Object Co-segmentation via Graph Optimized-Flexible Manifold Ranking

Rong Quan[1], Junwei Han[1], Dingwen Zhang[1], Feiping Nie[2]
1 School of Automation, 2 School of Computer Science and Center for OPTIMAL,
Northwestern Polytechnical University, Xi'an, 710072, P. R. China
rongquan0806, junweihan2010, zhangdingwen2006yyy, feipingnie@gmail.com

Abstract

Aiming at automatically discovering the common objects contained in a set of relevant images and segmenting them as foreground simultaneously, object co-segmentation has become an active research topic in recent years. Although a number of approaches have been proposed to address this problem, many of them are designed with the misleading assumption, unscalable prior, or low flexibility and thus still suffer from certain limitations, which reduces their capability in the real-world scenarios. To alleviate these limitations, we propose a novel two-stage co-segmentation framework, which introduces the weak background prior to establish a globally close-loop graph to represent the common object and union background separately. Then a novel graph optimized-flexible manifold ranking algorithm is proposed to flexibly optimize the graph connection and node labels to co-segment the common objects. Experiments on three image datasets demonstrate that our method outperforms other state-of-the-art methods.

1. Introduction

Given a set of images containing the same or similar objects from the same semantic class, the goal of object co-segmentation is to discover and segment out such common objects from all images, as shown in Figure 1. The problem of object co-segmentation is first proposed by Rother *et al.* [3], which demonstrates that simultaneously segmenting out the common objects in an image pair can achieve higher accuracy than segmenting in either single image alone. Following this work, a number of researchers make their efforts to develop more effective computational models [4-7] for co-segmenting objects in such image pairs. However, these methods only seek to co-segment objects in two images at a time, which results in direct limitations when extending beyond pairwise relations. By realizing this problem, more recent object co-segmentation approaches [8-16] propose to discover the common patterns of the co-occurring objects in group-level and thus can segment out them in more than two images. With such important progress, object co-segmentation has become to be more practical for the real-world problems because there are rich

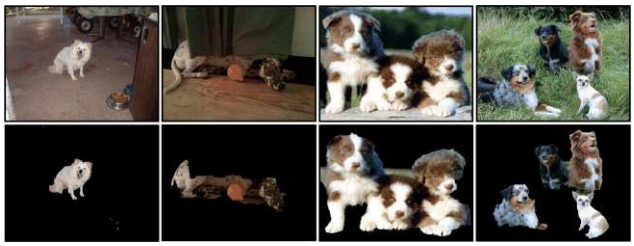
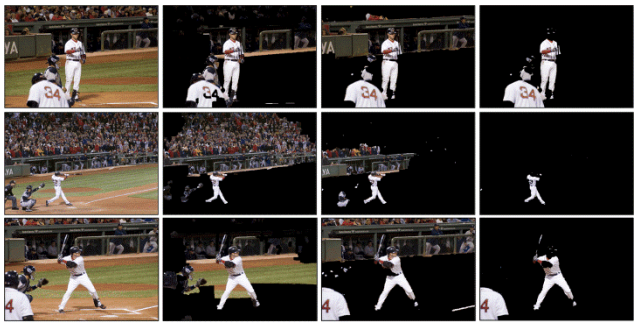


Figure 1: The first row: example images containing similar objects from the same semantic class. The second row: the co-segmentation results of our framework.



(a) (b) (c) (d)

Figure 2: (a) Several original images from the same object class. (b) The co-segmentation results of [1]. (c) The co-segmentation results of [2]. (d) Our results.

collections of multiple related pictures sharing the common objects or events in reality [17], such as the photo-sharing websites like Flickr and Facebook. However, when performing object co-segmentation in such real-world scenarios, the existing methods still suffer from certain limitations, which are mainly lied in the following aspects:

Misleading assumption: Some existing methods are based on a misleading assumption that the common regions contained by the given image group should be the objects of interest. However, lots of real-world image groups containing similar objects are collected in similar scenes and thus they also contain similar and co-occurring image background which may confuse these methods seriously. In this case, these methods always wrongly segment out the similar co-occurring image backgrounds, as shown in Figure 2. For example, Faktor *et al.* [2] proposed to discover the co-occurring regions firstly, and then perform co-segmentation by mapping between the co-occurring

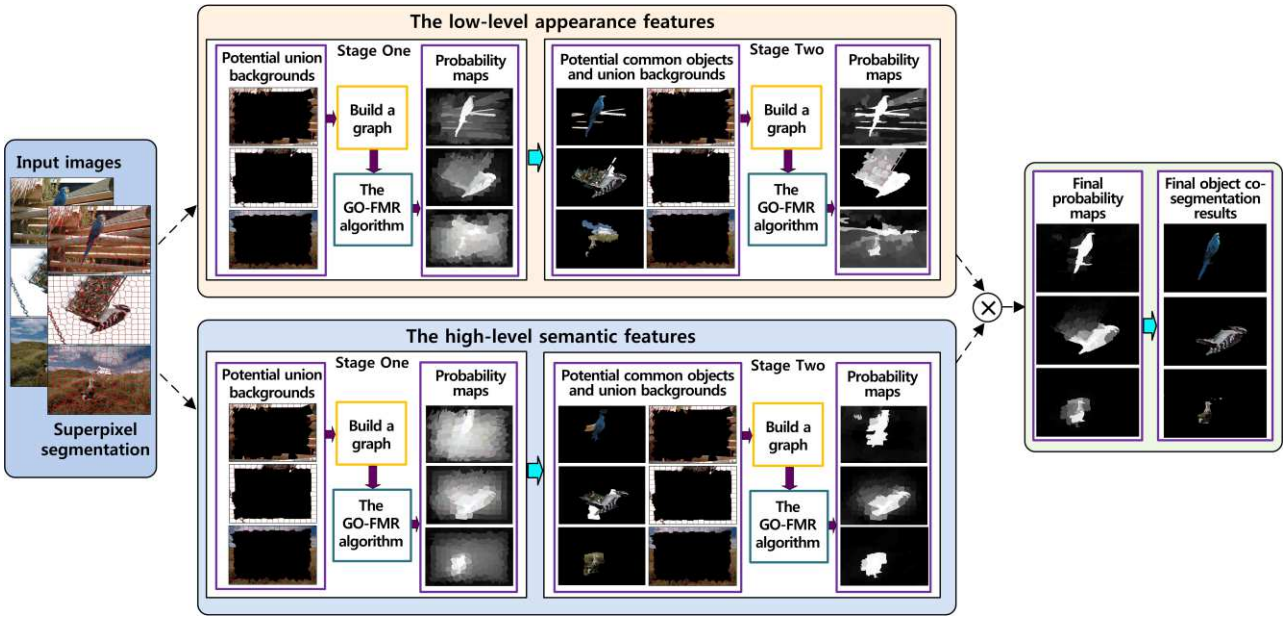


Figure 3: Overview of our two-stage object co-segmentation framework.

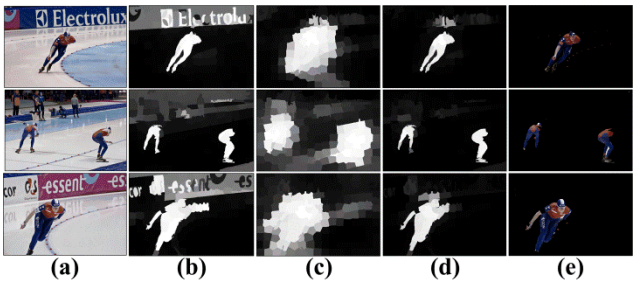
regions across the different images, whereas those extracted co-occurring regions also contain the common backgrounds. The models [7, 9] did not explicitly consider the common foreground and background in separate formulations, which leads to the failure results where arbitrarily shaped background regions (with similar appearance in both images) are segmented out as the common objects.

To solve this problem, we introduce a novel concept in this paper, i.e., the union background, which indicates the collections of image background regions in each image group, and then proposes to formulate the common objects and union background separately. Specifically, we represent the image group as a globally close-loop graph with superpixels as nodes. These nodes are ranked based on the similarity to background and foreground queries via a novel GO-FMR (graph optimized-flexible manifold ranking) algorithm which can infer both the optimal labels and connections of the nodes to precisely distinguish the common objects from the union background even when some background regions are similar and co-occurring in the image group (see Figure 2(d)).

Unscalable prior: For alleviating the confusion of common foreground and background, some approaches adopt certain prior knowledge, e.g. saliency and objectness in co-segmentation. For example, Rubinstein *et al*. [1] proposed to establish reliable correspondences between pixels in different images based on the extracted saliency regions. Vicente *et al.* [14] showed that requiring the foreground segment to be an object can significantly improve the co-segmentation performance and thus they introduced the objectness in their model via generating a pool of object-like proposal segmentations. However, such prior knowledge may not always guarantee to provide adequate and precise information when scaling up for the real-world scenarios due to the inestimable complexity and diversity in real world.

To solve this problem, we propose a novel two-stage framework in this paper, which can weaken the strong prior knowledge used in the previous work to a much more scalable prior, i.e., the background prior. As mentioned in [18, 19], the background prior comes from the basic rule of photographic composition, that is, most photographers do not crop objects of interest along the view frame. In other words, the image boundary is mostly background. Based on this prior, we initialize the graph in the first stage by connecting the superpixel nodes located in the image boundaries of the entire image group, i.e., the union background. By inferring via the proposed GO-FMR algorithm, we can obtain the image regions which are more different from the union background. It further provides informative knowledge for co-segmenting the common objects in the second stage.

Low flexibility: Some previous methods rely heavily on certain model configurations which are manually designed and kept fixed during the exploration of the common objects. However, such strategies are typically subjective and cannot generalize well to flexibly adapt to various real-world scenarios encountered in practice. Take the graph-based object co-segmentation approaches as the example. Joulin *et al.* [9] proposed a discriminative clustering algorithm, which combined the fixed Laplacian matrix and kernel matrix to formulate the spatial consistency and discriminative clustering, respectively. For better taking advantage of the information available from other images in



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

Figure 4: Illustration of the combination of the two kinds of image features. (a) The input images from the same object class. (b) The probability maps obtained from the low-level appearance features. (c) The probability maps obtained from the high-level semantic feature. (d) The probability maps obtained from the combination of low-level appearance features and high-level semantic feature. (e) The final binary masks of the input images.

the group, Kim *et al.* [20] carefully modelled the inter-image relationships via connecting the image regions of one image to all other images in an image cluster. As can be seen, these methods build the graph connections based on the certain aspects of human knowledge and keep them fixed during the optimization processes. However, the limited human knowledge we have may not always guarantee an optimal graph connection in various scenarios. Thus, such strategies keeping the graph connections unchanged during the optimization process may suffer from the low flexibility when dealing with the real-world scenarios.

To solve this problem, we propose a novel GO-FMR algorithm, which can alternatively optimize the superpixel labels as well as the connections in the image group. Specifically, given an initialized graph with the nodes connected based on certain human knowledge, the proposed GO-FMR algorithm can be guided by the human knowledge and further flexibly infer the optimal graph connection for the specific scenario rather than blindly trusting the human knowledge and keeping the manually designed graph connection fixed in all cases. Moreover, the prior information can also be easily incorporated into the proposed GO-FMR algorithm by initializing the seed nodes correspondingly. Thus, it can be adapted to the various real-world scenarios flexibly.

The concrete framework proposed in this paper to reduce above mentioned problems is shown in Figure 3. Given images within an image group, we first decompose each image into superpixels. Then, for each superpixel, we extract the low-level appearance features and high-level semantic features, respectively. Afterwards, the properties of the common objects are inferred via the GO-FMR based two-stage scheme to generate the probability maps for the different types of features separately. Finally, the obtained probability maps are integrated to generate the final object co-segmentation results. In summary, the contributions of this paper are three-folds:

- We make one of the earliest efforts to formulate the common object and union background separately, which can effectively suppress the co-segmentation of the common background in the real-world scenarios.
- We proposed a novel two-stage object co-segmentation scheme which relies on a much weaker background prior and thus can better scale up for more complex scenarios.
- We propose a novel GO-FMR algorithm to optimize the established globally close-loop graph, which can simultaneously infer both the labels of all the superpixel nodes and the optimal graph connection to best explore the relationships among all image regions.

2. Proposed Approach

For a set of images $\Omega = \{I_1, I_2, ..., I_m\}$ that contain a common object, our goal is to segment the common object instance in each image. As a pre-processing step, each image I_i in Ω is first over-segmented into n_i superpixels by the SLIC algorithm [21]. Then the whole image set Ω contains $n = \sum_{i=1}^{m} n_i$ superpixels. Our object co-segmentation process is performed on superpixel level.

2.1. The image features

In this paper, we adopt two kinds of image features, i.e., the low-level appearance features and the high-level semantic features, to capture different characteristics of each superpixel.

The low-level appearance features are sensitive to the appearance variations of images. The common objects detected based on this kind of features should have consistent appearance and well-defined object boundaries within each image, as shown in Figure 4(b). Three kinds of low-level appearance features are used in this work, including color, texture, and dense SIFT descriptors [22]. They are denoted by $\mathbf{c}_1$, $\mathbf{c}_2$ and $\mathbf{c}_3$, respectively.

As mentioned by Zhang *et al.* [17], the common objects in different images may share strong homogeneity in semantic level. Thus, we also apply the deep semantic feature in this paper. Specifically, we employ the 'CNN-S' [23] model which is pre-trained on the ImageNet [24] dataset to extract the high-level semantic representations. Firstly, we feed each image into the pre-trained CNN and extract the responses from the last convolutional layer as the higher-level image representations, which consists of 512 feature maps with size of 17×17. Next, we resize the feature maps to the size of the original image, and then use a max pooling operation on each superpixel to generate a 512-D CNN feature vector. Subsequently, an auto-encoder is further used to reduce the feature dimension to 24. The

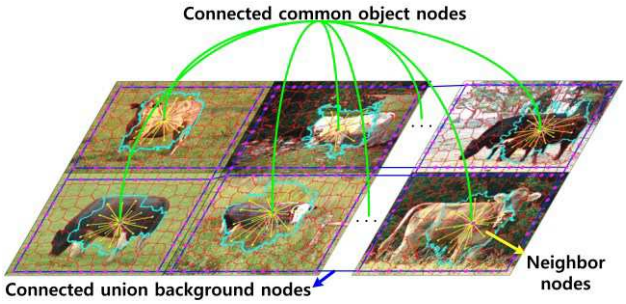


Figure 5: Illustration of the node connections in our graph. The green lines show that the potential common object nodes of all the images are connected together. The blue lines show that the potential union background nodes of all the images are connected together. The yellow lines within each image show that the neighbor nodes within each image are connected.

24-D CNN feature vector is denoted by $\mathbf{c}_4$. With such semantic feature, we can precisely locate the common objects in each image, as shown in Figure 4(c). However, as the original size of the feature maps is very small, even after the resizing operation, the feature maps are still very coarse. Thus the obtained co-segmentation results based on such CNN tend to be somewhat blurry.

As the low-level appearance features and the high-level semantic features describe different attributes of the images and play different roles in the co-segmentation process, we first perform our co-segmentation framework based on each of them separately. Then those two preliminary co-segmentation results are integrated into the final one, as shown in Figure 4(d).

2.2. The graph construction

We construct a globally close-loop graph $G=(V,E,\mathbf{A})$ on Ω, where each node in V corresponds to a superpixel in Ω, the edges in E connect all the related superpixels, and the affinity matrix $\mathbf{A}$ measures the similarities among all superpixels.

As spatially neighboring nodes with similar features tend to belong to the same class, we connect each node with not only its spatial neighbors, but also the neighbors of its neighbors to model the intra-image constraints in each image. In addition, as we formulate the common object and union background separately, we connect all the potential common object nodes together and union background nodes together to model their consistency relationships, respectively. The initial node connections in our graph can be illustrated in Figure 5.

The affinity matrix $\mathbf{A}=\left[a_{ij}\right]_{n\times n}$ measures the weights of $E=\left[e_{ij}\right]_{n\times n}$. For each edge e_{ij} that connects two nodes, we compute the similarity weight a_{ij} as follows:

$$a_{ij}=\exp\left(-\sum_{t=1}^{3}\lambda_t\left\|\mathbf{c}_t^i-\mathbf{c}_t^j\right\|^2\right) \quad (1)$$

or

$$a_{ij}=\exp\left(-\left\|\mathbf{c}_4^i-\mathbf{c}_4^j\right\|^2\right) \quad (2)$$

where $\mathbf{c}^i$ is the feature vector of node i. Eq. (1) is used for the low-level appearance features and Eq. (2) is used for the high-level semantic feature. Empirically, setting $\lambda_1=1$ and $\lambda_2=\lambda_3=0.7$ can produce good results in our experiments.

After obtaining $\mathbf{A}$, we further define the diagonal matrix $\mathbf{D}$ as the row sums of $\mathbf{A}$, and the graph Laplacian matrix $\mathbf{L}$ as:

$$\mathbf{L}=\mathbf{D}-\mathbf{A} \quad (3)$$

2.3. The GO-FMR algorithm

After modeling the similarity relationships among all superpixels in Ω through the graph $G=(V,E,\mathbf{A})$, we further partition all superpixels into the common object and the union background based on the graph. Here we formulate the graph labeling problem as a graph-based manifold ranking problem [25] and propose a novel semi-supervised learning technique called the graph optimized-flexible manifold ranking algorithm (GO-FMR), to infer the class labels of all superpixels. The graph-based manifold ranking problem [25] refers to the problem that given a node as query, all other nodes in the graph are ranked based on their correlations to the given query. Suppose that some superpixels in Ω have already been labelled as 1 (or used as queries). We use the GO-FMR algorithm to rank all superpixels based on their relevance to the labeled superpixels, where the ranking scores can be treated as the probabilities of these superpixels being labeled as 1, i.e. their prediction labels corresponding to the query superpixels.

Traditional graph labeling algorithms directly predict the class labels of all the superpixels based on the manually established affinity matrix $\mathbf{A}$. However, just depending on the manually established affinity matrix may not represent the real similarity relationships among all the superpixels. Even if it can, the process of computing the affinity matrix from the original image features itself will lose some information. Instead of just depending on the affinity matrix $\mathbf{A}$, the proposed GO-FMR algorithm additionally uses a projection to directly infer the predict labels from the original image features. It also automatically learns the optimal graph connection for specific scenario during the label prediction process to infer the final predict labels more accurately.

Let us denote Ω as a sample set $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n]\in\mathbb{R}$, where each sample $\mathbf{x}_i\in\mathbb{R}$ corresponds to a superpixel in Ω and t is the feature dimension. Suppose that some superpixels in Ω have already been labeled. We define a binary indicator vector $\mathbf{y}=[y_1,y_2,...,y_n]^T\in\mathbb{R}$, where $y_i=1$ means that $\mathbf{x}_i$ is labeled as 1, and $y_i=0$ means that $\mathbf{x}_i$ is unlabeled. Then, the prediction labels (or ranking scores) of all superpixels $\mathbf{f}=[f_1,f_2,...,f_n]^T\in\mathbb{R}$ can be computed by solving the following optimization problem:

$$\begin{aligned}(\mathbf{f}^*,\mathbf{w}^*,b^*,\mathbf{S}^*)&=\arg\min_{\mathbf{f},\mathbf{w},b,\mathbf{S}}\psi(\mathbf{f},\mathbf{w},b,\mathbf{S})\\&=\arg\min_{\mathbf{f},\mathbf{w},b,\mathbf{S}}\mathrm{tr}(\mathbf{f}-\mathbf{y})^T\mathbf{U}(\mathbf{f}-\mathbf{y})+\mathrm{tr}(\mathbf{f}^T\mathbf{L}_\mathbf{S}\mathbf{f})^T\\&\quad+\mu\left\|\mathbf{X}^T\mathbf{w}+\mathbf{1}b-\mathbf{f}\right\|_2^2+\eta\left\|\mathbf{S}-\mathbf{A}\right\|_F^2\end{aligned}\tag{4}$$

where $\mathbf{U}\in\mathbb{R}$ is a diagonal matrix with $u_{ii}=1$ for labeled superpixels and $u_{ii}=0.001$ for unlabeled ones, $\mathbf{w}\in\mathbb{R}$ is a projection vector, $b\in\mathbb{R}$ is a bias term, $\mathbf{1}\in\mathbb{R}$ is a vector with all elements being 1, and $\mathbf{X}^T\mathbf{w}+\mathbf{1}b$ is a linear projection that directly maps $\mathbf{X}$ to the prediction labels $\mathbf{f}$. $\mathbf{S}=[s_{ij}]_{n\times n}$ is the optimal affinity matrix used to infer the prediction labels $\mathbf{f}$, where each $s_{ij}\geq 0$. In addition, to simplify the computation, we make a constraint that $\sum_{i=1}^n s_{ij}=1$. $\mathbf{L}_\mathbf{S}$ is the graph Laplacian matrix computed from $\mathbf{S}$. The two parameters μ and η are used to balance different terms.

The first two terms in Eq. (4) constrain the label fitness (i.e., $\mathbf{f}$ should be close to the given labels of the labeled nodes) and manifold smoothness (i.e., $\mathbf{f}$ should be smooth on the entire graph of both the unlabeled and labeled nodes), respectively, which are normally used in traditional manifold ranking algorithms. In addition, the linear projection function $h(\mathbf{X})=\mathbf{X}^T\mathbf{w}+\mathbf{1}b$ is used to directly map $\mathbf{X}$ to the prediction labels $\mathbf{f}$. We use both the manifold ranking and the linear classification projection to predict the labels of all superpixels. The residue between the prediction results of these two methods is constrained to be as small as possible. $\mathbf{f}_0=\mathbf{X}^T\mathbf{w}+\mathbf{1}b-\mathbf{f}$ is the residue between $\mathbf{f}$ and $h(\mathbf{X})$, and $\|\mathbf{f}_0\|_2^2$ is a penalty term for the mismatch between them. In addition, the last term $\|\mathbf{S}-\mathbf{A}\|_F^2$ measures the difference between the learned optimal affinity matrix $\mathbf{S}$ and the human established affinity matrix $\mathbf{A}$. As we infer the optimal affinity matrix $\mathbf{S}$ under the guidance of the human established affinity matrix $\mathbf{A}$, this term ensures that $\mathbf{S}$ will not change too much from $\mathbf{A}$.

Algorithm 1 The GO-FMR algorithm

Input: A sample set $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n]\in\mathbb{R}$, a binary indicator vector $\mathbf{y}=[y_1,y_2,...,y_n]^T\in\mathbb{R}$, and a graph $G=(V,E,\mathbf{A})$ constructed on the sample set.

1: Set $\mathbf{U}\in\mathbb{R}$ as a diagonal matrix with $u_{ii}=1$ for labeled superpixels and $u_{ii}=0.001$ for unlabeled ones. Initialize the optimal similarity matrix $\mathbf{S}$ as $\mathbf{A}$.
2: **repeat**
3: Compute $\mathbf{f}$, $\mathbf{w}$, b by Eq. (5)
4: Compute $\mathbf{S}$ by Eq. (8)
5: **until** the prediction label vector $\mathbf{f}$ stops changing

Output: The prediction label vector $\mathbf{f}$

As those four variables in Eq. (4) cannot be solved simultaneously, we use an iterative optimization process to alternate between optimizing $\mathbf{S}$ and $\mathbf{f}$, $\mathbf{w}$, b. We first initialize the optimal affinity matrix $\mathbf{S}$ as $\mathbf{A}$. The detailed iterative process is as follows:

(1). Fix $\mathbf{S}$, and compute $\mathbf{f}$, $\mathbf{w}$, b. When $\mathbf{S}$ is fixed, this label prediction problem can be solved through the FME algorithm in [26]. As the objective function is proved to be jointly convex with respect to $\mathbf{f}$, $\mathbf{w}$, and b, there exists the optimal solutions for them, which are denoted as:

$$\begin{aligned}\mathbf{f}&=(\mathbf{U}+\mathbf{L}_\mathbf{S}+\mu\mathbf{H}_c+\mu\mathbf{N})^{-1}\mathbf{U}\mathbf{y}\\\mathbf{w}&=(\mathbf{X}_c\mathbf{X}^T)^{-1}\mathbf{X}_c\mathbf{f}\\b&=\frac{1}{n}(\mathbf{f}^T\mathbf{1}-\mathbf{w}^T\mathbf{X}\mathbf{1})\end{aligned}\tag{5}$$

where $\mathbf{N}=\mathbf{X}_c^T(\mathbf{X}_c\mathbf{X}^T)^{-1}\mathbf{X}_c\left[\mathbf{X}^T(\mathbf{X}_c\mathbf{X}^T)^{-1}\mathbf{X}_c-2\mathbf{I}\right]$, $\mathbf{X}_c=\mathbf{X}\mathbf{H}_c$, $\mathbf{H}_c=\mathbf{I}-(1/n)\mathbf{1}\mathbf{1}^T$, and $\mathbf{I}\in\mathbb{R}$ is an identity matrix. For more details about the computation process, please refer to [26].

(2). Fix $\mathbf{f}$, $\mathbf{w}$, and b, and compute $\mathbf{S}$. When $\mathbf{f}$, $\mathbf{w}$, b are fixed, with the prediction label vector $\mathbf{f}$ computed in the last step, $\mathbf{S}$ can be computed by solving the following optimization problem:

$$\mathbf{S}^*=\arg\min_{\mathbf{S}}\mathrm{tr}(\mathbf{f}^T\mathbf{L}_\mathbf{S}\mathbf{f})+\eta\|\mathbf{S}-\mathbf{A}\|_F^2\tag{6}$$

which can also be written as:

$$\mathbf{S}^*=\arg\min_{s_{ij}\geq 0,\sum_{i=1}^n s_{ij}=1}\sum_{i=1}^n\sum_{j=1}^n\|f_i-f_j\|_2^2 s_{ij}+\eta\sum_{i=1}^n\sum_{j=1}^n(s_{ij}-a_{ij})^2\tag{7}$$

Then, we can compute each column $\mathbf{s}_i\in\mathbb{R}$ of $\mathbf{S}$ via solving the following optimization problem:

$$\mathbf{s}_i^*=\arg\min_{\mathbf{s}_i\geq\mathbf{0},\mathbf{1}^T\mathbf{s}_i=1}\mathbf{d}_i^T\mathbf{s}_i+\eta\|\mathbf{s}_i-\mathbf{a}_i\|_2^2\tag{8}$$

where $\mathbf{a}_i \in \mathbb{R}$ is the i th column in $\mathbf{A}$, and $\mathbf{d}_i \in \mathbb{R}$ is a column vector that measures the label difference between each superpixel and the i th superpixel.

Afterwards, we can obtain the optimal affinity matrix $\mathbf{S}$ that corresponds to the current prediction label vector $\mathbf{f}$. Next, we update $\mathbf{f}$ based on $\mathbf{S}$. These two processes alternate until convergence. **Algorithm 1** summarizes the proposed GO-FMR algorithm.

2.4. Our two-stage co-segmentation framework

In our work, we define the common objects as the regions in different images that are similar to each other and different from the union backgrounds. We develop a two-stage co-segmentation framework to first detect some potential common objects by comparing with the union backgrounds, and then refine the obtained potential common objects by further considering the similarity between them, as shown in the middle part of Figure 3.

The first stage: The union backgrounds are first initialized as the superpixels on the image boundaries of all images. Then we construct a graph $G^{(1)}$ with all the initialized union background nodes connected together and all the spatially neighboring nodes connected. Next, by treating all the initialized union background nodes as the labeled nodes with label 1, we use the GO-FMR algorithm to compute all superpixel nodes' prediction labels $\mathbf{f}^{(1)}$.

As the potential common objects are defined as the superpixels that are different from the union backgrounds, we extract the potential common object superpixels from the prediction labels $\mathbf{f}^{(1)}$ as:

$$\mathbf{l}^{(1)} = \left(\overline{\mathbf{f}}^{(1)} >= \beta * mean\left(\overline{\mathbf{f}}^{(1)} \right) \right) \tag{9}$$

where $\overline{\mathbf{f}}^{(1)} = \mathbf{1} - \mathbf{f}^{(1)}$, and β controls the extraction of the common objects. A smaller β means that more superpixels will be extracted as potential common objects. $\mathbf{l}^{(1)} = \left[l_1^{(1)}, l_2^{(1)}, ..., l_n^{(1)} \right]^T$ is a binary indication vector with $l_i^{(1)} = 1$ means that the i th superpixel belongs to the potential common object.

The second stage. During this stage, we further compute the more accurate prediction labels for all superpixels by additionally considering the similarity among the common objects.

Specifically, we build a more comprehensive graph $G^{(2)}$, where all potential common object nodes are further connected except for the initialized union background nodes and the spatially neighboring nodes. Then, we treat all potential common object nodes as the labeled nodes with label 1, and use the GO-FMR algorithm to infer the final label predictions $\mathbf{f}^{(2)}$ of all superpixels. Finally, we generate a probability map for each image from $\mathbf{f}^{(2)}$, where each pixel value represents the likelihood of this pixel being the common object.

As shown in Figure 3, we perform our two-stage co-segmentation framework based on both the low-level appearance features and the high-level semantic features, respectively. Consequently, for each image, two probability maps are obtained based on these two kinds of image features. We multiply these two probability maps to obtain a more accurate one, where only the pixels detected as common objects by both the low-level appearance features and the high-level semantic features can have high probability of being labeled as common object. Finally, we apply a Grab-cut [27] algorithm to the final probability maps to obtain the final binary object co-segmentation maps.

3. Experiments

We evaluate our method on a widely used benchmark dataset, the iCoseg [28] dataset, and two more challenging datasets, the Internet [1] dataset and the PASCAL-VOC dataset [2]. Two widely used evaluation metrics are utilized: Precision, P (the percentage of correctly labeled pixels of both common object and union background), and Jaccard index, J (the intersection over union of the resulted co-segmentation map and the ground truth segmentation).

In our experiments, for each object class in the dataset, we first divide it into several smaller groups, where each group contains images with similar scenes, and then perform our co-segmentation framework on each group. Specifically, we use a k -means clustering algorithm to cluster the images based on their GIST [29] descriptors. Here, k is set to make sure that each group has about ten images.

The two parameters μ and η in Eq. (4) are set empirically: $\mu = 0.01$ and $\eta = 5$ for the low-level appearance features, and $\mu = 0.05$ and $\eta = 10$ for the high-level semantic features. The parameter β in Eq. (9) controls the extraction of the possible common object superpixels from $\mathbf{f}^{(1)}$. Empirically, we set β to 2 for the low-level appearance features and 1.5 for the high-level semantic features.

3.1. Experiments on the iCoseg dataset

The iCoseg [28] dataset is a widely used benchmark dataset for evaluating co-segmentation approaches. It contains 38 object classes of totally 643 images with human-given pixel-level segmentation ground-truth. The common objects of each class belong to the same object

instance under different viewpoints and illumination. The images in each object class have the same theme and similar backgrounds. Some example images are shown in Figure 6. We first evaluate our method on all these 38 object classes and then conduct experiments on a subset of the iCoseg dataset (16 classes of 122 images). This subset (sub-iCoseg) is also often used in previous works [1, 2, 14] to evaluate their co-segmentation approaches.

In Table 1, we show the comparison results of our method and two state-of-the-art co-segmentation algorithms of [2] and [30] on the whole iCoseg dataset. [30] is a supervised co-segmentation approach. Additionally, we also compare our results with [1, 2, 14] on the sub-iCoseg dataset, where [14] is a supervised co-segmentation method. The comparison results are shown in Table 2.

Table 1. Comparison results of the proposed method and two state-of-the-art co-segmentation methods on the iCoseg dataset in terms of average Precision (denoted by P) and Jaccard index (denoted by J). Because [30] does not provide their Jaccard index results, we thus do not present them here.

iCoseg	Ours	[2]	[30]
P	**93.3**	92.8	91.4
J	**0.76**	0.73	-

Table 2. Comparison results of the proposed method and three state-of-the-art co-segmentation methods on the iCoseg dataset in terms of average Precision and Jaccard index.

sub-iCoseg	Ours	[2]	[1]	[14]
P	**94.8**	94.4	89.6	85.4
J	**0.82**	0.79	0.68	0.62

As shown in Table 1 and Table 2, our method outperform all other co-segmentation methods. Compared with the second best approach [2], our method has slightly higher Precision and much higher Jaccard index. Note that the results of [1, 2, 14] are taken from Table 1 in [2]. Some results of our co-segmentation method are visualized in Figure 6. We can see that our co-segmentation method can accurately detect the common object instances in different images. Besides, almost no distracting background regions are detected as common objects any more.

3.2. Experiments on the Internet dataset

The Internet dataset [1] consists of thousands of images from the Internet through three query expansions: car, horse, and airplane. The common objects of each object class in this dataset are similar objects from the same semantic class. This dataset is a challenging dataset. Some example images are shown in Figure 6. We can see that the common objects in different images have quite different colors, scales, poses, and viewing-angles, and the backgrounds in each object class are also different from each other. In addition, each object class in this dataset contains some noisy images that do not contain the common objects. In our experiment, we follow [1] and [31] to utilize a subset of 100 images per class for evaluation. All those images in this subset are with human-given segmentation ground-truth.

We compare our method with four state-of-the-art approaches [1, 9, 31, 32] and Table 3 shows the comparison results of each object class. The comparison results show that our method outperforms all other methods on all three object classes. Note that the results of [1, 9, 31, 32] are taken from Table 2 in [31].

Table 3. Comparison results of the proposed method and four state-of-the-art co-segmentation methods on the subset of Internet dataset in terms of average Precision and Jaccard index.

	Car		Horse		Airplane	
	P	J	P	J	P	J
[9]	58.7	37.1	63.8	30.1	49.2	15.3
[32]	68.8	0.04	75.1	6.43	80.2	7.90
[1]	85.3	64.4	82.8	51.6	88.0	55.8
[31]	87.6	64.8	86.1	33.3	90.2	40.3
Ours	**88.5**	**66.8**	**89.3**	**58.1**	**91.0**	**56.3**

Figure 6 shows some sample results. As can be seen, our co-segmentation method can accurately detect the common objects of each class. However, for the noisy images of each object class, we cannot always successfully recognize them. For example, the last image of each object class shown in the third row of Figure 6 is a noisy image, which does not contain the common object. We successfully recognize the noisy image in the plane class, but fail in other two object classes.

Table 4. Comparison results of the proposed method and [2] on the PASCAL-VOC dataset in terms of the average Precision and Jaccard index.

PASCAL-VOC	Ours	[2]
P	**89**	84
J	**0.52**	0.46

3.3. Experiments on the PASCAL-VOC dataset

The PASCAL-VOC dataset formed in [2] is also a benchmark for evaluating co-segmentation approaches. It consists of 1037 images of 20 object classes from the well-known PASCAL-VOC 2010 dataset. This dataset is more challenging due to extremely large intra-class variability and distracting background clutter. Figure 6 shows some example images from this dataset. In this database, many common objects have similar colors with the backgrounds. For example, in the bird class (the last

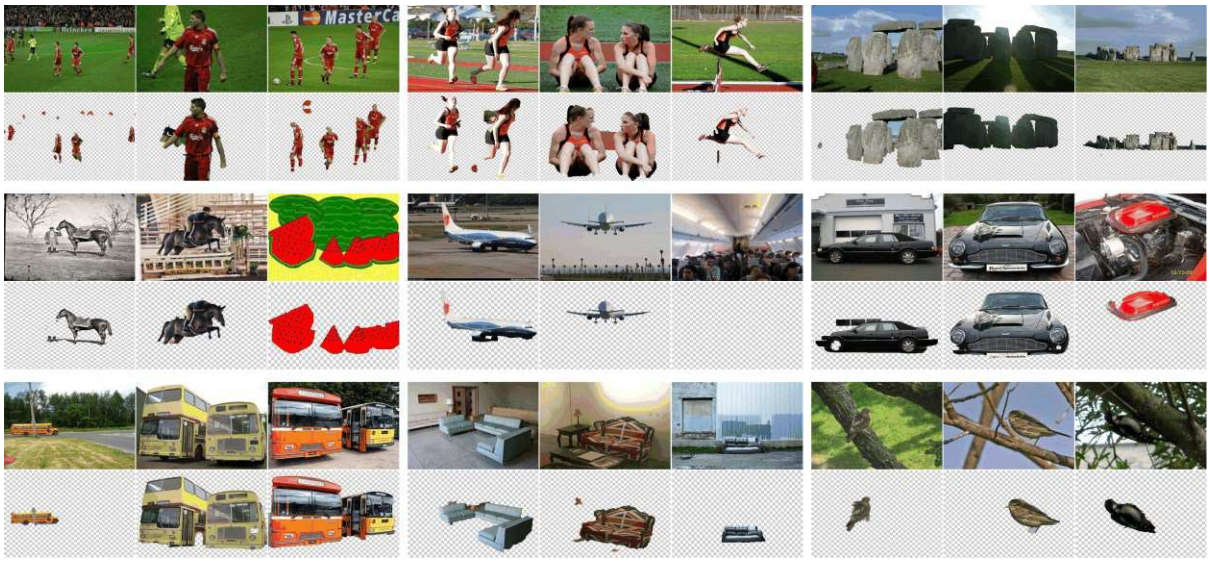


Figure 6. Some visualization results of our cosegmentation framework on the iCoseg (Rows 1-2), Internet (Rows 3-4), and PASCAL-VOC dataset (Rows 5-6), respectively. For each image set with three relevant images, we show the original images in the first row and the corresponding object segmentation results in the second row.

example in the fifth row of Figure 6), the branches have very similar colors with the birds. Table 4 presents the comparison results of our method and [2] on the PASCAL-VOC dataset. As can be seen from the results, our method outperforms [2] and both our average Precision and Jaccard index are much higher than [2].

3.4. Model component analysis

In this section, we implement experiments on the PASCAL-VOC dataset to further analyze some components in the proposed framework. Firstly, we demonstrate the effectiveness of integrating both the low-level appearance features and high-level semantic features. Specifically, we report the performance based on each individual feature in Table 5, respectively. As can be seen, the performance of integrating both features is significantly better than that of using either individual feature, which demonstrates the importance of leveraging the complementary information existed in different kinds of features as we pointed in Section 2.1.

Secondly, we demonstrate the effectiveness of the graph optimization step. Specifically, we report the performance of our method without graph optimization in Table 5. As can be seen, by executing graph optimization, the co-segmentation performance of our method improves a lot. This demonstrates that flexibly inferring the optimal graph connections to fit different scenarios can obtain much encouraging co-segmentation performance.

Lastly, we compare the performance of our method with a baseline method which adopts the conventional manifold ranking strategy as used in [25]. As shown in Table 5, the baseline method achieves performance much worse than the proposed method, which demonstrates the effectiveness of the proposed GO-FMR algorithm.

Table 5. The results for model component analysis. Note that 'Low', 'High', 'w/o OP', 'MR' and 'Ours' denote the image co-segmentation results that only based on the low-level appearance features, only based on the high-level semantic features, without graph optimization, the conventional manifold ranking strategy and our complete framework.

PASCAL-VOC	'Low'	'High'	'w/o OP'	'MR'	'Ours
P	80.8	85.0	86.9	78.9	**89.0**
J	0.45	0.49	0.45	0.41	**0.52**

4. Conclusion

In this paper, we have proposed a novel computational framework for object co-segmentation. By introducing a new concept of weak background prior, we constructed globally close-loop graphs to formulate the common object and union background separately. Afterwards, we designed a graph optimized-flexible manifold ranking algorithm to flexibly optimize the graph connection and node labels, which finally yielded the co-segmentation results. The comprehensive evaluations on three publically available benchmarks and comparisons with a number of state-of-the-art approaches have demonstrated the superiority of the proposed work.

Acknowledgements: This work was supported in part by the National Science Foundation of China under Grants 61522207 and 61473231.

References

[1] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. in *CVPR,* 2013.

[2] A. Faktor, and M. Irani. Co-segmentation by composition. in *ICCV,* 2013.

[3] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. in *CVPR,* 2006.

[4] D. S. Hochbaum, and V. Singh. An efficient algorithm for co-segmentation. in *ICCV,* 2009.

[5] L. Mukherjee, V. Singh, and C. R. Dyer. Half-integrality based algorithms for cosegmentation of images. in *CVPR,* 2009.

[6] Y. Mu, and B. Zhou. Co-segmentation of image pairs with quadratic global constraint in MRFs. in *ACCV,* 2007.

[7] S. Vicente, V. Kolmogorov, and C. Rother. Cosegmentation revisited: Models and optimization. in *ECCV,* 2010.

[8] K.-Y. Chang, T.-L. Liu, and S.-H. Lai. From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. in *CVPR,* 2011.

[9] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. in *CVPR,* 2010.

[10] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. in *CVPR,* 2012.

[11] L. Mukherjee, V. Singh, and J. Peng. Scale invariant cosegmentation for image groups. in *CVPR,* 2011.

[12] L. Mukherjee, V. Singh, J. Xu, and M. D. Collins. Analyzing the subspace structure of related images: Concurrent segmentation of image sets. in *ECCV,* 2012.

[13] J. C. Rubio, J. Serrat, A. López, and N. Paragios. Unsupervised co-segmentation through region matching. in *CVPR,* 2012.

[14] S. Vicente, C. Rother, and V. Kolmogorov. Object cosegmentation. in *CVPR,* 2011.

[15] H. Fu, D. Xu, S. Lin, and J. Liu. Object-based RGBD image co-segmentation with mutex constraint. in *CVPR,* 2015.

[16] F. Meng, H. Li, G. Liu, and K. N. Ngan. Object co-segmentation based on shortest path algorithm and saliency model. *TMM,*14(5):1429-1441, 2012.

[17] D. Zhang, J. Han, C. Li, and J. Wang. Co-saliency detection via looking deep and wide. in *CVPR,* 2015.

[18] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. in *ECCV,* 2012.

[19] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu. Background prior-based salient object detection via deep reconstruction residual. *TCSVT,*25(8):1309-1321, 2015.

[20] E. Kim, H. Li, and X. Huang. A hierarchical image clustering cosegmentation framework. in *CVPR,* 2012.

[21] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *TPAMI,*34(11):2274-2282, 2012.

[22] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *TPAMI,*33(5):978-994, 2011.

[23] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. in *CVPR,* 2009.

[25] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. in *CVPR,* 2013.

[26] F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *TIP,*19(7):1921-1932, 2010.

[27] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Trans on Graphics* 23(3):309-314, 2004.

[28] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. in *CVPR,* 2010.

[29] A. Oliva, and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Prog Brain Res,*155(1):23-36, 2006.

[30] D. Kuettel, M. Guillaumin, and V. Ferrari. Segmentation propagation in imagenet. in *ECCV,* 2012.

[31] X. Chen, A. Shrivastava, and A. Gupta. Enriching visual knowledge bases via object discovery and segmentation. in *CVPR,* 2014.

[32] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. in *ICCV,* 2011.