Image Co-Segmentation via Consistent Functional Maps

Fan Wang
Stanford University
fanw@stanford.edu

Qixing Huang
Stanford University
huangqx@stanford.edu

Leonidas J. Guibas
Stanford University
guibas@cs.stanford.edu

Abstract

Joint segmentation of image sets has great importance for object recognition, image classification, and image retrieval. In this paper, we aim to jointly segment a set of images starting from a small number of labeled images or none at all. To allow the images to share segmentation information with each other, we build a network that contains segmented as well as unsegmented images, and extract functional maps between connected image pairs based on image appearance features. These functional maps act as general property transporters between the images and, in particular, are used to transfer segmentations. We define and operate in a reduced functional space optimized so that the functional maps approximately satisfy cycle-consistency under composition in the network. A joint optimization framework is proposed to simultaneously generate all segmentation functions over the images so that they both align with local segmentation cues in each particular image, and agree with each other under network transportation. This formulation allows us to extract segmentations even with no training data, but can also exploit such data when available. The collective effect of the joint processing using functional maps leads to accurate information sharing among images and yields superior segmentation results, as shown on the iCoseg, MSRC, and PASCAL data sets.

1. Introduction

Co-segmentation, i.e., jointly segmenting a collection of similar images, has received a good deal of attention recently in the vision literature [10, 24, 19]. Compared with single image segmentation, co-segmentation has the potential of aggregating information from multiple images to improve the segmentation of individual images.

The crucial task in co-segmentation is to estimate and represent relations between different images consistently. So far this has been approached by computing point-based maps between pairs of images using local descriptors such as SIFT. Although this strategy works well on images that contain the same object instance at various viewpoints and scales, it is less effective when images exhibit large appearance or object variations.

Moreover, even though one can establish maps between pairs of images, in the context of multiple images, it is challenging to enforce global consistency among these maps — so that compositions of maps along cycles approximate the identity map or, equivalently, compositions of maps along different paths between two images are approximately the same. This cycle-consistency constraint is an essential regularizer for the problem. It alleviates the imperfections of the individual maps and the difficulty of making path choices when transferring information between images.

In this paper we present a novel framework, called *consistent functional maps*, for representing and computing consistent appearance relations among a collection of images. The proposed framework modifies the functional map framework [16] to use it on pairs of images instead of shapes, and further extends it to handle multiple images under consistency constraints. The basic idea of functional maps is to equip each image with a linear functional space, and represent relations between images as linear maps between these functional spaces. This functional representation is powerful because image descriptors and segmentations can be considered as functions on images, and their relations can thus be encoded as linear constraints on the linear map between the two spaces. In particular, with a properly chosen basis for each functional space, optimizing functional maps between images becomes optimization of the familiar transformation matrices. This allows us to apply rich matrix optimization and linear algebraic techniques.

Most importantly, for our purposes, in a network of images connected by such functional maps, the functional setting admits of an easy approach for enforcing the consistency of these maps. By introducing a latent basis for the functional space associated with each image, the consistency of the functional maps is equivalent to the fact that each functional map transforms the source image latent basis to the target image latent basis. This leads to a simple formulation of the cycle-consistency constraint, enabling us to compute consistent functional maps among multiple images by solving a tractable optimization problem.

Given the consistent functional maps computed between pairs of images, we jointly optimize segmentations of all images so that they (i) are consistent with each other when transported via the functional maps, and (ii) agree with segmentation boundary clues presented on each image (e.g., sharp edges). We note that this optimization procedure is easily modified to incorporate labeled images as input, in which case we simply let the labeled images provide additional clues for segmentation.

The proposed approach exhibits significantly improved performance on two standard cosegmentation data sets i-Coseg [3] and MSRC [21] compared with recent state-of-the-art methods. Moreover, we create a more challenging data set with a larger number of images and larger variance in object appearance using images from the PASCAL VOC data set [8]. Our method outperforms other techniques on this data set as well.

2. Related Work

Earlier work on joint segmentation mainly compared the visual features of image pairs, such as foreground color histogram [18], SIFT [14], saliency [5], and Gabor features [9]. Joulin et al. [10] formulated co-segmentation as a discriminative clustering problem such that the resulting foreground and background could be separated with the largest margin. Region matching was applied to exploit inter-image information by establishing correspondences between the common objects in the scene. This allows us to jointly estimate the appearance distributions of both the foreground and the background [19]. In the supervised setting, a pool of object-like candidate segmentations were generated and a random forest regressor was trained to score each pair of segmentations [24]. All these works succeeded in automatically generating co-segmentation results. However, only a few of them [10, 19, 24] focus on the challenging data sets iCoseg and MSRC which contain images with different viewpoints, illumination, and object deformation. Recently, segmentation masks were transferred from the training windows to similar windows in test images [11], and images were jointly segmented in a energy minimization framework with multiple unary potentials [12].

Functional maps are related to graph matching for feature correspondences in object categorization and image matching [4, 13, 7, 23]. In these methods, an image is usually represented as a graph whose nodes are regions in the image. The edges of the graph reflect the underlying spatial structure of the image, such as region proximity, and are used to guarantee the geometric consistency of nearby regions during matching. An objective function describing appearance similarity and geometric compatibility is maximized to establish visual correspondences [4, 13]. Since the graph matching problem is NP-hard in most versions, an important topic under this theme is to design efficient algorithms for approximately solving the assignment problem [7, 23]. In the proposed method, we also use a graph to model an image via its super-pixel decomposition. However, our framework solves the graph matching problem in a functional setting, which is fundamentally different from point-wise correspondences and leads to a linear system with an easily-obtained optimal solution for each pairwise functional map.

The cycle-consistency constraint has been applied in the vision community for obtaining consistent affine matches among multiple images [25, 17]. These approaches are typically formulated as solving constrained optimization problems, where the objective functions encode the score of maps, and the constraints enforce the consistency of maps along cycles. However, these approaches assume that correct maps are dominant in the graph so that the small number of bad maps can be identified through their participation in many bad cycles. Moreover, as there is an exponential number of cycles, how to effectively sample cycles remains an open question. The consistency property of functional maps is also related to diffusion maps [6] and vector diffusion maps [22]. However, only orthonormal transformation matrices were allowed in vector diffusion maps, while we allow arbitrary maps.

3. Problem Statement and Overview

Our input is a collection of N similar images $\mathcal{I} = \{I_1, \cdots, I_N\}$, with each image containing an object of the same class, e.g, a cow. The goal is to jointly segment the objects from all the input images. We distinguish between an unsupervised setting, where only the input images are given, and a semi-supervised setting, where objects in the first $L \ll N$ images are pre-segmented.

Without loss of generality, we assume that each image is over-segmented into $K = 200$ super-pixels. We represent each image I_i as the graph $(\mathcal{P}_i, \mathcal{E}_i)$ of the super-pixel subdivision, where $\mathcal{P}_i = \{p\}$ collects all the super-pixels, and edges in $\mathcal{E}_i$ connect adjacent super-pixels. Edges in this graph are weighted according to the boundary length shared by the corresponding super-pixels, With this setup, the goal can be formulated as jointly computing a subset of super-pixels $\mathcal{O}_i \subset \mathcal{P}_i$ from each image I_i that represents the underlying object (or objects).

3.1. Building blocks

To better explain the proposed approach, we present a brief introduction to functional maps adapted from [16] for mapping meshed 3D shapes to our super-pixel setting, and formulate the cycle-consistency constraint. We remark that the obvious analog of [16] in the image domain is to use functions over the image pixels, but in our experience this works much less well (as shown in supplemental material).

Functions on super pixels. We formulate image segmentation as computing an indicator functions on the super-

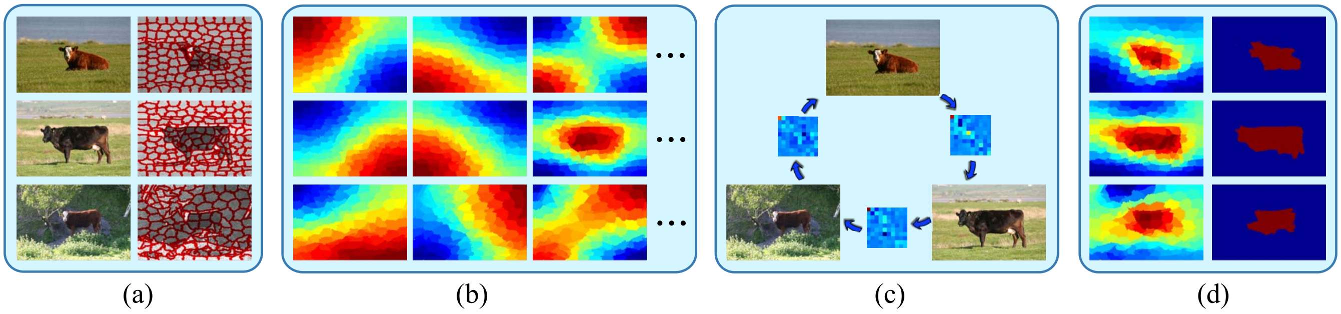


Figure 1: Overview of the proposed framework. (a) the original image and its super-pixel representation; (b) the first few Laplacian eigenfunctions, visualized on the image super-pixels; (c) the functional maps between images which satisfy cycle-consistency; (d) The final segmentation functions and the corresponding binary results after thresholding.

pixels. Define a function $f : \mathcal{P}_i \rightarrow \mathbb{R}$ that assigns each super-pixel $p \in \mathcal{P}_i$ to a real value $f_i(p)$. Let $\overline{\mathcal{F}}_i$ denote the space of all functions on $\mathcal{P}_i$. It is clear that $\overline{\mathcal{F}}_i \cong \mathbb{R}^K$ is a linear space of dimension K. Any segmentation $\mathcal{O}_i \in \mathcal{P}_i$ corresponds to a binary indicator function $f_{\mathcal{O}_i} \in \overline{\mathcal{F}}_i$ where $f_{\mathcal{O}_i}(p) = 1, \forall p \in \mathcal{O}_i$, and $f_{\mathcal{O}_i}(p) = 0, \forall p \in \mathcal{P}_i \setminus \mathcal{O}_i$. On the other hand, any function $f \in \overline{\mathcal{F}}_i$ induces a segmentation $\mathcal{O}_i = \{p | f(p) > t_i\}$, given a properly chosen threshold t_i.

Reduced functional space. To improve efficiency, we reduce the search space of segmentation indicator functions to a subspace $\mathcal{F}_i \subset \overline{\mathcal{F}}_i$ of dimension $M < K$ for each image I_i, spanned by a basis $B_i = (b_i^1, \cdots, b_i^M)$. In the following, we use $\mathbf{f}$ to denote the coefficients of f with respect to B_i. In other words, $f = \sum_{j=1}^{M} f_j b_i^j = B_i \mathbf{f}$. Please see §4 for details.

Functional map. Relations between images can be easily described as linear functional maps in the functional setting. Specifically, a functional map from $\mathcal{F}_i$ to $\mathcal{F}_j$ is given by a matrix $X_{ij} \in \mathbb{R}^{M \times M}$, where X_{ij} maps a function $f \in \mathcal{F}_i$ with coefficient vector $\mathbf{f}$ to the function $f' \in \mathcal{F}_j$ with coefficient vector $\mathbf{f}' = X_{ij}\mathbf{f}$. We refer the reader to [16] for a more detailed introduction and intuition. In §5, we show how to adapt this framework to the image setting.

Cycle consistency. In the functional setting, the cycle-consistency constraint can be described as the fact that a transported function along any loop should be identical to the original function. Suppose we are given a connected directed graph $\mathcal{G}$ that connects some pairs of images in $\mathcal{I}$. Denote $X_{ij} : \mathcal{F}_i \rightarrow \mathcal{F}_j$ as the functional map associated with edge $(i, j) \in \mathcal{G}$. Let $\mathcal{C}$ denote the space of all cycles in $\mathcal{G}$, then the cycle consistency constraint can be described as

$$\begin{array}{l} X_{i_k i_0} \cdots X_{i_1 i_2} X_{i_0 i_1} \mathbf{f} = \mathbf{f} \\ \forall (I_{i_0}, I_{i_1}, \cdots, I_{i_k}) \in \mathcal{C},\ \mathbf{f} \in \mathcal{F}_{i_0}. \end{array} \quad (1)$$

Computationally, it is difficult to account for all the loops and high-order constraints in Eq. 1. Instead, we introduce a latent basis $Y_i = (\mathbf{y}_i^1, \cdots, \mathbf{y}_i^M)$ for each image I_i. This latent space is expected to include functions that are consistent across multiple images, e.g., segmentation functions of the underlying objects. With this setup, we simply constrain that X_{ij} are consistent with these latent basis, i.e.,

$$X_{ij} Y_i = Y_j, \quad \forall (i, j) \in \mathcal{G}. \quad (2)$$

It is easy to see that Eq. 2 is equivalent to Eq. 1 since for any function $\mathbf{f}_g$ in the global coordinate system and its coefficient vector $\mathbf{f} = Y_{i_0}\mathbf{f}_g \in \mathcal{F}_{i_0}$, we have

$$\begin{aligned} X_{i_k i_0} \cdots X_{i_1 i_2} X_{i_0 i_1} \mathbf{f} &= X_{i_k i_0} \cdots X_{i_1 i_2} X_{i_0 i_1} Y_{i_0} \mathbf{f}_g \\ &= X_{i_k i_0} Y_{i_k} \mathbf{f}_g = Y_{i_0} \mathbf{f}_g = \mathbf{f}. \end{aligned}$$

3.2. Approach overview

The proposed approach proceeds in three stages (Fig. 1). The first stage computes a reduced functional space on each image. The second stage optimizes consistent functional maps between pairs of images. The objective function combines a term that quantifies the quality of pair-wise functional maps, and another term that enforces the consistency among all functional maps. Given these consistent functional maps, the final stage generates the segmentations by jointly optimizing segmentation functions that (i) align with the segmentation clues on each image and (ii) are consistent with neighboring image segmentations after transportation by functional maps. This optimization problem can easily incorporate supervision information when some of the images have ground-truth segmentations.

4. Reduced Functional Spaces

We choose $\mathcal{F}_i$ as the eigen-space spanned by the first M eigenvectors of the normalized graph Laplacian $L_i \in \mathbb{R}^{K \times K}$, motivated by some practical success of using these eigenvectors or combinations of them as segmentation indicator functions [20]. An important distinction of the proposed approach from previous methods is that we do not commit to any segmentation indicator function at this stage. Instead, these are jointly selected later over all input images using optimized functional maps.

The segmentation functions can be approximated well in the reduced eigen-space. Fig. 2 shows that when $M =$

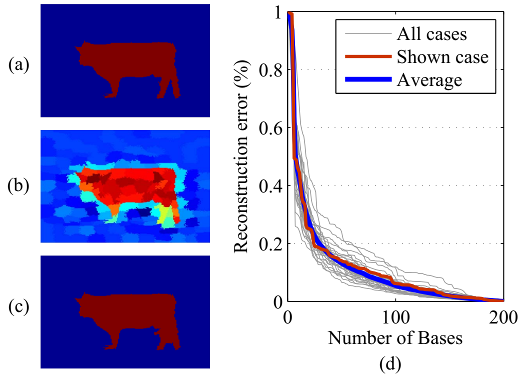


Figure 2: (a) Example of a binary segmentation function; (b) Approximation of (a) with 30 basis functions (18.2% error); (c) Binary version of (b) by thresholding (3% error); (d) The gray lines are reconstruction errors of typical segmentation functions; the red line shows the case in (b) and (c); the blue line is the average of all cases.

30, the normalized error between the original segmentation function and its projection to the reduced space is usually less than 20%. Furthermore, if we convert the projected function into a binary function by greedy thresholding (§6.2), the averaged error is further reduced to 3%. In other words, optimizing the segmentation functions in $\mathcal{F}_i$ is sufficient for our purposes.

5. Extracting Consistent Functional Maps

Here we describe how to construct a sparse graph of images (§5.1), and how to compute consistent functional maps for each edge in this graph, given the reduced functional spaces $\mathcal{F}_i, 1 \le i \le N$ (§5.2 - §5.5).

5.1. Similarity graph

We compute a sparse similarity graph $\mathcal{G}$ for the input image collection $\mathcal{I}$, and only compute functional maps between pairs of images specified by $\mathcal{G}$. In this paper, we simply connect each image with its $k = 30$ nearest images using their GIST [15] descriptors $\mathbf{g}_i, 1 \le i \le N$. We assign a weight to the edge between image pair $(i, j) \in \mathcal{G}$ via

$$w_{ij} = \exp(-\|\mathbf{g}_i - \mathbf{g}_j\|^2/2\sigma^2), \tag{3}$$

where $\sigma = \text{median}(\|\mathbf{g}_i - \mathbf{g}_j\|)$ is the median of image descriptor differences.

5.2. Aligning image features

When computing the functional map X_{ij} from image I_i to image I_j, it is natural to enforce that X_{ij} agrees with the features computed from the images. In the functional setting, this is equivalent to the constraint that $X_{ij}\mathbf{d}_i \approx \mathbf{d}_j$, where $\mathbf{d}_i$ and $\mathbf{d}_j$ are corresponding descriptor functions on the two images represented in the reduced functional space. Let D_i (D_j) collect all descriptors of image I_i (I_j) in columns; then the function preservation constraints can be written as

$$f_{ij}^{\text{feature}} = \|X_{ij}D_i - D_j\|_1, \tag{4}$$

where $\|\cdot\|_1$ denotes the element-wise L1-norm (sum of absolute values of all elements) to account for noise in image descriptors. In this paper, the features in D_i and D_j include 3 average RGB values, a 64-dimensional color histogram, and the bag-of-visual-words histograms with 300 visual words. In total there are 367 descriptor functions, yielding an over-determined problem.

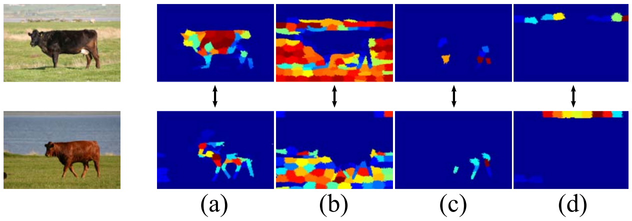


Figure 3: Visualization of some probe functions that are put in correspondence by the functional map. The four probe functions shown are the 1st (a) and the 27th (b) dimension of the color histogram, and the 17th (c) and the 291st(d) dimension of the bag-of-visual-words feature.

5.3. Regularization

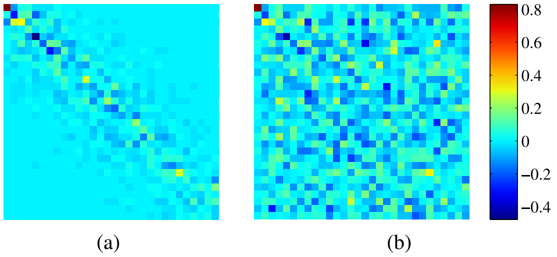
The regularizer used for X_{ij} is represented as:

$$f_{ij}^{\text{reg}} = \sum_{1\le s,s'\le M} \Big(|\lambda_i^s - \lambda_j^{s'}|X_{ij}(s,s')\Big)^2, \tag{5}$$

where λ_i^s ($\lambda_j^{s'}$) denotes the s-th (s'-th) eigenvalue of the graph Laplacian matrix L_i (L_j). Note that for similar images (i.e., similar spectra of eigenvalues), minimizing f_{ij}^{reg} essentially forces X_{ij} to be close to a diagonal matrix. This is an expected effect, since the magnitudes of eigenvalues reflect the frequencies of the corresponding eigenvectors, and eigenvectors of similar frequencies (corresponding to similar scales) are more likely to be related [20]. Fig. 4 shows an example of functional map with and without the regularization term — the one with regularization is closer to a diagonal matrix, meaning that eigenvectors are transported only to their counterparts with similar frequencies.

5.4. Incorporating map consistency

As described in Eq. 2, we formulate the cycle-consistency constraint of functional maps by introducing a latent basis Y_i for each $\mathcal{F}_i$, and force each functional map X_{ij} to transform Y_i into Y_j. Practically, we found that it is better to consider a reduced basis $Y_i \in \mathbb{R}^{M\times m}$. This is because $\mathcal{F}_i$ is generated by eigenvectors of L_i, and there is no guarantee that these eigenvectors are totally consistent, due to inter-image variability. We choose $m = 20$ for



(a) (b)

Figure 4: The functional map (a) with and (b) without commutativity regularization.

all experiments. With this setup, we formulate the map consistency term as

$$f^{\text{cons}} = \sum_{(i,j)\in\mathcal{G}} w_{ij} f_{ij}^{\text{cons}} = \sum_{(i,j)\in\mathcal{G}} w_{ij} \|X_{ij}Y_i - Y_j\|_{\mathcal{F}}^2 \,, \quad (6)$$

where $\|\cdot\|_{\mathcal{F}}$ denotes the Frobenius norm.

Note that merely minimizing Eq. 6 would force the Y_i to be zero matrices. We thus impose an additional constraint $Y^T Y = I_m$, where the latent basis matrix Y is simply $(Y_1^T, \cdots, Y_N^T)^T$. This ensures that the columns of Y are linearly independent, favoring solutions of Y_i that are not rank-deficient.

5.5. Optimization

Combining Eq. 4-6, we arrive at the following optimization problem for computing consistent functional maps:

$$\begin{aligned} \min \quad & \sum_{(i,j)\in\mathcal{G}} w_{ij} \left(f_{ij}^{\text{feature}} + \mu f_{ij}^{\text{reg}} + \lambda f_{ij}^{\text{cons}} \right) \\ s.t. \quad & Y^T Y = I_m, \end{aligned} \quad (7)$$

where λ and μ control the tradeoffs between different objective terms. For all the experiments, we set $\lambda = 10$ and $\mu = 40$. The effect of the consistency term is shown in Fig. 5 and a segmentation function transferred along a cycle is illustrated in Fig. 6.

To effectively solve Eq. 7, we use an alternating optimization strategy, which decouples the optimization of $\{X_{ij}, (i,j) \in \mathcal{G}\}$ from the optimization of Y, leading to subproblems that are much easier to solve.

Optimizing functional maps X_{ij}. When the latent basis matrix Y is fixed, X_{ij} can be optimized independently, i.e., the optimal value of X_{ij} is given by

$$X_{ij}^{\star} = \arg\min_{X} \left(f_{ij}^{\text{feature}} + \mu f_{ij}^{\text{reg}} + \lambda f_{ij}^{\text{cons}} \right) \,. \quad (8)$$

Eq. 8 is a quadratic program, and we use SeDuMi [1] to solve it efficiently. In the first iteration where the latent basis matrix Y is unknown, we set $\lambda = 0$.

Optimizing latent basis matrix Y. When the functional maps X_{ij} are fixed, Eq. 7 for solving the latent basis matrix Y becomes

$$\begin{aligned} \min \quad & \text{trace}(Y^T W Y) \\ \text{s.t.} \quad & Y^T Y = I_m, \end{aligned} \quad (9)$$

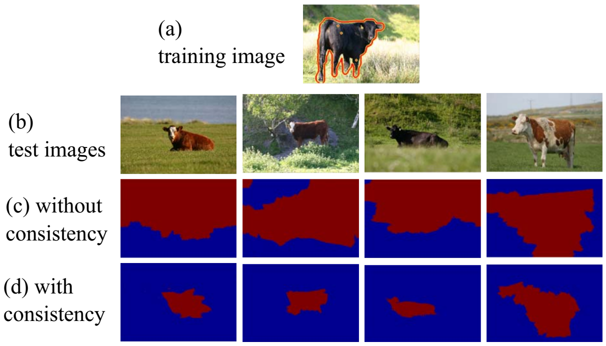
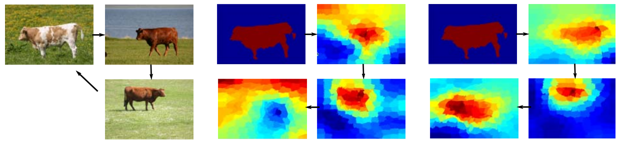


Figure 5: (a) Training image with ground truth segmentation; (b) test images; segmentation results transferred from (a) through the maps obtained by Eq. 7 without the consistency term in (c) and with the consistency term in (d).



(a) example image cycle (b) w/o consistency (c) w/ consistency

Figure 6: Given a cycle of 3 images in (a), the segmentation function of the first image is transferred along the cycle. The final function transferred back looks like the original one more in (c) when the maps are consistent than that in (b) when map consistency is not enforced.

where matrix $W \in \mathbb{R}^{NM \times NM}$ consists of $N \times N$ blocks, with the (i,j)-th block

$$W_{ij} = \begin{cases} \sum\limits_{(i,j')\in\mathcal{G}} w_{ij'}(I_m + X_{ij'}^T X_{ij'}) & i = j \\ -w_{ij}(X_{ji} + X_{ij}^T) & (i,j) \in \mathcal{G} \\ 0 & \text{otherwise.} \end{cases}$$

The following proposition provides the analytical solutions to Eq. 9:

Proposition 1 *Denote by $\sigma_1 \leq \cdots \leq \sigma_m$ the first m eigenvalues of W. Let $U = (\mathbf{u}_1, \cdots, \mathbf{u}_m)$ collect the corresponding eigenvectors. The optimal solution to Eq. 9 is given by*

$$Y = UV, \quad \forall\, V \in \mathcal{O}(m), \quad (10)$$

where $\mathcal{O}(m)$ denotes the space of all orthonormal matrices of dimension $m \times m$.

It is clear that the value of $\|X_{ij}Y_i - Y_j\|_{\mathcal{F}}^2$ is invariant for any orthonormal V. We simply set the optimal value $Y^{\star} = U$.

Stopping criterion. Let $X_{ij}^{(k)}$ denote the value of X_{ij} at iteration k. We alternate between the optimization of X_{ij} and Y until $\|X_{ij}^{(k+1)} - X_{ij}^{(k)}\|_{\mathcal{F}} / \|X_{ij}^{(1)}\|_{\mathcal{F}} < 10^{-6}, \forall (i,j) \in \mathcal{G}$.

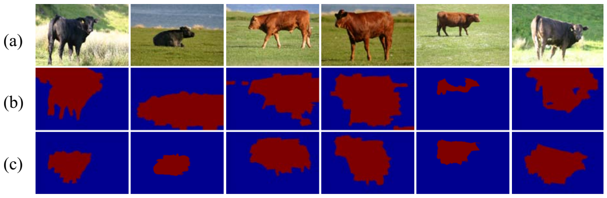


Figure 7: Generated segmentation function (c) compared with normalized cut results (b). Our results are more consistent. All results are shown in the granularity of super-pixels.

6. Generating Consistent Segmentations

Given the consistent functional maps $\{X_{ij}\}$, the final stage of the proposed approach jointly optimizes an approximate segmentation indicator function $f_i \in \mathcal{F}_i$ for each image. We then generate the final segmentation by rounding/binarizing f_i into a segmentation indicator function.

6.1. Joint optimization of segmentation functions

To optimize the coefficient vectors $\mathbf{f}_i$ of segmentation functions $f_i = B_i\mathbf{f}_i$, we minimize an objective function which consists of a map consistency term f^{map} and a segmentation term f^{seg}. The map consistency term f^{map} ensures the segmentation functions are consistent with the optimized functional maps:

$$f^{\text{map}} = \sum_{(i,j)\in\mathcal{G}} w_{ij}\|X_{ij}\mathbf{f}_i - \mathbf{f}_j\|_{\mathcal{F}}^2. \tag{11}$$

The segmentation term sums the alignment score between each segmentation function and segmentation clues provided on each image. We formulate each alignment score as a quadratic function $f_i^T L_i f_i = \mathbf{f}_i^T B_i^T L_i B_i \mathbf{f}_i$[20]. For unlabeled images, we set L_i as the normalized graph Laplacian. For labeled images, we re-define L_i as the normalized graph Laplacian of the graph that only connects super-pixels within the foreground or background segment. The segmentation term is then given by:

$$f^{\text{seg}} = \sum_{i=1}^{N} \mathbf{f}_i^T B_i^T L_i B_i \mathbf{f}_i. \tag{12}$$

Combing Eq.11 and 12, we arrive at the following constrained optimization problem:

$$\min \ f^{\text{seg}} + \gamma f^{\text{map}} \quad s.t. \sum_{i=1}^{N} \|\mathbf{f}_i\|^2 = 1, \tag{13}$$

where γ controls the importance of f^{map} with respect to f^{seg}. In this paper, we set $\gamma = 10$. In the same spirit as optimizing the latent basis Y in Eq. 9, we constrain the norm of $\mathbf{f}_i$ to prevent trivial solutions. The effect of the global consistency term is shown in Fig. 7 in comparison with Normalized Cut results.

It is easy to see that Eq. 13 is equivalent to $\min \bar{\mathbf{f}}^T Z \bar{\mathbf{f}}$ subject to $\|\bar{\mathbf{f}}\|_2^2 = 1$, in which Z can be obtained by adding $B_i^T L_i B_i$ to each diagonal block of W, i.e. $Z = \text{Diag}(B_i^T L_i B_i) + \gamma W$. The optimal solution is given by the second smallest eigenvector $\bar{\mathbf{f}}^\star = (\mathbf{f}_1^\star, \cdots, \mathbf{f}_N^\star) \in \mathbb{R}^{NM\times 1}$ of matrix Z. The corresponding segmentation function of each image I_i is then obtained by $s_i = B_i\mathbf{f}_i^\star$.

6.2. Rounding segmentation functions

The continuous functions s_i generated above already delineate the objects in the images well. To convert them into binary indicator functions, we simply sample 30 thresholds within the interval $[\min(s_i), \max(s_i)]$ uniformly, and choose the threshold whose corresponding segmentation has the smallest normalized cut score [20].

7. Experimental Results

7.1. Standard co-segmentation data sets

iCoseg data set We first evaluated our method on the iCoseg data set [3], which contains 38 object classes (643 images in total) with known pixel-level segmentation. Images in each class contain very similar objects, and within-class variability is relatively low. Segmentation accuracy is calculated as the percentage of correctly labeled pixels.

Table 1 shows the accuracy of our unsupervised joint segmentation method, two other state-of-the-art unsupervised co-segmentation algorithms [10, 19], and a supervised method [24]. The same number of images are used in all methods. The accuracy is averaged over 20 random selections for each class. Our method is significantly better than the state-of-the-art unsupervised methods in most of the cases, and comparable and sometimes even better than the supervised one [24]. Additionally, we also show the comparison of average accuracy with the segmentation transfer method [12] in Table 2. Note that even though [12] is a supervised method trained with the entire PASCAL VOC10 training set, our unsupervised method is better than their simplified version ("image + transfer" in [12]), and comparable with their full model.

MSRC data set We further evaluated the proposed method on the MSRC data set. It includes 591 pixel-wise labeled images of 23 object classes. The accuracy of our unsupervised joint segmentation method is shown in Fig. 8a, in comparison with other recent co-segmentation algorithms [10] and [19]. We select the same classes as reported by [10] and [19]. Our method is significantly better in most of the cases. It takes about 20 minutes to jointly segment 30 images with low memory usage. Please refer to supplemental material for details of computational cost.

Semi-supervised joint segmentation accuracy of our method is compared with [24] and [12] in Fig. 8b. Ten randomly selected images are used in testing for our method and [12] as in [24], and the remaining images are used in training. Our method again outperforms the state-of-the-art

class	[10]	[19]	[24]	FMaps-uns
Alaska Bear	74.8	86.4	90.0	**90.4**
Red Sox Players	73.0	90.5	90.9	**94.2**
Stonehenge1	56.6	87.3	63.3	**92.5**
Stonehenge2	86.0	88.4	88.8	87.2
Liverpool FC	76.4	82.6	87.5	**89.4**
Ferrari	85.0	84.3	89.9	**95.6**
Taj Mahal	73.7	88.7	91.1	**92.6**
Elephants	70.1	75.0	43.1	**86.7**
Pandas	84.0	60.0	92.7	88.6
Kite	87.0	89.8	90.3	**93.9**
Kite panda	73.2	78.3	90.2	**93.1**
Gymnastics	90.9	87.1	91.7	90.4
Skating	82.1	76.8	77.5	78.7
Hot Balloons	85.2	89.0	90.1	**90.4**
Liberty Statue	90.6	91.6	93.8	**96.8**
Brown Bear	74.0	80.4	95.3	88.1
Average	78.9	83.5	85.4	**90.5**

Table 1: Segmentation accuracy on the iCoseg data set.

Supervised		Unsupervised
image + transfer [12]	full model [12]	FMaps
87.6	91.4	90.5

Table 2: Average accuracy on the iCoseg data set.

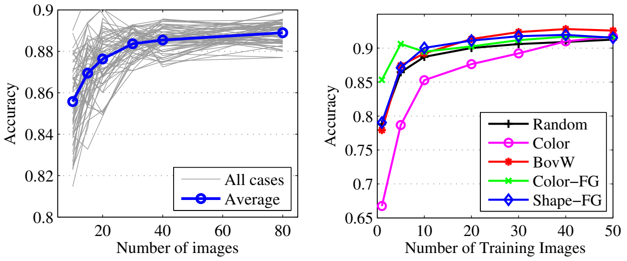
methods in most of the cases.

Discussion The functional maps are surprisingly effective in building natural correspondences between images. The superior performance shows the effectiveness of the consistent functional maps and the joint optimization framework in generating robust results from the image network, despite the imperfections of individual maps. Although exact pixel or region correspondences may exist, in the more general functional formulation cycle consistency can be easily enforced, yielding improved results. Example segmentation results are shown in supplemental material.

7.2. Larger data set

Besides the standard data sets, we investigated performance on a larger and more diverse data set, created by retrieving images with selected class labels from the PASCAL VOC 2012 data set. Images from the same object class are treated as a group for joint segmentation.

In the semi-supervised setting, the images from the "training" set and the "validation" set of PASCAL are used as labeled and unlabeled images, respectively. Our framework is compared with another state-of-the-art segmentation transfer method [12], and the results are in Fig. 8c. Our method significantly improves the segmentation performance. In the last column, we also include the performance of our unsupervised framework with all images in each class jointly segmented without any label information. Note that in quite a few cases, our unsupervised technique already outperforms the supervised method of [12].



(a) Unsupervised

(b) Supervised

Figure 9: Segmentation accuracy as a function of (a) unlabeled images in the unsupervised setting and (b) labeled images in the supervised setting.

For natural objects such as cat, cow, dog, horse, and sheep, the performance of our unsupervised method is very close to the supervised version. This suggests that, because the image set does not have much variation in object appearance, adding labeled images does not provide much additional information beyond the clues contained in the unlabeled images. On the other hand, for man-made objects such as bus and car, the training data are more helpful because the appearances of related objects differ a lot.

Sensitivity Analysis In the unsupervised case, we also investigated accuracy as a function of the number of images. Starting from a random subset of the "aeroplane" class in PASCAL, we add more and more images to the unsupervised set in random order. This improves accuracy in general because added images provide more segmentation cues to each other (Fig. 9a).

In the supervised case, we vary the number of training images selected from the "training set", and evaluate the performance on a separate "validation set" of 90 images (Fig. 9b). Ground truth segmentations are added for training according to their image similarities to the validation set using global color histogram (Color) or Bag-of-visual-Words (BovW). As the number of training images increases, the segmentation accuracy improves rapidly, then gradually saturates. The curve for adding training images in random order is also shown (averaged over 20 runs). The faster increase using BovW suggests that it may be used to actively select more helpful training images. To further confirm this observation, we pretend that the foregrounds of the test images are known, and select training images based on similarities in foreground color histogram (Color-FG) and foreground shape histogram (Shape-FG [2]). Their curves rise even faster than BovW, confirming the importance of selecting good training images.

8. Conclusion

We have proposed a framework for joint image segmentation, in which functional between images are jointly esti-

class	N	[10]	[19]	FMaps-uns
cow	30	81.6	80.1	**89.7**
plane	30	73.8	77.0	**87.3**
face	30	84.3	76.3	**89.3**
cat	24	74.4	77.1	**88.3**
car(front)	6	87.6	65.9	87.3
car(back)	6	85.1	52.4	**92.7**
bike	30	63.3	62.4	**74.8**

(a) MSRC, unsupervised

class	[24]	[12]	FMaps-s
cow	94.2	92.5	**94.3**
plane	83.0	86.5	**91.0**
car	79.6	88.8	83.1
sheep	94.0	91.8	**95.6**
bird	95.3	93.4	**95.8**
cat	92.3	92.6	**94.5**
dog	93.0	87.8	91.3

(b) MSRC, supervised

class	N	L	[12]	FMaps-s	FMaps-uns
plane	178	88	90.7	**92.1**	89.4
bus	152	78	81.6	**87.1**	80.7
car	255	128	76.1	**90.9**	82.3
cat	250	131	77.7	**85.5**	82.5
cow	135	64	82.5	**87.7**	85.5
dog	249	121	81.9	**88.5**	84.2
horse	147	68	83.1	**88.9**	87.0
sheep	120	63	83.9	**89.6**	86.5

(c) PASCAL

Figure 8: Performance comparison on the MSRC and PASCAL data sets. N and L denote the number of images and the number of labeled images in each class, respectively. FMaps-s and FMaps-uns are supervised and unsupervised versions of the proposed method, respectively.

mated and co-optimized to ensure better cycle-consistency in the image network. Using the obtained functional maps, segmentation functions of all images are jointly optimized so that they are consistent under functional transport and well-aligned with each image's own segmentation cues. The new method significantly outperforms the recent state-of-the-art methods on iCoseg, MSRC, and PASCAL VOC 2012.

Many topics remain for further exploration, including how the performance depends on the characteristics of the image set, and whether multiple object classes can be handled at once — possibly providing a tool for automated entity extraction from image collections. In addition, building the image network based on the GIST descriptor can be improved, and more sophisticated methods to assess image similarity can be brought to bear on the problem.

In effect, our consistently aligned network of images serves as an abstraction of the object class represented by the images. We believe that this approach, focussing on establishing transport mechanisms for image properties in a network setting and then using global analysis tools over the entire network, can be beneficial to other vision problems on joint analysis of image collections.

9. Acknowledgement

The authors acknowledge the support of grants NSF IIS 1016324 and CNS 0832820, AFOSR FA9550-12-1-0372, ONR N00014-13-1-0341, a Google research award, and the Max Planck VCC Center.

References

[1] *http://sedumi.ie.lehigh.edu/*. 5

[2] M. Ankerst, G. Kastenmller, H.-P. Kriegel, and T. Seidl. 3D shape histograms for similarity search and classification in spatial databases. *LNCS*, 1651:207–226, 1999. 7

[3] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. iCoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, 2010. 2, 6

[4] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *CVPR*, 2005. 2

[5] K.-Y. Chang, T.-L. Liu, and S.-H. Lai. From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In *CVPR*, 2011. 2

[6] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. 2

[7] O. Duchenne, A. Joulin, and J. Ponce. A graph-matching kernel for object categorization. In *ICCV*, 2011. 2

[8] M. Everingham, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88:303–338, 2010. 2

[9] D. S. Hochbaum and V. Singh. An efficient algorithm for co-segmentation. In *ICCV*, 2009. 2

[10] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, 2010. 1, 2, 6, 7, 8

[11] D. Kuettel and V. Ferrari. Figure-ground segmentation by transferring window masks. In *CVPR*, 2012. 2

[12] D. Kuettel, M. Guillaumin, and V. Ferrari. Segmentation propagation in ImageNet. In *ECCV*, 2012. 2, 6, 7, 8

[13] M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. In *ICCV*, 2005. 2

[14] L. Mukherjee, V. Singh, and C. R. Dryer. Scale invariant cosegmentation for image groups. In *CVPR*, 2011. 2

[15] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 4

[16] M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher, and L. Guibas. Functional maps: A flexible representation of maps between shapes. In *SIGGRAPH*, 2012. 1, 2, 3

[17] R. Roberts, S. N. Sinha, R. Szeliski, and D. Steedly. Structure from motion for scenes with large duplicate structures. In *CVPR*, pages 3137–3144, 2011. 2

[18] C. Rother, V. Kolmogorov, T. Minka, and A. Blake. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into MRFs. In *CVPR*, 2006. 2

[19] J. C. Rubio, J. Serrat, A. Lopez, and N. Paragios. Unsupervised co-segmentation through region matching. In *CVPR*, 2012. 1, 2, 6, 7, 8

[20] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22:888–905, 2000. 3, 4, 6

[21] J. Shotton, J. Winn, C. Rother, and A. Criminisi. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006. 2

[22] A. Singer and H.-T. Wu. Vector diffusion maps and the connection Laplacian. *CPAM*, 65(8):1067–1144, 2012. 2

[23] L. Torresani, V. Kolmogorov, and C. Rother. Feature correspondence via graph matching: Models and global optimization. In *ECCV*, 2008. 2

[24] S. Vicente, C. Rother, and V. Kolmogorov. Object cosegmentation. In *CVPR*, 2011. 1, 2, 6, 7, 8

[25] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *CVPR*, 2010. 2