels and $\mathcal{Y}$ still denotes the feature sets of object proposals. For superpixels feature sets in all views $\mathcal{X} = \{\mathbf{X}^{(1)}, ..., \mathbf{X}^{(V)}\}$, $\mathbf{X}^{(v)} = [x_1^{(v)}, x_2^{(v)}, ..., x_N^{(v)}] \in \mathrm{R}^{n_v \times N}$ contains the vth view feature data of all $N = N_q \times Q$ superpixels in subgroup $\mathcal{I}$, where $\mathbf{x}_i^{(v)} \in \mathrm{R}^{n_v \times 1}$is the$vth$ view feature data of the ith superpixel. Refer to Section III.B for the descriptions of $\mathcal{Y}$ and the only difference is the number of object proposals $M = Q \times N_{op}$.

Thirdly, the bipartite graph for a pair of $\mathbf{X}^{(v)}$ and $\mathbf{Y}^{(v)}$in each view is constructed by using Eq. (1). In the construction, we can obtain similarity matrices $\mathbf{W}_{\mathbf{X}}^{(v)}$ and $\mathbf{W}_{\mathbf{Y}}^{(v)}$according to the definitions in Eq.(2). As for co-occurring matrix $\mathbf{C}$measuring the importance of superpixels for their object proposals, the importance measure function$f(\cdot)$ is obtained as follows. The first step is to compute the objectness score $p(x)$ for each pixel x by summing the scores of all the object proposals that contain pixel x. Then, the objectness score $p(s)$ for each superpixel s is computed by averaging the scores of all the pixels that are included in superpixel s. Afterwards, the objectness scores of superpixels in the same object proposal are normalized to make the sum to be 1. Then, the normalized $p(s)$ is used to measure the importance value of superpixel s for its object proposal.

Finally, given the constructed multiple bipartite graphs, we can use Algorithm 1 to obtain the cluster indicator matrix $\mathbf{G}_2 \in \{0, 1\}^{(N+M)\times K}$ for both superpixels and object proposals. From the top N rows of $\mathbf{G}_2$, we can get K superpixel clusters $\{c_k\}_{k=1}^{K}$, where c_k is the cluster center vector of the kth cluster. Next, we describe how to measure the co-saliency for every superpixel cluster in details.

2) Cluster-Level Co-Saliency: In this subsection, we will detail the computation of co-saliency at cluster-level based on contrast cue. Contrast cue is widely used for saliency computation at pixel/region-level in a single image. The work in [20] further extended the traditional contrast computation to cluster-level and obtained satisfactory results. To enhance the cluster-level contrast computation, we introduce a weighted term by jointly considering the cluster's repetitiveness among multiple images and compactness in an image. The motivation is based on the observation that superpixels belonging to co-salient foreground cluster are not only widely distributed over the majority images in the subgroup but also compactly distributed within an individual image compared with superpixels of background cluster.

For a given cluster c_k, let $c_{k,q}$ be the set of superpixels that come from image I_q. We utilize the entropy to measure the

repetitiveness, which is defined as:

$$H(c_k) = -\sum_{q=1}^{Q} \pi(q) \log \pi(q) \tag{8}$$

where $\pi(q)$ is the proportion of $c_{k,q}$, *i.e.*, $\pi(q) = |c_{k,q}|/|c_k|$. $H(c_k)$ is maximized when cluster c_k is uniformly distributed and is 0 if its superpixels come from the same image. Larger $H(c_k)$ indicates that the cluster's superpixels appeared in more images in the group, which should be assigned with larger co-saliency value.

In a single image, superpixels of foreground cluster tend to be compactly distributed with a lower spatial variance than background superpixels. Hence we define the compactness by using its spatial variance:

$$\varphi(c_k) = \sum_{q=1}^{Q} \sum_{x_i \in c_{k,q}} ||\zeta_i - \mu_{k,q}||^2 \tag{9}$$

where ζ_i is the position of superpixel x_i and $\mu_{k,q}$ is the mean position of cluster c_k in image I_q.

By jointly considering the cluster's repetitiveness and compactness, the cluster's co-saliency can be computed from

$$S(c_k) = \sum_{c_i \neq c_k} \frac{\exp(H(c_i))}{\varphi(c_i)} d(c_i, c_k) \tag{10}$$

where $d(c_i, c_k)$ defines the Euclidean distance between cluster c_i and c_k. The co-saliency $S(c_k)$ is normalized to [0, 1].

According to Eq. (10), clusters containing salient objects appeared in majority of images are assigned with high weights. And the contrast of these clusters is enhanced. On the contrary, the contrast of background clusters is attenuated. Thus, the contrast difference between foreground and background clusters is effectively enlarged.

3) Pixel-Level Co-Saliency: Inspired by [39], we obtain the final pixel-level co-saliency map from the cluster-level map based on Gestalt laws, which suggests that the pixels which are close to the foci of attention should be more salient than far away pixels. Accordingly, we first obtain the foci of attention area Θ_q in image $I_q$by thresholding the normalized cluster-level co-saliency map S with $0.3 \times max(S)$ as suggested by [23]. Then, the final co-saliency for each pixel p in image I_q is defined as

$$S^*(p) = S(p) \cdot \exp(-\omega \cdot \min_{j \in \Theta_q} d(\tau_p, \tau_j)) \tag{11}$$

where τ_p is the spatial position of pixel p, $d(\tau_p, \tau_j)$ defines the spatial positional distance between pixel p and pixel j in Θ_q, and the scale parameter ω is set to 6 according to [23].

V. Experimental Results

In this section, we firstly introduce the experimental settings including the descriptions of two widely used benchmark datasets, the adopted evaluation metrics and the implementation details. Then, we compare the proposed framework with 8 state-of-the-art methods on two benchmark datasets. Finally, detailed evaluation of the proposed MV-SRCC in those two stages is given.

A. Experimental Settings

1) Datasets: In order to evaluate the performance of the proposed method, we conducted a set of qualitative and quantitative experiments on two benchmark datasets: the iCoseg dataset [40] and the MSRC dataset [41]. Specifically, the iCoseg dataset includes 38 image groups of totally 643 images, which is the largest public dataset for co-saliency detection until now. These images cover many semantic concepts such as air shows, pandas and hot balloons, etc. Manually labeled pixel-wise ground truth is available per image. The MSRC dataset includes 240 images belonging to 7 groups. Compared with iCoseg dataset, MSRC dataset appears to be more challenging as it contains co-salient objects with different colors or shapes.

2) Evaluation Metrics: To evaluate the clustering performance of the first stage, three standard measures consisting of Accuracy (ACC), Normalized Mutual Information (NMI), and Purity were exploited to measure the clustering performance [42]. ACC is defined as the number of correct clustering results divided by the number of all images. NMI assesses the level of agreement between two affinity matrices formed by the clustering labels and the ground truth. Purity is a simple evaluation measure. To compute this criterion, we first count the number of most frequent ground truth appearing in each cluster. Then this number is divided by the number of all image number to obtain the Purity.

To evaluate the performance of the generated final co-saliency maps, we adopted three widely used criteria: 1) the Precision-Recall (PR) curve which is drawn by using the precision rate versus the true positive rate (or the recall rate) at each threshold from 0 to 255 of co-saliency maps; 2) the Average Precision (AP) score which is obtained by calculating the area under the PR curve; 3) the F-measure which is defined by:

$$F_\beta = \frac{(1+\beta^2) \Pr ecision \times Recall}{\beta^2 \times \Pr ecision + Recall} \tag{12}$$

where we set $\beta^2 = 0.3$ as suggested in [21]. The Precision and Recall are calculated under the a self-adaptive threshold $T = \mu + \varepsilon$ that was also recommended in [43], where μ and ε are the mean value and the standard deviation of the co-saliency map, respectively.

3) Implementation Details : In the experiments, our method is implemented in MATLAB on a computer with an Intel i5-4590 CPU (3.3 GHz) and 16 GB RAM. For each image, the object proposals extraction was performed by objectness method [12] with the object proposals number N_{op} being 100 and the superpixel over segmentation was carried out by the SLIC method [13] with the superpixel number N_q being 200. To extract the multi-view features of the image and corresponding object proposals and superpixels, three convolutional layers (i.e., *conv1*, *conv3*, *conv5*) of CNN-S model were employed. To integrate multiple features, we chose $\alpha = 10$ as the weight in Eq. (5), which shows good performance in our empirical study.

Considering the problem investigated in our paper, the knowledge of which images belong to and how many images contained in a certain image group are unknown whereas the

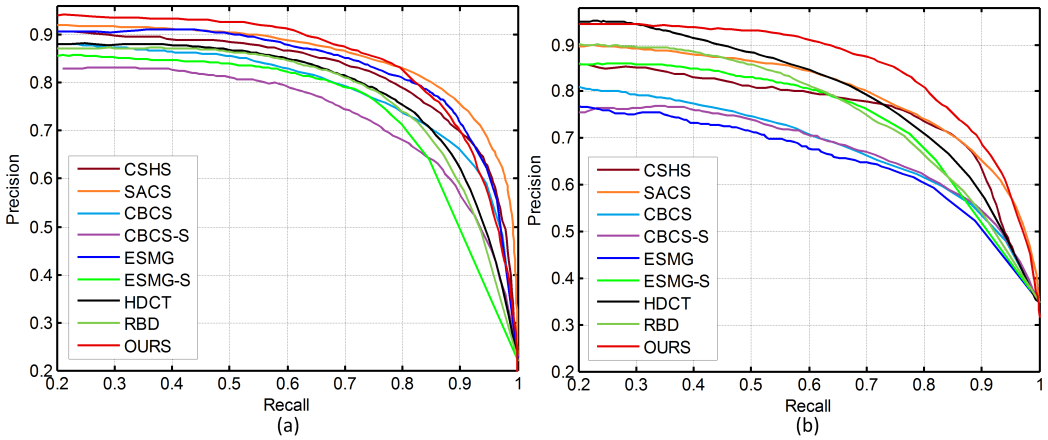

Fig. 4. Quantitative experimental results (in terms of the precision-recall curve) of the proposed approach and other state-of-the-art methods on the iCoseg (a) and MSRC (b) datasets.

TABLE I
AP AND F-MEASURE SCORES ON iCOSEG DATASET

Methods	HDCT[43]	RBD[44]	CSHS[25]	CBCS[20]	CSBS-S[20]	ESMG-S[27]	ESMG[27]	SACS[23]	OURS
AP	0.804	0.798	0.839	0.805	0.769	0.767	0.849	0.865	**0.868**
F-measure	0.770	0.780	0.754	0.740	0.715	0.752	0.801	0.793	**0.810**

TABLE II
AP AND F-MEASURE SCORES ON MSRC DATASET

Methods	HDCT[43]	RBD[44]	CSHS[25]	CBCS[20]	CSBS-S[20]	ESMG-S[27]	ESMG[27]	SACS[23]	OURS
AP	0.824	0.787	0.783	0.713	0.699	0.766	0.679	0.799	**0.853**
F-measure	0.720	0.712	0.711	0.588	0.620	0.706	0.616	0.711	**0.784**

number of categories in image sets is known. For example, there are 38 categories in iCoseg dataset including stonehenge, bear, elephants, etc. And the MSRC dataset includes 7 categories such as car, book, face, etc. This setting is more suitable in practice particularly for the large-scale image set obtained from the internet by inputting a keyword in social media portals such as Google and Flickr. In this context, for the first stage the cluster number is set to be the number of categories in image sets. For the second stage, we follow [20] to set the cluster number of superpixels asmin$\{3Q, 18\}$(where Q is the image number in each subgroup) because [20] has demonstrated the effectiveness of addressing co-saliency detection in large-scale data.

B. Comparison With Co-Saliency Detection Methods

To evaluate the performance of our framework for co-saliency detection, we compared our method with 8 state-of-the-art algorithms, i.e., HDCT [44], RBD [45], CSHS [25], CBCS [20], CBCS-S [20], ESMG [27], ESMG-S [27], and SACS [23], where the first two algorithms were proposed by the traditional saliency detection approaches while the last six algorithms were proposed by state-of-the-art co-saliency detection approaches. Moreover, CBCS-S and ESMG-S are the single image saliency detection method in CBCS [20] and ESMG [27]. It needs to further point out that our method was performed on the whole image dataset for simultaneously detecting multi-class co-salient objects whereas other co-saliency detection methods were performed on each subgroup of relevant images to detect each single class one by one. Those subgroups of images in methods [20], [23], [25], [27] may be pre-classified by human subjects.

1) Quantitative Performance Comparison: For quantitative evaluation, we compared the proposed approach with above 8 state-of-the-art methods. Fig.4 (a) and (b) show the corresponding PR curve performance of all the competing approaches on the iCoseg dataset and MSRC dataset separately. As can been seen, the proposed approach (OURS) performs the best at recall rates among [0, 0.8] on the iCoseg dataset and performs much better than other algorithms on the MSRC dataset. In addition, we also present the corresponding AP and F-measure scores of all state-of-the-art methods on the two benchmark datasets in Table 1 and 2. Clearly, our method can achieve the highest AP and F-measure values compared with other 8 state-of-the-art methods. On the iCoseg dataset, the second best co-saliency method (SACS) can obtain quite competitive performance. Our method only has a slight advantage compared with SACS, which achieves about 0.3% and 1.7% improvement in terms of AP and

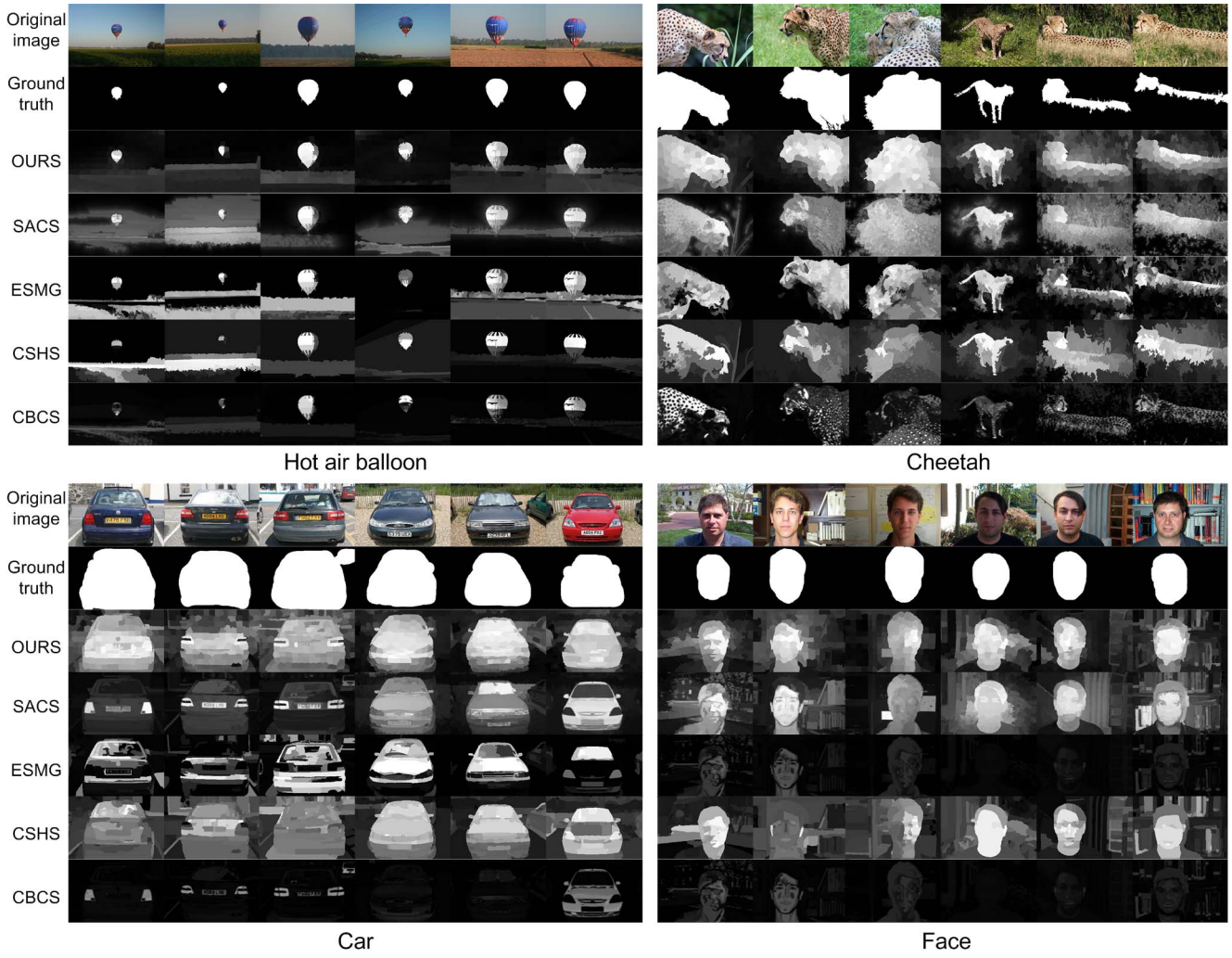

Fig. 5. Some examples of co-saliency detection on iCoseg and MSRC datasets. Hot air balloon and Cheetah in the first row belong to the iCoseg dataset. Car and Face in the second row belong to the MSRC dataset.

F-measure. Notice that the problem setting investigated in our paper is much complex than that investigated in other state-of-the-art methods. Our problem setting is to detect multi-class co-salient objects from "cluttered" image sets consisting of arbitrary number of categories whereas other state-of-the-art methods only perform detection from one "clean" image group containing common objects in one category. Even in this case, our method still achieves better performance than all other state-of-the-art methods, which clearly demonstrates the effectiveness of the proposed method. In addition, as shown in Table 2, our method provides a significant improvement for the more challenging MSRC dataset. On average, our method performs better than SACS by 5.4% and 7.3% in terms of AP and F-measure, as well as 2.9% and 6.4% compared with the second best method (HDCT) on the MSRC dataset. It is notable that our proposed multi-class co-salient object detection method even achieves much better performance than the best single-class co-salient object detection method on MRSC dataset.

2) Qualitative Performance Comparison: For an intuitive illustration, Fig.5 shows the co-saliency detection results of some examples in four image groups, i.e., Hot air balloon group and Cheetah group from iCoseg dataset, Car group and Face group from MSRC dataset. As can be seen, our method can obtain more accurate results than other co-saliency methods. For example, as shown in the Face group, our method still works well in the cases of complex backgrounds and can robustly suppress the background regions. The examples in the Cheetah group indicate that our method can accurately detect the co-salient objects even if they are in different viewpoints, scales and shapes.

C. Evaluation of MV-SRCC

In this subsection, we present the detailed evaluation of the proposed MV-SRCC algorithm for the two stages on iCoseg dataset. Two traditional clustering methods (i.e., k-means [46] and spectral clustering [47]) and SRCC algorithm proposed in this paper were employed as baseline clustering methods for performance comparisons.

1) Evaluation of MV-SRCC in Stage 1: To evaluate the effectiveness of the proposed MV-SRCC in the first stage for identifying subgroups, all these three clustering methods were used for clustering images based on the image-level features extracted from three convolutional layers (i.e., *conv1*,

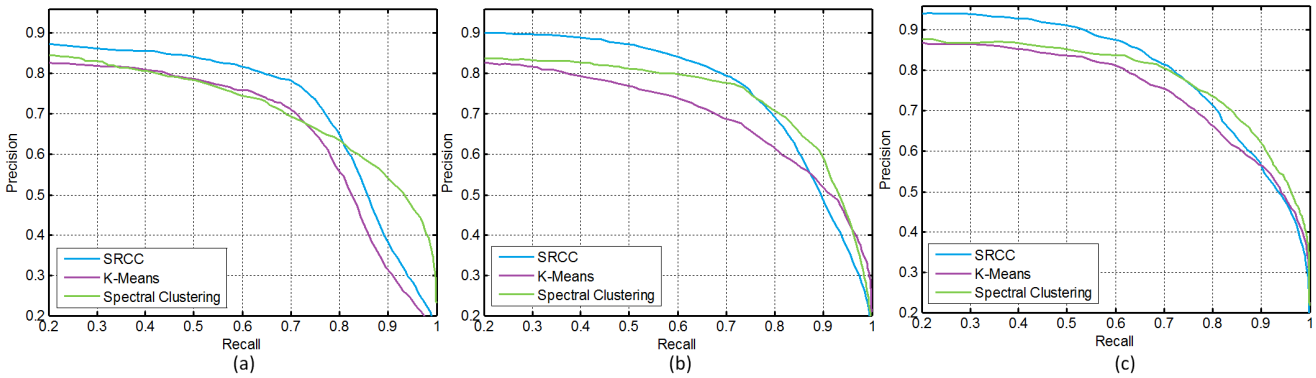

Fig. 6. Precision-recall curve of applying the baseline clustering methods used different features: (a) *Conv1*, (b) *Conv3* and (c) *Conv5*.

TABLE III
EVALUATION OF MV-SRCC IN STAGE 1 ON ICOSEG DATASET

Methods	**K-Means**			**Spectral Clustering**			**SRCC**			**MV-SRCC**
Features	*Conv1*	*Conv3*	*Conv5*	*Conv1*	*Conv3*	*Conv5*	*Conv1*	*Conv3*	*Conv5*	*Conv1,3,5*
ACC	0.622	0.717	0.753	0.745	0.795	0.824	0.785	0.879	0.903	**0.949**
NMI	0.802	0.818	0.846	0.838	0.871	0.892	0.886	0.924	0.963	**0.981**
Purity	0.701	0.779	0.806	0.765	0.841	0.866	0.857	0.919	0.941	**0.972**

TABLE IV
EVALUATION OF MV-SRCC IN STAGE 2 ON ICOSEG DATASET

Methods	**K-Means**			**Spectral Clustering**			**SRCC**			**MV-SRCC0**	**MV-SRCC**
Features	*Conv1*	*Conv3*	*Conv5*	*Conv1*	*Conv3*	*Conv5*	*Conv1*	*Conv3*	*Conv5*	*Conv1,3,5*	*Conv1,3,5*
AP	0.698	0.722	0.773	0.737	0.763	0.802	0.753	0.790	0.832	**0.879**	0.868
F-measure	0.537	0.625	0.743	0.593	0.665	0.699	0.651	0.711	0.757	**0.810**	0.809

conv3, *conv5*) . Table 3 gives the corresponding ACC, NMI and Purity for each clustering method on different features. As can be seen, the proposed MV-SRCC obtains the best clustering performance in all metric terms. By combining multiple features, MV-SRCC can improve about 4.6%, 1.8% and 3.1% in terms of ACC, NMI and Purity compared with SRCC. Furthermore, SRCC achieves much better clustering performance than k-means and spectral clustering, which indicates the superiority of incorporating the information from object proposals for clustering images in the first stage. Besides, clustering based on image-level features extracted from *conv5* performs better than *conv1* and*conv3* for all baseline methods. This indicates that higher level convolution layer can capture more semantic object information which forms more robust image representations.

2) Evaluation of MV-SRCC in Stage 2: We firstly evaluated the final co-saliency performance differences of the proposed MV-SRCC algorithm performed on well-organized subgroups and automatic-divided subgroups produced by the first stage. In Table 4, we denote the MV-SRCC running on well-organized subgroups as MV-SRCC0 for clarity. As can be seen, performing co-salient object detection on our automatic-divided subgroups achieves competitive results with only a decrease of 1.1% and 0.1% in terms of AP and F-measure, which clearly demonstrates the effectiveness of our proposed co-saliency detection framework.

We also investigated the co-saliency performance of k-means, spectral clustering and SRCC on different features. Note that for fair comparison, all the baseline clustering methods are performed in each well-organized subgroup to cluster superpixels. Fig.6 and Table 4 present the co-saliency results in terms of PR curve, AP and F-measure. Clearly, MV-SRCC0 achieves the best performance and provides a significant performance improvement over SRCC, which shows that multiple features are beneficial to improve performance. Generally, better performance can be achieved by using features from higher level convolution layer for all clustering methods.

D. Comparison With Other Multi-View Features

In this subsection, we investigated how much performance improvement can be gained by integrating multiple features

TABLE V

PERFORMANCE OF DIFFERENT MULTI-VIEW FEATURES IN STAGE 1 ON iCOSEG DATASET

Features	ACC	NMI	Purity
Color, LBP, SIFT	0.809	0.882	0.869
Conv1,3,5	**0.949**	**0.981**	**0.972**

TABLE VI

PERFORMANCE OF DIFFERENT MULTI-VIEW FEATURES IN STAGE 2 ON iCOSEG DATASET

Features	AP	F-measure
Color, LBP, SIFT	0.787	0.702
Conv1,3,5	**0.868**	**0.810**

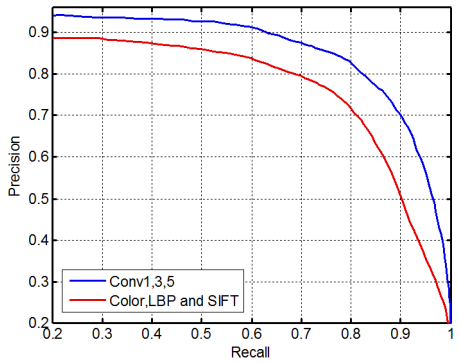

Fig. 7. Precision-recall curve of different multi-view features.

from multiple convolutional layers of CNN model compared with other multi-view features on iCoseg dataset.

For this comparison experiment, three types of features such as color histogram, LBP and SIFT are employed as the multi-view features. Specifically, we firstly use a fixed-size patch (16×16 pixels) with the spacing step to be 8 pixels to extract all the above features in each patch of the image. For color histogram, each channel of the RGB color space is quantized into 64 bins to form a total histogram feature length of 192 by simply concatenation of the three channels. For LBP feature, we use the common used 8 neighbors to get the binary values and convert the 8-bit binary values into a decimal value for each pixel in the patch, which result in a 256 dimensional LBP feature. For SIFT, the histogram of gradients is computed over a 4×4 grid and the gradients are then quantized into eight bins, which result in a 128 dimensional SIFT feature. Then, the 2×2 spatial grid cell partitioning scheme and max-pooling method are adopted as described in Section IV.A to obtain the final features of image, object proposals and superpixels.

Table 5 lists the ACC, NMI and Purity for the performance of the first stage. Fig.7 presents the precision-recall curve of MV-SRCC with the two different multi-view features and Table 6 gives the corresponding AP and F-measure. For the first stage, multi-view features from CNN achieves significantly better clustering performance in all metric terms as shown in Table 5. From Table 6 and Fig.7 for the performance of the second stage, we can observe that combination of features extracted from conv1,3,5 layers can improve 8.1% and 9.8% in term of AP and F-measure, and achieves much higher precision for all recall than multi-features of color histogram, LBP and SIFT. The results for both the two stages clearly demonstrate the superior of multi-view features extracted from CNN.

VI. CONCLUSION AND FUTURE WORK

This paper has proposed a novel multi-class co-salient detection framework which is among the earliest efforts to simultaneously detect multi-class co-salient objects from complex and practical image sets. Firstly, we formulated multi-class co-salient object detection as a co-clustering problem. Then, we proposed a novel MV-SRCC algorithm and developed the multi-class co-salient detection framework based on two stages of MV-SRCC. Even without the strong assumption to restrict the given image group to contain only one class of co-salient objects, the proposed approach was demonstrated to have even better performance than the traditional state-of-the-art co-salient object detection methods.

Since our proposed method attempts to perform co-salient object detection directly from complex image sets based on two-stage co-clustering algorithm, the final co-saliency performance heavily depends on the effectiveness of co-clustering in two stages. Although the information of image scenes and the inside object proposals are considered to identify subgroups of relevant images, the proposed MV-SRCC algorithm could not well separate similar image scenes such as the scenes of playing soccer with man and woman players both wearing red jersey.

In the future, we plan to use more powerful CNN model to extract robust features for improving the clustering performance. In addition, we intend to design an effective clustering-based contrast measure method that can handle the noise in the subgroup. Also, it is desirable to exploit the extra information such as ranking information obtained from social media portals such as Google and Flickr to improve the co-salient object detection performance. Moreover, it is interesting to leverage the proposed MV-SRCC algorithm to facilitate other tasks, e.g., object detection and annotation from high-resolution remote sensing images [48], [49] and video saliency detection [50].

APPENDIX

In order to solve this objective function in Eq. (5), firstly, we set the derivative of Eq. (5) with respect to $\mathbf{G}^{(v)}$ to zero. Then, we get

$$\mathbf{G}^{(v)} = \alpha(\mathbf{L}^{(v)} + \alpha\mathbf{I})^{-1}\mathbf{G}. \tag{13}$$

Further, we have

$$\begin{aligned}\mathbf{G} - \mathbf{G}^{(v)} &= \mathbf{G} - \alpha(\mathbf{L}^{(v)} + \alpha\mathbf{I})^{-1}\mathbf{G} \\ &= (\mathbf{I} - \alpha(\mathbf{L}^{(v)} + \alpha\mathbf{I})^{-1})\mathbf{G} \\ &= ((\mathbf{L}^{(v)} + \alpha\mathbf{I})(\mathbf{L}^{(v)} + \alpha\mathbf{I})^{-1} - \alpha(\mathbf{L}^{(v)} + \alpha\mathbf{I})^{-1})\mathbf{G} \\ &= \mathbf{L}^{(v)}(\mathbf{L}^{(v)} + \alpha\mathbf{I})^{-1}\mathbf{G} \end{aligned} \tag{14}$$

Substitute the resultant $\mathbf{G}^{(v)}$in Eq. (13) and $\mathbf{G} - \mathbf{G}^{(v)}$ in Eq. (14) into the second term of Eq. (5), and

we have

$$\begin{aligned}
&\alpha Tr((\mathbf{G}-\mathbf{G}^{(v)})^T(\mathbf{G}-\mathbf{G}^{(v)}))\\
&\quad=\alpha Tr(\mathbf{G}^T(\mathbf{L}^{(v)}+\alpha\mathbf{I})^{-1}\mathbf{L}^{(v)}\mathbf{L}^{(v)}(\mathbf{L}^{(v)}+\alpha\mathbf{I})^{-1}\mathbf{G})\\
&\quad=\alpha Tr(\mathbf{G}^T(\mathbf{L}^{(v)}+\alpha\mathbf{I})^{-1}(\mathbf{L}^{(v)}+\alpha\mathbf{I})L^{(v)}(\mathbf{L}^{(v)}+\alpha\mathbf{I})^{-1}\mathbf{G})\\
&\qquad-Tr(\alpha\mathbf{G}^T(\mathbf{L}^{(v)}+\alpha\mathbf{I})^{-1}\mathbf{L}^{(v)}\alpha(\mathbf{L}^{(v)}+\alpha\mathbf{I})^{-1}\mathbf{G})\\
&\quad=\alpha Tr(\mathbf{G}^T\mathbf{L}^{(v)}(\mathbf{L}^{(v)}+\alpha\mathbf{I})^{-1}\mathbf{G})-Tr((\mathbf{G}^{(v)})^T\mathbf{L}^{(v)}\mathbf{G}^{(v)})
\end{aligned}\tag{15}$$

Thus, Eq. (5) is equivalent to minimizing the following objective function

$$\min_{\mathbf{G}^T\mathbf{G}=\mathbf{I}}\sum_{v=1}^{V}\alpha Tr(\mathbf{G}^T\mathbf{L}^{(v)}(\mathbf{L}^{(v)}+\alpha\mathbf{I})^{-1}\mathbf{G}).\tag{16}$$

Also, since

$$\begin{aligned}
\mathbf{L}^{(v)}(\mathbf{L}^{(v)}+\alpha\mathbf{I})^{-1}&=(\mathbf{L}^{(v)}+\alpha\mathbf{I}-\alpha\mathbf{I})(\mathbf{L}^{(v)}+\alpha\mathbf{I})^{-1}\\
&=\mathbf{I}-\alpha(\mathbf{L}^{(v)}+\alpha\mathbf{I})^{-1}
\end{aligned}\tag{17}$$

and let $\mathbf{P}=\sum_{v=1}^{V}\mathbf{I}-\alpha(\mathbf{L}^{(v)}+\alpha\mathbf{I})^{-1}$, Eq. (5) is then transformed to the following optimization problem

$$\min_{\mathbf{G}^T\mathbf{G}=\mathbf{I}} Tr(\mathbf{G}^T\mathbf{P}\mathbf{G}).\tag{18}$$

REFERENCES

[1] L. Wang, G. Hua, R. Sukthankar, J. Xue, and N. Zheng, "Video object discovery and co-segmentation with extremely weak supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 640–655.

[2] D. S. Hochbaum and V. Singh, "An efficient algorithm for co-segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 269–276.

[3] H. Fu, D. Xu, B. Zhang, and S. Lin, "Object-based multiple foreground video co-segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3166–3173.

[4] Y. Yang, B. Geng, Y. Cai, A. Hanjalic, and X.-S. Hua, "Object retrieval using visual query context," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1295–1307, Dec. 2011.

[5] X. Guo, D. Liu, B. Jou, M. Zhu, A. Cai, and S.-F. Chang, "Robust object co-detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3206–3213.

[6] S. K. Divvala, A. Farhadi, and C. Guestrin, "Learning everything about anything: Webly-supervised visual concept learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3270–3277.

[7] X. Chen, A. Shrivastava, and A. Gupta, "NEIL: Extracting visual knowledge from Web data," in *Proc. IEEE Conf. Comput. Vis.*, Dec. 2013, pp. 1409–1416.

[8] X.-S. Hua and J. Li, "Prajna: Towards recognizing whatever you want from images without image labeling," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 137–144.

[9] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2001, pp. 269–274.

[10] L. Zhang, C. Chen, J. Bu, Z. Chen, D. Cai, and J. Han, "Locally discriminative coclustering," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1025–1035, Jun. 2012.

[11] B.-K. Bao, W. Min, T. Li, and C. Xu, "Joint local and global consistency on interdocument and interword relationships for co-clustering," *IEEE Trans. Cybern.*, vol. 45, no. 1, pp. 15–28, Jan. 2015.

[12] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.

[13] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[14] H.-T. Chen, "Preattentive co-saliency detection," in *Proc. IEEE Conf. Image Process.*, Sep. 2010, pp. 1117–1120.

[15] D. E. Jacobs, D. B. Goldman, and E. Shechtman, "Cosaliency: Where people look when comparing images," in *Proc. ACM Symp. User Interface Softw. Technol.*, 2010, pp. 219–228.

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

[17] Z. Tan, L. Wan, W. Feng, and C.-M. Pun, "Image co-saliency detection by propagating superpixel affinities," in *Proc. IEEE Conf. Acoust., Speech Signal Process.*, May 2013, pp. 2114–2118.

[18] 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. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2129–2136.

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

[20] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.

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

[22] X. Cao, Y. Cheng, Z. Tao, and H. Fu, "Co-saliency detection via base reconstruction," in *Proc. ACM Conf. Multimedia*, 2014, pp. 997–1000.

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

[24] 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. IEEE Conf. Multimedia Expo*, Jul. 2014, pp. 1–6.

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

[26] S. Du and S. Chen, "Detecting co-salient objects in large image sets," *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 145–148, Feb. 2015.

[27] Y. Li, K. Fu, Z. Liu, and J. Yang, "Efficient saliency-model-guided visual co-saliency detection," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 588–592, May 2014.

[28] S. Ningmin and L. Jing, "Improved structured sparse PCA for cluster-based co-saliency detection," in *Proc. ACM Conf. Int. Multimedia Comput. Service*, 2015, p. 30.

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

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

[31] Z. Li, X.-M. Wu, and S.-F. Chang, "Segmentation using superpixels: A bipartite graph partitioning approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 789–796.

[32] W.-S. Chu, Y. Song, and A. Jaimes, "Video co-summarization: Video summarization by visual co-occurrence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3584–3592.

[33] W. Liu, S. Li, X. Lin, Y. Wu, and R. Ji, "Spectral–spatial co-clustering of hyperspectral image data based on bipartite graph," *Multimedia Syst.*, vol. 22, no. 3, pp. 355–366, 2015.

[34] J. Huang, F. Nie, and H. Huang, "Spectral rotation versus K-means in spectral clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2013, pp. 431–437.

[35] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3626–3633.

[36] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 447–456.

[37] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

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

[39] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[40] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "iCoseg: Interactive co-segmentation with intelligent scribble guidance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3169–3176.
[41] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 1800–1807.
[42] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[43] Y. Jia and M. Han, "Category-independent object-level saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1761–1768.
[44] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 883–890.
[45] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2814–2821.
[46] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.
[47] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.
[48] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.
[49] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.
[50] D. Zhang, J. Han, L. Jiang, S. Ye, and X. Chang, "Revealing event saliency in unconstrained video collection," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1746–1758, Apr. 2017.

Xiwen Yao received his B.S. and Ph.D. degree from the Northwestern Polytechnical University, China, in 2010 and 2016, respectively.

He is currently a research assistant of Northwestern Polytechnical University. His research interests include saliency detection, remote sensing processing and object detection.

Junwei Han received the B.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 1999 and 2003, respectively.

He is currently a Professor with Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and multimedia processing. He is an Associate Editor of the IEEE Transactions on Human-machine Systems, Neurocomputing, and Multidimensional Systems and Signal Processing.

Dingwen Zhang received the B.E. degree from Northwestern Polytechnical University, Xi'an, China, in 2012, where he is currently pursuing the Ph.D. degree. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, USA. His research interests include computer vision and multimedia processing, especially on saliency detection, video understanding, and weakly supervised learning.

Feiping Nie received the Ph.D. degree in Computer Science from Tsinghua University, China in 2009. His research interests are machine learning and its applications, such as pattern recognition, data mining, computer vision, image processing and information retrieval. He is now serving as Associate Editor or PC member for several prestigious journals and conferences in the related fields.