Unsupervised Multi-Class Co-Segmentation via Joint-Cut Over L_1-Manifold Hyper-Graph of Discriminative Image Regions

Jizhou Ma, Shuai Li, Hong Qin, *Senior Member, IEEE*, and Aimin Hao

***Abstract*— This paper systematically advocates a robust and efficient unsupervised multi-class co-segmentation approach by leveraging underlying subspace manifold propagation to exploit the cross-image coherency. It can combat certain image co-segmentation difficulties due to viewpoint change, partial occlusion, complex background, transient illumination, and cluttering texture patterns. Our key idea is to construct a powerful hyper-graph joint-cut framework, which incorporates mid-level image regions-based intra-image feature representation and L_1-manifold graph-based inter-image coherency exploration. For local image region generation, we propose a bi-harmonic distance distribution difference metric to govern the super-pixel clustering in a bottom-up way. It not only affords drastic data reduction but also gives rise to discriminative and structure meaningful feature representation. As for the inter-image coherency, we leverage multi-type features involved L_1-graph to detect the underlying local manifold from cross-image regions. As a result, the implicit supervising information could be encoded into the unsupervised hyper-graph joint-cut framework. We conduct extensive experiments and make comprehensive evaluations with other state-of-the-art methods over various benchmarks, including iCoseg, MSRC, and Oxford flower. All the results demonstrate the superiorities of our method in terms of accuracy, robustness, efficiency, and versatility.**

***Index Terms*— Unsupervised co-segmentation, L_1-graph, hyper-graph joint-cut, bi-harmonic distance.**

I. Introduction and Motivation

CO-SEGMENTATION aims to jointly segment the co-occurring similar objects by exploiting mutual supervising information implied in image sets. And it facilitates many downstream applications, including object recognition, video segmentation, image-based modeling and analysis, etc. Following the pioneering co-segmentation work [1], various methods have been proposed by enhancing different technical foci, such as Markov random field (MRF) [1]–[5], discriminative clustering [6]–[8], sub-modular optimization and anisotropic diffusion [9], subspace clustering [10]–[12], hierarchical clustering [13], semi-supervised learning [14], [15], segmentation propagation [5], etc. Although co-segmentation methods have achieved growing successes and can accommodate more and more complex image sets, some common challenges still exist due to lacking enough flexibility, robustness, and efficiency. Specifically, the typical difficulties can be summarized as follows.

From the perspective of basic primitive generation, most of the methods explore local coherency based on low-level vision primitives such as pixel or super-pixel. It inevitably gives rise to less meaningful supervising information exploration, time-consuming calculation, and weak robustness. Some methods also employ patch-based mid-level primitives [13], [16]–[18] or object-based high-level primitives [5], [19] to facilitate the co-segmentation process. However, cross-image co-occurring contents may vary in shape, color, scale, occlusion, and local deformation. Thus, it is nontrivial to adaptively conduct meaningful pre-segmentation.

From the perspective of feature representation, most of the methods directly adopt color histogram [1], [3], [6] [7], [14] or bag-of-words [4], [6], [7], [20] related descriptors, which are hard to capture the intrinsic feature of co-occurring objects. Although some more advanced methods [13], [18], [21] begin to take complementary multi-features into account via histogram concatenation and/or its weighted superposition, it may cause unpredictable error accumulations for some basic primitives only having strict consistency in certain feature space.

From the perspective of global correlation analysis, simple $k-nearest-neighbor$ based graph construction [9] has become an off-the-shelf tool to explore the underlying intra-image and inter-image structures. However, such unbiased neighbor-sampling graph structure tends to weaken the anisotropy, because less discriminative coherency propagation may lead to overly-relaxed co-segmentation.

From the practical perspective, most of the state-of-the-art methods are hard to accommodate moderately large scale image sets due to computation overhead, such as building image correlation structure (e.g., pixel-wise kernels [6], [7]) and nonlinear optimization (e.g., the multi-task low-rank affinity pursuit [11]), which will increase in linear or even exponential time as the scale of image set grows. Moreover, except for low-level vision primitives based computation, most

Manuscript received November 20, 2015; revised June 20, 2016; accepted November 7, 2016. Date of publication November 21, 2016; date of current version January 30, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61672077, Grant 61672149, Grant 61602341, Grant 61532002, Grant 61190120, Grant 61190121, and Grant 61190125 and in part by the National Science Foundation of USA under Grant IIS-0949467, Grant IIS-1047715, and Grant IIS-1049448. *(Corresponding authors: Shuai Li and Aimin Hao.)*

J. Ma, S. Li, and A. Hao are with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China (e-mail: lishuai@buaa.edu.cn; ham@buaa.edu.cn).

H. Qin is with Stony Brook University, Stony Brook, NY 11794 USA.

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. The supplementary material contains a video file. The total size of the file is 35.3 MB.

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.2016.2631883

1057-7149 © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

of the time-consuming tasks in existing methods are hard to be parallelized.

To tackle the aforementioned challenges, based on the bi-harmonic distance definition in 2D image space and the graph clustering based global co-segmentation idea proposed in our previous work [8], we further propose some brand new technical elements and integrated them into a newly-designed L1-manifold hyper-graph based joint-cut framework. It gives rise to a novel and flexible unsupervised multi-class co-segmentation framework. Specifically, the salient contributions can be summarized as follows:

- We pioneer an L_1-manifold hyper-graph joint-cut framework for unsupervised multi-class co-segmentation, which gives rise to intrinsic exploitation of implied cross-image information in a bottom-up way, while affording light-weight GPU-parallel computation, and thus enables efficient, robust, and flexible co-segmentation.
- We suggest a meaningful local image region generation method based on the intra-image bi-harmonic distance distribution analysis, which not only affords the data reduction of the basic primitives to be processed, but also gives rise to soft invariant feature representation of flexible co-occurring objects.
- We formulate a coherency measurement based on L_1-manifold graph by integrating multi-type features, which defines the affinities of inter-image mid-level regions in a more meaningful way, and strengthens the anisotropy property of the coherency propagation of co-occurring candidates.

Besides, in our proposed framework, each involved technical element is well designed with full justification. In Section VIII-E, we quantitatively analyze the benefits and limitations of the key technical elements (bi-harmonic distance distribution based metric, mid-level regions, L_1-graph, and joint-cut) by intentionally disabling different elements one by one. Please refer to Section VIII-E for more details.

II. Related Work

A. Visual Primitives for Coherency Analysis

In the co-segmentation field, many methods directly employ pixels as the basic primitives for the coherency analysis [1]–[3], [12], [15]. To combat certain co-segmentation challenges due to multiple foregrounds, partially co-occurring objects, high-variability objects, and larger image set, Kim *et al.* [13], [22] adopted super-pixels as the primitives, and some others resort to object-level primitives [5], [10], [19]. Thus, researchers gradually form a consensus that the role of discriminative visual primitives is important. However, it is still hard to perform unsupervised meaningful object-level pre-segmentation. To deal with it, many methods in computer vision are proposed to employ mid-level regular-rectangle patches to serve as discriminative visual primitives [23]–[26], wherein the generated patches are expected to be structure-preserving as much as possible while avoiding segmenting an entire object into many overly-messy parts. To extend this concept and get better results, [18] took advantage of irregular mid-level regions via roughly pre-segmenting some candidate parts that may co-occur across the image set.

Given the basic visual primitives, a common way to compute intra-image and inter-image affinities is to measure the Euclidean distance or Chi-square distance between primitive-pair's histogram-like feature descriptors. Many works strive to find adapted descriptors to solve certain problem. Rubio *et al.* [4] encoded the graph matching information into the inter-image affinities based on MRF. Glasner *et al.* [27] proposed a novel region contour based descriptor. In addition, Kim and Xing [16] measured affinity based on Gaussian mixture model (GMM) and spatial pyramid matching (SPM), and Faktor and Irani [18] proposed a composed reconstruction error based affinity metric. By comparison, some physics-based metrics are more general to handle various complex cases. For example, anisotropic heat diffusion distance [28], commute time distance [4], [29], and geodesic distance [30], can enable more robust and informative affinity measurement for the flexible visual primitives with deformation, occlusion, and noise perturbation.

B. Correlation-Structure Construction

Upon the definition of affinity metric, it still needs to construct a correlation structure to facilitate the global coherency propagation towards co-segmentation. Currently, various graph based methods are widely used, such as $k-nearest-neighbor$ methods and $\epsilon-ball$ methods. And Hash table is also be applied to search nearest neighbor [31], [32]. To make the correlation-structure more anisotropic, many graph embedding based subspace learning methods are also proposed [12], [33]–[35]. As Mukherjee *et al.* [12] suggested, this kind of methods usually formulate co-segmentation problem as an elegant framework that permits general non-parametric appearance model compositions. Recently, sparse representation based correlation-graph [36] attracts more and more attentions because of its many built-in advantages, such as being robust to data noise, supporting subspace abstraction, and global sparsity. For example, Cheng *et al.* [36] experimentally demonstrated the superiority of L_1 graph in spectral clustering, subspace learning, and semi-supervised learning. Li *et al.* [37] devised an L_1-norm governed discriminative low rank matrix recovery algorithm for robust co-segmentation. And [10] and [11] further integrated the sparse representation based subspace clustering into a multi-feature co-segmentation framework. Moreover, being different from the aforementioned L_1-graph based works, Meng *et al.* [38] formulated a multi-feature selection based model, wherein the best feature combination can be learned adaptively by optimizing an L_1 regularized energy function.

C. Joint-Segmentation Models

Learning-based models are usually used to conduct co-segmentation, which usually employ priori knowledge to guide the labeling of co-occurring objects via iterative refinement. For instance, Wang and Liu [14] proposed a semi-supervised co-segmentation method to handle a large image set with only a few training-image foregrounds. Similarly, Kttel et al. [15] recursively propagated the already-segmented images to guide the segmentation of new images,

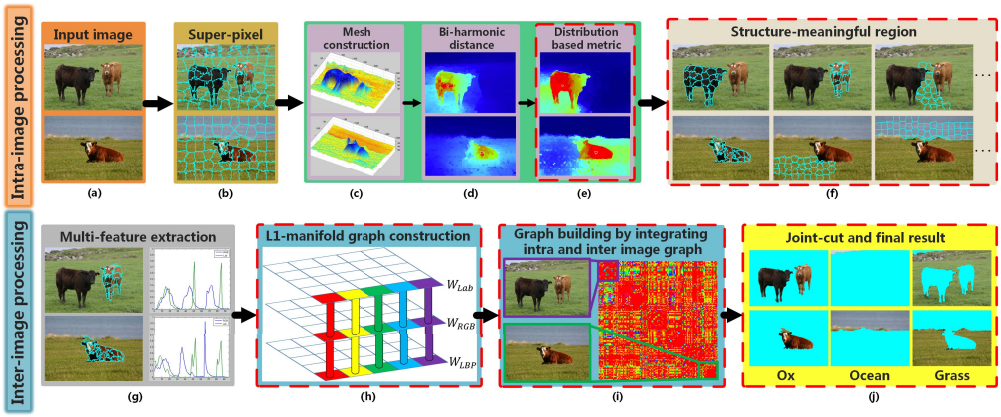


Fig. 1. The pipeline of our framework, wherein our salient contributions are highlighted with red dotted rectangles. (a) Input image set (only showing two images of the image set); (b) Over-segmented super-pixels; (c) Mesh construction and Laplacian matrix computation; (d) Bi-harmonic distance field over (c); (e) Bi-harmonic distance distribution difference based metric; (f) Generated mid-level structure-meaningful regions; (g) Multi-type features extraction for local regions; (h) Inter-image graph construction based on our defined L_1-manifold; (i) Hyper-graph construction by combining the inter-image L_1-manifold graph and intra-image coherency-structure graphs; (j) Final co-segmentation results.

wherein they designed an increasing pool with existing annotations to exploit the semantic hierarchy of ImageNet. As for noisy Web image collections, Wang *et al.* [39] trained their co-segmentation framework with dozens of top-ranked images obtained through text query. Besides, co-segmentation problem can also be casted as coherency-preserving spectral clustering or optimization based discriminative clustering problems [6], [8], [10], [13], [17], [27], [40]. Although all of these methods can be roughly classified into the clustering category, they all have unique technical highlights. For example, Rubinstein *et al.* [41] introduced the hypothesis that the co-occurring objects are cross-image salient ones. Meanwhile, to efficiently solve the co-clustering problem, Glasner *et al.* [27] formulated the problem as a quadratic semi-assignment problem. Joulin *et al.* [6], [7] imposed additional constraints to accommodate multi-class co-segmentation based on positive definite kernels. And Kim *et al.* [13] proposed hierarchical image clustering co-segmentation framework by taking into account the connections of multi-scale regions.

III. Method Overview

Fig. 1 shows the pipeline of our framework, wherein the novel parts corresponding to our contributions are highlighted with red dotted rectangles and also briefly described as follows:

A. Bi-Harmonic Distance Distribution Based Metric Definition

Following our previous work [8], we further define a new Laplace matrix by naturally incorporating 5D coordinates information. And then, we leverage the distribution differences of different super-pixels' bi-harmonic distance fields to define a physics-based intra-image coherency metric. Please refer to Section IV and Section V-A for details.

B. Structural Meaningful Mid-Level Region Generation

Based on the proposed metric, we iteratively build a preliminary anisotropic diffusion region for each super-pixel. The regions will be further used to guide the iterative merging and refinement of super-pixels towards meaningful image region generation. Please refer to Section V-B for details.

C. Multi-Type Features Involved L_1-Manifold Hyper-Graph Construction

We compute the affinities among inter-image regions by optimizing the L_1-norm constrained energy function, wherein the multi-type features are encoded to facilitate the subspace construction. Based on the inter-image L_1-manifold graph, we further adopt the proposed metric to only construct local spatially-adjacent intra-image graphes. Please refer to Section VI for details.

D. Unsupervised Hyper-Graph Joint-Cut Model

Following the bottom-up idea throughout this paper, which proceeds from pixel, to super-pixel, meaningful local regions, and hyper-graph segments gradually, we first adopt spectral clustering over this hyper-graph to roughly obtain co-segmentation candidates. And then, we generalize the conventional supervised GrabCut method [42] to an unsupervised case aided by a joint model for cross-image candidates. Finally, we conduct joint-cut over the hyper-graph to obtain the final co-segmentation results. Please refer to Section VII for details.

IV. Bi-Harmonic Distance Computation Over Super-Pixel Spanned Manifold

Bi-harmonic distance [43] is a type of intrinsic metric and has achieved great success in geometry processing, whose calculation is built upon the Laplacian matrix constructed on a manifold mesh. In our previous work [8], a mesh corresponding to a 2D image can be directly constructed by extruding the image plane along the pixel's gray value ($Z-axis$). As shown in Fig. 2(a), the flower image has been over segmented with super-pixels, and its corresponding 3D mesh is built by computing each super-pixel's average gray value to serve as the $Z-axis$ coordinate. And the top-row of Fig. 2(b) shows the 2D bi-harmonic distance distribution by projecting that of the 3D mesh (shown in the bottom-row of Fig. 2(b)) to 2D image.

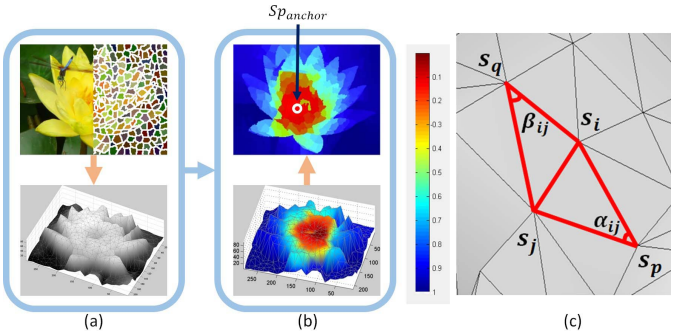


Fig. 2. Illustration of bi-harmonic distance based metric definition. (a) Super-pixel segmentation and corresponding manifold mesh construction; (b) Bi-harmonic distance distribution over the manifold, wherein the anchored super-pixel *Spanchor* is highlighted with white circle and the color bar shows the normalized distance; (c) The opposite angles α_{ij} and β_{ij} on the mesh.

Motivated by our previous work [8], we propose a new Laplacian matrix definition to better respect the image structural anisotropy based on Lab color features rather than gray values [8]. To begin with, we extrude the 2D image to a 3D mesh by assigning a 5D coordinate $(x,\ y,\ L,\ a,\ b)$ to each mesh vertex (super-pixel), wherein x and y denote an average spatial position of the pixels within certain super-pixel, and $L,\ a,\ b$ represent three channels of the Lab color. Then, the vertex set can be denoted as $S = \{s_1, s_2, \ldots, s_n\}$, wherein the mesh topology of these vertices (super-pixels) can be constructed via Delaunay triangulation. Next, we define the new bi-harmonic distance metric based on discrete Laplacian-matrix $\mathbf{L} = \mathbf{A}^{-1}\mathbf{M}$. Here $\mathbf{A}$ is a diagonal matrix to normalize the affinity matrix $\mathbf{M}$ (defined by opposite angles, and refer to Fig. 2(c)), and $\mathbf{A}_{ii}$ is proportional to the average area of the triangles sharing vertex s_i. $\mathbf{M}$ is formulated as

$$\mathbf{M}(i,j) = \begin{cases} \sum_{j=1}^{n} m_{ij} & \text{if } i = j \\ -m_{ij} & \text{if } s_i \text{ and } s_j \text{ are adjacent} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $m_{ij} = \cot\alpha_{ij} + \cot\beta_{ij}$. In particular, when computing the opposite angles α_{ij} and β_{ij} (Fig. 2(c)) involved in $\mathbf{M}$, we should take into account 5D information as

$$\begin{cases} \alpha_{ij} = \arccos\left(\dfrac{(\mathbf{v_i}-\mathbf{v_p})\ \cdot\ (\mathbf{v_j}-\mathbf{v_p})}{\left|(\mathbf{v_i}-\mathbf{v_p})\right|\ \cdot\ \left|(\mathbf{v_j}-\mathbf{v_p})\right|}\right) & \alpha_{ij} = \measuredangle s_i s_p s_j \\ \beta_{ij} = \arccos\left(\dfrac{(\mathbf{v_i}-\mathbf{v_q})\ \cdot\ (\mathbf{v_j}-\mathbf{v_q})}{\left|(\mathbf{v_i}-\mathbf{v_q})\right|\ \cdot\ \left|(\mathbf{v_j}-\mathbf{v_q})\right|}\right) & \beta_{ij} = \measuredangle s_i s_q s_j, \end{cases} \tag{2}$$

where $\mathbf{v_i} = (x_i,\ y_i,\ L_i,\ a_i,\ b_i)$ denotes the 5D components of vertex (super-pixel) s_i.

In sharp contrast, although our previous work [8] also uses the same 5D information, it considers them separately and takes $(L,\ a,\ b)$ as one dimension via deciding difference sign in gray space. This design can help the image mesh to be visualized in 3D space but cannot fully represent anisotropic property, because it makes the manifold shape be greatly affected and even disturbed by the gray value. As shown in Fig. 3, comparing with [8], the proposed Laplacian matrix definition gives rise to better anisotropic property. Specifically, the distance propagations of our method are almost along the local structures of "sky" and "ox" (See Fig. 3(c)). However, the structure-awareness results in [8] are more local under the same parameter setting with our current results. Moreover, since such distance in some sense represents dissimilarity, we can further transform the bi-harmonic distance matrix $\mathbf{d}_B$ to the similarity metric matrix $\mathbf{W}_B$ by the Gaussian function.

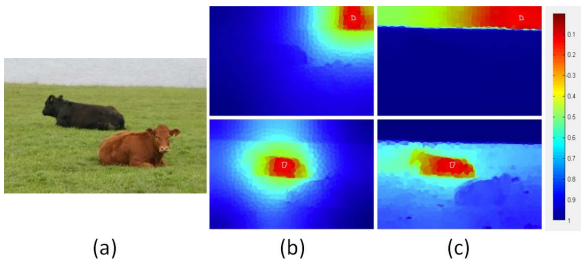


Fig. 3. Performance comparison between two types of Laplacian matrix definition. The color bar shows normalized distance projected from the manifold mesh to its corresponding image. (a) Original image; (b) Bi-harmonic distance distribution from the anchor super-pixel (white boundary) to others computed by the method of [8]; (c) Laplacian matrix defined by the proposed high-dimensional angles (Eq. 2) and the corresponding distance distribution from the anchor super-pixel (white boundary) to others.

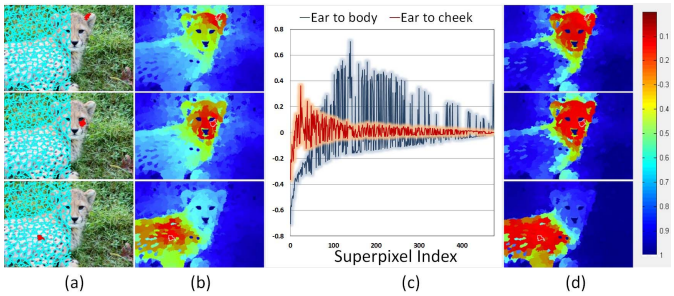


Fig. 4. Illustrations of why distribution differences can serve as intra-image coherency metric. The color bar shows normalized distance. (a) Super-pixel segmentation over original images, with anchor super-pixels (ear, cheek, and body) marked in red; (b) Bi-harmonic distance fields corresponding to different anchor super-pixels; (c) Illustration of the bi-harmonic distance distribution differences, red curve represents distribution differences from cheetah's ear to cheek $(\mathbf{W}_B(s_{ear},*) - \mathbf{W}_B(s_{cheek},*))$, and blue curve represents the distribution differences from cheetah's ear to body $(\mathbf{W}_B(s_{ear},*) - \mathbf{W}_B(s_{body},*))$. And super-pixel indices are sorted in the ascend order from the cheetah's ear to other regions of the image; (d) The dissimilarity measurement based on bi-harmonic distance distribution difference.

V. Structure-Meaningful Image Region Generation

A. Bi-Harmonic Distance Distribution Metric Definition

Comparing with bi-harmonic distance metric (Fig. 4(b)), we elaborate a more compact and local-global structure-aware metric (Fig. 4(d)) by exploring the bi-harmonic distance distribution differences among super-pixels. This metric goes beyond the pair-wise view of the distance to handle a more global case. It is motivated by two aspects. In the global sense, the super-pixels located within the same structure/object of an image should be not far from each other in the spatial and feature spaces. So, they should have similar bi-harmonic

distance distribution because of the intrinsic structure-aware property of this metric. Moreover, in the local sense, the distribution similarity between two super-pixels should be more expressive and compact compared with directly using the distance as measurement, because the encoded global information is able to eliminate the instability of local features to certain extent. In other words, although two parts belonging to one object may not be near under certain measurement, both of them are far from the background, which makes them close to each other in a global view.

The dissymmetric distribution difference matrix $\mathbf{d}_{DE}$ is defined as:

$$\mathbf{d}_{DE}(s_i, s_j) = \frac{F_{DE}(s_i, s_j) + \mathbf{W}_{Lab}(s_i, s_k) F_{DE}(s_k, s_j)}{1 + \mathbf{W}_{Lab}(s_i, s_k)}, \quad (3)$$

where $\mathbf{W}_{Lab}$ is the color similarity matrix, s_k is the adjacent super-pixel that has the highest Lab color affinity to s_i. And F_{DE} denotes the distribution difference via

$$F_{DE}(i, j) = ||\mathbf{W}_B(i, *) - \mathbf{W}_B(j, *)||_2^2. \quad (4)$$

Here $\mathbf{W}_B$ is a symmetric affinity matrix governed by bi-harmonic distance, $\mathbf{W}_B(i, *)$ is the i-th row of $\mathbf{W}_B$, and $|| * ||_2^2$ is the square of the Euclidean distance. In a nutshell, $\mathbf{W}_B(i, *) - \mathbf{W}_B(j, *)$ measures two anchor super-pixels' similarity by considering the differences of their resulted bi-harmonic distance distributions over the entire manifold.

It should be noted that, only F_{DE} can fully illustrate the proposed metric. The weighted average of Eq. 3 exactly follows the key idea "global" of the metric, and further improves the metric robustness. Here we adopt the traditional color affinity rather than the bi-harmonic distance governed affinity $\mathbf{W}_B$ to compute the weight, because bi-harmonic distribution may be unreliable in some fuzzy boundaries or narrow regions. Finally, we further convert the distribution difference matrix $\mathbf{d}_{DE}$ to the affinity matrix $\mathbf{W}_{DE}$ through a Gaussian function. And we perform different normalization for the convenience of parameter setting, which is one of the critical goals of this robust metric. In details, all the elements in $\mathbf{W}_B$ and $\mathbf{W}_{Lab}$ are normalized to be within 0 to 1 in a matrix-wise fashion, while those in $\mathbf{W}_{DE}$ are normalized to be within 0 to 1 in a row-wise fashion.

Fig. 4 illustrates why our idea is feasible over a complex cheetah image, wherein the texture of the cheetah makes the feature coherency very messy. As shown in Fig. 4(b), the bi-harmonic distance between the ear and cheek is not as near as that in Euclidean space, and both ear and bi-harmonic distance distribution differences by subtracting two super-pixels' distance distribution vectors ($\mathbf{W}_B(s_{ear}, *) - \mathbf{W}_B(s_{cheek}, *)$). In order to observe some implicit tendencies, we have sorted the super-pixels in an ascend order according to the bi-harmonic distance distribution of the "ear" super-pixel ($\mathbf{W}_B(s_{ear}, *)$, the first row of Fig. 4(c)). Although the beginning of red line in Fig. 4(c) shows weaker distribution differences around the "ear" region (from "ear" to "cheek"), the differences decrease gradually in the far regions. This explains why the super-pixels covering the head have high similarity (Fig. 4(d) shows that the head region's distributions corresponding to "ear" and "cheek" super-pixels are both red).

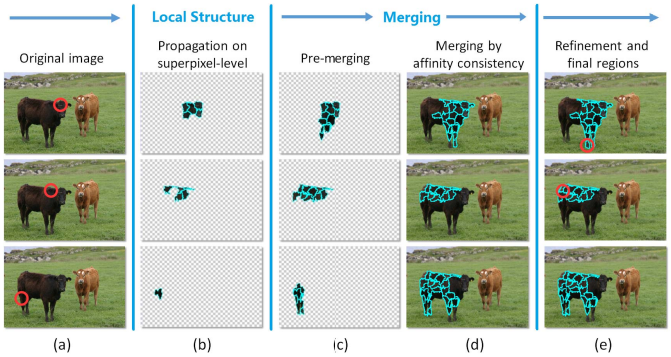


Fig. 5. The pipeline of meaningful region generation method. (a) Original images with anchored super-pixels marked by red circles; (b) Basic region generation; (c) Merge basic regions according to the overlapping ratio; (d) Merge the regions in (c) to form mid-level regions according to the proposed affinity consistency metric; (e) Refine the merging results to obtain the final regions, wherein the merged super-pixels during refinement are marked with red circles.

In contrast, the distribution difference of "ear" to "body" is large (the blue line in Fig. 4(c)), because they belong to different local structures. Thus, such metric can well reveal the meaningful structures of an image, and it is both globally stable and locally discriminative.

B. Meaningful Image Region Generation

We integrate the proposed affinity metric and a novel affinity consistency metric to facilitate the robust generation of meaningful image regions. These regions consist of some irregular overlapping super-pixels, which in some sense represent certain local structures. We prefer the obtained regions to be both structure-meaningful and relatively as local as possible. However, the commonly-used up-front region generation methods, such as super-pixel or graph based clustering, are not robust for complex image scenes. In contrast, we find that generating **local structures** (basic regions) and selectively **merging** them are superior to those up-front methods, because the accuracy and size of the regions can be conservatively and gradually controlled. As a result, if the generated **local structures** have been accurate enough, the final regions will have high possibility to be correct even if it may not be structurally integrated. And the proposed bi-harmonic distribution based metric could help us obtain accurate local structures.

For local structure generation, we initialize a basic region for each super-pixel via a diffusion process (Fig. 5(b)). These basic regions cannot only be easily controlled by the proposed structure-meaningful metric but also represent higher-level information compared to original super-pixels. Suppose the neighboring local structures will be overlapped or similar with each other, we merge them to form structure-meaningful regions according to their affinity consistency and the overlapping ratio (Fig. 5(c-d)). It should be noted that, the local structure may overlap with each other, and the final mid-level regions will keep this property. Besides, the dissection and discussion of this "soft" scheme can found in Section VIII-E. Fig. 5 illustrates the pipeline of our meaningful region generation method. Here it involves several bottom-up steps, and the details are described as follows.

Algorithm 1 Basic Region (Structure) Constructing

Input:
The super-pixels of an image $S = \{s_1, s_2, ..., s_n\}$;
The distance distribution difference based affinity matrix $\mathbf{W}_{DE}$;
The Lab color affinity matrix $\mathbf{W}_{Lab}$;
The super-pixel adjacent matrix $\mathbf{M}_{adj}$.

Output:
The basic regions $R = \{R_1, R_2 \dots R_n\}$.

1: **for** Each super-pixel $i \in [1, n]$ **do**
2: The super-pixel s_i is represented by the region $R_i = \{s_i\}$;
3: The candidate super-pixels set $R_{candi} = \{s_c \mid \mathbf{M}_{adj}(s_c, R_i) = 1\}$;
4: The score of the candidates $Score(s_c) = \mathbf{W}_{DE}(s_i, s_c),\ s_c \in R_{candi}$;
5: **while** $\exists Score(s_c) >= T_{near},\ s_c \in R_{candi}$ **do**
6: $s_{select} = \arg\max\ Score(R_{candi}),\ s_{select} \in R_{candi}$;
7: $R_i = R_i \cup s_{select}$;
8: Update the set of candidate super-pixels $R_{candi} = \{s_c \mid \mathbf{M}_{adj}(s_c, R_i) = 1\}$;
9: Update the score of the candidates $Score(s_c) = \mathbf{W}_{DE}(s_i, s_c) \cdot max(\mathbf{W}_{DE}(s_r, s_c))$ $s.t.\ \mathbf{M}_{adj}(s_r, s_c) = 1,\ s_c \in R_{candi},\ and\ s_r \in R_i$;
10: $s_{remove} = \arg\min\ \mathbf{W}_{Lab}(s_i, R_{candi}),\ s_{remove} \in R_{candi}$;
11: Remove s_{remove} from R_{candi};
12: **end while**
13: **end for**
14: **return** The basic regions $R = \{R_1, R_2 \dots R_n\}$.

1) Basic Region Initialization: Our strategy can be summarized as *"pick the best one while abandoning the worst one"*. This discreet strategy aims to avoid trapping in local wrong results for those complex image scenes, which also conforms with the motivation of our local-to-global, progressive framework, and more details can be found in Algorithm 1. Generally speaking, each basic region begins with an anchored super-pixel (Step 2) and a candidate super-pixel set (Step 3). In each loop, the current region merges one super-pixel (Step 7) and updates the candidate set by considering a larger spatial range (Step 8). Meanwhile, the candidate super-pixel with the weakest color affinity is removed from the candidate set (Step 11), which can avoid possible border-crossing risk due to the diffusion on some fuzzy boundaries. Moreover, the merging criterion is determined by the weighted average of $\mathbf{W}_{DE}$ ($Score$ in Step 9), which in fact decides the propagation direction.

The key steps of this algorithm have been marked in blue. During the iterative propagation, in order to merge a new super-pixel into the current incomplete region, the most significant problem is how to decide the weights. We consider this problem mainly from two aspects. First, the structure-aware propagation should not be far away from the anchor super-pixel (In $\mathbf{W}_{DE}(s_i, s_c)$, s_i is the anchor super-pixel). Second, as for a diffusion process, we need to find the best diffusion trajectory that has the highest structural affinity (In $max(\mathbf{W}_{DE}(s_r, s_c))$, s_r is actually the super-pixel on the boundary of R_i). Benefiting from the proposed compressive and convincing metric in Section V-A, we can set a fairly high value for T_{near}, whose threshold value is fixed to be 0.9 for all cases.

2) Affinity Consistency Definition: Although the neighboring local structures have high possibility to overlap with each other, **overlapping rate** is insufficient to guide the merging process for two reasons. First, the region is defined by super-pixels, thus it is hard to find an appropriate overlapping rate threshold when two regions have significant difference in size. Second, the image structures may gradually change. It means that, the overlapping rate may mistakenly merge those different but fuzzy connected structures. Therefore, we propose a region-wise structure-meaningful metric to evaluate the merging result based on $\mathbf{W}_{DE}$.

In detail, if region R_A can be merged with region R_B, they must share the same structure, and the change of the interior affinity distribution should not be drastic after merging. Therefore, we define an affinity consistency error metric $E_{merge}(R_A, R_B)$ as

$$|F_{std}(\mathbf{W}_{DE}(R_A \cup R_B)) - F_{std}(\mathbf{W}_{DE}(R_B))|. \tag{5}$$

Here $\mathbf{W}_{DE}$ is an affinity matrix of super-pixels, and $*$ denotes the super-pixels covered by the region. Thus, $\mathbf{W}_{DE}(*)$ means a sub-matrix describing region's inner affinity distribution. And F_{std} measures the standard deviation or average value of the matrix elements.

3) Affinity Consistency Guided Merging: We have explained why **affinity consistency** metric E_{merge} and **overlapping rate** can be used as merging criteria. During merging, we still follow the strategy of (*"pick the best one while abandoning the worst one"*) to merge only one pair of regions in each loop. And the method is detailed in Algorithm 2. To begin with, we find the merging candidates with similar local structures (Step 1) according to overlapping rate. Then, the best candidate, which is really needed to be merged rather than just overlapped, can be found easily. Meanwhile, since we expect to merge small-size fragments into integrated large-size region, we sort the existing regions $R_1, R_2 \dots R_n$ in an ascend order according to their sizes (Step 7). In each loop, the structure-meaningful region is initialized by an anchored region R_{end} at the tail of the region queue, which is the biggest one. Then we orderly search the appropriate regions along the region queue, and merge them in R_{end} (Step 10). It means that R_{end} keeps changing during the search process, which prevents it from becoming too large to conform with its mid-level size.

The key steps of Algorithm 2 are marked in blue. Specifically, if two regions have the same structures ($\mathbf{W}_{or}(R_k, R_{end}) >= T_{stru}$) and low-affinity consistency error ($E_{merge}(R_k, R_{end}) < T_{err}$), we merge the fragment R_k into the largest anchor region R_{end}. Since the pre-merged regions produced in Step 1 themselves are already very similar if they belong to the same local structure, the reasonable changes of the involved two fixed parameters can only effect the region size. Specifically, T_{stru} is set to be 0.5 in this paper, which means that two regions may be in the same local structure. Besides, T_{err} is designed to measure the affinity consistency changes after the merging procedure. We fix it to be 0.2 in this paper.

4) Affinity Consistency Based Refinement: We have obtained many meaningful regions at this point but some fragmental regions still exist. It is mainly because of the complexity of natural images, such as abrupt change of illumination, tunes, fine details, and so on. The reason why we do not solve it in Algorithm 2 is that we cannot judge the best

Algorithm 2 Basic Region Merging

Input:
The basic regions of an image $R = \{R_1, R_2 \dots R_n\}$;
The overlapping rate matrix of basic regions $\mathbf{W}_{or}$;
The threshold value T_{stru}, T_{err}.

Output:
The structure-meaningful region set $P = \{p_1, p_2 \dots p_m\}$.

1: **for** Each $i \in [1, n]$ and each $j \in [1, n]$ **do**
2: **if** $\mathbf{W}_{or}(R_i, R_j) >= T_{stru}$ **then**
3: $R_i = R_i \cup R_j$;
4: **end if**
5: **end for**
6: Initialize region set $P = \phi$;
7: Sort R by regions' size (R_{end} is the biggest);
8: **while** $R \neq \phi$ **do**
9: **for** Each $k \in [1, size(R) - 1]$ **do**
10: Merge the fragmented regions
$R_{end} = R_{end} \cup R_k$
$s.t.1\ E_{merge}(R_k, R_{end}) < T_{err}$
and $\mathbf{W}_{or}(R_k, R_{end}) > T_{stru}$;
11: Remove R_k from R;
12: **end for**
13: Update region set $P = P \cup R_{end}$;
14: Remove R_{end} from R;
15: **end while**
16: **return** The structure-meaningful region set $P = \{p_1, p_2 \dots p_m\}$.

Algorithm 3 Merging Refinement

Input:
The regions of an image $P = \{p_1, p_2 \dots p_m\}$.

Output:
The final mid-level regions $P = \{p_1, p_2 \dots p_k\}$.

1: Find the fragment set $F = \{p_i \mid p_i \in P \ and \ size(p_i) <= 2\}$;
2: **while** $F \neq \phi$ **do**
3: $[p_i, p_j] = \arg\min\ E_{merge}(p_i, p_j)$,
$s.t.\ p_i \in F,\ p_j \in P \ and \ p_i \neq p_j$;
4: $p_j = p_j \cup p_i$;
5: Remove p_i from F;
6: **end while**
7: **return** The final mid-level regions $P = \{p_1, p_2 \dots p_k\}$.

ownership of these "noise regions" if we have not obtained all the meaningful regions. Therefore, we need an additional refinement process to merge them into some already-obtained meaningful regions (see Algorithm 3). The merging process is still decided by the proposed affinity consistency error metric E_{merge}. It should be noted that, the algorithm searches the best merging pairs (Step 3) globally here until there are no fragmental regions (Step 2).

Finally, as shown in Fig. 5(e), the black ox comprises three overlapped meaningful regions. The merged fragmental patches via refinement procedure are marked with red circles. And the quantitative analysis for the superiority of the proposed metric and meaningful image regions is detailed in Section VIII-E.

VI. L_1-Manifold Hyper-Graph Construction

For the obtained structure-meaningful mid-level region, we should further extract proper region-level features to represent it in alternative feature space. In this paper, we use simple average Lab/RGB color features, uniform local binary pattern histogram (uniform LBP), and bag-of-words descriptors. In practical implementation, we parallelize LBP algorithm on CUDA, wherein each CUDA thread is responsible for single sample point. According to our experiments, we can extract one million of LBP features and organize them properly within 3 seconds.

A. Multi-Feature Involved Inter-Image L_1-Manifold Graph

Following the key idea of [10] and [11], we employ graph-based methods to exploit the implicit coherency across images. From the perspective of manifold learning, the common idea is that, one point on a non-linear high-dimensional manifold can be approximated as a sparse linear combination of other data points.

Assume N denotes the total number of the generated mid-level regions, given H types of feature $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_H\}$, $\mathbf{X}_h$ is a $D \times N$ matrix whose column denotes a D-dimension feature vector corresponding to certain region in the $h-th$ feature space, and $\mathbf{W}_h$ denotes an $N \times N$ coefficient matrix representing the region affinities in the h-th feature space. And let $\mathbf{W} = \{\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_H\}$ be the coefficient matrix set of all types of features, we can reformulate the objective function as

$$\begin{aligned}\bar{\mathbf{W}} = \arg\min_{\mathbf{W}} & \sum_{h=1}^{H}(||\mathbf{X}_h\mathbf{W}_h - \mathbf{X}_h||_F^2 + \lambda||\mathbf{W}_h^T\mathbf{W}_h||_{1,1}) \\ & + \alpha||\mathbf{Z}||_{2,1} + \beta||\mathbf{Z}||_{1,1} \\ & s.t.\ diag(\mathbf{W}_h) = 0 \quad and\ \mathbf{W}_h > 0. \end{aligned} \tag{6}$$

Here the minimization of $||\mathbf{X}_h\mathbf{W}_h - \mathbf{X}_h||_F^2$ denotes that, the h-th feature of a region should be approximated by others with the least possible error. The minimization of $||\mathbf{W}_h^T\mathbf{W}_h||_{1,1}$ requires $\mathbf{W}_h$ to be as sparse as possible. And λ is used to control the weight of the sparsity of the reconstruction coefficients. In addition, though $||\mathbf{W}_h||_{1,1}$ is usually employed as a penalty term to get a sparse coefficient matrix $\mathbf{W}_h$, its optimization is expensive when N is large. For this reason, we change it to $||\mathbf{W}_h^T\mathbf{W}_h||_{1,1}$ (note that, it was initially proposed by Wang *et al.* [35]). Benefiting from this, the involved subspace clustering is much faster than traditional sparse subspace clustering (SSC) methods [34]. Moreover, $\mathbf{Z} = [\mathbf{W}_1', \dots, \mathbf{W}_H']^T$ is used to seek the cross-feature consistency, wherein $\mathbf{W}_i'$ denotes the vectorized $\mathbf{W}_i$ (transforming $N \times N$ matrix $\mathbf{W}_i$ to a $N^2 \times 1$ vector $\mathbf{W}_i'$). In Eq. 6, the minimization of $||\mathbf{Z}||_{2,1} = \sum_{j=1}^{N^2} ||\mathbf{Z}(*, j)||_2$ is the key, which intrinsically integrates the multi-type features. Specifically, each column $\mathbf{Z}(*, j)$ denotes the affinity of a region pair in different feature space. This penalty term enforces some columns of $\mathbf{Z}$ to be zero to pursue *column-wise* sparsity, so that the regions should be far away from each other in each feature space. However, single $L_{2,1}$-norm cannot exploit more mutual information among different feature spaces, because H-feature elements of each column are considered as a whole due to L_2-norm. In contrast, the penalty term $||\mathbf{Z}||_{1,1}$ is the sum of all elements in matrix $\mathbf{Z}$, which controls the overall sparsity of $\mathbf{Z}$. Accordingly, combining $||\mathbf{Z}||_{1,1}$ with $||\mathbf{Z}||_{2,1}$ is designed for *row-wise* sparsity, which is expected to control the number of features involved in each column, and thus gives rise to certain

type of "feature selection". Besides, the L_1-norm governed two penalty terms enforce the obtained affinities to be both sparse and consistent in multi-feature space. Here we employ the spectral project gradient method [44] to solve Eq. 6, and get the optimal region reconstruction coefficient set for each feature space $\bar{\mathbf{W}} = \{\bar{\mathbf{W}}_1, \bar{\mathbf{W}}_2, \cdots, \bar{\mathbf{W}}_H\}$. Finally, we sum the affinities generated from multi-type features to define the edge weights of the L_1 graph as

$$a_{ij} = \frac{1}{2}\left(\sqrt{\sum_{h=1}^{H}(\bar{\mathbf{W}}_h)_{ij}^2} + \sqrt{\sum_{h=1}^{H}(\bar{\mathbf{W}}_h)_{ji}^2}\right). \tag{7}$$

B. Local Structure Coherency Based Intra-Image Graph

To further strengthen the intra-image region affinities to be the L_1 graph, we define the weights of the intra-image sub-graphs according to their structural differences under our bi-harmonic distance distribution based metric. The affinity of two intra-image regions p_k and p_l is defined as follows:

$$a_{kl} = \begin{cases} \sum_{s_i \in p_k} \sum_{s_j \in p_l} \mathbf{W}_{DE}(s_i, s_j) & \text{if } p_k \text{ and } p_l \text{ are adjacent} \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where s_i means the i-th super-pixel in the region p_k, and s_j is the j-th super-pixel of region p_l. Since $\mathbf{W}_{DE}$ is decided by the local structure and can be explained as a type of affinity diffusion, it confirms well with our understandings about intra-image graph. Thus, we replace the intra-image parts (diagonal blocks) of Eq. 7 with the newly-defined Eq. 8 to form the final hyper-graph. And the entire elements in both inter-image graph matrix and each intra-image sub-graph are normalized to [0, 1].

VII. Hyper-Graph Joint Cut Based on Unsupervised GrabCut

We first perform Normalized cuts over the meaningful region hyper-graph to obtain a set of sub-graphs. Each of them is roughly a candidate of co-segmentation results comprising of some regions. After that, we manually assign which sub-graph result represents the foreground in single-foreground co-segmentation cases and this step is not necessary for most multi-class cases. Benefiting from our meaningful regions and robust graph design, these candidates can segment out the foreground perfectly in some simple cases and give rise to roughly correct results in most commonly-encountered cases. However, the foreground generated from meaningful regions still may be coarse and incomplete. Therefore, we will further conduct a pixel-level refinement via hyper-graph joint-cut aided by candidate results. We design pixel-level hyper-graph joint-cut model by extending the popular GrabCut algorithm [42] in the cross-image sense. The state-of-the-art GrabCut requires supervised stroke guidance and ignores inter-image information which is apparently not suitable for the multi-image co-segmentation problem. Therefore, we improve it to handle co-segmentation refinement mainly in two aspects. First, our candidate results can serve as a reasonably good initialization to the GrabCut algorithm compared with the bounding box based supervised approach. Second, our method can put candidate foregrounds from all images together to define a joint object model during iterations. It can be seen as a "the minority is subordinate to the majority" way.

Algorithm 4 Hyper-Graph Joint-Cut

Input:
The candidate foreground segments of each image $C = \{c_1, c_2 \ldots c_n\}$;
The background segments of each image $B = \{b_1, b_2 \ldots b_n\}$;
The maximum iterations N_{iter}.

Output:
The final refined foreground segments $\bar{C} = \{\bar{c_1}, \bar{c_2} \ldots \bar{c_n}\}$.

Initialize mask set of candidates $M = \{m_1, m_2 \ldots m_n\}$;
for Each $i \in [1, n]$ **do**
  Initialize mask m_i to be zero matrix;
  Label c_i covered region to be "possible foreground" ; $m_i(c_i) = 1$;
  Label c_i two super-pixel-spacing neighborhood to be "possible foreground" ; $m_i(neigh(c_i)) = 1$;
end for

Initialize refined foreground set $\bar{C} = C$;
for $t \in [1, N_{iter}]$ **do**
  Update foreground GMM by all current candidates $GMM_f = GMM(\bar{C})$;
  for Each $i \in [1, n]$ **do**
    Update image background GMM $GMM_b = GMM(b_i)$;
    Run original GrabCut $[\bar{c_i},\ b_i] = GrabCut(\bar{c_i}, GMM_f, GMM_b)$;
  end for
end for

return The final foreground segments $\bar{C} = \{\bar{c_1}, \bar{c_2} \ldots \bar{c_n}\}$.

Generally speaking, the macro-view on the pipeline of our hyper-graph joint-cut algorithm is very similar to that of the original GrabCut algorithm. But there are also some different nontrivial details, which can be found in Algorithm 4 (marked in blue). First, we initialize Gaussian mixed model (GMM) using our candidate co-segmentation result instead of supervised stroke and bounding box. Second, we use the pixels in the cross-image candidate regions to build a foreground GMM. As for the background GMMs, we compute them in an image-wise fashion because of the flexibility of the image set. Our pixel-level hyper-graph only covers the nearest neighborhood of the candidate regions that locate in one or two super-pixel spacing distance, which not only avoids error accumulation but also limits the possible labels of the foregrounds to be assigned in the image set. And our methods can automatically accommodate multi-class co-segmentation in a completely unsupervised way, because we can run the proposed joint-cut algorithm for each candidate region belonging to different classes.

Fig. 6(a) demonstrates a typical instance, wherein the geese with white body and yellow mouth swim on the dark water. Fig. 6(b) shows the inaccurate co-segmentation results in the regions of stones, because the colors of white geese are not distinctive enough under different view point and illumination conditions. And some state-of-the-art methods also produce the similar inaccurate results. Fig. 6(d) shows the single-image Grabcut iterative results. The algorithm takes both stone and geese as the foreground due to the fact of loosing extra supervised information during iterations. However, our hyper-graph joint-cut algorithm can adopt the inter-image affinity to solve this problem. As shown in Fig. 6(c), joint foreground model can wipe off the false-alarmed distinctive

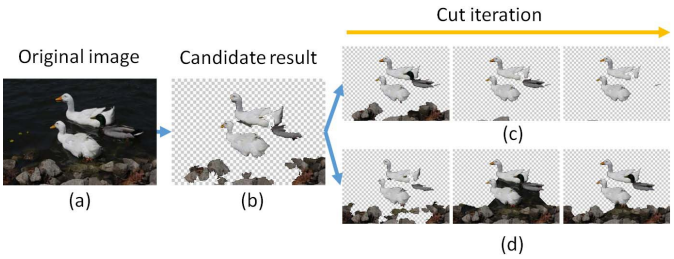


Fig. 6. Hyper-graph Joint-cut. (a) One of original images; (b) Our co-segmentation candidates with wrong foreground regions; (c) The gradual changes of foreground segments during joint-cut iterations; (d) Foreground segment changes during original GrabCut iterations.

foreground parts. Specifically, as shown in the first column of Fig. 6(c), even though the algorithm may integrate the wrong foreground regions during the first few iterations, it converges to the correct results quickly. In all of our experiments, we empirically set the iteration number to be 10. And quantitative evaluation between our joint-cut and GrabCut is detailed in Section VIII-E.

VIII. Experimental Results and Evaluations

We have implemented our method on a PC with Geforce GTX 660 GPU, Intel Core I7 CPU and 24G memory using C++ and MATLAB R2013a. We demonstrate the accuracy and efficiency advantages of our method via extensive experiments on the popular iCoseg, MSRC, and Oxford flower datasets. As for evaluation, we name our method as "**MCJLH** " (Multi-class Co-segmentation via Joint-Cut over L_1-Manifold Hyper-Graph) and compare our method with the state-of-the-art unsupervised co-segmentation methods, including Kim11 [9], Joulin12 [6], Faktor13 [18], Rubin13 [41] and Li14 [5]. And the well-known Jaccard similarity (J) is employed to conduct quantitative accuracy analysis ($J = \frac{GT_i \bigcap R_i}{GT_i \bigcup R_i}$, where i is the foreground index and GT means the ground truth).

A. Parameter Settings of Different Methods

1) Our Method (MCJLH): The number of super-pixels is empirically set to be 200 in all the experiments, which can well facilitate our physics-based affinity measurement and structure-meaningful region generation. Three threshold values T_{near}, T_{stru}, and T_{err} mentioned in region generation Algorithm 1, and Algorithm 2 are assigned the same values for all the experiments as explained in Section V. Meanwhile, for meaningful region's representation, we select Lab / SIFT / CSIFT combination for iCoseg dataset, SIFT / CSIFT for MSRC dataset, and Lab / CSIFT for Oxford flower dataset. We denote this feature combination of our method as "**MCJLH (Bow)**". Furthermore, LBP, as a weaker but efficient feature, is combined with Lab for all datasets to verify the robustness of our method "**MCJLH (LBP)**". The number of Lab color histogram bins of each channel is set to be 22. And the uniformly-sampled bag-of-words SIFT / CSIFT descriptors are extracted with the toolbox [45], and then are constructed via LLC coding [46] with 100 codebook words. As for LBP, we use 2×2 window based 16-ring uniform sampling to build 243 $16-bit$ pattern. Then, we employ these

TABLE I

The Accuracy Comparison for Two-Class Co-Segmentation. "-": The Method's Source Code / Website Cannot Provide Corresponding Data for This Experimental Setting

Method	iCoseg	MSRC (full)	MSRC (subset)	Oxford flower
Kim11	0.417	0.390	0.489	0.600
Joulin12	0.589	0.553	0.615	0.529
Rubin13	0.685	**0.884**	-	-
Faktor13	0.736	0.681	0.646	0.716
Li14	0.658	0.615	-	-
MCJLH (LBP)	0.734	0.689	0.756	0.789
MCJLH (BoW)	**0.785**	0.706	**0.768**	**0.818**

histogram-formed feature to weigh the edges of L_1-manifold graph in Eq. 6, wherein the involved parameters are set as $\lambda = 1000$, $\alpha = 0.01$, $\beta = 0.1$. Besides, we make the cluster number range from K to $2K$ (K is the number of objects to be segmented), and the same setting is also applied to Kim11 and Joulin12. For the joint-cut step, we set the maximum number of iterations to be 10. And the refining domain is set as the regions covered by two super-pixel spacing neighborhoods of the candidate segments.

2) Kim11: Since its central idea is to orderly select the color-coherency area as large as possible based on the greedy algorithm, which is very color-sensitive. To compensate it, we allow tuning its Gaussian parameter within the range of $0.25 \leqslant \beta \leqslant 0.6$ (default) when Kim11 produces unsatisfactory results.

3) Joulin12: Joulin12 involves two types of features, which are SIFT and color histogram, we finally select the better results obtained from the two-type feature selection for comparison. Moreover, we respect its default option that initializes Joulin12 with Joulin10 [7] for the best results. Besides, we allow slightly tuning the parameter "Laplacian weight" to be 0.1 (default), 1, and 10 when their results are less promising.

4) Faktor13: The only parameter in their source codes is the feature type, which can be chosen from the color histogram and color HOG descriptors. We run the experiment based on these selections and choose the results with the best accuracy for comparison.

5) Rubin13 and Li14: We directly take their published results for comparison. But Li14 focuses on evaluation and repairing of the state-of-the-art methods' co-segmentation results. It means that Li14 improves accuracy to certain extents when using the co-segmentation results from different methods as the initialization. Therefore, we just use Li14's best results for comparison.

B. Accuracy Evaluation for Two-Class Co-Segmentation

According to the quantitative accuracy evaluations listed in Table I, our method outperforms other methods in most cases. In a nutshell, Faktor13, Rubin13, and our method are better than Kim11, Joulin12 and Li14. Besides, since the iCoseg, MSRC, and Oxford flower datasets have their own characteristics, these compared methods have different advantages in different datasets respectively, which are detailed as follows.

1) iCoseg: iCoseg is one of the best datasets to verify the initial motivation of co-segmentation problem. It is challenging for all the methods, because the objects' co-occurring in

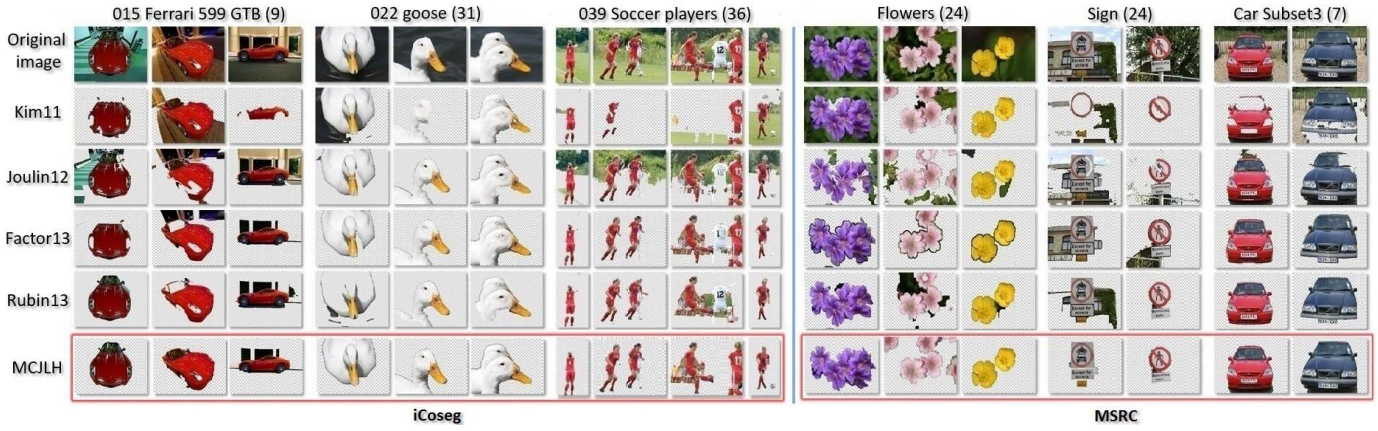


Fig. 7. Two-class co-segmentation comparison over iCoseg and MSRC datasets.

the images shows variation in perspective, tone, illumination, complex around environments, etc. From Table I, we can see that **MCJLH (Bow)** outperforms the competitors. And even when **MCJLH (LBP)** is the third best, it almost approaches the second best one (Faktor13). As shown in Fig. 7, our proposed meaningful region can better guarantee the integrity of the foreground, such as the cases of "Ferrari 599 GTB", and "goose". Kim11 produces the most amount of false results because of two reasons. First, it measures similarities by direct Euclidean distance in single color feature space. Second, the greedy algorithm based orderly selection of the color-coherency areas is too specific to handle most of situations without obvious similar color-coherency areas among different images. Thus, the method lacks an intrinsic meaningful feature representation and measurement to express various complex cases. It illustrates a sharp contrast to our bi-harmonic distance distribution based metric and multi-feature based L1-manifold graph. Joulin12 outperforms others in very few cases, wherein the foregrounds always comprise fragmented parts that have no clear affiliations. We think that the segment-size related constraints in Joulin12 play key roles to enforce the fragments together. Rubin13 is good but not the best, which depends on matching and saliency detection in their framework. On iCoseg dataset, it cannot detect clear saliency distribution in many images, while we can clearly notice messy saliency distribution in the whole image sets. Faktor13, which also employs a variant of joint models based on GrabCut (similar to our method), is the second-best method because each image set in the iCoseg dataset contains common foreground object. However, since our hyper-graph joint-cut process is initialized with candidate regions, which have been as good as possible, we can limit the potential area of the possible foreground in a small local region. Thus, it makes our joint-cut method fast and eliminate redundant parts easily.

2) MSRC: The foregrounds in the image groups of MSRC dataset may be totally different in colors and textures, which should heavily affect the accuracy of co-segmentation. For the co-segmentation framework with limited power, it will generate an undesirable inter-image graph in some cases. In contrast, our structure-meaningful region gives rise to less-accurate affinities among cross-image regions. Compared with Kim11, Joulin12, Faktor13 and Li14, our method shows superiority on most of subsets of MSRC dataset (see Fig. 7). However, Rubin13 outperforms all other methods over the full image set (not a regrouped subset) in Table I, because it relies on strong single image classifier and its inter-image terms affect the results less. This phenomenon can also be observed from their own experimental analysis [41]. Another reason for Rubin13's good performance is that the foreground in MSRC dataset is always the main body of an image, which shows good saliency property, unlike those in iCoseg dataset. Although Li14 can propagate good segmentations by consistency evaluation and completeness evaluation, Li14's results are not outstanding in iCoseg and MSRC datasets. The reason is that their framework depends on uncontrollable pre-segmentation and many criteria of quality evaluation. In some sense, Li14 provides a practical way to select good objective-like segments. However, since its evaluation energy function has nine terms, it is complicated and unstable for various segmentation initializations. And such instability may effect its following propagation step.

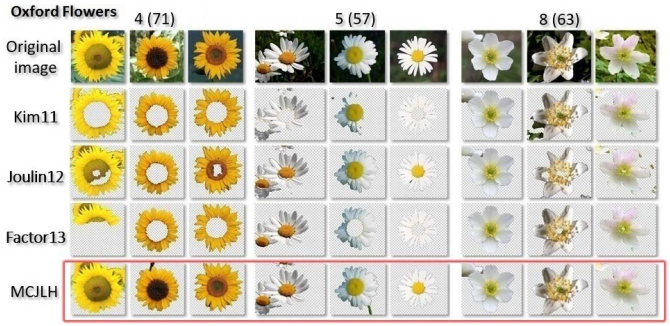


Fig. 8. Two-class co-segmentation comparison over Oxford flowers dataset.

3) Oxford Flower: All of the image sets in Oxford flower dataset are obviously larger than others. They are readily available to test the robustness and efficiency of co-segmentation methods because more images bring more cross-image information with large variation, it becomes extremely challenging to build robust affinities. Fig. 8 shows some results. Even if there is a distinguishable color similarity, different images in an image set also vary drastically in tone, illumination, and details. Although this difficulty will be amplified with the increase of image number, our method outperforms other

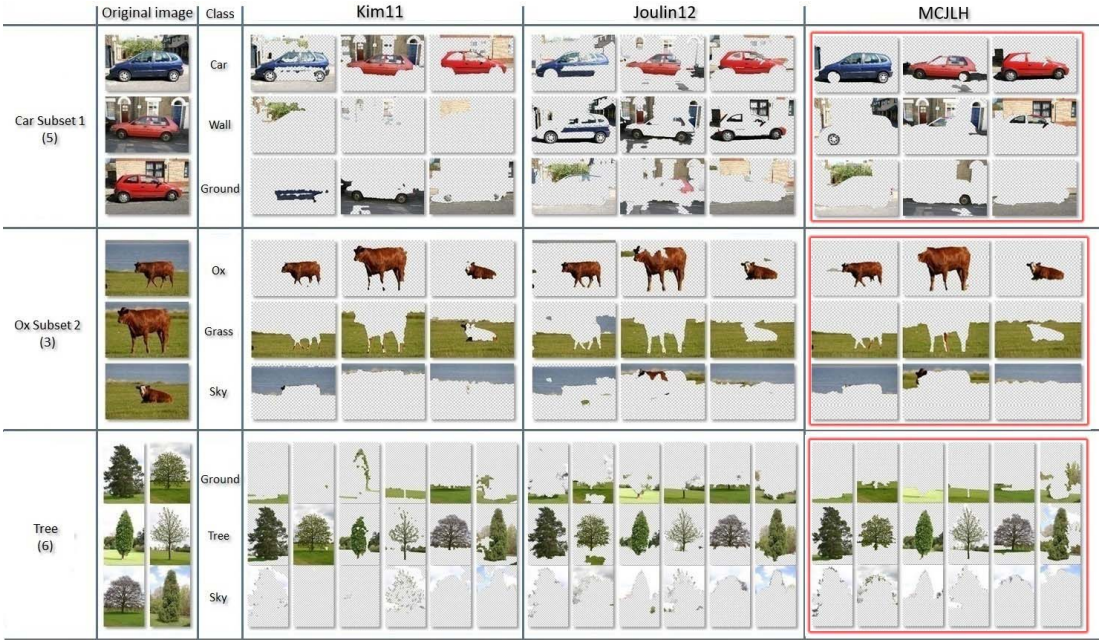


Fig. 9. Multi-class co-segmentation comparison over MSRC dataset.

methods in accuracy (see in Table I). In this dataset, a flower always comprises only $2-6$ meaningful regions, and about 20 regions for each image. Combining the distinguishable color coherency and L_1-manifold based affinity measurement over a small number of meaningful regions, as shown in Fig. 8, it is not hard for our method to separate the regions into different feature space, and further pick good candidate foreground regions.

C. Accuracy Evaluation for Multi-Class Co-Segmentation

The multi-class co-segmentation results are shown in Table II and Fig. 9. For each image set, we compute the Jaccard similarity (J) for each object class. The most difficult challenge in multi-class cases is to distinguish different classes in an enormous feature space. Directly benefiting from our meaningful image region generation, our method outperforms Kim11 and Joulin12. The final step of all three methods are based on certain type of co-clustering, wherein Kim11 uses a cross-image greedy algorithm, Joulin12 optimizes an energy function, and we make use of the hyper-graph joint-cut framework. In contrast, our proposed meaningful image region can effectively express the underlying structural object and facilitate the grouping of relevant features. However, Kim11's super-pixel-wise clustering and Joulin12's pixel-wise clustering (even if they also take the advantage of super-pixel in a rather different way) tend to easily mix up the foreground parts. Besides, our multi-feature based L_1-manifold graph is another reason to guarantee better results. As shown in Fig. 9, Kim11 obtains good results on "ox subset 2" image set mainly because of the high color coherency. Except for color information, Joulin12 uses SIFT descriptor, and our method adopts LBP descriptor. However, we can obtain fairly good affinities when there is no color similarity between foregrounds. For Joulin12, it performs a bit worse than ours on the image sets of "car subset 1", "ox subset 2", and "tree", because solely using the SIFT descriptor cannot guarantee meaningful intra-image measurement.

TABLE II

THE ACCURACY COMPARISON FOR MULTI-CLASS CO-SEGMENTATION OVER THE REGROUPED SUB-SET OF MSRC DATASET

Image Set	class	Kim11	Joulin12	MCJLH (LBP)
car subset 1	**average**	0.392	0.453	**0.498**
	car	0.457	0.461	**0.705**
	grass	**0.349**	0.246	0.345
	ground	0.370	**0.652**	0.443
car subset 2	**average**	0.335	0.384	**0.656**
	car	0.302	0.423	**0.595**
	wall	0.287	0.287	**0.639**
	ground	0.415	0.442	**0.733**
ox subset 1	**average**	0.309	0.396	**0.710**
	ox	0.196	0.451	**0.693**
	grass	0.577	0.511	**0.845**
	sky	0.154	0.227	**0.592**
ox subset 2	**average**	0.651	0.487	**0.721**
	ox	0.487	0.463	**0.733**
	grass	**0.841**	0.587	0.802
	sky	0.626	0.411	**0.629**
plane	**average**	0.530	**0.714**	0.679
	plane	0.335	0.461	**0.557**
	sky	0.714	**0.804**	0.768
	grass	0.542	**0.877**	0.712
tree	**average**	0.492	0.448	**0.726**
	ground	0.466	0.401	**0.755**
	tree	0.668	0.478	**0.720**
	sky	0.341	0.466	**0.704**
building	**average**	0.264	0.415	**0.513**
	building	0.396	0.530	**0.576**
	grass	0.145	0.263	**0.332**
	sky	0.251	0.451	**0.630**

D. Efficiency Evaluation

Table I and Table II have proved that our method is still capable of obtaining competitive accuracy under weaker LBP features. Taking this fact as the prerequisite, Table III shows the average time cost according to the statistical results.

TABLE III

THE AVERAGE RUNTIME (S) COMPARISON PER IMAGE FOR TWO-CLASS CO-SEGMENTATION. N: THE NUMBER OF IMAGES

Dataset	N	Kim11	Joulin12	Faktor13	MCJLH (LBP)
iCoseg	642	1.42	335.84	338.56	**1.12**
MSRC	218	1.59	338.78	267.74	**1.17**
Oxford flower	527	2.51	641.11	227.79	**1.31**

TABLE IV

THE AVERAGE PERCENTAGE (%) OF EACH STEP'S TIME COST IN OUR METHOD

Pipeline		Percentage (%)
Intra-image processing steps (45.62)	Bi-harmonic distance	22.17
	Distribution difference	7.62
	Mid-level region generation	0.05
	Region feature	15.78
Inter-image processing steps (54.38)	L_1 graph	6.41
	Clustering	0.20
	Joint-cut	47.77

It shows that our method outperforms others in efficiency. The most important reason is that, an image is represented by fewer primitives (i.e., meaningful regions). Kim11 is the second most efficient one but has lower accuracy. Generally speaking, Faktor13 is better than Joulin12 in efficiency, but both of them cannot handle larger image sets rapidly. In contrast, our method can segment each image in about one second on average.

Moreover, Table IV shows the average percentage of each step's time cost of our method, which takes into account all the designed experiments. According to our experiments, the time cost of our method has approximately linear relationship with the image set scale. And the statistical results show that the intra-image processing (45.62%) and inter-image processing (54.38%) have a similar influence on efficiency, while the final hyper-graph joint-cut step (47.77%) is the most time consuming step in our framework. Therefore, when less accurate results enabled by the candidate segments are enough in some applications, we could probably bypass the joint-cut step completely to gain better efficiency. Yet, the time expense of the joint-cut step is relatively mild in spite of different number of inter-image processing tasks, because the time complexity of updating foreground model is $O(N)$ for each iteration. As for the time cost of intra-image processing, "Region feature" could have been a time-consuming task, but our CUDA-based parallel implementation makes it much more efficient, especially for LBP descriptor. On the other hand, k-means and coding steps in "**MCJLH (BoW)**" are relatively slow, but it is almost irrelevant to other steps.

E. Component Dissection and Limitation Discussion of Our Framework

As shown in Table V, our accuracy improvements mainly benefit from three key technical elements, including bi-harmonic distance distribution based metric, structure-meaningful image region generation, and hyper-graph joint-cut, which is quantitatively analyzed by intentionally disabling different steps of our method one-by-one. And each of such

TABLE V

THE ACCURACY COMPARISON FOR DIFFERENT SETTINGS OF OUR FRAMEWORK. EACH ROW REPRESENTS AN ALTERNATIVE SOLUTION DETAILED IN SECTION VIII-E

Method	iCoseg	MSRC	Oxford flower
W_{Lab}+Clustering	0.682	0.680	0.812
W_{DE}+Clustering	0.726	0.710	0.801
Big super-pixel	0.731	0.701	0.805
GrabCut	0.746	0.718	0.799
MCJLH	**0.785**	**0.744**	**0.818**

well-designed competitor only has one different step compared with our full-version framework "**MCJLH**". For the competitors, $\mathbf{W_{Lab}}$**+Clustering** represents: super-pixel level clustering results based on Lab color affinities instead of the structure-meaningful regions; $\mathbf{W_{DE}}$**+Clustering** represents: super-pixel level clustering results based on the proposed metric instead of the structure-meaningful regions; **Large super-pixel** represents: super-pixels with large size instead of the structure-meaningful regions (the number of super-pixels is set to be same as the number of the structure-meaningful regions); **GrabCut** represents: original GrabCut algorithm instead of our joint-cut method but with the same initialization as "**MCJLH**".

According to the experiment results, if the step of our structure-meaningful region generation is disabled, the three competitors ($\mathbf{W_{Lab}}$**+Clustering**, $\mathbf{W_{DE}}$**+Clustering**, and **Big super-pixel**) all tend to ignore the local constraints, and thus cannot achieve satisfactory results for many details due to the lack of a powerful metric. In sharp contrast to these three competitors, the "propagation" property involved in our structure-meaningful region generation plays a very important role in guaranteeing the obtained regions to be both structure-meaningful and relatively as local as possible. Meanwhile, benefitting from a series of bottom-up steps, the accuracy and size of the regions can be conservatively and gradually controlled. Moreover, the proposed robust metric makes the corresponding parameters' setting more easy, which can absolutely be fixed for all cases. Although the final accuracy gaps between the competitors and "**MCJLH**" may be reduced by the same joint-cut step in some sense, "**MCJLH**" still outperforms them at about 5.4% to 10.3% precision, because the joint-cut step also depends on the quality of the candidate segments. As for the joint-cut step, according to our experiments, the wrongly-assigned image regions to certain candidates have more negative effect than those missed regions, which can be filled up by a proper cross-image joint model. Thus, to reduce such negative influence as much as possible, we resort to some local constraints to avoid the over-boundary problems during propagation. Besides, we should notice the fact that, $\mathbf{W_{DE}}$**+Clustering** is obvious better than $\mathbf{W_{Lab}}$**+Clustering** except for Oxford flower dataset (high color affinity). We believe with confidence that the gain is from our structure-meaningful metric $\mathbf{W_{DE}}$, because there are no other variables being involved, just as our aforementioned analysis for Fig. 3 in Section IV and Fig. 4 in Section V-A. Besides, our hyper-graph joint-cut can accommodate co-segmentation problem better than the original **GrabCut** method because of the full utility of implied

cross-image supervised information, which has also been proved in Section VII (Fig. 6).

Overlapping and **structure-meaningful** are two salient properties of our mid-level regions, which are worth to be further discussed. We elaborately make the regions overlap with each other via Step 1 in Algorithm 2. Since the overlapped parts of the regions involve multi-type features, it will bring extra connection to enhance the intra-image affinities. For inter-image case, this scheme also contributes to sparse-coding and manifold learning based processing, because the underlying key idea is "local linear representation" in certain feature space while the overlapped parts definitely have a common feature representation. As shown in Fig. 11(a), the obtained regions always overlap with each other and show good structure-propagation. Although the competitor $\mathbf{W_{Lab}}$+**Clustering** also performs well for this simple case, its resulted regions are too local to be structure-meaningful (Fig. 11(c)). Fig. 12 clearly illustrates the **structure-meaningful** property when one object only comprises fewer simple parts. Benefiting from the propagated merging based on our concise metric, here our method can better produce the four relatively integrated local structures (Fig. 12(a)). Specifically, the tiny black words on the board are uniformly assigned to a meaningful region. Moreover, with the fixed parameter settings (Section VIII-A), our method can commonly segment an image to about 20 regions, which gives rise to drastic and reliable data reduction, so it is not surprising that we can achieve outstanding efficiency. Taking the Oxford flower dataset as an example, each of its image sets is relatively larger than those of iCoseg and MSRC datasets, and on average it has about $70-100$ images. According to Table III, our method can achieve the co-segmentation of such image sets in few minutes while having higher accuracy like Faktor13 and Joulin12, however, Faktor13 and Joulin12 would need spend about 4 to 17 hours to finish the same tasks. Therefore, in unsupervised co-segmentation field, we are confident that our framework could have greater potential to handle large-scale image sets.

The L_1-manifold hyper-graph makes our method more flexible to deal with various applications by way of appropriate feature combinations. Although multi-type features can be adaptively encoded in the graph, how to select features with great care is still worth of discussion. First, single-type feature is insufficient for complex real-world cases. Fig. 10 shows a case where the co-segmentation results are respectively computed based on single Lab color feature (Fig. 10(b)), single SIFT feature ((Fig. 10(c))) and the combination of Lab/SIFT features (Fig. 10(d)). Although the original images are not complicated, obviously single-type feature cannot capture enough information to support correct co-segmentation. For example, the dark shadows (dissimilar with the Christ statue) and white clouds (similar with the Christ statue) are incorrectly segmented due to the fuzzy Lab color affinity. Moreover, there are some incorrectly-segmented parts in the SIFT-driven result (Fig. 10(c)), because SIFT feature performs badly in the smooth region without distinct gradient changes. In sharp contrast, the combination of Lab and SIFT features can produce the best results. Second, unsuitable multi-features

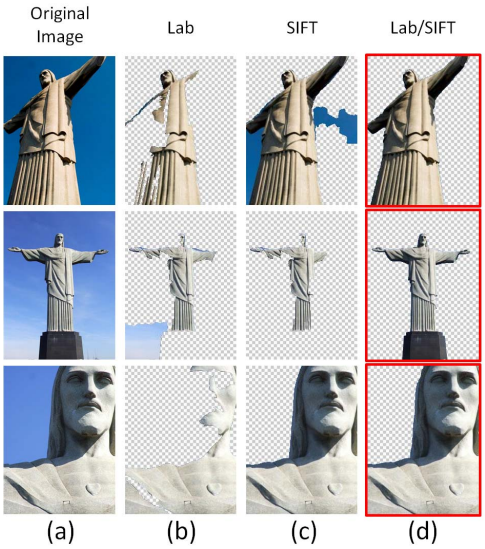


Fig. 10. Co-segmentation results based on different types of features. (a) Original images; (b) Single Lab color feature; (c) Single SIFT feature; (d) The combination of Lab and SIFT features.

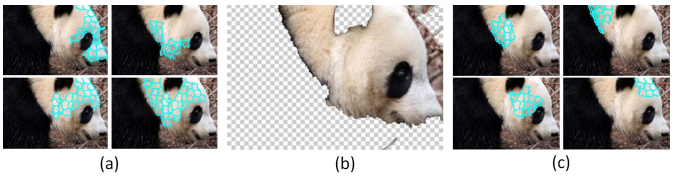


Fig. 11. Illustration of a failure case. (a) Four generated region examples over a white panda's head; (b) The incorrect foreground (panda) co-segmentation result generated by our method; (c) Four region examples generated by $\mathbf{W_{Lab}}$+Clustering.

Fig. 12. Illustration of the structure-meaningful property of our mid-level regions. (a) A sign board is divided into four relative integrated structures by our structure-meaningful mid-level region generation scheme; (b) Region generation results of W_{Lab}+**Clustering**; (c) Region generation results of **Big super-pixel**; (d) Region generation results of W_{Lab}+**Clustering**.

or naive combination of massive feature types may also confuse the global optimization of Eq. 6, because it cannot guarantee that dissimilar regions will be far away from each other in *each feature space*. For example, for some image sets in MSRC dataset (Fig. 7), the penalty terms ($||\mathbf{Z}||_{1,1}$ and $||\mathbf{Z}||_{2,1}$ in Eq. 6) are hard to meet the expectation that, SIFT feature should play a more important role than color-like features when many color features are involved in the optimization.

Yet, our framework still has some limitations. For example, as shown in Fig. 11, the foreground "panda" has two distinguishable color parts (black and white), which are hard to be integrated (see in Fig. 11(b)), because the intra-image coherency is too weak in all images, and the implied

cross-image supervised information loses its role in nature. Specifically, although Eq. 8 is expected to depict the structural meanings by defining the intra-image connections based on bi-harmonic distribution difference governed affinities $\mathbf{W}_{DE}$, we must recognize that this kind of structure essentially depends on color coherency because of the Laplacian matrix definition in Eq. 2. Meanwhile, although clustering by Normalized-Cuts step is naturally suitable for multi-class co-segmentation cases. However, due to the globally-unsupervised characteristics underlying the definition of $\mathbf{W}_{DE}$, it is still hard to segment certain foreground objects with some discriminative parts partitioned by clear boundaries, because the algorithm tends to regard them as different classes (clusters). Therefore, our structure-meaningful regions can facilitate to alleviate this problem to some extent by extending conventional local difference to mid-level difference. Obviously, in most cases, the foregrounds comprise of some different parts that are not too discriminative with each other, for example, the "panda" case shown in Fig. 11. When handling these cases, benefitting from the global perspective of inter-image mid-level coherency exploitation, our parallel normalization for each intra-image sub-graph facilitates to amplify such weak intra-image coherency (see Section VI-B).

IX. Conclusion

In this paper, we have detailed a novel and powerful method to address a suite of research challenges in the unsupervised co-segmentation problem with multiple foregrounds and high variability. The extensive experiments and accompanying evaluations verify the versatility and superiority of our method. In particular, the critical and novel technical components of our approach include: bi-harmonic distance distribution based new metric design, structure-meaningful mid-level region generation, L_1-manifold hyper-graph construction involving multi-type features, and unsupervised hyper-graph joint-cut model. Specially, the proposed metric can also contribute to other physics-based affinity measurement, and the structure-meaningful mid-level regions should give rise to better performance than the traditional primitives for sparse-coding and manifold learning, while the hyper-graph based joint-cut model provides a good scheme to improve certain supervised or semi-supervised co-segmentation methods.

Since our framework lays emphasis on the leverage of the mutual effects among different features, our ongoing efforts are geared towards finding a self-tuning or adaptive way to determine the involved parameters in a more intuitive way, so that they can be more easily understood and expressed comparing with pure mathematic subspace based method. Besides, exploring other relevant applications also deserves our immediate research endeavor. For example, we are planning to generalize our key idea to handle group saliency detection, image annotation, image retrieval, and context-aware image editing.

References

[1] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching—Incorporating a global constraint into MRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 993–1000.

[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] L. Mukherjee, V. Singh, and C. R. Dyer, "Half-integrality based algorithms for cosegmentation of images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2028–2035.

[4] J. C. Rubio, J. Serrat, A. M. Lopez, and N. Paragios, "Unsupervised co-segmentation through region matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 749–756.

[5] H. Li, F. Meng, B. Luo, and S. Zhu, "Repairing bad co-segmentation using its quality evaluation and segment propagation," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3545–3559, Aug. 2014.

[6] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 542–549.

[7] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1943–1950.

[8] J. Ma, S. Li, A. Hao, and H. Qin, "Unsupervised co-segmentation of complex image set via bi-harmonic distance governed multi-level deformable graph clustering," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2013, pp. 38–45.

[9] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 169–176.

[10] R. Hu, L. Fan, and L. Liu, "Co-segmentation of 3D shapes via subspace clustering," *Eurograph. Symp. Geometry Process.*, vol 31, no. 5, pp. 1703–1713, 2012.

[11] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan, "Multi-task low-rank affinity pursuit for image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2439–2446.

[12] L. Mukherjee, V. Singh, J. Xu, and M. D. Collins, "Analyzing the subspace structure of related images: Concurrent segmentation of image sets," in *Proc. 12th Eur. Conf. Comput. Vis.*, vol. 7575. 2012, pp. 128–142.

[13] E. Kim, H. Li, and X. Huang, "A hierarchical image clustering cosegmentation framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 686–693.

[14] Z. Wang and R. Liu, "Semi-supervised learning for large scale image cosegmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 393–400.

[15] D. Kuettel, M. Guillaumin, and V. Ferrari, "Segmentation propagation in imagenet," in *Proc. Eur. Conf. Comput. Vis.*, vol. 7578. 2012, pp. 459–473.

[16] G. Kim and E. P. Xing, "On multiple foreground cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 837–844.

[17] O. Sidi, O. van Kaick, Y. Kleiman, H. Zhang, and D. Cohen-Or, "Unsupervised co-segmentation of a set of shapes via descriptor-space spectral clustering," *ACM Trans. Graph.*, vol. 30, no. 6, 2011, Art. no. 126.

[18] A. Faktor and M. Irani, "Co-segmentation by composition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1297–1304.

[19] S. Vicente, C. Rother, and V. Kolmogorov, "Object cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2217–2224.

[20] 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.

[21] L. Mukherjee, V. Singh, and J. Peng, "Scale invariant cosegmentation for image groups," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1881–1888.

[22] 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.

[23] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2013, pp. 923–930.

[24] A. Jain, A. Gupta, M. Rodriguez, and L. S. Davis, "Representing videos using mid-level discriminative patches," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2013, pp. 2571–2578.

[25] S. Maji and G. Shakhnarovich, "Part discovery from partial correspondence," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2013, pp. 931–938.

[26] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 73–86.

[27] D. Glasner, S. N. P. Vitaladevuni, and R. Basri, "Contour-based joint clustering of multiple segmentations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2385–2392.

[28] F. Moreno-Noguer, "Deformation and illumination invariant feature point descriptor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1593–1600.

[29] R. Behmo, N. Paragios, and V. Prinet, "Graph commute times for image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[30] A. Criminisi, T. Sharp, and A. Blake, "Geos: Geodesic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 99–112.

[31] X. Liu, J. He, and B. Lang, "Multiple feature kernel hashing for large-scale visual search," *Pattern Recognit.*, vol. 47, no. 2, pp. 748–757, Feb. 2014.

[32] X. Liu, C. Deng, B. Lang, D. Tao, and X. Li, "Query-adaptive reciprocal hash tables for nearest neighbor search," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 907–919, Feb. 2016.

[33] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[34] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2009, pp. 2790–2797.

[35] S. Wang, X. Yuan, T. Yao, S. Yan, and J. Shen, "Efficient subspace segmentation via quadratic programming," in *Proc. 25th AAAI Conf. Artif. Intell. (AAAI Publications)*, 2011, pp. 519–524.

[36] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, "Learning with ℓ^1-graph for image analysis," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 858–866, Apr. 2010.

[37] Y. Li, J. Liu, Z. Li, Y. Liu, and H. Lu, "Object co-segmentation via discriminative low rank matrix recovery," *ACM Multimedia*, 2013. pp. 749–752.

[38] F. Meng, H. Li, K. N. Ngan, L. Zeng, and Q. Wu, "Feature adaptive co-segmentation by complexity awareness," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4809–4824, Dec. 2013.

[39] L. Wang, G. Hua, J. Xue, Z. Gao, and N. Zheng, "Joint segmentation and recognition of categorized objects from noisy Web image collection," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4070–4086, Sep. 2014.

[40] 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.

[41] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2013, pp. 1939–1946.

[42] C. Rother, V. Kolmogorov, and A. Blake, "'Grabcut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

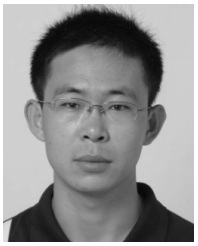
[43] Y. Lipman, R. Rustamov, and T. A. Funkhouser, "Biharmonic distance," *ACM Trans. Graph.*, vol. 29, no. 3, 2010, Art. no. 27.

[44] E. G. Birgin, J. M. Martínez, and M. Raydan, "Nonmonotone spectral projected gradient methods on convex sets," *J. Optim. Soc. Ind. Appl. Math.*, vol. 10, no. 4, pp. 1196–1211, 2000.

[45] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.

[46] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.

Jizhou Ma received the B.S. degree in computer science from the Hefei University of Technology, in 2012. He is currently pursuing the Ph.D. degree in technology of computer application with Beihang University, Beijing, China. His research interests include computer vision, pattern recognition, and image processing.



Shuai Li received the Ph.D. degree in computer science from Beihang University. He is currently an Assistant Professor with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. His research interests include computer graphics, pattern recognition, computer vision, physics-based modeling and simulation, and medical image processing.

Hong Qin (SM'08) received the B.S. and M.S. degrees from Peking University, and the Ph.D. degree from the University of Toronto, all in computer science. He is currently a Professor of Computer Science with the Department of Computer Science, Stony Brook University. His research interests include geometric and solid modeling, graphics, physics-based modeling and simulation, computer-aided geometric design, visualization, and scientific computing.

Aimin Hao received the B.S., M.S., and Ph.D. degrees from Beihang University all in computer science. He is currently a Professor with the Computer Science School and an Associate Director of the State Key Laboratory of Virtual Reality Technology and Systems with Beihang University. His research interests include virtual reality, computer simulation, computer graphics, geometric modeling, image processing, and computer vision.