Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Coarse-to-fine Image Co-segmentation with Intra and Inter Rank Constraints

Lianli Gao, Jingkuan Song, Dongxiang Zhang, Heng Tao Shen*
Center for Future Media and School of Computer Science and Engineering,
University of Electronic Science and Technology of China, Chengdu 611731, China
{lianli.gao, zhangdo}@uestc.edu.cn, jingkuan.song@gmail.com, shenhengtao@hotmail.com

Abstract

Image co-segmentation is the problem of automatically discovering the common objects co-occurring in a set of relevant images and segmenting them as foreground simultaneously. Although a bunch of approaches have been proposed to address this problem, many of them still suffer from certain limitations, e.g., supervised feature learning and complex models, which hinder their capability in the real-world scenarios. To alleviate these limitations, we propose a novel coarse-to-fine co-segmentation (CFC) framework, which utilizes the coarse foreground and background proposals to learn a robust similarity measure of the features in an unsupervised way, and then devises a simple objective function based on the definition of image co-segmentation. Specifically, we first generate superpixels for all the images and extract their features. Instead of using existing distance metrics, we utilize object proposal methods to generate coarse foreground and background to learn a similarity measure of superpixels to construct a robust feature similarity graph. Then we design an intuitive objective function to learn a segmentation similarity graph which should be consistent with feature similarity graph and also be able to co-segment the superpixels in the images into either foreground and background. This objective function can be further reformulated as a graph learning problem with intra and inter rank constraints. Experiments on two commonly used image datasets (iCoseg and MSRC) demonstrate that CFC outperforms other state-of-the-art methods. Notably, this performance is achieved by using only HSV feature.

1 Introduction

Co-segmentation, i.e., jointly segmenting the common objects from a collection of similar images, has attracted an increasing attention recently in computer vision community [Joulin *et al.*, 2010; Yuan *et al.*, 2017; Mukherjee *et al.*, 2018; Maninis *et al.*, 2017; Song *et al.*, 2016]. Compared with single image segmentation [Gao *et al.*, 2016], co-segmentation has the advantage of utilizing the very weak prior that the set of images contain the same objects to improve the segmentation of individual images. At the meanwhile, co-segmentation is a challenging problem, which faces several major issues: 1) The classical segmentation methods are designed for a single image while co-segmentation deals with multiple images. How to design a co-segmentation model and minimize the model is crucial for image co-segmentation. 2) The common object extraction depends on the foreground similarity measurement. But the foreground usually varies in color, shape and scale, which makes the foreground similarity measurement difficult. Moreover, the model minimization is highly related to the selection of similarity measurement. Therefore, the foreground similarities measurement is another important issue.

A straight-forward way for co-segmentation is to extend classical single image based segmentation models. In general, the extended models can be represented as

$$\ell = \ell_{\mathbf{s}} + \ell_{\mathbf{g}} \tag{1}$$

where $\ell_{\mathbf{s}}$ is the single image segmentation term, which guarantees the smoothness and the distinction between foreground and background in each image, and $\ell_{\mathbf{g}}$ is the co-segmentation term, which focuses on evaluating the consistency between the foregrounds among the images.

Many classical single image segmentation models can be used to form $\ell_{\mathbf{s}}$, e.g, MRF segmentation models [Lee *et al.*, 2015] and discriminative clustering [Joulin *et al.*, 2010]. The co-segmentation term $\ell_{\mathbf{g}}$ is used to evaluate the multiple foreground consistency, which is introduced to guarantee the common object segmentation. Due to the variance of the foreground, the similarity measure using visual features is not accurate enough.

On the other hand, segmentation proposal selection based approaches [Vicente *et al.*, 2011; Meng *et al.*, 2012; 2013] have recently drawn more attention. These methods aim at select the real common objects from the segmentation proposals, and the assumption that the common targets should be real-world objects makes co-segmentation much more efficient. Compared with co-segmenting images using pix-

*Corresponding author: Lianli Gao, Heng Tao Shen

els, segmentation proposal selection based approaches have their advantages. First, instead of starting from scratch, they incorporate either class-specific or class-independent object proposal methods to generate a set of candidates, and then look for a pool of regions that have high probability to cover the objects by some of the proposals. The proposals are usually more robust, and they contain rich information than pixels. Therefore, they usually achieve better accuracy than methods based on pixels. Secondly, they are more efficient compared with pixel-based methods, because the number of proposals is usually much small than the number of pixels in images. However, their disadvantages are obvious as well. They usually reply heavily on the performance of object proposals. Also, due to the variety of the common objects, it usually requires manually selecting features [Meng *et al.*, 2012] or supervised feature learning performed beforehand [Vicente *et al.*, 2011; Meng *et al.*, 2013].

To address the above issues, in this work, we propose a novel coarse-to-fine image co-segmentation framework, which introduces graph-based image co-segmentation using superpixels based on intra and inter rank constraints. To take advantage of proposal, we utilize the coarse foreground and background proposals to learn a robust similarity measure of the superpixels. It is worthwhile to highlight the following aspects of our method here: 1) We propose a novel coarse-to-fine image co-segmentation framework. To take advantage of object proposals but avoid heavily depending on the proposal quality, we utilize multiple image proposal strategies and generate a set of coarse foreground and background. These incomplete but with few noise coarse segments are further used to learn a similarity measurement of the raw features to obtain a fine image co-segmentation. 2) We design an intuitive graph-based objective function for image co-segmentation based on its definition. Taking the feature similarity graph as input, we aim to learn a segmentation similarity graph which should be consistent with feature similarity graph and also be able to co-segment the superpixels in images into either foreground or background. We devise an efficient solution for this objective function. 3) Experiments for image co-segmentation on real-world datasets demonstrate the superiority of our method over existing methods. Notably, this performance is achieved using HSV feature only.

2 Our Approach

In this section, we introduce our method which consists of two phases (see Fig. 1). Firstly, we further introduce how to construct an original similarity graph using the coarse foreground and background proposals. Then, a graph-based image co-segmentation algorithm is proposed based on the original similarity graph of the superpixels.

We first introduce the notations which will be used in the rest of the paper. Given a set of M images $\mathbf{I} = \{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_M\}$ for an object class, we first generate a set of N oversegmenting superpixels $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_N\}$ as a preprocessing, and then extract features $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ from these N superpixels. A superpixel is an image segment consisting of pixels that have similar visual characteristics. The goal is to learn the similarity matrix $\mathbf{S}$ between each superpixel based on the features $\mathbf{X}$, and all the superpixels are segmented to either foreground or background.

2.1 Coarse Foreground and Background Proposal for Original Similarity Graph Construction

Instead of segmenting common objects from an object class from scratch, we start by generating a set of coarse foreground and background proposals. The motivation is that if we can generate pure foreground and background with high precision but low recall, we can use these regions as coarse proposals to construct an accurate original similarity graph. Inspired by recent progress in object proposal and salient object detection, we find that those two techniques have potential to propose incomplete by with few noise objects in an image. This is because salient object detection (e.g., PCA Saliency [Margolin *et al.*, 2013]) is originally a task of predicting the eye-fixations on images, and recently has been extended to detecting regions with salient attention-grabbing objects, while object proposal (e.g.,MCG [Arbeláez *et al.*, 2014]) aims to generate a set of regions to delineate candidate objects in an image. However, current generated saliency maps are typically blurry, especially near the boundary of salient objects, while object proposal methods can provide accurate object boundaries but it is still difficult to obtain complete objects by merging together sets of regions, thus they are complementary to each other.

Therefore, in this study we take the advantages of both salient object detect [Margolin *et al.*, 2013] and object proposal [Arbeláez *et al.*, 2014] to produce coarse foreground and background, which respectively contain incomplete but relatively pure foreground and background, to refine our similarity graph. Here, we introduce the construction details. Given an image $I_g \in \mathbb{R}^{w \times h}$, where w and h are the width and length.

Saliency map and candidate mask generation. At beginning, PCA Saliency [Margolin *et al.*, 2013] takes I_g as input and produce a saliency map $s\_map \in \mathbb{R}^{w \times h}$, while MCG [Arbeláez *et al.*, 2014] takes I as input and produces a ranked list of object candidates. Here, we merge the top-5 ranked object candidates to output a candidate mask $c\_mask \in \mathbb{R}^{w \times h}$. If $c\_mask_{i,j} = 1$, it stands that the pixel $I_g\,(i, j)$ is an object pixel otherwise it is a background pixel.

Refine candidate with UCMs. After we obtained the initialized c_Mask, we apply ucms to refine c_Mask. Firstly, we extract superpixesl with $ucm \leq 0.5$ and then remove the superpiexles containing border pixels, except those superpixels meets one of the following conditions: having less than $\frac{1}{8}\,(w + h)$ border superpixels; 2) only containing left or right boarder pixels; and 3) the number of top/bottom boarder pixels is less than $\frac{3}{4}$ of the corresponding boarder length. Next, we project these superpixels to $ucm \leq 0.18$ and then merge with c_Mask to obtain a new c_Mask. Note that, all the parameter values are empirically set as above. Next, we move to coarse foreground and background generation.

Coarse background generation. At first, we compute a threshold th_b by averaging the values of s_Map that are lower than the mean of s_Map. Next, comparing each element of s_map with th_b to initialize the coarse background mask

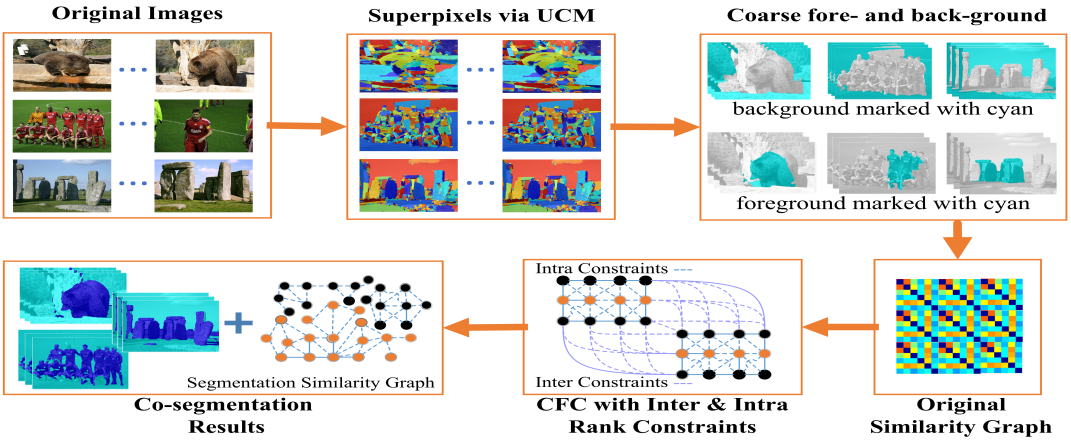

Figure 1: The overview of our proposed coarse-to-fine image co-segmentation method.

b_mask by setting the larger elements as 1, otherwise setting as 0. Finally, we sum the initialized b_mask with c_mask to update b_mask. If an element of b_mask is less than 2, it stands for the corresponding pixel is a pure background pixel.

Coarse foreground generation. First, we look for elements of s_map that are larger than s_Map mean and then set them as 1, otherwise set as 0, to form an initial coarse foreground mask f_mask. Next, we joint f_mask and c_mask to obtain the coarse foreground mask f_mask. If the element of f_mask is 1, the corresponding pixel is pure foreground.

2.2 Image Co-segmentation with Intra and Inter Rank Constraints

Instead of following the conventional pipeline for image co-segmentation, in our framework (see Fig. 1), we propose a novel perspective to solve the graph-based image co-segmentation problem. Based on the definition of image co-segmentation, all the images have two segments (foreground and background), and each image also has two segments. Conventional methods usually construct a similarity graph $\mathbf{A}$ based on the features of superpixels, and all the superpixels are connected as just one connected component. Ideally, if the superpixels of all the images (and also for each image) are connected as exactly 2 components, the superpixels are naturally divided into 2 segmentations and this segmentation is the optimal solution for different criteria, e.g., normalized cut [Shi and Malik, 2000]. Then we can get our objective function:

$$\begin{aligned} &\min_{\mathbf{S}} \|\mathbf{S}-\mathbf{A}\|_F^2 \\ &st. \begin{cases} s_{ij} \geq 0, \mathbf{s}_i\mathbf{1} = 1, \\ \mathbf{S} \text{ has 2 connected components}, \\ \mathbf{S}_m \text{ has 2 connected components} \end{cases} \end{aligned} \tag{2}$$

where $\mathbf{S}$ is the segmentation similarity graph to be learned, and $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the original similarity graph. $\mathbf{S}_m$ is a block of $\mathbf{S}$ and it is the segmentation similarity matrix of image $\mathbf{I}_m$. The object function aims to learn a $\mathbf{S}$ which is similar to $\mathbf{A}$. To avoid the case that some rows of $\mathbf{S}$ are all zeros, we further constrain the $\mathbf{S}$ such that the sum of each row of $\mathbf{S}$ is one. Under these constraints, we learn a $\mathbf{S}$ that best approximates the initial affinity matrix $\mathbf{A}$. The last two constraints require that all the images have two segments (foreground and background), and each image also has two segments, which is the definition of image co-segmentation.

The problem in (2) is not easy to solve because of the constraints. To tackle this, we start from the following theorem. For a nonnegative similarity matrix $\mathbf{S}$, there is a Laplacian matrix $\mathbf{L}$ associated with it. According to the definition of Laplacian matrix, suppose each superpixel $\mathbf{x}_i$ is assigned a random value as $\mathbf{f}_i \in \mathbb{R}^{1\times K}$, then $\mathbf{L}$ can be calculated as:

$$\sum\nolimits_{ij} \|\mathbf{f}_i - \mathbf{f}_j\|_2^2 s_{ij} = 2tr\left(\mathbf{F}^T\mathbf{L}\mathbf{F}\right) \tag{3}$$

where $\mathbf{F} \in \mathbb{R}^{N\times K}$ with the i-th row formed by $\mathbf{f}_i$, $\mathbf{L} = \mathbf{D} - \frac{\mathbf{S}^t+\mathbf{S}}{2}$ is called the Laplacian matrix in graph theory, the degree matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$ is defined as a diagonal matrix where the i-th diagonal element is $\sum_j (s_{ji} + s_{ij})/2$. The Laplacian matrix $\mathbf{L}$ has the following property.

Theorem 1 *The number K of the eigenvalue 0 of the Laplacian matrix $\mathbf{L}$ is equal to the number of connected components in the graph with the similarity matrix $\mathbf{S}$ if $\mathbf{S}$ is nonnegative.*

Theorem 1 indicates that if $rank(\mathbf{L}) = N - 2$, then the superpixels have 2 connected components based on $\mathbf{S}$. Thus, (2) can be reformulated as:

$$\begin{aligned} &\min_{\mathbf{S}} \|\mathbf{S}-\mathbf{A}\|_F^2 \\ &st. \begin{cases} s_{ij} \geq 0, \mathbf{s}_i\mathbf{1} = 1 \\ \text{rank}\left(\mathbf{L}\right) = N-2 \\ \text{rank}\left(\mathbf{L}_m\right) = N-2 \end{cases} \end{aligned} \tag{4}$$

where $\mathbf{L}$ is the laplacian matrix of $\mathbf{S}$, and $\mathbf{L}_m$ is the laplacian matrix of $\mathbf{S}_m$. It is still difficult to solve the problem (4). Because $\mathbf{L} = \mathbf{D} - \left(\mathbf{S}^T + \mathbf{S}\right)/2$ and $\mathbf{D}$ also depends on $\mathbf{S}$. The constraint $\text{rank}(\mathbf{L}) = N-2$, $\text{rank}(\mathbf{L}_m) = N-2$ are not easy to tackle. Suppose e_i is the i-th smallest eigenvalue of $\mathbf{L}$ and $(e_m)_i$ is the i-th smallest eigenvalue of $\mathbf{L}_m$, we know $e_i, (e_m)_i \geq 0$ since L and L_m are positive semi-definite. It

can be seen that the problem (4) is equivalent to the following problem for a large enough value of α and β [1]:

$$\begin{aligned} &\min_{\mathbf{S}} \|\mathbf{S}-\mathbf{A}\|_F^2 + 2\alpha \sum_{i=1}^{2} e_i + 2\beta \sum_{i=1}^{2} (e_m)_i \\ &s.t.\ s_{ij} \geq 0, \mathbf{s}_i\mathbf{1} = 1 \end{aligned} \tag{5}$$

When α and β are set to a large enough value, $\sum_i^2 e_i$ and $\sum_i^2 (e_m)_i$ will be imposed to be close to 0, which results in $rank\left(\mathbf{L}\right) = N-2$ and $rank\left(\mathbf{L}_m\right) = N-2$.

According to the Ky Fan's Theorem [Fan, 1949], we have:

$$\sum_{i=1}^{2} e_i = \min_{\mathbf{F}\in\mathbb{R}^{N\times 2}, \mathbf{F}^T\mathbf{F}=\mathbf{I}} tr\left(\mathbf{F}^T\mathbf{L}\mathbf{F}\right) \tag{6}$$

Therefore, (5) is further equivalent to:

$$\begin{aligned} &\min_{\mathbf{S},\mathbf{F},\mathbf{F}_m} \|\mathbf{S}-\mathbf{A}\|_F^2 + 2\alpha tr\left(\mathbf{F}^T\mathbf{L}\mathbf{F}\right) + 2\beta \sum_{m=1}^{M} tr\left(\mathbf{F}_m^T\mathbf{L}_m\mathbf{F}_m\right) \\ &st. \begin{cases} s_{ij} \geq 0, \mathbf{s}_i 1 = 1 \\ \mathbf{F} \in \mathbb{R}^{N\times 2}, \mathbf{F}^T\mathbf{F} = \mathbf{I}_N \\ \mathbf{F}_m \in \mathbb{R}^{N_m\times 2}, {\mathbf{F}_m}^T\mathbf{F}_m = \mathbf{I}_{N_m} \end{cases} \end{aligned} \tag{7}$$

Compared with the original problem 2, the problem 7 is much easier to solve.

2.3 Iterative Optimization

We propose an iterative method to minimize the above objective function in Eq.7. We first formally define the relationships between $\mathbf{S}$ to $\mathbf{S}_m$, and $\mathbf{F}$ to $\mathbf{F}_m$. Recall that $\mathbf{S}$ is the original similarity graph of all images, and $\mathbf{S}_m$ is the similarity graph of the m-th image. Therefore, $\mathbf{S}_m = \mathbf{S}\left(ids_m : ide_m, ids_m : ide_m\right)$, where ids_m and ide_m are the index for the starting and ending position of superpixels in M images set. Formally, we define a selection matrix $\mathbf{U} \in \mathbb{R}^{N\times N_m}$ so that $\mathbf{S}_m = \mathbf{U}_m^T\mathbf{S}\mathbf{U}_m$, where $\mathbf{U}$ is a matrix with all zeros, but the ids_m to ide_m rows are filled by an identity matrix $\mathbf{I}_{N_m}$. Similarly, we can get $\mathbf{F}_m = \mathbf{U}_m^T\mathbf{F}$.

By taking the original similarity matrix $\mathbf{A}$ as input, in each iteration, our algorithm first updates $\mathbf{F}$ and $\mathbf{F}_m$ given $\mathbf{S}$, and then updates $\mathbf{S}$ by fixing $\mathbf{F}$ and $\mathbf{F}_m$. These steps are described below.

Update F and $\mathbf{F}_m$

By fixing $\mathbf{S}$, we can obtain $\mathbf{F}$ and $\mathbf{F}_m$ by optimizing Eq.7. This is equivalent to optimize the following objective function:

$$\begin{aligned} &\min_{\mathbf{F},\mathbf{F}_m} \alpha tr\left(\mathbf{F}^T\mathbf{L}\mathbf{F}\right) + \beta \sum_{m=1}^{M} \mathbf{F}_m^T\mathbf{L}_m\mathbf{F}_m \\ &s.t.\ \mathbf{F}^T\mathbf{F} = \mathbf{I}_N, \mathbf{F}_m^T\mathbf{F}_m = \mathbf{I}_{N_m} \end{aligned} \tag{8}$$

which can be further reformulated as:

$$\begin{aligned} &\min_{\mathbf{F}} tr\left(\mathbf{F}^T\left(\alpha\mathbf{L} + \beta\sum_{m=1}^{M}\mathbf{U}_m\mathbf{L}_m\mathbf{U}_m^T\right)\mathbf{F}\right) \\ &s.t.\ \mathbf{F}^T\mathbf{F} = \mathbf{I}_N, \mathbf{F}_m^T\mathbf{F}_m = \mathbf{I}_{N_m} \end{aligned} \tag{9}$$

$\mathbf{F}$ can be solved by obtaining the two smallest eigenvector of $(\alpha\mathbf{L} + \beta\sum_{m=1}^{M}\mathbf{U}_m\mathbf{L}_m\mathbf{U}_m^T)$, and $\mathbf{F}_m = \mathbf{U}^T\mathbf{F}$.

[1] In the real implementation, we initialize with 1000, and increase α, β to $\alpha\times 2, \beta\times 2$ if the current number of connected components is less than 2, and decrease α, β to $\alpha/2, \beta/2$ if the current number of connected components is larger than 2

Update S

By fixing $\mathbf{F}$ and $\mathbf{F}_m$, we can obtain $\mathbf{S}$ by optimizing Eq.7. It is equivalent to optimize the following objective function:

$$\begin{aligned} &\min_{\mathbf{S},\mathbf{S}\geq 0,\mathbf{S}\mathbf{1}=1} \|\mathbf{S}-\mathbf{A}\|_F^2 + \alpha\sum_{i,j=1}^{N}\|\mathbf{f}_i - \mathbf{f}_j\|^2 s_{ij} \\ &+\beta\sum_{m=1}^{M}\sum_{p,q=1}^{N_m}\left\|\left(\mathbf{f}_m\right)_p - \left(\mathbf{f}_m\right)_q\right\|^2 \left(\mathbf{s}_m\right)_{pq} \end{aligned}, \tag{10}$$

and we can rewrite the second part as:

$$\sum_{m=1}^{M}\sum_{p,q=1}^{N_m}\left\|\left(\mathbf{f}_m\right)_p - \left(\mathbf{f}_m\right)_q\right\|^2 \left(\mathbf{s}_m\right)_{pq} = \sum_{i,j=1}^{N}\sum_{m=1}^{M}\left(b_m\right)_{ij} s_{ij} \tag{11}$$

where $\mathbf{B}_m = \mathbf{U}_m\mathbf{V}_m\mathbf{U}_m^T$ and $\left(v_m\right)_{pq} = \left\|\left(\mathbf{f}_m\right)_p - \left(\mathbf{f}_m\right)_q\right\|^2$. Then (10) becomes:

$$\begin{aligned} &\min_{\mathbf{S}} \|\mathbf{S}-\mathbf{A}\|_F^2 + \sum_{i,j=1}^{N}\left(\alpha\|\mathbf{f}_i - \mathbf{f}_j\|^2 + \beta\sum_{m=1}^{M}\left(b_m\right)_{ij}\right)s_{ij} \\ &s.t.,\ s_{ij}\geq 0, \mathbf{s}_i\mathbf{1} = 1 \end{aligned} \tag{12}$$

Denote $g_{i,j} = \alpha\|\mathbf{f}_i - \mathbf{f}_j\|^2 + \beta\sum_{m=1}^{M}\left(b_m\right)_{ij}$, and (12) can be reformulated as:

$$\begin{aligned} &\min_{\mathbf{S}, s_{ij}\geq 0, \mathbf{s}_i\mathbf{1}=1} \sum_{i,j=1}^{N}\left(s_{ij} - a_{ij}\right)^2 + \sum_{i,j=1}^{N} g_{ij}s_{ij} \\ &\Rightarrow \min_{\mathbf{S}, s_{ij}\geq 0, \mathbf{s}_i\mathbf{1}=1} \sum_{i=1}^{N}\left\|\mathbf{s}_i - \left(\mathbf{a}_i - \tfrac{1}{2}\mathbf{g}_i\right)\right\|_2^2 \end{aligned} \tag{13}$$

The problem in Eq.13 is simplex [Nie *et al.*, 2016] and the critical step of the projected gradient method is to solve the following proximal problem:

$$\min_{\mathbf{x}\geq 0, \mathbf{x}^T\mathbf{1}=1} \frac{1}{2}\|\mathbf{x} - \mathbf{c}\|_2^2 \tag{14}$$

This proximal problem can be solved by an efficient iterative algorithm [Nie *et al.*, 2016]. Then each $\mathbf{s}_i$ can be efficiently solved, and we can get the updated graph $\mathbf{S}$.

We update $\mathbf{F}$ and $\mathbf{S}$ iteratively until the objective function Eq.7 converges.

3 Experiments

We evaluate our algorithm on the task of image co-segmentation. Firstly, we compare our results with state-of-the-art algorithms on two standard datasets. Then, we study the influence of different features for our algorithm.

3.1 Experimental Settings

Datasets

We evaluated the proposed algorithm on two public benchmark datasets: the iCoseg dataset [Batra *et al.*, 2010] and the MSRC dataset [Winn *et al.*, 2005]. iCoseg dataset contains 38 image classes of totally 643 images with manually labeled pixel-wise ground-truth masks. We follow [Lee *et al.*, 2015] to use a subset of 16 object classes. The later includes 591 pixelwise labeled images of 23 object classes. We select the same classes as reported by [Joulin *et al.*, 2010;

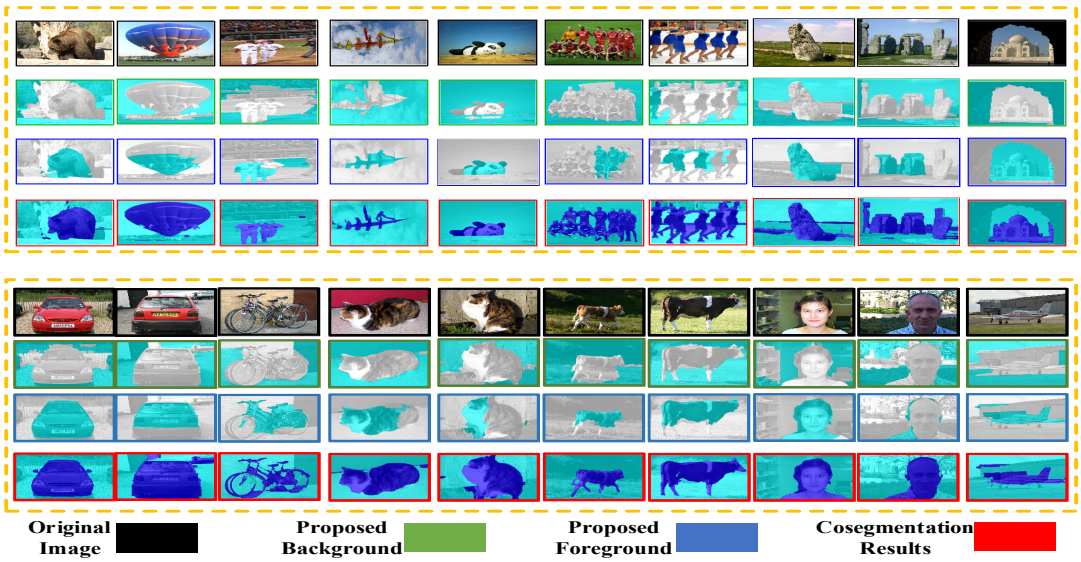

Figure 2: Qualitative results for co-segmentation on two datasets.

Lee *et al.*, 2015]. The complex background of the MSRC dataset makes it more challenging, and it involves considerable changes in viewpoints and illumination.

Following [Lee *et al.*, 2015], we evaluate accuracy, precision and recall for the segmentation. A segmentation accuracy is defined as the percentage of correctly labeled pixels, while precision and recall are defined as the percentage of correctly labeled foreground.

Features

We utilize two kinds of image features, i.e., the shallow features and the deep features, to capture different characteristics of each superpixel. Specifically, we used HSV feature in our work, and we further test another three kinds of shallow features, including LAB, texture, and dense SIFT descriptors. Software from [van de Sande *et al.*, 2011] is used, and the dimensions for each feature is 75-D for HSV, 75-D for LAB, 240-D for texture, and 1024-D for SIFT.

For deep feature, we employ the CNN-S [Chatfield *et al.*, 2014] model which is pre-trained on the ImageNet dataset. Firstly, we feed each image into the pre-trained CNN and get the responses from the last convolutional layer, 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.

We run our algorithm using single image (IS) and co-segmentation (CS). Instead of running our method on the proposed non-background regions, we also test the performance of our algorithm using the full image (FI).

3.2 Results on iCoseg Dataset

To evaluate the performance of our method, we compare it with different image co-segmentation algorithms, e.g., MRW [Lee *et al.*, 2015], RM [Rubio, 2012], OC [Vicente *et al.*, 2011], CFM [Wang *et al.*, 2013], CFO [Li *et al.*, 2016].

	Class	RM	OC	CFM	CFO	MRW	Our		
						MRW	FI	IS	CS
1	Alaskan bear	86.4	90.0	90.4	91.6	87.3	92.0	86.7	**93.5**
2	Balloon	89.0	90.1	90.4	94.9	97.7	99.1	95.2	**99.1**
3	Baseball	90.5	90.9	94.2	97.1	97.1	97.1	97.7	**97.2**
4	Bear	80.4	**95.3**	88.1	93.8	93.7	86.3	72.3	94.1
5	Elephant	75.0	43.1	86.7	93.0	**93.1**	82.1	85.8	88.7
6	Ferrari	84.3	89.9	**95.6**	91.7	91.9	91.6	83.2	91.7
7	Gymnastics	87.1	91.7	90.4	95.0	96.1	97.8	86.0	**97.6**
8	Kite	89.8	90.3	93.9	**96.7**	95.7	97.1	85.8	96.3
9	Kite panda	78.3	90.2	93.1	93.9	96.0	94.6	90.5	**96.2**
10	Liverpool	82.6	87.5	89.4	93.3	88.5	**96.2**	95.4	94.0
11	Panda	60.0	**92.7**	88.6	76.9	84.8	82.2	70.4	89.7
12	Skating	76.8	77.5	78.7	**98.9**	91.6	77.7	77.8	89.9
13	Statue	91.6	93.8	**96.8**	93.7	94.5	99.2	94.6	95.8
14	Stonehenge	87.3	63.3	92.5	91.6	95.9	93.5	87.5	**96.6**
15	Stonehenge2	88.4	88.8	87.2	87.1	**90.7**	86.7	72.7	87.2
16	Taj Mahal	88.7	91.1	92.6	88.0	95.2	97.8	96.3	**97.4**
	Average	83.5	85.4	90.5	92.3	93.1	92.0	86.1	**94.1**

Table 1: Accuracy of state-of-the-art image co-segmentation algorithms on iCoseg dataset

Table 1 shows the accuracy of state-of-the-art image co-segmentation algorithms on iCoseg dataset and Table 2 depicts the precision and recall for the different methods. We also show some qualitative results in Fig. 2. From these figures, we have the following observations:

Compared to state-of-the-art image co-segmentation algorithms, our method achieves better performance in terms of segmentation average accuracy. This indicates that our graph-based coarse-to-fine image co-segmentation strategy can result in a good segmentation. MRW [Lee *et al.*, 2015] and [Li *et al.*, 2016] are strong competitors, and they outperform our method for some object classes, e.g., 'elephant', 'skating' and 'stonehenge2'. One reason is that the proposed non-background regions of our method is not that accurate, and

	Class	P MRW [Lee *et al.*, 2015]	R MRW [Lee *et al.*, 2015]	P Our CS	R Our CS
1	Alaskan bear	45.0	97.2	79.4	91.4
2	Balloon	66.2	94.3	96.3	96.4
3	Baseball	84.5	73.6	87.7	74.3
4	Bear	75.9	98.5	76.8	97.8
5	Elephant	71.9	97.1	77.2	74.9
6	Ferrari	81.4	92.4	98.6	68.0
7	Gymnastics	99.6	75.4	97.2	88.7
8	Kite	95.2	94.6	89.7	83.3
9	Kite panda	94.5	93.1	90.8	98.0
10	Liverpool	56.2	79.4	66.8	96.7
11	Panda	95.4	71.5	86.7	93.5
12	Skating	95.1	93.6	85.7	91.4
13	Statue	99.1	77.4	98.7	83.4
14	Stonehenge	74.0	97.9	88.7	92.6
15	Stonehenge2	82.3	96.6	89.1	86.1
16	Taj Mahal	80.8	92.8	89.2	95.4
	Average	81.1	89.1	87.4	88.2

Table 2: Precision and Recall of state-of-the-art image co-segmentation algorithms on iCoseg dataset

some foreground are not covered by the non-background regions, e.g., 'elephant' and 'stonehenge2'. Another reason is that HSV is a simple visual feature, and is not powerful enough to segment visually different superpixels into one part, e.g., the black skirts and white pants for 'skating'.

Compared with FI, the performance for CS is improved by 2.1%. This indicates that our proposed non-background regions can get rid of some noise, and improve our graph-based segmentation algorithm. On the other hand, the improvement of CS over IS is more significant, and it is 8.0%. IS attempts to separate an object from its background in a single image, but without consideration of multiple images, the definition of 'object' is ambiguous. For instance, given the single image of 'baseball', it is not clear whether the object should consist of the people in white only or all the people. Thus, IS may fail to delineate desired objects, especially in the cases of several objects. In contrast, CS overcomes the ambiguity, by extracting regions that occur repeatedly across images.

Comparing with MRW [Lee *et al.*, 2015] in terms of precision and recall, our method CS achieves much better performance for precision with 7.4%, but slightly worse performance for recall with 0.7% decrease. These results show that our CS can generate relative smaller foreground than MRW, but these regions are more precise.

The qualitative results are demonstrated as the top-rows in Fig. 2 . The results show that our CS can successfully co-segment images for simple background (e.g., 'balloon'), complex background (e.g., 'alaskan bear'), single object (e.g., 'panda') and multiple objects (e.g., 'skating'). In general, better performance is achieved for single object segmentation with simple background. In cases of complex background and foreground, e.g., 'skating' and 'Stonehenge2', our segmentation performance is unsatisfactory, and the proposed foreground and background have more noise than other cases.

3.3 Results on MSRC Dataset

On MSRC dataset, we compare our method to state-of-the-art image co-segmentation algorithms, e.g., MRW, DC [Joulin *et al.*, 2010], RM [Rubio, 2012], CFM [Wang *et al.*, 2013]. Table 3 shows the accuracy of state-of-the-art image co-segmentation algorithms on MSRC dataset and Table 4 depicts the precision and recall for the different methods. We also show some qualitative results in Fig. 2. We have the following observations:

	Class	RM	DC	CFM	MRW MRW	Our FI	Our IS	Our CS
1	car(front)	65.9	87.6	87.3	84.3	86.0	66.6	**88.7**
2	car(back)	52.4	85.1	**92.7**	81.7	64.7	66.8	85.7
3	bike	62.4	63.3	74.8	70.1	61.6	60.7	**76.2**
4	cat	77.1	74.4	88.3	82.2	83.7	71.6	**90.1**
5	cow	80.1	81.6	89.7	93.1	92.4	86.2	**93.2**
6	face	76.3	84.3	**89.3**	81.0	74.8	73.2	83.6
7	plane	77.0	73.8	87.3	**93.0**	67.8	61.1	85.1
	Average	70.8	78.6	**87.1**	83.7	76.0	69.2	**86.2**

Table 3: Accuracy of state-of-the-art image co-segmentation algorithms on MSRC dataset

	Class	P MRW [Lee *et al.*, 2015]	R MRW [Lee *et al.*, 2015]	P Our CS	R Our CS
1	car(front)	99.3	68.4	95.7	79.8
2	car(back)	94.4	68.3	85.8	85.7
3	cat	57.2	89.1	64.3	76.2
4	cow	74.1	53.5	83.0	80.9
5	dog	95.6	77.1	91.1	83.7
6	plane	66.1	58.4	57.2	72.1
7	sheep	55.1	88.7	59.5	78.1
	Average	77.4	71.9	76.7	79.5

Table 4: Precision and Recall of state-of-the-art image co-segmentation algorithms on MSRC dataset

Compared to state-of-the-art image co-segmentation algorithms, our method CS achieves comparable performance in terms of segmentation average accuracy. This indicates that our CS can result in a good segmentation for cases of complex background. [Wang *et al.*, 2013] has better accuracy than our CS, and it outperform our method for object classes of 'car(back)' and 'face' by 7.0% and 5.7%. However, they use multiple features, and our CS achieves this performance by using HSV feature only. By using sophisticated features or multiple features, the performance can be potentially improved. On the other hand, our CS obtains the best performance for 'car(front)', 'bike', 'cat' and 'cow', and it consistently outperforms MRW,[Rubio, 2012] and [Joulin *et al.*, 2010] except for the class of 'plane'.

Compared with FI, the performance for CS is improved by 10.2%, which is greater than iCoseg dataset. The indicates that for complex background, our proposed non-background regions can better get rid of noise. On the other hand, the improvement of CS over IS is much more significant, and it is 17.0%. The reason for the improvement of CS over IS is similar to that of iCoseg dataset. CS overcomes the ambiguity of IS by extracting regions that co-occur in all the images.

Comparing with MRW [Lee *et al.*, 2015] in terms of precision and recall, our method CS achieves much better performance for recall, but slightly worse performance for precision. These show that our CS can segment more regions to foreground than MRW, without decreasing the precision.

The qualitative results are demonstrated as the bottom-rows in Fig. 2. The qualitative results show that our CS can successfully co-segment images for simple background (e.g., 'cow'), complex background (e.g., 'face'), single object (e.g., 'face') and multiple objects (e.g., 'bike'). In general, bet-

ter performance is achieved for single object segmentation with simple background. In cases of complex background and foreground, e.g., 'car(front)' and 'airplane', the proposed foreground and background have more noise then other cases. Therefore, our segmentation performance is unsatisfactory.

4 Conclusions

In this work, we propose a novel unsupervised coarse-to-fine co-segmentation (CFC) framework, which introduces graph-based image co-segmentation based on intra and inter rank constraints, and also utilizes the coarse foreground and background proposals to construct a robust original similarity graph. Experiments on two commonly used image datasets (iCoseg and MSRC) demonstrate that our method achieves comparable or even better performance than state-of-the-art methods by using only HSV feature.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities (Grant No. ZYGX2014J063, No. ZYGX2014Z007) and the National Natural Science Foundation of China (Grant No. 61772116, No. 61502080, No. 61632007, No. 61602049).

References

[Arbeláez *et al.*, 2014] Pablo Andrés Arbeláez, Jordi Pont-Tuset, Jonathan T. Barron, Ferran Marqués, and Jitendra Malik. Multiscale combinatorial grouping. In *CVPR*, pages 328–335, 2014.

[Batra *et al.*, 2010] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, pages 3169–3176, 2010.

[Chatfield *et al.*, 2014] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

[Fan, 1949] Ky Fan. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences of the United States of America*, 35(11):652, 1949.

[Gao *et al.*, 2016] Lianli Gao, Jingkuan Song, Feiping Nie, Fuhao Zou, Nicu Sebe, and Heng Tao Shen. Graph-without-cut: An ideal graph learning for image segmentation. In *AAAI*, pages 1188–1194, 2016.

[Joulin *et al.*, 2010] Armand Joulin, Francis R. Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, pages 1943–1950, 2010.

[Lee *et al.*, 2015] Chulwoo Lee, Won-Dong Jang, Jae-Young Sim, and Chang-Su Kim. Multiple random walkers and their application to image cosegmentation. In *CVPR*, pages 3837–3845, 2015.

[Li *et al.*, 2016] Kunqian Li, Jiaojiao Zhang, and Wenbing Tao. Unsupervised co-segmentation for indefinite number of common foreground objects. *IEEE Trans. Image Processing*, 25(4):1898–1909, 2016.

[Maninis *et al.*, 2017] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. *CoRR*, abs/1711.09081, 2017.

[Margolin *et al.*, 2013] Ran Margolin, Ayellet Tal, and Lihi Zelnik-Manor. What makes a patch distinct? In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 1139–1146, Washington, DC, USA, 2013. IEEE Computer Society.

[Meng *et al.*, 2012] Fanman Meng, Hongliang Li, Guanghui Liu, and King Ngi Ngan. Object co-segmentation based on shortest path algorithm and saliency model. *IEEE Trans. Multimedia*, 14(5):1429–1441, 2012.

[Meng *et al.*, 2013] Fanman Meng, Hongliang Li, King Ngi Ngan, Liaoyuan Zeng, and Qingbo Wu. Feature adaptive co-segmentation by complexity awareness. *IEEE Trans. Image Processing*, 22(12):4809–4824, 2013.

[Mukherjee *et al.*, 2018] Prerana Mukherjee, Brejesh Lall, and Snehith Lattupally. Object cosegmentation using deep siamese network. *CoRR*, abs/1803.02555, 2018.

[Nie *et al.*, 2016] Feiping Nie, Xiaoqian Wang, Michael I. Jordan, and Heng Huang. The constrained laplacian rank algorithm for graph-based clustering. In *AAAI*, pages 1969–1976, 2016.

[Rubio, 2012] José C. Rubio. Unsupervised co-segmentation through region matching. In *CVPR*, pages 749–756, 2012.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[Song *et al.*, 2016] Jingkuan Song, Lianli Gao, Mihai Marian Puscas, Feiping Nie, Fumin Shen, and Nicu Sebe. Joint graph learning and video segmentation via multiple cues and topology calibration. In *ACM Multimedia*, pages 831–840, 2016.

[van de Sande *et al.*, 2011] Koen E. A. van de Sande, Jasper R. R. Uijlings, Theo Gevers, and Arnold W. M. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, pages 1879–1886, 2011.

[Vicente *et al.*, 2011] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov. Object cosegmentation. In *CVPR*, pages 2217–2224, 2011.

[Wang *et al.*, 2013] Fan Wang, Qixing Huang, and Leonidas J. Guibas. Image co-segmentation via consistent functional maps. In *ICCV*, pages 849–856, 2013.

[Winn *et al.*, 2005] John M. Winn, Antonio Criminisi, and Thomas P. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, pages 1800–1807, 2005.

[Yuan *et al.*, 2017] Ze-Huan Yuan, Tong Lu, and Yirui Wu. Deep-dense conditional random fields for object co-segmentation. In *IJCAI*, pages 3371–3377, 2017.